

# Bonding and Structure in Solids

**J. C. Phillips**
*Lucent Technologies*

I. Introduction: Molecules and Solids
II. Molecular Crystals
III. Ionic Crystals and Electronegativity
IV. Covalent Crystals and Directed Valence Bonds
V. Mixed Covalent and Ionic Bonding
VI. Metallic Bonding
VII. Quantum Structural Diagrams
VIII. Complete Quantum Structure Analysis
IX. Chemical Bonding in Solids in the Third Millennium

## GLOSSARY

**Atom** Smallest unit of an element.
**Bond** Electronic configuration that binds atoms together
**Covalent bond** Chemical bond formed by electron sharing.
**Crystals** Solids in which the atoms are arranged in a periodic fashion.
**Electronegativity** Measure of the ability of an atom to attract electrons.
**Glass** Solid in which the atoms are not arranged in a periodic fashion and which melts into a supercooled liquid when heated rapidly.
**Ionic bond** Chemical bond caused by charge transfer.
**Metallic** Material with high electrical conductivity at low frequency.
**Molecule** Bonded atoms in a gas.
**Valence** Number of electrons used by an atom to form chemical bonds.

**THE RELATIVE POSITIONS** of atoms in molecules and solids are described and explained in terms of the arrangements of their nearest neighbors. Together with the chemical valences of the atoms as given by the periodic table, these arrangements of the bonding determine the structure and physical properties of solids. Both structure and properties can be used to separate solids into various classes where further quantitative trends can be systemically described by structural diagrams.

## I. INTRODUCTION: MOLECULES AND SOLIDS

The combinations of atoms found in the vapor phase are called molecules. Molecules containing a small number of atoms have been studied accurately and extensively. Most of our knowledge of chemical bonding between atoms comes from these studies. When atoms are condensed to

form solids, the atomic density is much greater, as reflected by the number of atoms that are nearest neighbors of any given atom. This number is called the coordination number. An example is the molecule NaCl, in which each atom has one nearest neighbor. In solid NaCl each atom has six nearest neighbors.

Solids in their pure forms are nearly always crystalline. A crystal is a periodic arrangement of atoms along lines, which in turn is repeated periodically along planes. Finally, the planes are repeated periodically to form the crystal lattice.

Most of our knowledge of crystal structures comes from the diffraction of waves of photons, electrons, or neutrons by lattice planes. Usually all the atomic positions in the crystal can be determined this way. By comparing chemical trends in bond lengths in crystals with those in molecules, one can often infer the nature of the electronic charge distribution responsible for chemical bonding in the crystal. From this it may be possible to predict the nature of chemical bonding at crystalline defects or even in noncrystalline solids, which are amorphous or glassy.

The structures of millions of solids are known by diffraction. To understand these structures one begins by studying the simplest cases and classifying them into groups. The main groups are characterized as molecular, metallic, ionic, and covalent. In most solids the actual bonding is a mixture to some degree of these different kinds of chemical interaction. While most solids are complex, the inorganic solids, which are best understood because they have had the widest technological applications, are usually either simple examples from a main group or are closely related to them. In contrast, organic and biologically important molecules may be quite complex. The chemical and structural simplicity of technologically important inorganic solids stems from the requirement of availability of techniques for production in bulk.

Certain general techniques are widely used for describing bonding and structure in solids. Tables of atomic radii are available for ionic, covalent, and metallic bonding. Deviations of bond lengths from values predicted by these radii of order 1–3% often reveal critical structural features of importance to material fabrication and properties. The cohesion of solids can be connected to the cohesion of the elements. A binary solid $A_mB_n$ is said to have heat of formation $\Delta H_f$, which is the difference between $m$ times the cohesive energy of A plus $n$ times that of B minus the cohesive energy of $A_mB_n$. This heat of formation can be estimated with often remarkable accuracy from Pauling's table of elemental electronegativities $X(A)$. This is probably the most widely used table in science apart from the periodic table of the elements, and it is shown here as Table I.

## II. MOLECULAR CRYSTALS

We now turn to the differently bonded main groups of solids. The molecular crystals are the simplest case, because the intermolecular forces are typically much weaker than the intramolecular ones. As a result the structure of the molecules, as reflected, for example, by bond lengths and vibration frequencies, is almost the same in the solid as in the gas phase. Some examples of materials that form molecular solids are the inert gases, diatomic halogens, closed-shell molecules such as methane, and many planar aromatic molecules such as benzene. Typically, in molecular crystals the heat of fusion per molecule per bond is at least 10 times smaller than the bond dissociation energy.

The binding forces that hold molecular crystals together may arise from electric dipoles if the molecules carry permanent dipole moments (e.g., HCl). When the molecules have no permanent moment, binding arises from mutually induced dipole moments (van der Waals interactions).

**TABLE I Electronegativity Table of the Elements According to Pauling**

| | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Li | Be | B | | | | | | | | | | | C | N | O | F |
| 1.0 | 1.5 | 2.0 | | | | | | | | | | | 2.5 | 3.0 | 3.5 | 4.0 |
| Na | Mg | Al | | | | | | | | | | | Si | P | S | Cl |
| 0.9 | 1.2 | 1.5 | | | | | | | | | | | 1.8 | 2.1 | 2.5 | 3.0 |
| K | Ca | Sc | Ti | V | Cr | Mn | Fe | Co | Ni | Cu | Zn | Ga | Ge | As | Se | Br |
| 0.8 | 1.0 | 1.3 | 1.5 | 1.6 | 1.6 | 1.5 | 1.8 | 1.8 | 1.8 | 1.9 | 1.6 | 1.6 | 1.8 | 2.0 | 2.4 | 2.8 |
| Rb | Sr | Y | Zr | Nb | Mo | Tc | Ru | Rh | Pd | Ag | Cd | In | Sn | Sb | Te | I |
| 0.8 | 1.0 | 1.2 | 1.4 | 1.6 | 1.8 | 1.9 | 2.2 | 2.2 | 2.2 | 1.9 | 1.7 | 1.7 | 1.8 | 1.9 | 2.1 | 2.5 |
| Cs | Ba | La–Lu | Hf | Ta | W | Re | Os | Ir | Pt | Au | Hg | Tl | Pb | Bi | Po | At |
| 0.7 | 0.9 | 1.1–1.2 | 1.3 | 1.5 | 1.7 | 1.9 | 2.2 | 2.2 | 2.2 | 2.4 | 1.9 | 1.8 | 1.8 | 1.9 | 2.0 | 2.2 |
| Fr | Ra | Ac | Th | Pa | U | Np–No | | | | | | | | | | |
| 0.7 | 0.9 | 1.1 | 1.3 | 1.5 | 1.7 | 1.3 | | | | | | | | | | |

The structures of molecular crystals are determined primarily by packing considerations and thus vary from material to material according to molecular shape. Molecular solids are generally poor conductors of electricity, and even the photoconductivity is generally small unless metallic impurities are added to "sensitize" the material.

## III. IONIC CRYSTALS AND ELECTRONEGATIVITY

Before discussing the structure of ionic crystals in detail, we shall familiarize ourselves with the concept of electronegativity, defined by Pauling as "the ability of atoms in the bonded state to attract electrons to themselves." Atoms in solids are in a variety of bonded states, and it is due to Pauling's insight that we have come to realize that the atomic electronegativity that he defined in terms of heat of formation (Section I) is indeed nearly constant for each element. His idea is that in solids charge flows from cations with smaller electronegativity to anions with greater electronegativity and that the heat of formation resulting from this charge flow is proportional to $(X_c - X_a)_v^2$, where $X_c$ and $X_a$ are the cation and anion electronegativities, respectively.

Ionic crystals are composed of cations and anions with very large electronegativity differences, such as alkali metals and halides, columns I and VII of the periodic table, respectively. In this case the charge transfer of valence electrons is almost complete, so that the core configurations become isoelectronic to those of inert-gas atoms (e.g., $Na^+$ to Ne, $Cl^-$ to Ar). While some energy is required to ionize the cations and transfer electrons to the anions, this energy is more than recovered thanks to the larger electronegativity of the anions and the mutual attraction of cations by their anion neighbors. In the case of the alkali halides, the cohesive energies can be estimated within a few percentage points by assuming complete charge transfer and evaluating the electrostatic energies (including ion polarization energies). A core–core repulsive energy, required by the exclusion principle, completes the calculation, which was first sketched around 1910.

As one might expect, the overall features of the crystal structures of ionic crystals are given quite well by packing spherical cations and anions in the appropriate proportions indicated by their chemical formulas. However, the ions are not quite the incompressible spheres suggested by their isoelectronic analogy to inert-gas atoms. If they were, one could use simple geometrical arguments (originating around 1930) to predict a coordination number of 8 (CsCl structure), 6 (NaCl structure), or 4 (ZnS structure). These correspond to packing cations and anions of nearly equal size (CsCl structure), and then successively larger anion/cation size ratios lead to increasing anion–anion contacts, thus reducing coordination numbers. These "radius ratio" rules do not actually describe the crystal structures, as shown in Fig. 1. What this means is that the ions should not be regarded as hard spheres, but rather as centers of quantum mechanically determined electronic charge distributions. Additional evidence for the breakdown of classical electrostatic models is contained in the elastic constants of the alkali halides. If these models were correct, the elastic constants would satisfy certain relations (the Cauchy relations) valid for central force interactions. These relations are not satisfied for most of the alkali halides, another indication of quantum mechanical interactions. Some simplified modern treatments of these and related problems are discussed in the following sections.

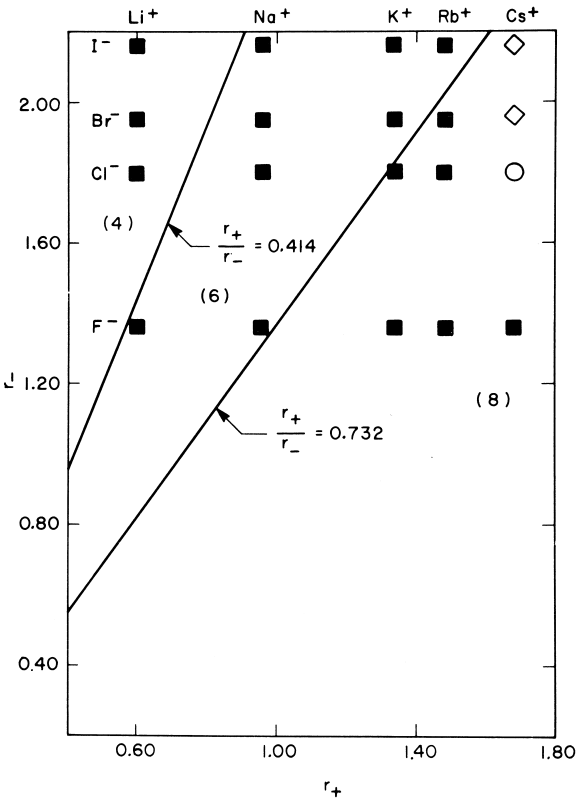



**FIGURE 1** The structures of the alkali halides $M^+X^-$ as functions of classical ionic radii $r_+$ and $r_-$, respectively. (Coordination numbers in parentheses are those predicted by the classical ionic model.) In the upper left corner, for example, $Li^+I^-$ is predicted by the classical model to have coordination number 4, but the symbol indicates it is actually sixfold coordinated. Key: ■, sixfold coordinated; ◇, eightfold coordinated; ○, six- or eightfold coordinated. [From Phillips, J. C. (1974). In "Solid State Chemistry" (N. B. Hannay, ed.), Vol. 1, Plenum, New York.]

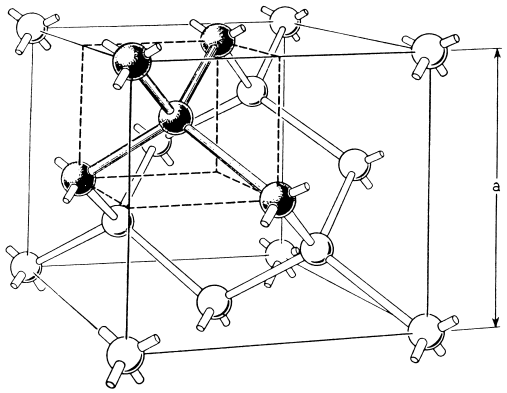



**FIGURE 2** The tetrahedrally coordinated diamond structure, which describes many technologically important semiconductors such as silicon. [From Phillips, J. C. (1970). Phys. Today 23 (Feb.), 23.]

## IV. COVALENT CRYSTALS AND DIRECTED VALENCE BONDS

Whereas ionic crystals can be (at least roughly) described in classical terms, structure and bonding in covalent crystals can be understood only in terms of quantum mechanical electron orbital wave functions. Prototypical covalent crystals have the diamond structure. Many technologically important semiconductors such as silicon and germanium have this structure or a closely related one, the zinc blend or wurtzite structure. In these structures each atom is tetrahedrally coordinated (Fig. 2).

The structure shown in Fig. 2 can be explained simply in terms of directed valence electron orbitals. The valence configuration of the atom is $ns^2\,np^2$, with $n = 3$ for silicon. In the crystal this becomes $ns\,np^3$ so that (counting electron spin twofold degeneracy) both the $ns$ and $np$ levels are half-filled. These four states can be combined to form four directed valence orbitals with tetrahedral geometry. The wave functions on nearest neighbors can be combined in phase to form bonding states or out of phase to form antibonding states. Then wave-function overlap produces an energy gap between these states (Fig. 3). This energy gap is the basis of the technologically important electronic and optical properties of semiconductors.

The covalent energy gained by wave-function overlap or interference is much more sensitive to structural perfection than is the energy associated with classical ionic interactions. A very important consequence of this sensitivity is that it is possible to produce semiconductor crystals such as silicon in far purer and far more structurally perfect states than has been possible with any other solid. It is possible to add impurities with designed concentrations and locations to tailor the chemical and mechanical design of the solid with far greater precision and ease than for any other solid. Thus, the quantum mechanical nature of structure and bonding in covalent silicon is the key to its technological significance.

## V. MIXED COVALENT AND IONIC BONDING

Most semiconductors and insulators have neither purely covalent nor purely ionic bonding, but their bonding is described as a mixture of covalent and ionic effects. The

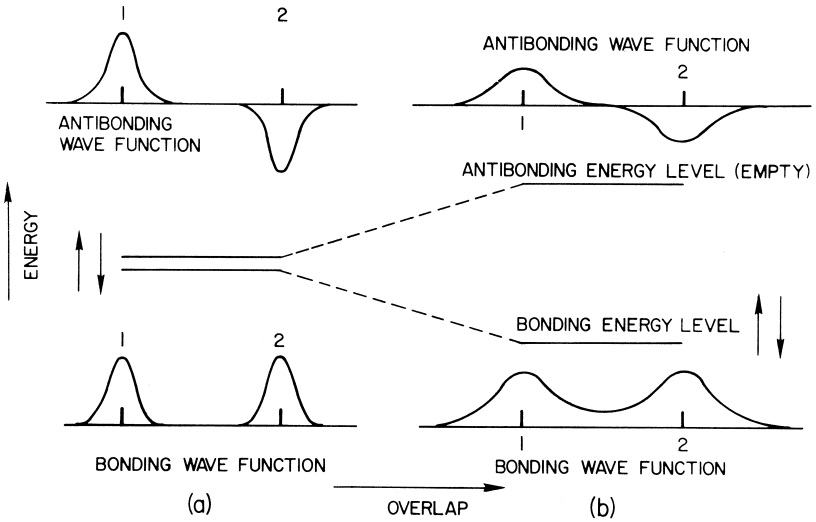



**FIGURE 3** Sketch of electronic interactions between directed valence orbitals that produce an energy gap between bonding states (electrons shared between nearest neighbors) and antibonding states (no electron sharing). [From Phillips, J. C. (1970). Phys. Today 23 (Feb.), 23.]

way in which the mixture occurs is of great importance both scientifically and technologically. We shall discuss several important examples.

The simplest case occurs for the tetrahedrally coordinated covalent structure shown in Fig. 2. This structure contains two kinds of atomic sites: site A with only B neighbors, and vice versa. In silicon and gemanium both sites are occupied by the same atom, which has four valence electrons. However, one can occupy the two sites with different atoms, such that the total number of valence electrons is eight per atom pair (formally represented by $A^N B^{8-N}$). Many compounds of this type with $N = 3$ and $N = 2$ are known. Two examples with $N = 3$ are GaAs and InP and ternary and quaternary alloys (Ga, In) (As, P). These materials are transparent in the near infrared, and their optical properties can be adjusted by "band gap engineering." They are important as high intensity, low power monochromatic light sources or as light amplifiers (lasers).

The next interesting case is the triatomic material $SiO_2$ (silica). The electronegativity difference between silicon and oxygen is large, so the bonding here contains a large ionic component. At the same time each oxygen atom contains six valence electrons while silicon has four, so the total number of valence electrons per molecular unit is $16 = 8 + 8$. This favors covalent bonding. In the solid each silicon atom has tetrahedral oxygen neighbors, while each oxygen atom has two silicon neighbors, which is again the coordination characteristic of covalent bonding. Silica is chemically stable and can be made very pure, for much the same reasons that silicon can. This high purity is essential to technological applications in the context of optical fibers for communications.

Another feature of silica is that it can easily be cooled into a solid state that is not crystalline but more like a frozen liquid. The state is called a glass. The ductility of glasses at high temperatures is essential to the manufacture of optical fibers. However, glasses are also ductile on a molecular scale and so do not form molecular "cracks," which would be arrays of broken bonds that might be electrically active and destructive to the electronic capabilities of semiconductor devices such as transistors. It is one of nature's most felicitous accidents that silicon electronic devices can be packaged by simply oxidizing the surfaces of solid silicon to form a protective coating of silica, $SiO_2$. The silica coating is not only chemically stable (because of its covalent-ionic bonding), but is also mechanically stable because of the ductility of vitreous silica at the molecular level. Thus, the interface between the crystalline silicon electronic device and the silica coating is itself almost perfect. It does not store fixed charge, even when the thickness of the silica is only a few molecular layers. This greatly enhances the performance of microelectronic devices (integrated circuits on silicon "chips."

The last case is primarily ionic materials with a covalent component. Many oxides in which the oxygen atoms are three- to sixfold coordinated fall in this class, and this includes many ceramic materials. These materials can have high melting points and good chemical stability, but they are brittle and for this reason their range of technological applications is limited.

## VI. METALLIC BONDING

Broadly speaking, three kinds of elements are found in metals. They are the simple $s$–$p$ metallic elements from the left-hand side of the periodic table, such as lithium, aluminum, and lead; the rare-earth and transition elements with $f$ and $d$ valence electrons, such as titanium, iron, and nickel; and metalloid elements, such as carbon, silicon, and phosphorus, which may also in certain combinations form covalent solids. In metals the coordination number (or number of nearest neighbors) is much larger (usually twice as large or more) than the number of valence electrons. This means that the directed valence bonds found in molecules or in covalent crystals are much weaker (although not completely absent) in metals.

The high electrical and thermal conductivity of metals is a result of the absence of a gap in the energy spectrum between filled and empty electronic states. This high electrical conductivity in turn reduces the contribution to cohesion associated with charge transfer because the internal electric fields are limited by electronic redistribution or charge flow on an atomic scale. Thus, ionic interactions are reduced in metals compared with ionic crystals.

The reduction of covalent or molecular bonding as well as ionic bonding in metals presents a paradox. If neither of these bonding mechanisms is fully effective, to what forces do metals owe their cohesion? Modern quantum theory shows a complex correlation of the motion of metallic valence electrons, which reduces the Coulomb repulsive energy between these like charges while leaving almost unchanged the attractive Coulomb interaction between negatively charged electrons and positively charged atom cores. It is this correlation energy that is primarily responsible for metallic cohesion.

From studies of the structure and cohesion of metals it appears that $d$ valence electrons (as in the transition metals) contribute almost as effectively to metallic cohesion as $s$ and $p$ electrons. The $f$ electrons in rare-earth metals, on the other hand, play a minor role in metallic cohesion but occasionally have magnetic properties. Transition metals are notable for their strong magnetic properties (iron, cobalt, and nickel), as well as their high melting points and refractory properties, which result from the large number of combined $s$, $p$, and $d$ valence electrons.

The compound with the highest known melting point is tungsten carbide (WC), an interesting combination of a transition element whose *d* levels are half-full with a metalloid element whose *s* and *p* valence levels are half-full. Also, here tungsten is very large and carbon is very small, which makes possible an ionic contribution to the cohesive and refractory properties.

## VII. QUANTUM STRUCTURAL DIAGRAMS

The description of structure and bonding in solids given in the preceding sections is largely qualitative, but it is a fair (although abbreviated) account of most of what was generally known as a result of quantum mechanical analysis in the period from 1930 to 1960. Starting in 1960 a more quantitative description was developed that enables us to inspect systematic trends in structure and bonding with the aid of quantum structural diagrams.

With a structural diagram one assigns to each element certain characteristics and then treats these characteristics as configuration coordinates, which are used to construct structural maps. The natural classical configuration coordinates are atomic size and electronegativity, as defined by Pauling (see Table I). To these we may add the number of valence electrons per atom. One then takes a class of binary compounds, say $A^NB^M$, with the same value of $P = N + M$ and uses size differences (or ratios) as well as electronegativity differences as Cartesian coordinates. If the characteristics or configuration coordinates have genuine value for describing structure and bonding, compounds composed of different elements A and B but with similar values of their Cartesian coordinates, should have the same crystal structure. Put somewhat differently, the structural map should separate into simple regions, with each region containing compounds with the same crystal structure.

Early attempts to construct structural maps of this kind using classical coordinates were only partially successful; as many as 10 or 20% or more of the compounds were misplaced. From this failure most workers concluded that the problem of structure and bonding in solids, and especially in metals where the number of known compounds exceeds $10^4$, was simply too complex to solve in any simple way. Finding a solution was left to the indefinite future, when computers became large enough and quantum mechanical methods accurate enough to predict structures on a case-by-case basis.

Recent research has shown that the idea of structural diagrams is itself valid but that previous failures arose from the use of largely classical coordinates. In addition to the number of valence electrons per atom (a quantum concept), one must also use other quantum variables to replace the classical variables of atomic size and electronegativity. This has been done in several ways, which are substantially equivalent. The simplest case is $A^NB^M$ compounds where A and B have only *s* and *p* valence electrons and $N + M = P = 8$, which means that the *s* and *p* valence levels are half-full. In this case one can separate ionic and covalent crystal structures by separating the average energy gap between occupied and empty electronic states into ionic and covalent components, represented by $C$ and $E_h$, respectively. Both NaCl (ionic) and diamond, silicon, and germanium (covalent) crystals (Fig. 2) belong in this group, with $C/E_h = 0$ in the latter and $C/E_h$ largely in the former. The quantum structural diagram for $A^NB^{8-N}$ nontransition-metal compounds shown in Fig. 4 not only is a huge improvement on the classical structure diagram shown in Fig. 1 but also is an exact separation of covalent and ionic crystal structures.

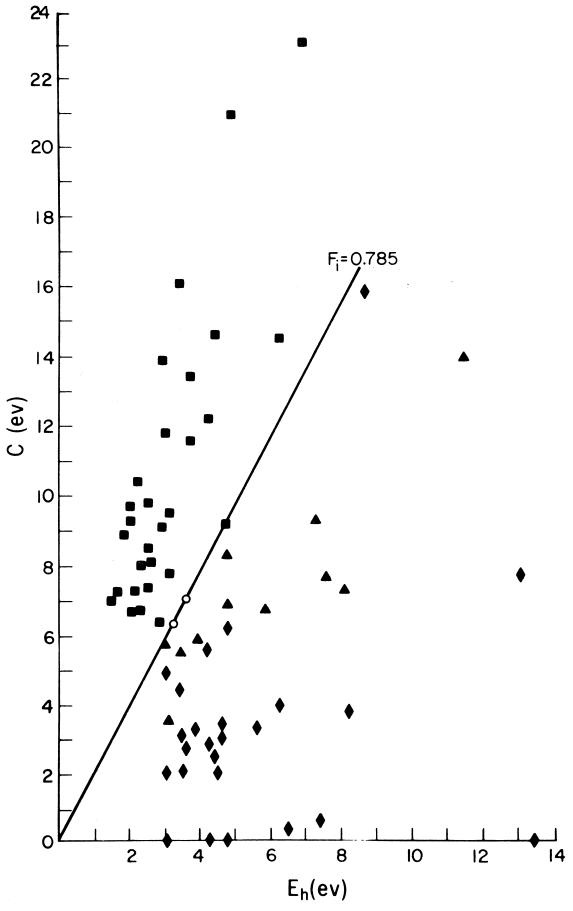



**FIGURE 4** The separation of the energy gap shown in Fig. 3 into covalent and ionic components ($E_h$ and $C$, respectively) generates a structural map that separates fourfold- and sixfold-coordinated $A^NB^{8-N}$ crystals perfectly (no transition or rare-earth elements). The structures and coordination numbers (in parentheses) are as follows: ♦, diamond, zinc blend (4); , wurtzite (4); ■, rock salt (6); ○, rock salt/wurtzite (6, 4). [From Cohen, M. L., Heine, V., and Phillips, J. C. (1982), Sci. Am. 246 (6), 82.]

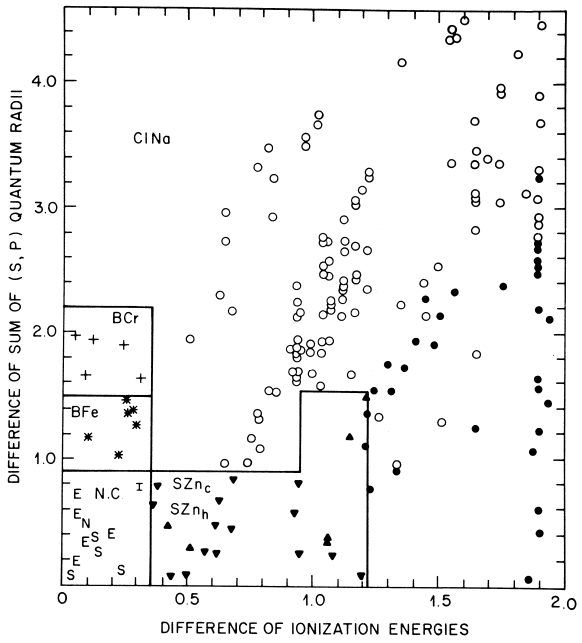



**FIGURE 5** A general separation of $A^NB^{8-N}$ crystal structures utilizes quantum coordinates defined for all elements including rare-earth and transition metals. Compounds containing the latter are indicated by open symbols. [From Villars, P. (1983). J. Less-Common Met. 92, 215.]

To extend this analysis to transition and rareearth metals as well as compounds in which the valence shell is not exactly half-full is a monumental taks that includes ~1000 AB compounds, ~1000 $AB_2$ compounds, and more than 1000 $AB_3$ and $A_3B_5$ compounds, as well as more than 7000 ternary compounds. The correct quantum coordinates for these 10,000 compounds have been identified from a field of 182 candidate coordinates, some classical and some quantum coordinates. All the best coordinates are found to be quantum coordinates, and these turn out to be the atomic ionization potential and a suitably defined quantum core size. The result for $A^NB^{8-N}$ compounds (where A or B or both may be transition or rare-earth elements) is shown in Fig. 5. It is representative of the best global analysis of structure and bonding in solids available in 1992. This structural map is 97% successful.

In addition to binary compounds one can use diagrams to analyze ternary compounds. Ternary ionic compounds usually contain two kinds of cations, and their structures are determined by cation radius ratios. Ternary metallic compounds are more complex, and their structures are determined by valence electron numbers, size differences, and electronegativity differences, much as for the binary compounds in Fig. 5. Many structure–property relationships can be recognized with these diagrams which conveniently display general trends in both binaries and ternaries.

## VIII. COMPLETE QUANTUM STRUCTURE ANALYSIS

On a case-by-case basis a full discussion of structure and bonding in a given solid can be achieved using the most advanced computational techniques combined with the most sophisticated computers. Work with sufficient precision and flexibility to describe the structure of solid surfaces, point defects, and solid transitions under high pressures became available in selected cases in the 1980s. An excellent example is shown in Fig. 6, which gives the total energy of crystalline silicon in different crystal structures as a function of volume. From these curves transition pressures and volumes can be obtained from the tie-line (common tangent) construction due (~100 years ago) to Gibbs. It is interesting that all the results shown in Figs. 4, 5, and 6 are based on a particular approach to the quantum structure of solids that is known as the pseudopotential method.

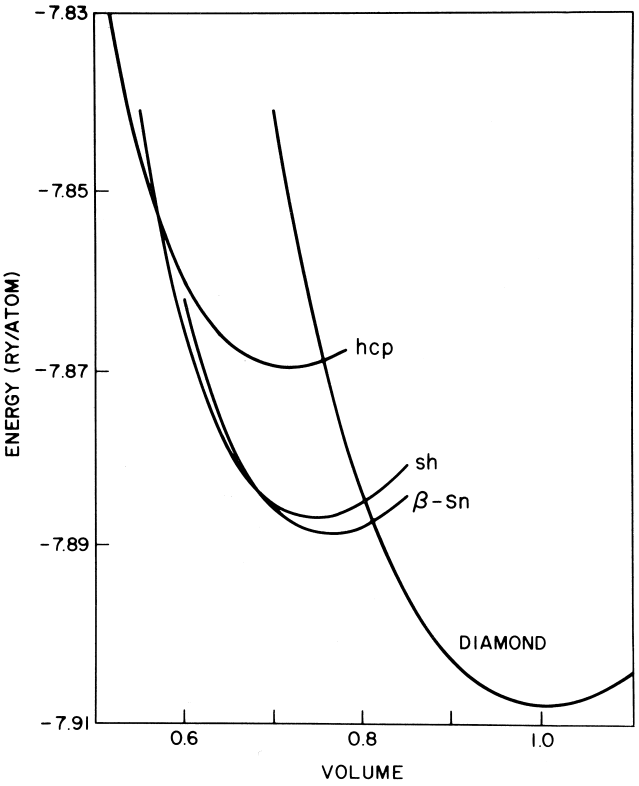



**FIGURE 6** A plot of the total energy of silicon crystals in different crystal structures as a function of atomic volume. At atmospheric pressure the diamond structure has the lowest energy, but at pressure of hundreds of thousands of atmospheres silicon is more stable in other structures. Such high pressures can be produced in the laboratory, and they are also found at great depths below the earth's surface. [From Chang, K. J., and Cohen, M. L. (1984). Phys. Rev. B 30, 5376.]

## IX. CHEMICAL BONDING IN SOLIDS IN THE THIRD MILLENNIUM

The evolution of microelectronic devices towards smaller and smaller dimensions will soon reach the level of 2.5 nm (25A) or less, which is basically the molecular level. At this level the concepts of chemical bonding discussed here cease to be only theoretical abstractions and become valuable tools for guiding microelectronic device design and manufacture. A remarkable aspect of much recent research is that it demonstrates that both macroscopic and quantum ideas of materials can be implemented at the molecular level when the processes involved are well controlled.

## SEE ALSO THE FOLLOWING ARTICLES

CRYSTALLOGRAPHY • EXCITONS, SEMICONDUCTOR • FERROMAGNETISM • GLASS • QUANTUM MECHANICS • SOLID-STATE CHEMISTRY • SOLID-STATE ELECTROCHEMISTRY • SUPERCONDUCTIVITY • VALENCE-BOND THEORY • X-RAY ANALYSIS

## BIBLIOGRAPHY

Adams, D. M. (1974). "Inorganic Solids," Wiley, New York.
Chang, K. J., and Cohen, M. L. (1984). *Phys. Rev. B* **30,** 5376.
Cohen, M. L., Heine, V., and Phillips, J. C. (1982). *Sci. Am*. **246(6),** 82.
Pauling, L. (1960). "Nature of the Chemical Bond," Cornell Univ. Press, Ithaca.
Phillips, J. C. (1970). The chemical bond and solid state physics, *Phys. Today* **23** (February), 23.
Phillips, J. C. (1974). In "Solid State Chemistry" (N. B. Hannay. ed.), Vol. 1: The Chemical Structure of Solids, Plenum, New York.
Tosi, M. P. (1964). *Solid-State Phys.* **16,** 1.
Villars, P. (1983). *J. Less-Common Met.* **92,** 215.
Villars, P. (1985). *J. Less-Common Met*. **109,** 93.
Villars, P., and Phillips, J. C. (1988). *Phys. Rev. B* **37,** 2345.
Wigner, E. P., and Seitz, F. (1955). *Solid-State Phys*. **1,** 1.



# Chemical Physics

**Richard Bersohn**
**Bruce J. Berne**
*Columbia University*

I. Properties of Individual and Pairs of Molecules
II. Collective Properties

## GLOSSARY

**Born–Oppenheimer approximation** A quantum mechanical explanation for the approximate separation of molecular energy into electronic, vibrational, and rotational energies.

**Electric multipole moment** If the charge density of a system is $\rho(r, \theta, \phi)$ where $r, \theta, \phi$ are spherical polar coordinates, then the $l$th multipole moments are the set of averages $\int \rho(r, \theta, \phi) r^l Y_l^m(\theta, \phi)\, dV$. The moments of a spherically symmetric charge distribution are zero.

**Green–Kubo relations** Expressions for transport coefficients such as viscosity, thermal conductivity, and rate constants in terms of time correlation functions.

**Molecular dynamics method** A method for simulating the properties of many-body systems based on solving classical equations of motion.

**Monte Carlo method** A method for simulating the equilibrium properties of many-body systems based on random walks.

**Normal coordinates** The coordinates of a vibrating system that oscillate with a single frequency.

**Partition function** A sum over quantum states used to determine thermodynamic properties from the quantum mechanical energy levels.

**Path integral methods** A formulation of quantum mechanics and quantum statistical mechanics developed by Feynman.

**Radial distribution function** The average density of fluid atoms as a function of distance from a given fluid atom.

**Raman scattering** An inelastic scattering of a photon by a molecule; the difference in energy between the incident and scattered photon is a difference of molecular energy levels.

**Spectroscopy** The measurement of energy levels.

**Statistical mechanics** A general theory of many particle systems that relates bulk properties to microscopic properties.

**Time correlation functions** A function that describes the correlation between properties of a system at different times.

**CHEMICAL PHYSICS** is the physics of the individual and collective properties of molecules. However, the distinction between chemistry and chemical physics is largely a matter of emphasis. The approach of the chemical physicist is theoretical. He searches for underlying theoretical principles, and the molecules that he uses are often a means to an end, whereas the synthetic chemist usually considers the molecules that he synthesizes and their reactions as ends in themselves. This article on chemical

physics is divided into two sections, one on phenomena which depend primarily on the properties of individual and pairs of molecules and the other on phenomena which are primarily collective.

## I. PROPERTIES OF INDIVIDUAL AND PAIRS OF MOLECULES

Studies in chemical physics can be loosely classified as spectroscopic, structural, and dynamic. Spectroscopy is concerned with the determination of molecular energy levels. Structural studies are aimed at finding the distribution of particles within a molecule and molecules within a liquid or solid. The location of the nuclei defines the structure of the molecule, that is, the distances between nuclei and the angles between internuclear vectors. The distribution of electrons is intimately connected with the forces that hold the atoms together. Dynamics involves the relation of the rate of molecular transformations and changes of state caused by collision to the intra- and intermolecular forces.

### A. Molecular Spectroscopy

Spectroscopy is the measurement of energy level differences. This is most usually accomplished by measuring the frequencies of light absorbed or emitted by a molecule, but it is sometimes done by measurements of the energy of an incident photon or particle together with a measurement of a scattered photon or particle. For example, the frequency of scattered light may differ from that of the incident light. The absolute value of the frequency difference is a difference of energy levels of the molecule divided by Planck's constant. This phenomenon, called Raman scattering, has many analogs. Electron loss spectroscopy is extensively used to measure vibrational frequencies of surfaces. The difference in energy between incident and scattered electrons is, in general, a quantized energy left in the solid. When very slow ("cold") neutrons are scattered by a warm liquid or solid, the scattered neutrons move faster than those in the incident beam.

In some spectroscopies the scattered particle whose energy is measured is not the same as the incident particle. For example, in photoelectron spectroscopy an incident photon with known energy whose wavelength is in the XUV (<100 nm) or X-ray region produces a photoelectron. Careful measurements of the kinetic energy of the electron yield energy level spacings in the positive ion. In photodissociation spectroscopy, when a molecule is dissociated into fragments by light, measurement of the kinetic energy of a fragment yields the internal energy of the fragments.

Before discussing spectroscopy, it might be appropriate to warn and console the reader about the multitude of units of energy used to describe spectroscopic phenomena. The SI unit of energy is the joule ($1\ \mathrm{J} = 10^7$ ergs), but this unit is rarely used by spectroscopists. Just as the nuclear spectroscopist uses one million electron volts ($1\ \mathrm{MeV} = 9.65 \times 10^{10}$ J) as a unit of energy, the X-ray spectroscopist uses electron volts ($1\ \mathrm{eV} = 9.65 \times 10^4$ J). The ultraviolet, visible, and infrared spectroscopists use the $\mathrm{cm}^{-1}$ unit, which is not an energy unit but a reciprocal of a wavelength ($1\ \mathrm{cm}^{-1} \times h/c = 11.96$ J). The microwave and radiofrequency spectroscopists use the megahertz ($1\ \mathrm{MHz} \times h = 3.99 \times 10^{-4}$ J), and nuclear magnetic resonance (NMR) spectroscopists often use a dimensionless quantity, the relative frequency shift in parts per million. Some spectroscopists use units of energy, others units of reciprocal wave length, and still others frequency units. Indeed the reference to spectral features by their wave lengths is still a very common practice, although it is to be deplored. There are natural reasons for these different choices of energy units but they are bewildering to the beginner. In a way, these units are like the light year, a nonstandard unit convenient to the astronomer but to nobody else.

Molecular spectroscopy is generally divided into four branches, each corresponding to a different type of motion and, in general, to a different frequency range. Electronic spectroscopy is the study of the differences in electronic energy levels which occur in atoms and molecules; the corresponding frequencies are $10^{14}$–$10^{17}$ Hz. Vibrational spectroscopy is the study of molecular vibrations, whose frequencies are typically of the order of $10^{12}$–$10^{14}$ Hz. Rotational spectroscopy is the study of the rotational energy levels of molecules; corresponding rotational frequencies are typically in the range of microwaves, $10^{10}$–$10^{12}$ Hz. Finally, NMR spectroscopy is the study of the magnetic fields acting on a nucleus.

The given divisions are not quite as sharp as represented. Frequency ranges overlap, and electronic spectroscopy is often a source of information of vibrations and rotations as well as electronic motion. Nevertheless, this classification of spectra is of fundamental importance. The theoretical justification for this classification is the Born–Oppenheimer approximation. This is an approximation that exploits the fact that the electron is $10^4$–$10^5$ times lighter than a typical nucleus. Classically speaking, the electron revolves around the molecule so rapidly that during an electron period the nuclei do not have time to react to the different positions of the electron and barely move. Thus, electronic energies can be calculated with the assumption that the nuclei are stationary. They are not stationary, of course, but the electronic energy levels can (in principle) be calculated for any given arrangement of the nuclei. The arrangement or configuration that produces the least electronic energy is the equilibrium structure of the molecule.

Let us pause to consider the coordinates used to describe the shape and location of an isolated set of $N$ atoms. The configuration of a molecule with $N$ atoms is determined by $3N$ coordinates. Of these, six are external coordinates in that they describe the location of the center of mass and the spatial orientation of the molecule. However, the electronic energy is independent of these coordinates, for example, the position and orientation of the molecule. Thus, there are $3N-6$ internal coordinates ($3N-5$ for a linear molecule) which are, in general, vibrations.

The electronic energy, considered as a function of the nuclear coordinates, behaves as a potential energy function for the nuclear motion. In a stable molecule, the atoms execute small vibrations about the position of the minimum of electronic energy. This statement implies that, in general, for each electronic state the molecule will have a different equilibrium structure. In mathematical terms, the potential in the $n$th electronic state can be expanded in powers of the displacement about the equilibrium position,

$$V_n(\mathbf{Q}) = V_n(\mathbf{Q}_0) + \sum_i \left(\frac{\partial V}{\partial Q_i}\right)_0 (Q_i - Q_{i0}) + \frac{1}{2}\sum_i \sum_j \left(\frac{\partial^2 V_n}{\partial Q_i \partial Q_j}\right)_0 \times (Q_i - Q_{i0})(Q_j - Q_{j0})$$

where $\mathbf{Q}$ is the set of vibrational coordinates and $\mathbf{Q}_0$ is their equilibrium value. If $\mathbf{Q}_0$ is truly the equilibrium position, then all the first derivatives in the preceding equation will vanish. If the $Q_i$ are chosen to be normal coordinates, then all the cross derivatives vanish as well and the following equation is obtained:

$$V_n(\mathbf{Q}) = V_n(\mathbf{Q}_0) + \frac{1}{2}\sum_{i=1}^{3N-6} \frac{\partial^2 V_n}{\partial Q_{i0}^2}(Q_i - Q_{i0})^2 + \cdots$$

A normal coordinate is a coordinate that classically will vary sinusoidally in time with a single frequency. In contrast, an arbitrary nonnormal coordinate will vary in time with a number (up to $3N-6$) of frequencies. The quantities $(\partial^2 V_n/\partial Q_i^2)_0$ are the force constants, the curvatures of the potential with respect to the normal coordinates $i$. From this definition, it is clear that the force constant may be different in different electronic states and indeed even the normal coordinates may differ. The simplest example is a diatomic molecule where the equilibrium internuclear distance is different in two different electronic states.

### 1. Electronic Spectroscopy

The electronic spectroscopy of atoms is clearly separated by energy domains into transitions of valence electrons and transitions of inner shell electrons. When two or more atoms combine to form a molecule, to a good approximation, core electrons remain core electrons but valence electrons occupy three types of states: bonding, nonbonding, and antibonding. For example, in the water molecule there are two core electrons in the oxygen inner shell, two pairs of bonding electrons, and two pairs of nonbonding electrons. Lower energy electronic transitions are often explained by promotion of electrons from nonbonding to antibonding states. At somewhat higher energies, electrons in bonding states may be promoted to antibonding states. At still higher energies, one has Rydberg series, which are a set of atomiclike transitions in which valence electrons are excited to progressively higher energy states where the electron circum navigates the ion core at, typically, a considerable distance. Excitations of inner shell electrons require photons with energies in the X-ray region.

Electronic spectra are identified not only by their energies but also by two closely related quantities: their intensity and their polarization. Electromagnetic fields cause transition in molecules by exciting resonances in the electronic or nuclear motion. The interaction energy is almost invariably expanded in multipoles, that is, in powers of the size of the molecule divided by the wavelength of light. This ratio is so small that for practical purposes only the electric and magnetic dipoles interacting with the external electric and magnetic fields, respectively, are considered. There is a fairly sharp distinction between the domain of spectroscopy in which the electric and magnetic dipole moments are important in causing transitions. The electric dipole moment matrix elements are typically of the order of $ea_0$, the electronic charge multiplied by the Bohr radius. The magnetic dipole moment matrix elements are typically on the order of $e\hbar/mc$, the electronic charge multiplied by the Compton wavelength. However, $\hbar/mc$ is a factor of $e^2/\hbar c \simeq \frac{1}{137}$ smaller than $a_0 = \hbar^2/me^2$. Transition probabilities are classically proportional to the square of the time derivative of the appropriate dipole moment, and quantum mechanically to the square of the matrix element between the initial and final states of that dipole moment operator. Hence, magnetic dipole transition probabilities are typically four to five orders of magnitude weaker than electric dipole transitions. Magnetic dipole transitions are therefore rarely observed in molecular spectroscopy except in paramagnetic systems where only the magnetic dipole interaction can accomplish magnetic moment reorientation.

The polarization of an electric dipole transition is the direction along which the external time-dependent electric field must lie in order to cause the transition. Transitions in an atom placed in a potential, independent of angle, would be unpolarized in the sense that the transition probability would be independent of direction of the time-dependent optical field. In all molecules except those with tetrahedral or octahedral symmetry, the absorption is anisotropic; in some directions there will be no absorption. The vector $\boldsymbol{\mu}_{if}$, the dipole moment matrix element between the

initial and final states, is often called the transition dipole. In a linear molecule, the transition dipole must lie either parallel or perpendicular to the molecular axis. In a planar molecule, typical of the majority of dyes, the transition dipole must lie either in the plane in some particular direction or perpendicular to the plane. The polarization of an absorption is most often determined by measuring the absorption of polarized light by an oriented set of molecules as in a single crystal.

### 2. Vibrational Spectroscopy

Infrared and Raman spectroscopy are used to determine the $3N-6$ vibrational frequencies of a molecule with $N$ atoms ($3N-5$ if the molecule is linear). There are both experimental and theoretical problems. On the experimental side, the intensity of a vibrational transition depends on the square of the rate of change of the dipole moment during the vibration. However, certain modes have no accompanying change in dipole moment and therefore electric dipole transitions are forbidden; examples of these modes are the symmetric stretch of $CO_2$ or the vibration of $N_2$. Raman spectroscopy often can supply the missing information. On the theoretical side, the $3N-6$ vibrational frequencies once determined must be interpreted in terms of $(3N-6)/(3N-7)/2$ force constants, that is, second derivatives of the potential function. One instructive point has emerged: Vibrational frequencies of most molecules display very little differences in the condensed phase and in the gas phase. Obvious exceptions to this rule are vibrations in such strongly interacting molecules as the water molecule.

One of the frontiers of chemical physics is the vibrationally highly excited "hot" molecule. Vibrations are no longer neatly divided into a set of harmonic normal mode oscillations. Vibrations are strongly anharmonic. Vibrations and rotations are no longer clearly separated and the nature of the states are difficult to describe. Yet it is precisely these energy-rich molecules that are the precursors to chemical reactions, whether they are unimolecular or bimolecular.

### 3. Radiofrequency Spectroscopy

Radiofrequency spectroscopy comprises many branches including nuclear magnetic resonance, nuclear quadrupole resonance, electron spin resonance, atomic beam resonance, optical pumping of atoms, and microwave spectroscopy of rotating molecules. The results of the last technique are described in the section on molecular structure. Nuclear magnetic resonance has contributed vastly more to chemistry than the other techniques.

The development of nuclear magnetic resonance is an extraordinary chapter in the history of science in which the original goal, while eventually achieved, was dwarfed by an unforseen discovery. Nuclei with odd numbers of neutrons or protons have a net quantized angular momentum $I\hbar$ where $I = \frac{1}{2}, 1, \frac{3}{2}, \ldots$. The nuclear angular momentum implies a rotation of positive charge which in turn produces a magnetic moment. The existence of nuclear magnetic moments was manifested by the hyperfine structure of atoms, an interaction energy between the nuclear magnetic moment and a magnetic field generated by the electrons. Nuclear magnetic moments had also been detected by atomic beam experiments in which atomic beams were deflected in different directions by inhomogeneous magnetic fields; radio-frequency transitions changed the relative intensities. However, the nuclear magnetic moment could be determined only if the effective magnetic field of the electron could be calculated; in the late 1940s, when NMR was developed, computers adequate to calculate these fields did not exist. Nuclear magnetic resonance measurements on solid or liquid samples seemed like the perfect solution. The moments $\mu$ could be derived from the simple equation $h\nu = \mu B$ where $h$ is Planck's constant, $B$ is the magnetic induction, and $\nu$ the frequency used to excite the transition between the $2I+1$ magnetic sublevels of the nucleus.

In practice, it was found that the apparent magnetic moment varied with the chemical environment of the nuclei. The effect was due to a local magnetic susceptibility, which, in general, increased with the number of atomic electrons. For the particular case of the protons ${}^{1}_{1}\mathrm{H}$ and ${}^{13}_{6}\mathrm{C}$, resonance fields (or frequencies) differ at various nonequivalent positions and this fact serves as a decisive structural tool for organic molecules. As important as the structure determination is, the really unique feature is the ability to measure exchange rates between different environments. For example, cyclohexane has six axial and six equatorial hydrogen atoms that interchange as the molecule converts from one chair form through the boat form to a second equivalent chair form. At room temperature, only a single resonance line is seen because the two sites interchange at a rate much faster than the frequency difference between them. When the liquid is cooled, the rate decreases and two lines are seen. The contributions of NMR to solid-state physics, all branches of chemistry, biochemistry, and even medicine through the imaging of internal organs have been over-whelming. The original goal targets, the magnetic moments of the stable nuclei, have all been measured and tabulated in reference books but a theory has not been developed to explain them.

## B. Molecular Structures

Molecular structure is most often determined by scattering methods. A beam of photons or material particles (chosen

so that their wave-lengths are comparable to close internuclear distances) is directed at the sample. At particular scattering angles, constructive interference will occur between the waves scattered by pairs of atoms or, in a crystal, planes of atoms. By measuring the relative scattered intensities at the various angles at which constructive interference occurs, a structure may be inferred. The structure of a condensed phase is probed by X-rays or neutrons; the structure of a gas molecule or a surface is commonly determined by electron scattering.

In several types of spectroscopy, molecular rotational energies are determined from which the three moments of inertia can be deduced. By measuring the spectra of the same molecule with different nuclear isotopes, additional sets of moments of inertia can be obtained.

Molecular structure is the distribution of nuclei and electrons within a molecule. The location of the nuclei, often just called the structure, is the molecular architecture. Closely connected to this architecture is the distribution of electrons circumnavigating the heavy nuclei. The techniques for determining the nuclear positions are based on interferences in scattering (diffraction) or spectroscopic measurements of moments of inertia.

### 1. Diffraction Methods

When an incident wave is scattered by a molecule, the scattered wave consists of a sum of contributions, large and small, from each of the electrons and nuclei in the molecule. The scattered intensity, which is the square of the scattered amplitude, will consist of a sum of squares of the individual amplitudes plus a sum of products of scattered amplitudes originating from different particles. The former sum is independent of the distances between particles, whereas the latter is structure dependent. The constructive interference between the different amplitudes gives rise to significant diffraction peaks when the interparticle distances are comparable to the wavelength of the incident wave. In practice, this means that the wavelength must be of the order of an angstrom, (1 Å $= 10^{-10}$ m). The most common "particles" used for scattering have been the X-ray ($\lambda \sim 1$–2 Å) and the velocity-selected slow neutron ($\lambda \sim 1$ Å). In a few cases, especially for diffraction from surfaces, hydrogen atoms and helium atoms have been used.

X-ray diffraction from single molecular crystals has been our most important source of knowledge of molecular structure. The assumption, in most cases solidly confirmed by experiment, is that the forces that hold molecules together in the solid are much weaker than the interatomic forces that determine the molecular structure. Therefore, the structure in the solid will be negligibly different in the solid, liquid, and gas phases. There are, of course, exceptions to this rule such as $PCl_5$, which is $PCl_4^+$ $PCl_6^-$ in the solid state, but the rule is generally valid.

Nuclei scatter X-rays to a negligible extent because they are so heavy that they do not react to the time-varying electric field of the photon. The electrons will contribute a total scattering amplitude proportional to:

$$f = \int \rho(\mathbf{r}) e^{i\mathbf{s}\cdot\mathbf{r}}\, dV$$

where $\mathbf{r}$ is a vector whose origin is at the nucleus, $\rho(\mathbf{r})$ is the electron density at the point $\mathbf{r}$, $\mathbf{s} = 2\pi(\mathbf{k} - \mathbf{k}_0)$, and $dV$ is a volume element. The wave vectors $\mathbf{k}$ and $\mathbf{k}_0$ are of the scattered incident X-rays, respectively. The atomic scattering factor $f$ varies with the atom, being larger for atoms with more electrons. The amplitude for scattering from an entire unit cell of a crystal is represented (approximately) as a sum over atomic scattering amplitudes:

$$F = \sum_j f_j e^{i\mathbf{s}\cdot\mathbf{R}_j}$$

where $f_j$ is the scattering amplitude of the $j$th atoms whose mean position is at the point $\mathbf{R}_j$ within the unit cell. There will be constructive interference (that is, Bragg peaks) whenever $\mathbf{s} \cdot (\mathbf{R}_j - \mathbf{R}_k)$ is an integral multiple of $2\pi$. There will be a large number of planes of atoms in the crystal for which this Bragg condition will be satisfied and therefore the crystal structure is in principle overdetermined.

As is shown by the previous equation, the atomic scattering amplitude is roughly proportional to the number of electrons in the atom. Thus, it is particularly difficult to observe scattering interferences from hydrogen atoms. Neutron-scattering amplitudes depend on nuclear cross sections which are of comparable magnitudes for all nuclei. An identical expression for $F$ is obtained except that the $f'$s are now nuclear scattering amplitudes. Because these amplitudes are small, neutron diffraction is carried out with condensed phases and not with gases.

Electron diffraction has also been used for determination of molecular structure. Electrons interact far more strongly with matter than X-rays and can penetrate only a few hundred angstroms into a crystal. They have been used, therefore, for structural measurements of solid surfaces and gaseous molecules. The fact that the gaseous molecules exhibit all orientations to an incoming electron beam greatly reduces the information content of the experiment. In other words, only systems with a small number of structural parameters can be effectively studied. The scattering amplitudes for individual atoms include a nuclear as well as an electronic term, but the essence of the structurally dependent interference remains the same.

### 2. Structure Inferred from Moments of Inertia

At very high resolution, electronic spectra exhibit rotational structure from which the moments of inertia of both the ground and excited electronic states can be extracted. Similarly, vibrational spectra taken at very high resolution will yield the moments of inertia in the ground and the excited vibrational states. Rotational Raman spectra have yielded moments of inertia of small symmetric molecules. The most extensive and accurate source by far of moment of inertia data is microwave spectroscopy. This technique is applicable to any polar gaseous molecule.

Frequencies in microwave spectroscopy that are directly proportional to inverse moments of inertia are usually measured to five or six significant figures. It would thus appear that bond distances are determined to better than 0.001 Å. In reality, there are usually more structural parameters required than measured moments of inertia. For example, in $CH_3I$, a pyramid-shaped molecule, the structure is determined by two bond distances and a bond angle. Because of the threefold symmetry axis, only two moments of inertia are measurable. The usual solution is an isotopic substitution (for example, deuterium for hydrogen). It is often found, however, that a unique structure cannot be found consistent with all the moments of inertia. The reason is that the different vibrational wave functions in the isotopically substituted molecules yield slightly different average structures.

## C. Dynamics of Molecular Processes

A system at equilibrium is characterized by a balance of opposing forces and reactions so that its macroscopic properties are independent of time. Time-dependent processes are caused by deviations from thermal equilibrium. For example, if there is a spatial gradient in temperature, electric potential, momentum, or concentration, equilibrium is restored by the transport of energy, charge, momentum, or identity, respectively. The ratios between the fluxes of these properties and the corresponding gradient are called the transport coefficients—namely, the coefficients of thermal conductivity, electrical conductivity, viscosity, and diffusion, respectively.

Even in a spatially uniform system there can be substantial departures from equilibrium. The system may not be in an equilibrium distribution over its internal states and often approaches equilibrium ("relaxes") exponentially with a rate characteristic of each degree of freedom, which is the reciprocal of the relaxation time. For example, there are electronic relaxation times, vibrational relaxation times, rotational relaxation times, and even nuclear or electronic spin lattice relaxation times. There are great variations in the rate of approach to equilibrium of these various degrees of freedom, varying from $10^{12}$ $\sec^{-1}$ to $10^{-3}$ $\sec^{-1}$.

When the chemical composition of a system changes with time, it is less clear how to express this change mathematically. When dealing with a reaction, for example, of the form:

$$\mathrm{A} + \mathrm{B} \rightleftarrows \mathrm{X}_1 \rightleftarrows \mathrm{X}_2 \cdots \rightleftarrows \mathrm{X}_n \rightarrow \mathrm{C} + \mathrm{D}$$

where A and B are reactants, C and D are products, and the $X_i$ are intermediates, known or unknown, it can often fit the observations to an empirical formula with one or more terms of the following form:

$$d(C)/dt = k(\mathrm{A})^a(\mathrm{B})^b(\mathrm{C})^c(\mathrm{D})^d$$

where $a$, $b$, $c$, and $d$ are not necessarily integer constants and the parentheses indicate concentrations of the enclosed species. The constant $k$ is called a rate constant. The complexity of this equation is due to the complexity of the reaction mechanism. If the reaction were really elementary (that is, if there were no intermediates), then the equation,

$$d(C)/dt = k(\mathrm{A})(\mathrm{B})$$

would express the fact that the reaction takes place as a consequence of collisions of A and B molecules.

Except for radiative and nonradiative decay of electronically excited molecules, the approach to equilibrium of a set of molecules is always accomplished by intermolecular interactions of some sort. In a low-pressure gas, the central event is a bimolecular collision. A collision changes the states of the pair of molecules involved in it. The most gentle change of state is the elastic scattering that involves only a rotation of the relative momentum of the pair. The probability of such an event is expressed by a differential cross section, $d\sigma/d\Omega$, which gives the number of the pairs whose relative momentum is scattered into $d\Omega$ (at a certain orientation, $\Omega$) per unit time, per unit incident flux. The integral of this differential cross section over all solid angles is called the total scattering cross section.

In perhaps the most famous experiment of atomic physics, Rutherford demonstrated that the deflection of $\alpha$ particles could be connected with the Coulomb law of force between nuclei and $\alpha$ particles. If the potential energy was known in advance, classical or quantum equations of motion could be used to determine the differential scattering cross section. In principle, the process can be reversed; that is, the theoretical potential energy acting between the two scatterers can be deduced from the experimental differential cross section. Indeed in principle if the scatterers are small enough, preferably atomic in nature, the potential energy can be calculated using the Schrödinger equation of quantum mechanics. A large amount of research in atomic scattering may be briefly summarized as follows. Atoms interact with a very steeply rising short-range potential and a relatively weak long-range attraction. The long-range attraction gives rise to

very large scattering at small angles superimposed on a much less intense isotropic scattering due to the short-range repulsion.

Elastic collisions can relax translational energies, producing a Boltzmann distribution of translational energies. However, relaxations of internal degrees of freedom can only be accomplished by inelastic collisions in which energy is exchanged between internal degrees of freedom and translation or between internal degrees of freedom of both molecules. The efficiency of these inelastic collisions expressed by the magnitudes of the inelastic cross sections depends on the nature and strength of the interaction between the molecules. Specifically, the intermolecular potential must depend on the molecular orientations in order to cause changes of rotational state and on an internal normal coordinate in order to cause changes in vibrational states. As a general rule, rotational equilibrium is usually achieved with only five to ten collisions, but vibrational equilibrium requires hundreds to thousands of collisions. The reason is simply that intermolecular potential energies are much more dependent on molecular orientation than on the relatively small vibrational amplitudes. A more subtle point is the following. The probability of an inelastic collision transferring an amount of energy $\hbar\omega$ between a more or less random translation and an internal degree of freedom is proportional to the spectral density of the random interaction at the frequency $\omega$. The typical frequency associated with translational motion during a collision would be the reciprocal of a collision duration. This frequency is just right for rotations, too small for vibrations, and too large for nuclear magnetic energy level separations. Because $V \rightarrow T$ (vibration to translation) relaxation is slow, another form of vibrational relaxation occurs earlier which is called $V \rightarrow V'$. These $V \rightarrow V'$ transitions between vibrational states of nearly the same energy, for example,

$$2HBr(v = 1) = HBr(v = 0) + HBr(v = 2)$$

have much larger cross sections than a $V \rightarrow T$ process such as:

$$HBr(v = 1) + M = HBr(v = 0) + M$$

where M is another molecule.

There is a hierarchy of relaxation problems in dilute gases. One begins with macroscopic transport processes whose transport coefficients can be expressed in terms of averages of scattering cross sections. These in turn can, in principle, be calculated from an intermolecular potential energy. Spectroscopically the evolution of an initially nonequilibrium distribution over internal states can be followed toward a Boltzmann distribution. The relaxation rate can be expressed in terms of inelastic scattering cross sections that can, again in principle, be calculated from the intermolecular potential energy function.

The chemical reaction itself can be treated as a collision problem in mechanics. The measured temperature-dependent rate constant for a chemical reaction, $k(T)$, is an average over the Boltzmann distributions for both translation and internal energies. There is not enough information in the function $k(T)$ to permit a potential energy function, a surface to be extracted by an inversion. An area of research, chemical dynamics is the study of reactions in which some form of state selection of the reactants or the products has been carried out. The state selection can take many forms. If the reactant molecules are formed into molecular beams, the direction and sometimes also the magnitude of the relative velocity will be known and the differential reaction cross section can be determined. A reactant can be excited to a specific vibrational state and the effect on the cross section can be measured. Similarly the vibrational and rotational state distribution of product molecules can sometimes be determined by laser-induced fluorescence, multiphoton ionization, or infrared emission. When sufficient information is obtained, it should ultimately be possible to extract a potential surface on which the reactants move and are transformed into products. The assumption here is that the set of reactant molecules and the set of product molecules correspond to different potential minima on the same surface. If this is not the case, the reaction is said to be nonadiabatic. The theoretical problem is more complicated; two potential surfaces must be extracted as well as a perturbing potential that couples them.

## D. Lasers in Chemical Physics

The laser is a light source possessing intensity, directionality, coherence, monochromaticity, and (often) tunability. In the past few decades, lasers have altered the face of most branches of experimental chemical physics. The resulting enhanced knowledge of the physical world has greatly extended the applicability of theory.

Lasers are of two types, cw (steady state) and pulsed. The cw laser can be made extremely narrow in frequency, down to KHz and even less. This narrow line width permits resolution of the rotational structure of all but the heaviest gas-phase molecules and the hyperfine structure of molecules with nonzero spin. The laser line width is typically narrower than the Doppler-broadened spectral line of a gas-phase molecule or atom. This gives rise to a new spectroscopy called Doppler spectroscopy, which measures the distribution of the velocity component in the direction of the probing laser. In solution, the narrow line width permits measuring the width of the Rayleigh scattered light, thus permitting rapid measurement of the molecular diffusion coefficient and molecular mass.

Pulsed lasers have durations in the nanosecond, picosecond, and even femtosecond domains. Strong lamps, both cw and pulsed, have been used in photochemistry for a long time, and their average power can be comparable with that of a laser. However, the short pulse duration of a ns laser enables the generation by photodissociation of a much higher instantaneous concentration of reactive intermediates such as atoms or radicals. These or their subsequent reaction products are in sufficient concentration that their nascent quantum state distribution can be probed. This measurement must be done under single-collision conditions in a time much shorter than the time required for thermalization of the rovibrational distribution. No such measurements are possible with a conventional light source.

The simultaneous or nearly simultaneous absorption of two or more quanta is a phenomenon unique to the laser. When a strong infrared laser is focused on a molecule, sometimes not just two but a large, somewhat indeterminate number of quanta result in a super hot molecule. Focusing a visible or ultraviolet laser on a molecule often results in ionization, nonresonant or resonant. The latter process is called resonance-enhanced multiphoton ionization (REMPI). With its aid, a molecule in a given rovibrational ground state, which is extremely difficult to detect, is converted to an easily detectible ion. A technique of ion imaging has been developed in which a molecular fragment or reaction product is ionized and then projected down a tube by a strong electric field. An array detector generates a pattern of rings. The radius of the ring yields the transverse velocity of the detected molecule, and the variation of intensity around the ring is a measure of the anisotropy of molecules with this particular speed.

Another nonlinear optical phenomenon unique to the laser is second harmonic generation in which strong light at frequency $\omega$ is partially converted to light at frequency $2\omega$. This is observed primarily in media lacking a center of symmetry and is especially suited to probing the interface between two media. A variation of the technique, sum frequency generation, permits measurement of the infrared absorption of the interface layer.

Picosecond lasers make possible the measurement of extremely fast processes, such as the lifetime of an electron in a conduction band in a semiconductor. The use of femtosecond (fs) pulses allows the observation in real time of even faster processes, such as the separation of fragments following an initial excitation. The separation time is typically $<100$ fs for a direct allowed transition, but it can be longer if the fragments are both heavy or if the dissociation is indirect (i.e., from a lower electronic state than the one originally excited). Consistent with the uncertainty principle, a fs pulse is broad in frequency. For example, a 50-fs pulse is about 1000 $cm^{-1}$ wide. Recently, it has been shown that by varying the relative phases of the different modes of which the fs pulse is composed a multiphoton dissociation of a metal carbonyl can be tailored to alter the relative yields of different photodissociation channels.

We finish with two examples of accepted verities which the laser has demolished. The statement "A light beam cannot resolve objects whose separation is less than its wavelength," is true in the farfield region where $r \gg \lambda$ but is not true in the opposite nearfield region where $r \ll \lambda$. It is now possible to excite molecules by laser light and then observe emission from individual molecules. Instead of observing an average excited state lifetime, the lifetimes of individual molecules are measurable. The action of one enzyme molecule can be measured instead of the average behavior of an ensemble of enzyme molecules.

Another truism is that "proteins cannot fly." In other words, large biomolecules cannot be studied by mass spectroscopy because they have no vapor pressure and if heated will be destroyed. Embedding the protein or polynucleotide between thin crystal layers and then pulse evaporating the crystallite with a ns laser produces gas-phase macromolecules. These can be ionized and their masses accurately determined. An important new window on biochemistry has been opened.

## II. COLLECTIVE PROPERTIES

Liquids and solids are systems that contain on the order of Avagodro's number ($6 \times 10^{23}$) of strongly interacting molecules. How do the macroscopic properties of such systems depend on the properties of the constituent isolated molecules? The properties of collective systems can be roughly divided into structural properties and dynamic properties. The theoretical study of such collective properties falls within the province of statistical mechanics.

The macroscopic properties of condensed systems are of central importance to almost all fields of science. These properties are described by phenomenological theories such as thermodynamics, hydrodynamics, electrodynamics, and chemical kinetics. Nevertheless, it should be clear that the laws of thermodynamics, the thermodynamic equation of state for given materials, the laws of hydrodynamics and electrodynamics, and the properties germane to these as well as all of chemical kinetics should be derivable from the underlying classical or quantum mechanical theory of atoms and molecules (in fact, the whole of chemistry should be predictable from mechanics.)

### A. Statistical Mechanics

Statistical mechanics is a molecular theory of macroscopic systems. It provides the bridge between the microscopic

world of nuclei and electrons and the macroscopic world of the phenomenological theories. Starting with the Hamiltonian of the system, the laws of equilibrium thermodynamics, hydrodynamics, electrodynamics, and chemical kinetics can be derived. The Hamiltonian H is the classical energy as a function of the momenta and positions of all the particles in the system. It consists of the sum of the kinetic and the potential energies.

Statistical mechanics can be subdivided into equilibrium and nonequilibrium statistical mechanics. The former deals with systems in thermodynamic equilibrium, whereas the latter deals with the time evolution of macroscopic systems.

### 1. Equilibrium Statistical Mechanics

The thermodynamic equation of state of a system can be derived from an expression that relates the Helmholtz free energy $A_N(T, V)$ to the microscopic properties of the system,

$$A_N(T, V) = -kT \ln Q_N(T, V)$$

where $T$ is the thermodynamic temperature (degrees Kelvin), $V$ is the volume, and

$$Q_N(T, V) = \sum_n \exp(-E_n/kT)$$

is the canonical partition function. Thus, to derive the thermodynamic equation of state (that is, the relationship between the free energy, the number of particles, the temperature $T$, and the volume $V$), $Q_N(T, V)$ must be calculated. In these equations $k$ is Boltzmann's constant, the sum goes over all quantum mechanical states of the system, and $\{E_n\}$ is the set of energy eigenvalues corresponding to these states. These energy levels are found by solving the Schrödinger equation of quantum mechanics, namely,

$$H\Phi_n(q_1, \ldots, q_N) = E_n\Phi_n(q_1, \ldots, q_N)$$

where $\Phi_n(q_1, \ldots, q_N)$ is an energy eigenfunction of the system and $\{q_1, \ldots, q_N\}$ are the coordinates specifying the positions of the particles. Thus, to predict the thermodynamic properties the allowed energies for the system at the given volume have to be determined. In very dilute gases it is possible to calculate the thermodynamic properties very accurately. This is already an impressive achievement, but liquids are more difficult to treat. In quantum mechanics, every particle has wavelike properties. It follows that the atoms and molecules that liquids are composed of occupy a region in space characterized by a diameter proportional to the particle's thermal deBroglie wavelength, a quantity inversely proportional to $(mT)^{1/2}$. Only at very low temperatures does the matter wave of one molecule interfere with the matter wave of another molecule. This interference can give rise to astounding properties as it does in liquid He which displays superfluidity (a purely quantum mechanical phenomenon). Because most liquids exist at high temperatures, it is possible to ignore these quantum effects and to treat them using classical statistical mechanics—that is, statistical mechanics based on classical mechanics. Although the treatment of strongly interacting systems is still very difficult in classical statistical mechanics, progress has been made on a variety of fronts. Statistical mechanical perturbation theory allows the properties of fluids to be calculated using the exact results for a simple reference system such as the hard sphere fluid. This is a fluid consisting of spheres that are not allowed to interpenetrate. The structure and thermodynamic properties of fluids can be determined by integral equations. The behavior of systems undergoing phase transformations can be determined by renormalization group techniques. The properties of real imperfect gases can be determined by diagrammatic techniques.

What is meant by the term "the structure of liquids"? Imagine an instantaneous photograph of a liquid. The atoms are packed together in a noncrystalline arrangement; that is, there is no long-range order. Nevertheless, the positions of nearby atoms are correlated. Because the atoms cannot overlap, the minimum distance from one atomic center to another is the atomic diameter. Thus, around an atom there will be an exclusion sphere outside of which near neighbors are found, but these neighbors once again define a region of space which excludes next nearest neighbors and so on. The radial distribution function gives the average density of atoms as a function of distance from a given atom in the system. Due to the previously stated packing effects, this function will exhibit peaks and troughs as a function of distance. At very large distances it will be equal to the bulk density because the presence of an atom at one place will not be felt by another atom very far from it. These functions can be measured by diffraction methods such as neutron scattering and X-ray scattering. They can also be determined from statistical mechanics.

### 2. Nonequilibrium Statistical Mechanics

In hydrodynamics, the time evolution of fluid flow depends not only on thermodynamic properties, but also on such transport properties as shear and bulk viscosities, thermal conductivity, mutual diffusion coefficients, etc. These characterise the transport of momentum, heat, and mass. In electrodynamics, the response of systems to the imposition of electric and magnetic fields depends on dielectric

response functions of the system, and in chemical kinetics the approach to equilibrium is described by empirical rate laws and chemical reaction rate constants. Expressions for the transport coefficients in terms of the microscopic dynamics of the quantum or classical systems can be derived by modern statistical mechanics. Moreover, it is now possible to derive the equations of macroscopic rate processes such as hydrodynamics and chemical kinetics starting with the Hamiltonian of the system and using the methods of nonequilibrium statistical mechanics.

Transport coefficients can be expressed in terms of time correlation functions or covariance functions of spontaneous fluctuations in equilibrium systems. For example, the self diffusion coefficient can be expressed as:

$$D = \frac{1}{3}\int_0^x dt\,\langle \mathbf{v}(0)\cdot\mathbf{v}(t)\rangle$$

where $\langle \mathbf{v}(0)\cdot\mathbf{v}(t)\rangle$ is the autocorrelation function of the velocity of a labeled particle, $\mathbf{v}$. Such expressions for transport coefficients are called Green–Kubo relations. All transport coefficients and chemical rate constants can be expressed as time integrals of time correlation functions; that is, as Green–Kubo relations. In addition, spectroscopic band shapes, NMR, and ESR line shapes and differential cross sections for the scattering of light and thermal neutrons can be expressed as space–time Fourier transforms of appropriate time correlation functions. The theory of such processes is called linear response theory. To treat nonlinear processes it is necessary to use mode–mode coupling theory.

## B. Numerical Statistical Mechanics

One of the great developments in statistical mechanics during the past several decades is the development of methods for simulating strongly interacting systems such as liquids, liquid crystals, solids, and glasses on computers. The formalism of statistical mechanics provides exact analytical expressions. Because it is often impossible to evaluate these expressions analytically, numerical methods must be used. In fact, prior to the development of computers very little progress was made in understanding liquids and amorphous solids. There are two major techniques used in computer simulation—molecular dynamics (MD) and Monte Carlo (MC)—and these have contributed enormously to the theory of condensed systems.

### 1. Molecular Dynamics

A starting point for the simulation of a classical liquid is the Hamiltonian of the system. If there are $N$ particles in the system, $6N$ equations of motion must be solved: one for each position coordinate and one for each momentum coordinate. Starting with a given set of positions and momenta, these equations are solved by finite difference techniques on a computer. At the end of each time interval the new positions and momenta of all the particles are recorded. After very many time steps this gives a trajectory of the whole many-body system. Thermodynamic properties are found by time averaging dynamical properties over the trajectory. Likewise, time correlation functions and thereby transport coefficients, rate constants, differential scattering cross sections, spectral line shapes, etc. can be determined from the trajectory. This method is called molecular dynamics. Recently, it has been possible to simulate nonequilibrium systems by clever applications of constraints and boundary conditions. Because simulations are done on finite systems, periodic boundary conditions are usually adopted to bypass unusual surface effects. Nevertheless, it is also possible to simulate surface properties. In order to use molecular dynamics, there must be knowledge of the forces that exist between the molecules. Only in the simplest systems is it possible to determine the forces between two molecules (either by experiment or by *ab initio* quantum chemical calculations). Thus, molecular modeling is required before molecular dynamics can be applied.

### 2. Monte Carlo Simulations

If there is interest only in equilibrium properties, an alternative to molecular dynamics is the Monte Carlo technique. This ingenious method is based on the theory of finite Markov chains. In classical statistical mechanics, the probability distribution for finding the particles at a given position in configuration space is proportional to the Boltzmann factor, $\exp(-U/kT)$, where $U$ is the potential energy of the system in that configuration. In Monte Carlo, an initial configuration is chosen. A particle is next moved to a new position. This move is either accepted or rejected based on a certain criterion. This process is repeated for each particle. The criterion for acceptance or rejection is based on a random number generator in such a way that the configurations thus generated are distributed according to the Boltzmann factor. These sampled configurations give a trajectory in configuration space that looks like a random walk. Averages of position-dependent properties over this random walk trajectory then give thermodynamic properties.

Simulation techniques have been applied to a wide variety of many-body systems. For example, hydrogen bonding in water and aqueous solutions, including protein solutions, has been studied in this way. Such methods are often used as experimental tests of theoretical assumptions. In many cases, totally new phenomena were first observed using computer simulations and only later observed in real

systems. The methods of statistical mechanics and numerical statistical mechanics are applicable to many fields outside of chemistry. For example, the question of quark confinement in high-energy physics can be formulated as a problem in statistical mechanics that can be treated by Monte Carlo techniques using the Feynman path integral representation. Another example involves the fragmentation of nuclei that can be regarded as a liquid droplet of nucleons.

With the advent of supercomputers these methods will continue to play a very large role in the study of collective phenomena. One day very complex reactions will be simulated on computers.

## SEE ALSO THE FOLLOWING ARTICLES

ATOMIC PHYSICS • ATOMIC SPECTROMETRY • CHEMICAL KINETICS, EXPERIMENTATION • KINETICS (CHEMISTRY) • LASERS • NUCLEAR MAGNETIC RESONANCE • PHYSICAL CHEMISTRY • POTENTIAL ENERGY SURFACES • STATISTICAL MECHANICS • SUPERCOMPUTERS • TRANSITION PROBABILITIES AND ATOMIC LIFETIMES

## BIBLIOGRAPHY

Berne, B. J., and Pecora, R. (1976). "Dynamic Light Scattering: With Applications to Chemistry, Biology and Physics," Wiley (Interscience), New York.

Berne, B. J., ed. (1976). "Modern Theoretical Chemistry," Vol. 6, Parts A and B, Plenum, New York.

Bernstein, R. B., and Levine, R. D. (1987). "Molecular Reaction Dynamics and Chemical Reactions," Oxford University Press, London.

Chandler, D. (1987). "Introduction to Modern Statistical Mechanics," Oxford University Press, New York.

Forster, D. (1975). "Hydrodynamic Fluctuation, Broken Symmetry and Correlation Functions," W. A. Benjamin, Reading, MA.

McQuarrie, D. A. (1975). "Statistical Mechanics," Harper & Row, New York.



# Cohesion Parameters

**Allan F. M. Barton**
*Murdoch University*

I. Main Classes of Cohesion Parameters
II. Limitations of Cohesion Parameters
III. Cohesion Parameters and Other Solvent Scales
IV. Applications of Cohesion Parameters
V. Evaluation of Cohesion Parameters
VI. Selected Values
VII. Current Status

## GLOSSARY

**Chameleonic behavior** Capacity of the molecules of a compound to assume a character similar to their environment by intermolecular or intramolecular association.

**Cohesion parameter** Quantity with dimensions of square root of pressure used to characterize the cohesive and adhesive properties of materials.

**Cohesive energy** Thermodynamic quantity describing the sum of molecular effects in a material that cause it to remain in a condensed state.

**Cohesive pressure (cohesive energy density)** Ratio of the cohesive energy to the volume for a material.

**Dispersion forces** Intermolecular forces present in all materials, whether polar or not, arising from the fluctuating molecular dipoles that result from the positive nucleus and negative electrons.

**Hansen parameters** Cohesion parameters resulting from the subdivision of the Hildebrand parameter into dispersion (nonpolar), polar, and hydrogen bonding contributions.

**Hildebrand parameter, solubility parameter** Square root of the cohesive pressure of a liquid.

**Homomorph** The homomorph of a polar molecule is a nonpolar molecule having the same size and shape.

**Interaction cohesion parameters** Cohesion parameters resulting from the subdivision of the Hildebrand parameter into dispersion (nonpolar), orientation, induction, acid, and base contributions.

**Regular solution** A solution (solvent plus solute) that has a completely random molecular distribution despite possible specific interactions between solute and solvent molecules.

**THE TERM COHESION PARAMETER** came into use in the early 1980s as a general term for a class of quantities with dimensions of $(\text{pressure})^{1/2}$. In the simplest application of the concept, the cohesion parameter values of each of two pure materials are multiplied together to yield a numerical value for the cohesive pressure expected when the materials are mixed or the adhesive pressure expected when they are in contact.

Cohesion parameters provide a simple method of correlating and predicting the cohesive and adhesive properties of substances from a knowledge of the properties of the individual components only. There are, of course, numerous more sophisticated theories and techniques for this purpose, but none is as easy to use in practical applications.

Cohesion parameters provide estimates of the enthalpy changes on mixing and do not explicitly treat any deviations from ideal entropy changes. They are based on J. H. Hildebrand's definition of a special class of "regular" solutions, those involving an ideal entropy change on mixing despite the existence of enthalpy changes, but are now applied far more widely.

The simplest cohesion parameter is the Hildebrand parameter, commonly called the *solubility parameter*, values of which are now being included by manufacturers of polymers and solvents in the specification sheets of their products.

## I. MAIN CLASSES OF COHESION PARAMETERS

Pure materials exist as condensed phases (e.g., liquids, crystals, glasses, rubbers) over certain ranges of temperature and pressure because in some circumstances these states are more stable than the gaseous state: There are energetic advantages in the molecules being packed closely together. In these condensed phases, strong attractive forces exist between the molecules. The *cohesive energy* $-U$ is positive (equal and opposite in sign to the internal energy), and the cohesive energy per unit volume $V$ of the material is defined as the *cohesive pressure* or cohesive energy density, given the symbol $c$.

Similarly, many mixtures also exist as homogeneous, condensed phases for the same reason: There is a significant cohesive pressure maintaining that state in existence. The basis of the cohesion parameter approach to mixtures is that a material with a high cohesive pressure requires more energy for dispersal than is gained by mixing it with a material of low cohesive pressure, so immiscibility (separation of phases) results, but two materials with similar cohesive pressure values gain cohesive energy on dispersal, so mixing occurs.

The change in cohesive pressure associated with the process of mixing two components $i$ and $j$ with their respective cohesive pressures ${}^{i}c$ and ${}^{j}c$ is given by the *interchange cohesive pressure*, ${}^{ij}A$,

$$ {}^{ij}A = {}^{i}c + {}^{j}c - 2^{ij}c, \tag{1} $$

where ${}^{ij}c$ is the cohesive pressure characteristic of the intermolecular forces acting between molecules of type $i$ and type $j$. This equation can be understood in a simple way by considering what happens when unit volumes of components $i$ and $j$ are mixed: Two $i$–$j$ interactions are formed for each pair of $i$–$i$ and $j$–$j$ interactions broken.

### A. Hildebrand Parameter

At pressures below atmospheric pressure (i.e., for temperatures below the boiling point), the molar cohesive energy $-U$ of a liquid can be taken as equal to the molar enthalpy of vaporization $\Delta H$ less the pressure–volume work (which for a vapor with ideal gas behavior is $RT$ per mole, where $R$ is the molar gas constant and $T$ the absolute temperature). This is the basis of the original definition by Joel H. Hildebrand and Robert L. Scott of the solubility parameter, hereafter called the *Hildebrand parameter* $\delta$:

$$ \delta = c^{1/2} = (-U/V)^{1/2} = [(\Delta H - RT)/V]^{1/2}. \tag{2} $$

This parameter was intended for use only with nonpolar, nonassociating liquid systems that form regular solutions, but its use has been extended to all types of material. For *all* liquids, the value of the Hildebrand parameter may be determined directly from thermodynamic data by means of this equation, but it is only for regular solutions that predictions using the Hildebrand parameter are reliable.

A regular solution has an ideal entropy of formation, that is, a random molecular distribution, despite the existence of interactions that lead to a nonideal enthalpy of formation (heat of mixing). This means that regular mixtures are restricted to those systems in which only dispersion forces are acting. (Dispersion forces, or London forces, arise from the fluctuating dipoles that result from the positive nucleus and negative electron "cloud" in every atom. They occur in all systems and are distinct from forces associated with molecular polarity and "chemical" interactions between molecules.) For systems like this, without the orientation and ordering effects of polar molecules, the cohesive pressure between unlike molecules is given to a good approximation by the geometric mean of the cohesive pressures of the individual components,

$$ {}^{ij}c = ({}^{i}c\,{}^{j}c)^{1/2}. \tag{3} $$

From a combination of Eqs. (1), (2), and (3), it follows that the interchange cohesive pressure is given by Eq. (4):

$$ {}^{ij}A = ({}^{i}c^{1/2} - {}^{j}c^{1/2})^2 = ({}^{i}\delta - {}^{j}\delta)^2. \tag{4} $$

By means of this fundamental equation, it is possible to describe the thermodynamics of the mixing process in terms of Hildebrand parameters. For example, the mole fraction activity coefficient $f_x$ at infinite dilution of component $j$ with molar volume ${}^{j}V$ in component $i$ is given by

$$RT \ln {}^{j}f_{x}^{\infty} = {}^{j}V^{ij}A = {}^{j}V({}^{i}\delta - {}^{j}\delta)^2. \quad (5)$$

Equations such as this hold exactly for regular solutions and are good approximations for many other useful systems.

On the basis of the assumtions made in the derivation of cohesion parameter expressions, the effective Hildebrand parameter $\bar{\delta}$ of a binary solvent mixture is volumewise proportional to the Hildebrand parameter values of its components, so

$$\bar{\delta} = {}^{i}\phi^{i}\delta + {}^{j}\phi^{j}\delta, \quad (6)$$

where $\phi$ is the volume fraction, defined by

$${}^{i}\phi = \frac{{}^{i}V{}^{i}x}{{}^{j}V{}^{i}x + {}^{i}V{}^{i}x}, \quad (7)$$

where $x$ is the mole fraction and $V$ is the molar volume.

## B. Component Cohesion Parameters

In addition to the dispersion forces described above for regular solutions, most chemical systems also exhibit polar interactions and specific interactions (e.g., hydrogen bonding). To be generally useful, theories and models aiming to systematize and predict the behavior of matter must deal with molecular interactions on the basis of their natures or origins as well as their strengths. The cohesive properties characteristic of the condensed states of matter are produced by various intermolecular forces, and the cohesive pressures ${}^{i}c$, ${}^{j}c$, and ${}^{ij}c$ represent the resultant effect of all these forces acting between molecules of types $i$ and $j$.

For this reason, five interaction cohesion parameters have been introduced to describe the properties of materials in greater detail than is possible with the Hildebrand parameter. These are dispersion, orientation, induction, Lewis acid, and Lewis base cohesion parameters.

Dispersion forces, occurring in all molecules, whether polar or not, give rise to a dispersion cohesive pressure ${}^{i}c_{d}$ and a corresponding dispersion cohesion parameter ${}^{i}\delta_{d}$ in a pure material $i$:

$${}^{i}c_{d} = {}^{i}\delta_{d}^{2}. \quad (8)$$

The nonpolar, dispersive interactions between unlike molecules of type $i$ and type $j$ provide a contribution to the cohesive pressure that is based on the geometric mean of the individual values:

$${}^{ij}c_{d} = \left({}^{i}c_{d}{}^{j}c_{d}\right)^{1/2} = {}^{i}\delta_{d}{}^{j}\delta_{d}. \quad (9)$$

A simple interpretation of this geometric-mean behavior is that the interaction is of a "symmetrical" nature: Each member of a pair of molecules interacts by virtue of the same molecular property (the polarizability). It follows that

$${}^{ij}A_{d} = {}^{i}\delta_{d}^{2} + {}^{j}\delta_{d}^{2} - 2{}^{i}\delta_{d}{}^{j}\delta_{d} = \left({}^{i}\delta_{d} - {}^{j}\delta_{d}\right)^2. \quad (10)$$

Orientation effects result from dipole–dipole (or Keesom) interactions and occur between molecules that have permanent dipole moments. The orientation cohesive pressure of a pure material $i$ is denoted ${}^{i}c_{o}$, and the corresponding orientation cohesion parameter ${}^{i}\delta_{o}$ is defined by

$${}^{i}c_{o} = {}^{i}\delta_{o}^{2}. \quad (11)$$

Like dispersion forces, these are symmetrical interactions, depending on the same property of each molecule, which in this case is the dipole moment. It follows that the geometric-mean rule is obeyed closely for orientation interactions between unlike molecules. For "ideal" polar molecules, which may be represented by spherical force fields with small ideal dipoles at their centers, this contribution to the cohesive pressure in mixtures of $i$ and $j$ molecules is

$${}^{ij}c_{o} = \left({}^{j}c_{o}{}^{j}c_{o}\right)^{1/2} = {}^{i}\delta_{o}{}^{j}\delta_{o} \quad (12)$$

and the interchange cohesive pressure due to orientation is

$${}^{ij}A_{o} = \left({}^{i}\delta_{o} - {}^{j}\delta_{o}\right)^2. \quad (13)$$

Dipole induction effects arise from dipole-induced forces (Debye interactions) occurring between molecules with permanent dipole moments and any other neighboring molecules, whether polar or not, and resulting in an induced nonuniform charge distribution. In contrast to dispersion and orientation interaction, dipole induction interactions are "unsymmetrical," involving the dipole moment of one molecule and the polarizability of the other. Thus, the cohesive pressure term for induction in a pure material $i$ involves the product ${}^{i}\delta_{i}{}^{j}\delta_{d}$, where ${}^{i}\delta_{i}$ is the induction cohesion parameter, and in a mixture of $i$ and $j$,

$${}^{ij}c_{i} = {}^{i}\delta_{i}{}^{j}\delta_{d} + {}^{j}\delta_{i}{}^{i}\delta_{d}. \quad (14)$$

It can be shown, therefore, that the interchange cohesive pressure due to induction is

$${}^{ij}A_{i} = 2{}^{i}\delta_{i}{}^{i}\delta_{d} + 2{}^{j}\delta_{i}{}^{j}\delta_{d} - 2{}^{j}\delta_{i}{}^{i}\delta_{d} - 2{}^{i}\delta_{i}{}^{j}\delta_{d} \quad (15)$$

$${}^{ij}A_{i} = 2\left({}^{i}\delta_{d} - {}^{j}\delta_{d}\right)\left({}^{i}\delta_{i} - {}^{j}\delta_{i}\right). \quad (16)$$

Lewis acid–base or electron donor–acceptor interactions can be denoted

$$\underset{\substack{\text{Lewis acid}\\(\text{electron pair}\\\text{acceptor})}}{\text{A}} + \underset{\substack{\text{Lewis base}\\(\text{electron pair}\\\text{donor})}}{:\text{D}} \rightleftarrows \overset{\delta^- \quad \delta^+}{\text{A}\cdots\text{D}} \quad (17)$$

The Lewis acid–base complex is formed by an overlap between a filled electron orbital of the donor with a vacant orbital in the acceptor and differs from a "normal" chemical bond in that only one molecule supplies the pair of electrons. These interactions are unsymmetrical, involving donor and acceptor with different roles, so it is necessary to use two separate parameters to characterize these interactions, a Lewis acid cohesion parameter $\delta_a$ and a Lewis base cohesion parameter $\delta_b$. The acid–base interchange cohesion pressure is

$$ {}^{ij}A_{ab} = 2\left({}^{i}\delta_a - {}^{j}\delta_a\right)\left({}^{i}\delta_b - {}^{j}\delta_b\right). \tag{18} $$

Hydrogen bonding interactions are a special type of Lewis acid–base reactions with the electron acceptor being a Brönsted acid (proton acid):

$$ \underset{\substack{\text{Electron pair}\\ \text{acceptor;}\\ \text{proton donor}}}{\text{—X—H}} + \underset{\substack{\text{Electron pair}\\ \text{donor; proton}\\ \text{acceptor}}}{\text{: Y—}} \rightleftarrows \text{—X—H} \cdots \text{Y—} \tag{19} $$

One of the assumptions central to the cohesion parameter approach to the properties of materials is that the various contributions to the cohesive pressure of a substance (either pure or mixed) are additive, so the interchange cohesive pressure for a mixing process is

$$ {}^{ij}A = {}^{ij}A_d + {}^{ij}A_o + {}^{ij}A_i + {}^{ij}A_{ab}. \tag{20} $$

For a pure substance, the total cohesion parameter is

$$ {}^{i}\delta^2 = {}^{i}\delta_d^2 + {}^{i}\delta_o^2 + 2{}^{i}\delta_i{}^{i}\delta_d + 2{}^{i}\delta_a{}^{i}\delta_b. \tag{21} $$

It is clear that this total cohesion parameter is identical to the Hildebrand parameter, which can be determined from Eq. (2).

## C. Hansen Parameters

A much simpler three-component cohesion parameter method was developed by C. M. Hansen on an empirical basis, the parameters being determined either experimentally or by semiempirical equations. This assumes that the cohesion pressure is made up of a linear combination of contributions from nonpolar or dispersion interactions ($\delta_d^2$), polar interactions ($\delta_p^2$), and hydrogen bonding or similar specific association interactions ($\delta_h^2$). The Hildebrand parameter (which can be determined from thermodynamic properties) for any material $i$ is related to the Hansen parameters by

$$ {}^{i}\delta^2 = {}^{i}\delta_d^2 + {}^{i}\delta_p^2 + {}^{i}\delta_h^2. \tag{22} $$

The interchange cohesive pressure associated with the mixing of $i$ and $j$ is

$$ {}^{ij}A = \left({}^{i}\delta_d - {}^{j}\delta_d\right)^2 + \left({}^{i}\delta_p - {}^{j}\delta_p\right)^2 + \left({}^{i}\delta_h - {}^{j}\delta_h\right)^2. \tag{23} $$

It is important to note that the Hansen parameter method ignores the unsymmetrical nature of the induction and acid–base components of the cohesive pressure. In particular, with hydrogen bonding there is no means of separating the proton donor and proton acceptor capabilities of any material. (Of course, this is true also of the Hildebrand parameter, but Hansen parameters *appear* to take hydrogen bonding into account, in a manner that is often misleading; see Section II.C.)

Despite these theoretical shortcomings, Hansen parameters have been fairly widely used in the polymer and coatings industries. Frequently, the three-component parameters have been plotted on three mutually perpendicular coordinates. A solubility "volume" in Hansen space is then drawn up for each solute and compared with the point locations in this space of each solvent. Equation (23) has been modified by doubling the scale on the dispersion axes with the aim of providing "spheres" of solubility for each solute. The distance of the solvent coordinates $({}^{i}\delta_d, {}^{i}\delta_p, {}^{i}\delta_h)$ from the center point $({}^{j}\delta_d, {}^{j}\delta_p, {}^{j}\delta_h)$ of the solute sphere of solubilities then is

$$ {}^{ij}R = \left[4\left({}^{i}\delta_d - {}^{j}\delta_d\right)^2 + \left({}^{i}\delta_p - {}^{j}\delta_p\right)^2 + \left({}^{i}\delta_h - {}^{j}\delta_h\right)^2\right]^{1/2} \tag{24} $$

or

$$ {}^{ij}A = \left({}^{i}\delta_d - {}^{j}\delta_d\right)^2 + 0.25\left[\left({}^{i}\delta_p - {}^{j}\delta_p\right)^2 + \left({}^{i}\delta_h - {}^{j}\delta_h\right)^2\right]. \tag{25} $$

This distance ${}^{ij}R$ can be compared with the radius ${}^{i}R$ of the solute solubility sphere, and if ${}^{ij}R < {}^{j}R$ the likelihood of the solvent $i$ dissolving the solute $j$ is high. The incorporation of the numerical factor 4 in Eq. (24) does not appear to be necessary to provide a spherical interaction volume, and an equation based on Eq. (23) is just as satisfactory:

$$ {}^{ij}r = {}^{ij}A^{1/2} = \left[\left({}^{i}\delta_d - {}^{j}\delta_d\right)^2 + \left({}^{i}\delta_p - {}^{j}\delta_p\right)^2 + \left({}^{i}\delta_h - {}^{j}\delta_h\right)^2\right]^{1/2}. \tag{26} $$

In some applications, only two of the three Hansen parameters are used, so that the location of solvents can be displayed on two-dimensional maps (e.g., Fig. 1) and compared with solute solubility regions.

## D. Other Cohesion Parameters

There are several variations on Hildebrand parameters, Hansen parameters, and interaction cohesion parameters, introduced in attempts to achieve a compromise between simplicity of operation and validity of prediction.

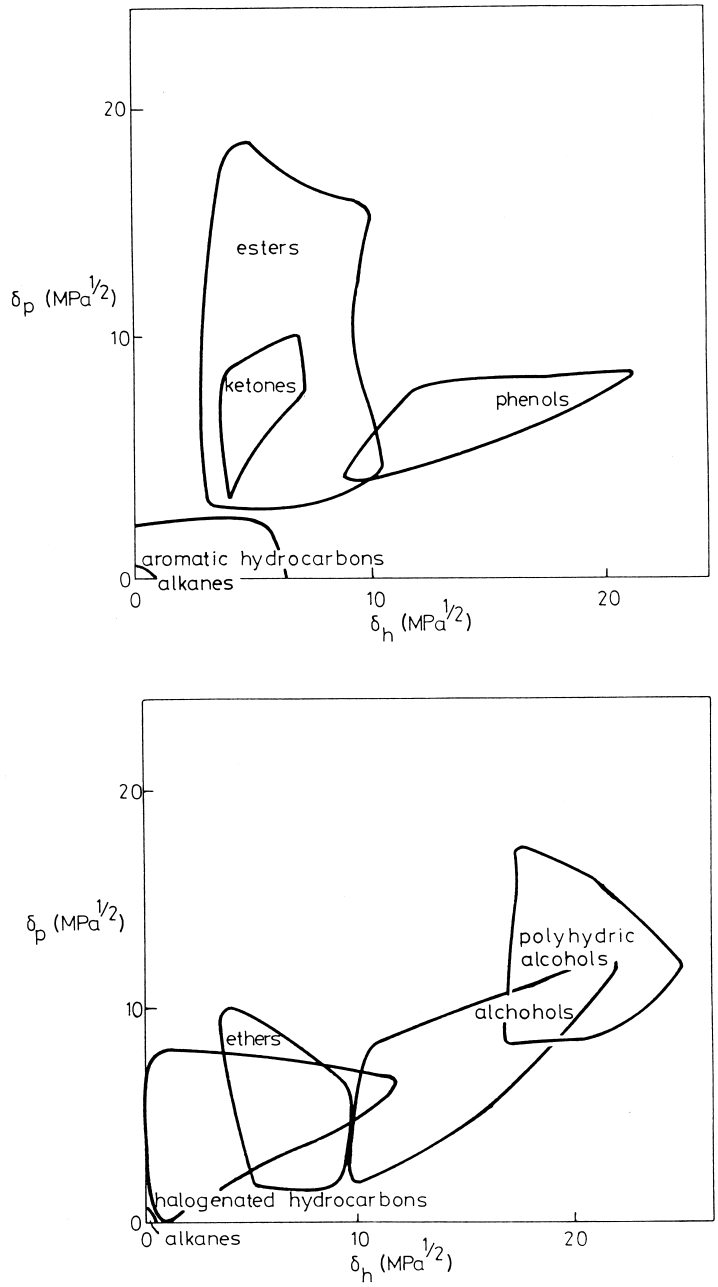



**FIGURE 1** Hansen parameter $\delta_p$–$\delta_h$ locations for various classes of organic compound. [Adapted from Klein, E., Eichelberger, J., Eyer, C., and Smith, J. (1975). *Water Res.* **9,** 807.]

Hansen parameters must be represented in three dimensions, but it is possible to use fractional cohesive pressures plotted on a triangular chart,

$$c_d = \delta_d^2/\delta^2; \quad c_p = \delta_p^2/\delta^2; \quad c_h = \delta_h^2/\delta^2 \tag{27}$$

where $\delta^2$ is given by Eq. (22). It is also possible to use triangular representation with fractional cohesion parameters,

$$f_d = \frac{\delta_d}{\delta_d + \delta_p + \delta_h};$$
$$f_p = \frac{\delta_p}{\delta_d + \delta_p + \delta_h}; \tag{28}$$
$$f_h = \frac{\delta_h}{\delta_d + \delta_p + \delta_h}$$

Triangular representations make the excessively simplifying assumption that the total or Hildebrand parameter

$\delta$ is constant for all materials and that it is the relative magnitude of the three contributions (dispersion forces, polar interactions, hydrogen bonding) that determines the extent of miscibility.

The Hildebrand parameter has been subdivided in other ways. One approach was to divide it into two main components, defining a nonpolar cohesion parameter $\delta_d$ and a polar parameter $\delta_o$. These are related by

$$\delta^2 = \delta_d^2 + \delta_o^2. \tag{29}$$

This approach neglects both induction interactions and, more important, specific interactions. The induction "correction" can be taken into account to some extent by means of the factor ${}^{ij}b$, as shown in the following expression for the interchange cohesive pressure:

$${}^{ij}A = \left({}^{i}\delta_d - {}^{j}\delta_d\right)^2 + {}^{ij}b\left({}^{i}\delta_o - {}^{i}\delta_o\right)^2. \tag{30}$$

This is equivalent to Eq. (25) with the $\delta_h$ term omitted.

## II. LIMITATIONS OF COHESION PARAMETERS

Cohesion parameters provide one of the simplest methods of correlating and predicting the cohesive and adhesive properties of interacting materials from a knowledge of the properties of the individual components. It is therefore to be expected that there will be severe limitations on their use. What is surprising is that this simple correlation method works as well as it does in practice. It is really a semiquantitative version of the statement "like dissolves like." In fact, of course, whether any particular correlation or prediction method is seen to "work" depends on the precision that is expected in the application being considered.

### A. Burrell Hydrogen Bonding Classes

In the use of Hildebrand parameters, the existence of hydrogen bonding is the most obvious cause of discrepancies from "regular" behavior. H. Burrell was one of the first to attempt to deal with the hydrogen bonding factor in the application of Hildebrand parameters to practical systems by the simple expedient of dividing solvents into three classes according to their hydrogen bonding capacities, on the assumption that complete miscibility can occur only if the degree of hydrogen bonding is comparable in the components:

1. Liquids with weak or poor hydrogen bonding capacity, including hydrocarbons, chlorinated hydrocarbons, and nitrohydrocarbons
2. Liquids with moderate hydrogen bonding capacity, including ketones, esters, ethers, and glycol monoethers
3. Liquids with strong hydrogen bonding capacity, such as alcohols, amines, acids, amides, and aldehydes.

This classification is still widely used in practical applications, and a few typical examples are presented in Tables I and II (Section VI).

### B. Geometric-Mean Corrections

One of the specific assumptions in the development of cohesion parameter expressions is the geometric-mean approximation [Eq. (3)]. There are several equations that permit some correction to be made for derivations from this behavior, for example,

$${}^{ij}c = ({}^{i}c\,{}^{j}c)^{1/2}(1 - {}^{ij}l), \tag{31}$$

where ${}^{ij}l$ is a dimensionless constant of the order of 0.01 to 0.1, characteristic of a given pair of materials. When this is incorporated, the empirical expression for interchange cohesive pressure corresponding to Eq. (4) becomes

$${}^{ij}A = ({}^{i}\delta - {}^{j}\delta)^2 + 2^{ij}l\,{}^{i}\delta\,{}^{j}\delta. \tag{32}$$

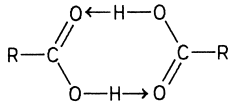
The value of the correction term can be estimated from either liquid- or gas-phase data.

### C. Chameleonic Behavior

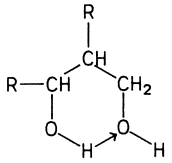
The most important situation requiring caution in the use of cohesion parameters, particularly Hildebrand parameters and Hansen parameters, is that in which electron donor–acceptor interactions within components are very different from those between components. Common examples are systems involving hydrogen bonding: alcohols (particularly methanol), carboxylic acids, water, primary and secondary amines, and glycol ethers. Most commonly, it is hydrogen bonding within the pure component that causes a user the most problems, as one tends to overlook this, whereas one is more on the alert for new interactions when components are mixed.

Although this has long been understood in a general way, it was put most clearly by K. L. Hoy, who proposed the term *chameleonic* for those compounds that have the capacity to assume a character similar to their environment.

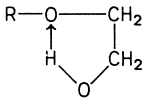
By dimerization or intramolecular association, what would otherwise be a polar material can behave in a nonpolar manner, thus minimizing the energy. Examples are carboxylic acids (structure I), glycol ethers (II and III), and diols (IV). It is clear that the cohesion parameters



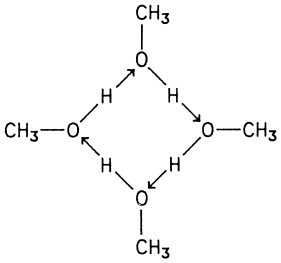
I.



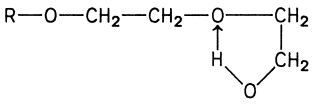
IV.



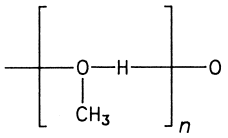
II.



V.



III.



VI.

**SCHEME 1**

of associated and dissociated forms are very different, so the particular value exhibited will depend on the situation in which the compound occurs. Structures such as tetramers (V) and chains (VI) have been proposed for methanol and methanol–ethanol mixtures.

In the case of water, it is necessary to distinguish between systems of water with low-permittivity organic liquids (i.e., associated water, with limited interaction between water and organic) and systems of water with hydrogen bonding organic liquids, where there is more intermolecular association at the expense of intramolecular association.

The difficulty in all such examples is that Hildebrand parameters and Hansen parameters are trying to indicate two or more things with one number or one set of numbers:

1. The extent of cohesion within the pure compound
2. The potential for cohesion in a mixture.

When one says that the Hansen hydrogen bonding parameter $\delta_h$ of ethanol is 20 $MPa^{1/2}$, it means that within pure ethanol $\delta_h^2 = 400$ MPa is the extent of cohesion due to hydrogen bonding, but it is also inferred that $\delta_h$ reflects the hydrogen bonding capability in a mixture.

It is clear that this will work best when the two components of the mixture are *similar*, and as the differences increase, situations arise where the effective $\delta_h$ may be much less than or much greater than that in a pure liquid.

As far as hydrogen bonding is concerned, some pure compounds are both proton donors and proton acceptors. Here, the extent of cohesion due to association in a

mixture generally can have any value up to that in a pure compound. Examples are alcohols, carboxylic acids, water, and primary and secondary amines.

In contrast, many compounds have a dominant capacity to accept protons: ketones, aldehydes, esters, ethers, tertiary amines, aromatic hydrocarbons, and alkenes. Here, the potential for cohesion in mixtures due to hydrogen bonding is *greater* than that in pure compounds. There are also some that are proton donors only, such as trichloromethane.

It is clear that neither Hildebrand parameters nor Hansen parameters are adequate to handle this problem in a quantitative way. There are two possible approaches:

1. To treat each associated species in both the pure liquids and the mixtures as new compounds, with formation constants that can be evaluated. This type of approach is traditional and well established, but rather cumbersome.
2. To use a full set of interaction cohesion parameters with separate Lewis acid and base contributions as well as dispersion, induction, and orientation terms. This method has the greater practical potential.

On reference to Eq. (20), it can be seen that when ${}^{ij}A_{ab}$ is large and negative, exothermic mixing (with evolution of heat) may be explained, in contrast to the restriction to athermic or endothermic processes, when only dispersion and polar forces exist. Thus, the answer to the criticism that Hildebrand parameters cannot cope with molecular association or exothermic interactions is that they should be expressed in the form of interaction cohesion parameters. The price to be paid is greater complexity in both evaluation and use.

## III. COHESION PARAMETERS AND OTHER SOLVENT SCALES

Parameters describing and correlating the solvent capacities of liquids have been based on a great variety of chemical and physical properties. Some are measures of liquid "basicity." Others are direct determinations of the solubility of a representative solute in a range of liquids, for example, the solubility of hydrogen chloride in the solvents at 10°C.

Once a solubility scale has been established, it is necessary to determine the position on it of any required solvent. If the solubility scale has a theoretical basis, it may be possible to calculate values from information on other properties, but if it is an empirical scale, direct testing is usually required.

Because Hildebrand parameter values can be determined readily and are widely available, they have been used in conjunction with various other theoretical or empirical parameters to provide more effective predictions of solvent properties. Several examples of the parameters that have been used in this way to form "hybrid" cohesion parameter scales are now described.

### A. Hydrogen Bonding Parameters

The Burrell hydrogen bonding classification (Section II.A) has been developed further by assigning quantitative values to the hydrogen bonding capabilities of liquids and plotting graphs of Hildebrand parameters against hydrogen bonding capacities.

An alternative quantitative measure of the capacity of organic liquids to donate and accept hydrogen bonds is sound velocity in solvent–paper systems. Paper fibers are held together largely by hydrogen bonds, and on wetting paper with a liquid, most disruption of the fiber bonding occurs in the presence of those solvents that preferentially form fiber–liquid hydrogen bonds. As a result, the velocity of sound through the paper, which depends on the degree of bonding, decreases as the liquid hydrogen bonding capability increases. Water, as the most common solvent for disrupting fiber–fiber bonds in paper, was chosen as a reference standard, and the hydrogen bonding parameter was defined as

$$\frac{100 \times \text{sound velocity in water-soaked paper}}{\text{sound velocity in liquid-soaked paper}}$$

### B. Electrostatic Parameters

The cohesive properties of materials are closely related to the various electrostatic properties, such as dipole moment, relative permitivity, polarizability, ionization potential, and refractive index. Both dipole moment and relative permittivity are of particular importance: In the absence of specific association, the dipole moment tends to determine the orientation of solvent molecules around molecular solutes, while dissolution of ions is promoted by high relative permittivity of the solvent. The *electrostatic factor*, which is defined as the product of dipole moment and relative permittivity, takes both effects into account and provides a basis for the classification of solvents. The *fractional polarity* is also noteworthy: The natures of molecular interactions have been discussed in terms of the fraction of the total interactions due to dipole–dipole or orientation effects. This fractional polarity can be calculated from the dipole moment, polarizability, and ionization potential once certain assumptions have been made.

### C. Spectroscopic Parameters

Numerous parameters have been developed on the basis of spectroscopic measurements. One direct hydrogen

bonding parameter is based on the effect that a liquid has on a small amount of alcohol introduced into that liquid: The greater the extent of hydrogen bonding between the liquid and the alcohol hydroxyl groups, the weaker the O—H bond and the lower the frequency of the infrared radiation absorbed by that bond. If deuterated methanol or ethanol is used, the O–D stretch band is in a spectroscopic region with little interference, permitting detection at low alcohol concentrations. The extent of the shift to lower frequencies of the O–D stretching infrared absorption of deuterated alcohol in the liquid under study thus provides a measure of its hydrogen bond acceptor capability. The spectrum of a solution of deuterated methanol or ethanol is compared with that of a solution in benzene or other reference liquid, and the hydrogen bonding parameter is defined as 10% of the O–D absorption shift expressed in wave numbers. The choice of the reference liquid is important. Benzene is not an inert solvent, the aromatic electron system having some hydrogen acceptor properties, and there is even the possibility of some hydrogen bonding between methanol and tetrachloromethane, another reference solvent. It appears that alkanes such as cyclohexane, heptane, and isooctane may be preferable as standards.

Spectroscopic hydrogen bonding parameters form a special case of a more general type of parameter described by such names as "electron donating power" and "electron accepting power."

### D. Empirical Solvent Scales

There are many empirical tests in common use for quantifying solvent behavior. One group of tests describes the "solvent power" or "strength" of liquid hydrocarbons. The *aniline point* or *aniline cloud point*, which is based on the fact that aniline is a poor solvent for aliphatic hydrocarbons and an excellent one for aromatics, is defined as the minimum equilibrium solution temperature for equal volumes of aniline and solvent. An approximately quantitative correlation exists between Hildebrand parameters and the aniline point for hydrocarbons, but it is of limited value. The *kauri–butanol number* (KB) is a measure of the tolerance of a standard solution of kauri resin in 1-butanol to added hydrocarbon diluent. There is an approximately linear relationship between the Hildebrand parameter and the KB number for hydrocarbons with KB $> 35$:

$$\delta/\mathrm{MPa}^{1/2} = 0.040\,\mathrm{KB} + 14.2. \qquad (33)$$

The solvent power of "chemical" or "oxygenated" liquids (such as alcohols, ketones, esters, and glycol ethers) is much greater than that of hydrocarbons, and different scales are necessary for their description. The dilution ratio is widely used since it is a direct measure of the tolerance of a solvent–resin mixture to added diluent, and qualitative correlations exist between dilution ratios and Hildebrand parameters. The heptane number of hydrocarbon solvents is a measure of the relative solvent power of high-solvency hydrocarbons in the presence of resins not soluble in heptane. The wax number and other miscibility numbers may also be correlated with the solvent Hildebrand parameter.

Another type of empirical solvent classification scheme has been developed in connection with chromatography, where it is useful to distinguish solvent strength or "polarity" from solvent "selectivity." Gas–liquid chromatographic methods are particularly convenient for quantitative characterization of the solvent properties of the stationary phase, whether this is a liquid or a polymer (see Section V.A). One approach is to define a polarity index, a measure of the capacity of a liquid to dissolve or interact with various polar test solutes. There is, in general, a good correlation between the polarity index values and Hildebrand parameter values, but liquids such as diethyl ether and triethylamine that are strong proton acceptors but have no proton donor capacity have Hildebrand parameter values similar to those of alkanes, although they show up as moderately polar on the polarity index scale. This is because Hildebrand parameters are based on pure liquid properties, while the polarity index is based on the interactions between different liquids.

### E. Hybrid Scales

Often cohesion parameters are included in hybrid solvent scales which incorporate several kinds of parameters. For example, a "universal solubility" treatment includes the Hildebrand parameter as a measure of the energy necessary to create a cavity in the bulk liquid to accommodate a solute molecule. This contribution is combined with parameters allowing for solute–solvent dipole interactions, hydrogen-bond donor acidity, hydrogen-bond acceptor basicity, and an empirical coordinate covalency parameter.

## IV. APPLICATIONS OF COHESION PARAMETERS

### A. Liquids

For a pair of regular liquids, the infinite dilution activity coefficient expression in terms of Hildebrand parameters is given by Eq. (5). For dilute solutions with specific interactions and size effects, resulting in nonzero enthalpies of mixing and nonideal entropies of mixing, the expressions become more complex but are still useful.

For systems composed of two liquids that have substantial but incomplete mutual miscibility, cohesion parameters may be used in combination with other empirical expressions. Although satisfactory data correlation is possible, prediction of the exact extent of mutual solubility is difficult.

Cohesion parameter methods have been used for the correlation of activity coefficients of a wide range of systems. One of the most flexible approaches to the correlation and prediction of activity coefficients has been to include terms for some or all of the following:

- Regular enthalpy of mixing from Hildebrand parameters
- Polar orientation and induction effects
- Entropy effects from hydrogen bonding equilibria
- Gibbs free energy term for the breaking of hydrogen bonds in association interactions

Each component is thus characterized by several parameters. The overall average error in prediction for 845 literature data points for ~300 systems was 25% in the infinite dilution activity coefficient, falling to better than 9% for all saturated hydrocarbons in all solvents, adequate for screening and for some design purposes.

One of the early applications of regular-solution theory was the discussion of activity coefficient ratios and equilibrium constants for complex formation in terms of Hildebrand parameters. There has been considerable debate as to the correct basis of the equilibrium constant (concentration, mole fraction, volume fraction, or molality), different methods proving preferable for various systems. Extension to complexes of ionic species (as in solvent extraction systems) has also occurred.

A common application of solvents is in the separation of the components of mixtures, by countercurrent liquid–liquid extraction or by extractive distillation or by azeotropic distillation. These solvent-aided separation processes can be planned with the aid of cohesion parameters. The selectivity ${}^{k}S_{ij}$ of a solvent $k$ toward a dilute mixture of $i$ and $j$, derived from relationships such as Eq. (5), is given by

$$RT \ln {}^{k}S_{ij} = {}^{i}V^{ik}A - {}^{j}V^{jk}A, \qquad (34)$$

with contributions from dispersion, orientation, induction, entropy, and acid–base terms. In general, the choice of a noninteracting solvent with a cohesion parameter that differs significantly from the cohesion parameters of the components to be separated enhances the solvent selectivity, although it reduces the solvent capacity. It is therefore necessary to reach a compromise such that the liquid has an adequate capacity but retains good selectivity and immiscibility between the phases. However, if a liquid can be obtained that specifically interacts with the component to be extracted, then both the selectivity and the solvent capacity are increased. For example, in the case of an alkane $i$ (for which both ${}^{i}\delta_a$ and ${}^{i}\delta_b$ are negligible), an aromatic hydrocarbon $j$ (${}^{j}\delta_a$ negligible), and an electron-accepting liquid $k$ (${}^{k}\delta_b$ negligible), the term for the specific interaction is given by $2{}^{j}V^{j}\delta_b{}^{k}\delta_a$.

Another common situation in which cohesion parameters have been employed involves solvent extraction of an electrically neutral ion pair or complex used to transport ionic species distributed in relatively dilute solution between two very dissimilar phases such as hydrocarbon and water. As in all other applications of cohesion parameters, it should not be expected that distribution ratios can be predicted in detail by means of cohesion parameters, and even if the infinite dilution values are approximately correct, these have limited applicability. Rather, the emphasis is on formulating broad trends and establishing correlations.

## B. Gases

In many situations, such as the dissolution of oxygen in water, there are negligible specific interactions when gases dissolve in liquids, and it is reasonable to expect cohesion parameters, even simple Hildebrand parameters, to provide methods of correlation and prediction.

J. H. Hildebrand and others have long pointed out the close relationship that exists between the logarithm of the mole fraction gas solubility, $\log {}^{j}x_s$, and the solvent Hildebrand parameter ${}^{i}\delta$. For example, the values of $\log {}^{j}x_s$ for nitrogen, argon, methane, ethylene, and ethane were found to be linear functions of ${}^{i}\delta$ for the normal primary alcohols as solvents, but it was subsequently shown that some curved lines became straight when ${}^{i}\delta^2$ was used in place of ${}^{i}\delta$:

$$-\ln {}^{j}x_s = -\ln {}^{j}x_{\text{ideal}} + ({}^{j}\bar{V}/RT)({}^{i}\delta - {}^{j}\delta)^2, \qquad (35)$$

where $\bar{V}$ is the average partial molar volume of the gas in a range of liquids. There are several similar equations, all tacitly or explicitly assuming that the gaseous solute is condensed to a hypothetical "liquid" state (with hypothetical liquid–state Hildebrand parameter and molar volume) before mixing with the solvent.

The solubilities of gases such as tetrafluoromethane, sulfur hexafluoride, and carbon dioxide deviate from straight lines because of various properties.

Another simple application of cohesion parameters to gases is the plotting of gas solubility directly against solvent Hildebrand parameter values. A relatively sharp maximum is usually obtained at a Hildebrand parameter value corresponding to that of the hypothetical liquid form of the dissolved gas.

As well as this situation of a gas possessing a hypothetical liquid-like molar volume when it dissolves in a liquid at low pressures, there is also the case of gases that liquefy or achieve liquid-like molar volumes because of low temperatures and/or high pressures. The solvent properties of compressed gases, especially carbon dioxide, for biochemicals and polymers have been receiving more attention recently because of their application in high-pressure gas chromatography and supercritical fluid chromatography.

Cohesion parameter concepts are being utilized for studies of cryogenic liquids, refrigerants, and aerosol propellants, as well as fundamental dense gas and vapor–liquid properties. For example, it is reported that even at 25 K above their critical temperatures, compressed gaseous helium ($\delta \approx 8\ \mathrm{MPa}^{1/2}$) and xenon ($\delta \approx 16\ \mathrm{MPa}^{1/2}$) at 200 MPa pressure separate into two phases because of their divergent cohesion properties.

## C. Solids

The regular-solution theory and the Hildebrand parameter are based on the enthalpy changes occurring when liquids are mixed. In order to extend these concepts to solutions of crystalline solids in liquids, it is necessary to estimate the thermodynamic activity of the solid referred to a hypothetical liquid subcooled below its melting point. If this is done, expressions such as Eq. (5) can be used to evaluate cohesion parameters of nonvolatile solutes or to estimate their solubilities.

The Hildebrand parameter, defined in Eq. (2), has a liquid state basis so this hypothetical liquid reference state is necessary when the method is extended to solids. Use of the vaporization enthalpy or sublimation enthalpy of a solid at 25°C in Eq. (2) uncorrected for the crystal–liquid transition enthalpy change does not yield a Hildebrand parameter. It is another type of cohesion parameter that should not be confused with any of those defined in Section I.

Despite this complication, the solubility (${}^{j}x_{\mathrm{s}}$ on the mole fraction scale) of a solid $j$ in a liquid $i$ has been shown to vary regularly with the Hildebrand parameter of the solvent, plots of $\log {}^{j}x_{\mathrm{s}}$ against ${}^{i}\delta$ being nearly linear or only slightly curved.

It is possible to derive more informative thermodynamic expressions that involve various assumptions, and several of these predict that, for solutions that approach regular behavior, plots of $\log {}^{j}x_{\mathrm{s}}$ against $({}^{i}\delta - {}^{j}\delta)^2$ should be approximately linear, and this relatively simple method is probably the most widely used method of correlating solid solubilities in terms of Hildebrand parameters.

If ${}^{i}\delta$, ${}^{j}\delta$, and ${}^{j}V$ (for the subcooled liquid $j$ corresponding to the solid of interest) as well as the melting point and entropy of melting are known for approximately regular systems, it is possible to calculate the data necessary for phase diagrams, either temperature against ${}^{i}x_{\mathrm{s}}$ plots for binary systems or triangular diagrams with temperature contours for ternary systems.

The more detailed and informative interaction cohesion parameters have not been widely used in the correlation of solid–liquid solubilities. Hansen parameters have been extended to ionic solids, but the determination of their values has not been pursued to any great extent. Any extension of interaction cohesion parameters to ionic systems in the future would be very valuable in dealing with solid–liquid systems.

## D. Polymers

The parameter most commonly used in the discussion of polymer solutions is the polymer–liquid interaction parameter $\chi$. This parameter has been associated with various theoretical treatments, but it can be considered a general, dimensionless parameter reflecting intermolecular forces between a particular polymer and a particular liquid. It was introduced as an enthalpy of dilution with no entropy component but has been considered subsequently to be a Gibbs free energy parameter. As originally formulated, the polymer–liquid interaction parameter was expected to be inversely dependent on absolute temperature and independent of polymer concentration, but as now empirically defined it depends in an unspecified way on temperature, solution composition, and polymer chain length.

It can be shown that the enthalpy part of the interaction parameter, $\chi_{\mathrm{H}}$, is related to the interchange cohesive pressure by

$$\chi_{\mathrm{H}} = {}^{ij}A\,{}^{i}V/RT. \tag{36}$$

If Hildebrand parameters are used, from Eq. (4),

$$\chi_{\mathrm{H}} = ({}^{i}V/RT)({}^{i}\delta - {}^{j}\delta)^2 \tag{37}$$

and in terms of Hansen parameters

$$\chi_{\mathrm{H}} = ({}^{i}V/RT)\left[\left({}^{i}\delta_{\mathrm{d}} - {}^{j}\delta_{\mathrm{d}}\right)^2 + \left({}^{i}\delta_{\mathrm{p}} - {}^{j}\delta_{\mathrm{p}}\right)^2 + \left({}^{i}\delta_{\mathrm{h}} - {}^{j}\delta_{\mathrm{h}}\right)^2\right]. \tag{38}$$

The enthalpy component of the polymer–liquid interaction parameter is therefore closely related to the cohesion parameters of the components of the polymer solution.

Cohesion parameter predictions of thermodynamic property values must be corrected to allow for the substantial size differences between polymer and solvent molecules. However, if only semiquantitative "compatibility" information is required, cohesion parameters of amorphous polymers may be used in the same way as those of liquids. When crystalline polymers are considered, it is

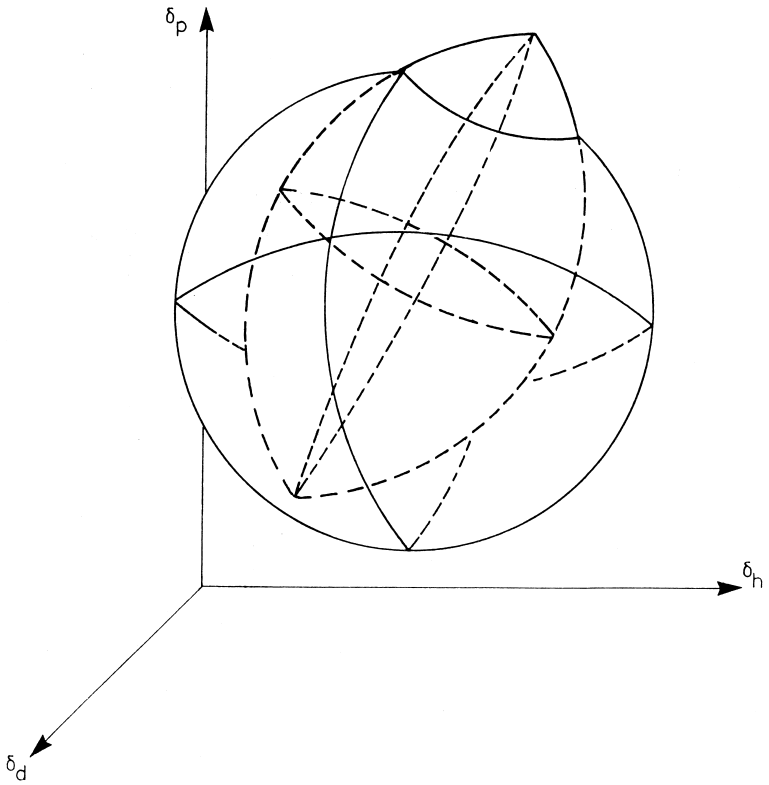



**FIGURE 2** Hansen parameter three-dimensional model showing the extent of overlap between poly(phenylene oxide phosphonate ester) (spherical) and cellulose acetate. [Adapted from Cabasso, I., Jaguer-Grodzinski, J., and Vofsi, D. (1977). In "Polymer Science and Technology" (D. Klempner and K. C. Frisch, eds.), Vol. 10, p. 1, Plenum, New York.]

in principle necessary to calculate the activity of the crystalline solid relative to the real or hypothetical amorphous material at the same temperature. In practice, the distinction between amorphous and crystalline polymers is not clear-cut, and nonthermodynamic empirical methods are frequently used in the estimation of polymer cohesion parameters.

It is not correct to assume that the "best" solvent for a polymer is necessarily the corresponding monomer or low molecular weight liquid polymer ("oligomer") made up of the same repeating units. Such liquids usually have cohesion parameter values lower than those of the polymers because their molar volumes are higher. Rather, the best solvent is another liquid with a cohesive energy rather higher than that of the monomer but with a cohesive pressure the same as that of the polymer.

In one of the original applications of cohesion parameters to polymer solutions, the liquid Hildebrand parameter was combined with the Burrell hydrogen bonding classification (Section II.A). Hansen's three-component cohesion parameter system was also first developed for polymer–liquid systems.

As well as the assumption of spherical Hansen parameter "volumes" of polymer–liquid interaction, as described in Section I.C and illustrated in Fig. 2, it is possible to present this information in the form of irregular volumes. Three-dimensional models, stereographs, sets of projections, and contour maps have all been used. In many cases the solubility behavior of a polymer can be adequately represented in two dimensions by $\delta$ and $\delta_h$ or by $\delta$ and $\delta_p$. Another useful type of plot involves the subdivision of the Hildebrand parameter into a "volume-dependent" part $\delta_v = (\delta_d + \delta_p)^{1/2}$ and a "residual" or hydrogen bonding part $\delta_h$. Triangular diagrams also provide a method of conveying information in three properties on two-dimensional plots by making the excessively simple assumption that the Hildebrand cohesion parameter is uniform for all materials and that it is the relative magnitude of the three contributions that determines the extent of miscibility. This is illustrated in Fig. 3.

Also widely used for polymers are hybrid maps of cohesion parameters with other quantities such as hydrogen bonding parameter, dipole moment, and fractional polarity.

Cohesion parameters are widely used to predict polymer solubility or swelling in liquids, but some liquids also affect the mechanical properties of polymers by environmental stress cracking or crazing. Whatever the cause of

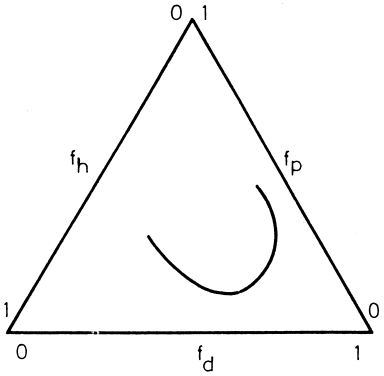



**FIGURE 3** Limiting boundary of solubility for cellulose nitrate in terms of fractional cohesion parameters for dispersion ( $f_d$), polar ($f_p$), and hydrogen bonding ($f_h$) effects. [Adapted from Gardon, J. L., and Teas, J. P. (1976). In "Treatise on Coatings" (R. R. Myers and J. S. Long, eds.) Vol. 2, Chap. 8, Dekker, New York.]

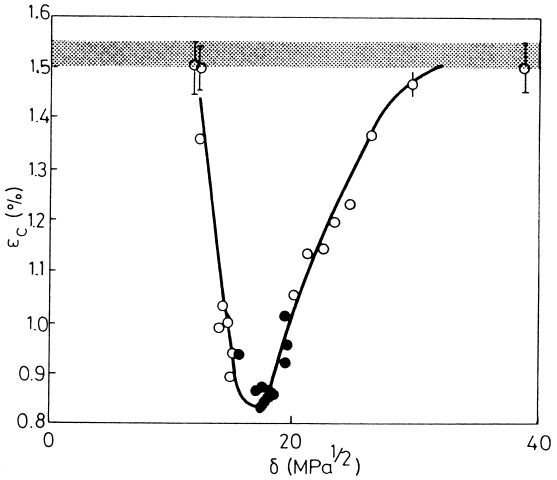



**FIGURE 4** Critical strain $\varepsilon_c$ for crazing (open circles) or cracking (filled circles) of poly(2,6-dimethyl-1,4-phenylene oxide) as a function of the Hildebrand parameter $\delta$ of the solvent. The $\varepsilon_c$ values in air are in the band near $\varepsilon_c = 1.5\%$. [Adapted from Bernier, G. A., and Kambour, R. P. (1968). *Macromolecules* **1,** 393.]

these effects, there is a surprisingly close correlation with the liquid Hildebrand parameter (Fig. 4).

## E. Rate and Transport Properties

Many varied theoretical and empirical scales are available for the description of the effect of solvents on the rates of chemical processes in solution, cohesion parameters forming only one. Because the transition state theory can be interpreted in terms of pseudothermodynamic properties, including volume and enthalpy, it is also possible to attribute cohesion parameter values to the activated state. Cohesion parameters are closely related to the internal pressures in the system, and for nonpolar reactions (as well as for polar reactions in nonpolar solvents) the internal pressure of solvents influences reaction rates in the same direction as does external pressure.

The activated complex in a chemical reaction is considered to have properties that approach those of the products of the reaction, and there is a general rule that, if the reaction is one in which the products are of higher cohesion than the reactants, it is accelerated by solvents of high cohesive pressure. Conversely, if the solvent is similar to the reactants in cohesion properties, the rate tends to be lower. For polar or ionic processes the effect of the solvent is more complex, and the cohesion parameter concept is less useful.

Viscous flow can be regarded as a rate process in which molecules move into holes or voids in the liquid. According to this model, the activation energy for viscous flow is related to the energy required to form such a hole and hence to the cohesion parameter of the liquid. One successful application was a relationship between Hildebrand parameters and the initial slopes of plots of dilute solution viscosities against mole fraction composition. In a very different kind of system, the dropping points of greases made from lithium stearate soap and oils have been shown to be a function of the Hildebrand parameters of the oils.

The limiting viscosities of dilute polymer solutions have also received attention from this point of view. The viscosity of a dilute solution is a maximum in the best solvent, that is, in one in which the cohesion parameters of solvent and polymer are comparable. In a "good" solvent the polymer molecules are "unfolded" or "uncoiled," obtaining to the maximum extent the more favorable polymer–liquid interactions (and therefore resulting in the greatest viscosity), while in a "poor" solvent the polymer molecules remain folded because of the more favorable intramolecular interactions. Viscosity–cohesion parameter correlations have proved successful for such dilute polymer solutions.

Although the viscosity of a dilute polymer solution is a minimum in poor solvents, as the polymer concentration is increased there is a changeover in behavior: The viscosity exhibits a maximum in poor solvents at higher concentration. There is an aggregation or clustering of polymer molecules as a preliminary to phase separation when solvents are used that have cohesion parameters near the limits of the miscibility range for the polymer. This rather complex behavior can be described in terms of entropy and free volume effects, which are not present in nonpolymer systems, and it should not be expected that it can be fully correlated by means of cohesion parameters that provide direct information only on the enthalpy aspect of the interactions.

Other transport properties have also been correlated by means of cohesion parameters, including gas, liquid, and solid diffusion, permeation in polymers, reverse osmosis in membranes, and the time-dependent mechanical properties of polymers.

### F. Surfaces

As a result of the close relationship between solubility or phase separation and surface activity, cohesion parameters are useful for the characterization of heterogeneous as well as homogeneous systems. Interfacial free energy ("surface tension") is the excess free energy due to the existence of an interface, arising from imbalanced molecular forces. These forces in the bulk of the material are the origin of cohesive properties, so the close link between adhesion and cohesion is easy to understand.

Various empirical or semiempirical equations have been used to link cohesion parameters with surface tension. There is considerable theoretical and experimental justification for subdividing surface free energy into additive components analogous to the cohesion parameter components on the basis of the types of molecular interaction. In fact, Hansen considered the characterization of surfaces in terms of the Hansen parameters of the liquids that spread spontaneously on them. Ideas such as these have been extended to practical problems of wetting, dewetting, adhesion, lubrication, adsorption, colloids, emulsions, and foams.

### G. Biological Systems

Solubility is of major importance in biochemical processes, and correlations with cohesion properties have been explored for such purposes as transport of molecules through biological tissues, rationalization of physicochemical influences on biological responses, formulation of drugs in liquid dosage form, design of insecticides, and properties of biocompatible materials.

One particularly interesting example is the action of fluorinated ethers as inhalation convulsants, on the one hand, or as anesthetics, on the other. Those such as hexafluorodiethyl ether with Hildebrand parameters of less than 15 $MPa^{1/2}$ are powerful convulsants, whereas those with $\delta > 15$ $MPa^{1/2}$ such as methoxyflurane are anesthetics. Although all the ethers dissolve equally well in bulk lipids and have similar octanol–water partition values, they dissolve differentially into specific local microenvironments or subregions, which can be considered to have Hildebrand parameter values different from that of the bulk. This indicates that the cohesion parameter concept is valuable both in models concerned with general partition, concentration, or activity and in models that assume more specific mechanisms.

## V. EVALUATION OF COHESION PARAMETERS

### A. Thermodynamic Calculations and Inverse Gas Chromatography

From the definition of the Hildebrand parameter [Eq. (2)], it is apparent that it is necessary to determine both the enthalpy of vaporization and the molar volume for its evaluation. There is rarely much difficulty in finding a reliable value for the molar volume of a liquid, and solids can be treated as subcooled liquids with molar volumes extrapolated from the liquid state values. Frequently, the main problem is obtaining the enthalpy of vaporization at the temperature of interest, usually 25°C. Direct experimental information is frequently unavailable, and extrapolation methods or even empirical calculations are often necessary, based on such properties as boiling point, corresponding states, activity coefficients, and association constants. The "$RT$" correction in Eq. (2) assumes that the vapor is ideal, and although gas law corrections may be applied, even at the normal boiling point the correction is usually negligible.

Polymers and solids pose particular problems because the enthalpy of vaporization is unavailable. Interactions of polymers or noncrystalline solids (particularly those used as plasticizers) with liquids can be studied conventionally by using the polymer or solid as the stationary phase in gas chromatography columns. The activity coefficients at infinite dilution of volatile liquids in the polymer may be determined by this "inverse gas chromatography" using mobile phases to investigate the properties of stationary phases, rather than the reverse. From these activity coefficients, cohesion parameters may be estimated for polymers and some organic solids and liquids.

Hildebrand parameters, Hansen parameters, and the more detailed interaction cohesion parameters can be evaluated from inverse gas chromatography results, but to date this has not been widely practiced.

The thermodynamic quantity internal pressure, given by

$$\pi = (\partial U/\partial V)_T = (\partial p/\partial T)_V - p \qquad (39)$$

and directly accessible from experiment for such nonvolatile materials as polymers [to which Eq. (2) cannot be applied] as well as mixed systems, can also provide cohesion parameter values.

### B. Empirical Methods

A list of liquids can be compiled with a gradation of Hildebrand parameter values to form a "solvent spectrum." In its most common form it includes subdivision

into categories of hydrogen bonding capacity, as indicated in Table I. The Hildebrand parameter of a solute can then be taken as the midpoint of the range of solvent Hildebrand parameters that provides complete miscibility or the particular value that provides maximum solubility, or maximum swelling in the case of a cross-linked polymer. An ASTM test method for polymer solubility ranges uses mixtures of solvents to provide a spectrum of closely spaced Hildebrand parameters. Other physical properties that can be used as well as solubility and swelling include viscosity and related properties such as grease dropping points. A range of semiempirical equations is also available for correlation and prediction of cohesion parameters.

### C. Homomorph Methods for Hansen Parameters

In any multicomponent cohesion parameter system such as Hansen parameters, there arises the problem of evaluating the components of the Hildebrand parameter separately. One obvious approach is to compare the properties of compounds that differ only in the presence or absence of a certain group. Here, the homomorph concept is useful: The homomorph of a polar molecule is a nonpolar molecule having very nearly the same size and shape.

For liquids, the Hansen dispersion parameter obtained by homomorph methods can be subtracted from the total cohesion pressure using Eq. (22), with the remainder being split into Hansen hydrogen bonding and polar parameters so as to optimize the description of the solubility and swelling behavior of a range of liquids and polymers. Both empirical methods and group methods (see next section) can be used. Once the three Hansen parameters for each liquid are evaluated, the Hansen parameters for each polymer can be obtained.

This method may distort the relative magnitudes of the intermolecular forces, but as pointed out in Section I.C, the theoretical bases of Hansen parameters are not good in any case.

Interaction cohesion parameters could, in principle, be evaluated in a similar way, but there has been little activity in this area.

### D. Group Contribution Methods

Many properties of materials change in a regular way with increasing chain length in a homologous series, and some properties are conveniently linear. The miscibility behavior of materials depends to a large extent on the cohesive and volume properties, specifically the molar cohesive energy $-U$ and the molar volume $V$, and these quantities, together with their ratios and products, can be estimated in terms of standard contributions from groups of atoms.

The molar cohesive energy can be represented by the summation of atomic or group contributions:

$$-U = -\sum_z {}^zU. \tag{40}$$

Hildebrand parameters can be calculated from

$$\delta = \left(\frac{-U}{V}\right)^{1/2} = \left(-\sum_z {}^zU \Big/ \sum_z {}^zV\right)^{1/2}. \tag{41}$$

Also useful are the group molar attraction constants ${}^zF$ defined by

$$\sum_z {}^zF = -\sum_z {}^zU \sum_z {}^zV = \delta V \tag{42}$$

so

$$\delta = \sum_z {}^zF \Big/ \sum_z {}^zV. \tag{43}$$

Values of ${}^zF$ and ${}^zV$ have been tabulated for the most common organic molecular groups.

## VI. SELECTED VALUES

It is neither appropriate nor practicable to provide here a comprehensive compilation of values, but exhaustive tables are included in the *Handbook of Solubility Parameters and Other Cohesion Parameters* listed in the bibliography. Rather, listed in Table I are typical values for some liquids whose Hildebrand parameter values, Hansen parameter values, and interaction cohesion parameter values are known with reasonable reliability. The Burrell hydrogen bonding classification (Section II.A) is also included. There is considerable variation in the Hansen parameters reported for water. A study of the solubilities of a range of *organics in water* suggests $\delta_d = 20$, $\delta_p = 18$, $\delta_h = 18$, and $\delta_t = 32\ \mathrm{MPa}^{1/2}$ rather than the results in Table I, which are more consistent with the behavior of *water in organic liquids*. This variability in Hansen parameter values is a fundamental problem associated with the use of the single parameter $\delta_h$, rather than the pair of acid and base parameters.

For polymers, interaction cohesion parameters are yet to be determined in any detail, so the values given in Table II are restricted to ranges of polymer Hildebrand parameters (for use with solvents of specified Burrell hydrogen bonding class) and sets of Hansen parameters [together with the interaction radius ${}^iR$ of Eq. (25)]. Table III presents preferred Hildebrand parameter values for some well-studied

**TABLE I Typical Hildebrand Parameter, Hansen Parameter, and Component Cohesion Parameter Values**

| Liquid | Hildebrand parameter ($\delta_t$/MPa$^{1/2}$) | Component cohesion parameters (MPa$^{1/2}$) | | | | | Hansen parameters (MPa$^{1/2}$) | | | Burrell hydrogen bonding class | Molar volume (cm$^3$ mol$^{-1}$) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $\delta_d$ | $\delta_o$ | $\delta_i$ | $\delta_a$ | $\delta_b$ | $\delta_d$ | $\delta_p$ | $\delta_h$ | | |
| Pentane | 14.5 | 14.5 | 0 | 0 | 0 | 0 | 14.5 | 0.0 | 0.0 | Poor | 115 |
| Hexane | 14.9 | 14.9 | 0 | 0 | 0 | 0 | 14.9 | 0.0 | 0.0 | Poor | 131 |
| Diethyl ether | 15.3 | 13.7 | 5 | 1 | 0 | 6 | 14.5 | 2.9 | 5.1 | Moderate | 105 |
| Cyclohexane | 16.8 | 16.8 | 0 | 0 | 0 | 0 | 16.8 | 0.0 | 0.2 | Poor | 108 |
| Ethyl acetate | 18.2 | 14.3 | 8 | 2 | 0 | 6 | 15.8 | 5.3 | 7.2 | Moderate | 98 |
| Toluene | 18.2 | 18.2 | 0 | 0 | 0 | 1 | 18.0 | 1.4 | 2.0 | Poor | 107 |
| Tetrahydrofuran | 18.6 | 15.5 | 7 | 2 | 0 | 8 | 16.8 | 5.7 | 8.0 | Moderate | 82 |
| Benzene | 18.8 | 18.8 | 0 | 0 | 0 | 1 | 18.4 | 0.0 | 2.0 | Poor | 89 |
| Acetone | 19.6 | 13.9 | 10 | 3 | 0 | 6 | 15.5 | 10.4 | 7.0 | Moderate | 74 |
| Chlorobenzene | 19.8 | 18.8 | 4 | 0.6 | 0 | 2 | 19.0 | 4.3 | 2.0 | Poor | 102 |
| Bromobenzene | 20.2 | 19.6 | 3 | 0.4 | 0 | 2 | 20.5 | 5.5 | 4.1 | Poor | 105 |
| 1,4-Dioxane | 20.7 | 16.0 | 11 | 2 | 0 | 2 | 19.0 | 1.8 | 7.4 | Moderate | 86 |
| Pyridine | 21.7 | 18.4 | 8 | 2 | 0 | 10 | 19.0 | 8.8 | 5.9 | Strong | 81 |
| Acetophenone | 21.7 | 19.6 | 6 | 1 | 0 | 7 | 19.6 | 8.6 | 3.7 | Moderate | 117 |
| Benzonitrile | 21.9 | 18.8 | 7 | 2 | 0 | 5 | 17.4 | 9.0 | 3.3 | Poor | 103 |
| Propionitrile | 22.1 | 14.1 | 14 | 4 | 0 | 4 | 15.3 | 14.3 | 5.5 | Poor | 71 |
| Quinoline | 22.1 | 21.1 | 4 | 0.6 | 0 | 9 | 19.4 | 7.0 | 7.6 | Strong | 118 |
| *N*,*N*-Dimethylacetamide | 22.1 | 16.8 | 10 | 3 | 0 | 9 | 16.8 | 11.5 | 10.2 | Moderate | 92 |
| Nitroethane | 22.5 | 14.9 | 12 | 5 | 0 | 2 | 16.0 | 15.5 | 4.5 | Poor | 71 |
| Nitrobenzene | 22.7 | 19.4 | 7 | 2 | 0 | 2 | 20.0 | 8.6 | 4.1 | Poor | 103 |
| *N*,*N*-Dimethylformamide | 24.1 | 16.2 | 13 | 5 | 0 | 9 | 17.4 | 13.7 | 11.3 | Moderate | 77 |
| Dimethylsulfoxide | 24.5 | 17.2 | 13 | 4 | 0 | 11 | 18.4 | 16.4 | 10.2 | Moderate | 71 |
| Acetonitrile | 24.7 | 13.3 | 17 | 6 | 0 | 8 | 15.3 | 18.0 | 6.1 | Poor | 53 |
| Nitromethane | 26.4 | 14.9 | 17 | 6 | 0 | 3 | 15.8 | 18.8 | 5.1 | Poor | 54 |
| Water (see text) | 48 | 13 | 31 | 21 | 34 | 22 | 16 | 16 | 42 | Strong | 18 |

polymers. For solid materials, in general only Hildebrand parameters are available (Table IV).

## VII. CURRENT STATUS

Theoretical development of this topic appears to have reached something of a plateau, but there is an increasing number of practical applications of cohesion parameters and of computational methods simplifying the process.

An excellent example of the determination and application of Hildebrand parameters to a "new" solvent and its compatibility with polymers is provided by 1,8-cineole. This compound, present at high levels in the leaf oils of some eucalypts, is proposed as a replacement for the solvent 1,1,1-trichloroethane (which is now known to cause

**TABLE II Typical Hildebrand Parameter and Hansen Parameter Values for Polymers, $\delta$ (MPa$^{1/2}$)**

| Polymer (manufacturer) | Hansen parameter | | | | Hildebrand parameter ranges in liquids of Burrell hydrogen bonding class | | |
|---|---|---|---|---|---|---|---|
| | $\delta_d$ | $\delta_p$ | $\delta_h$ | $^jR$ | Poor | Moderate | Strong |
| Pentalyn® 255 alcohol-soluble resin (Hercules) | 17.6 | 9.4 | 14.3 | 10.6 | 18–21 | 15–22 | 21–30 |
| Pentalyn® 830 alcohol-soluble rosin resin (Hercules) | 20.0 | 5.8 | 10.9 | 11.7 | 17–19 | 16–22 | 19–23 |
| Cellulose nitrate, 0.5 sec | 15.4 | 14.7 | 8.8 | 11.5 | 23–26 | 16–30 | 26–30 |
| Cellolyn® 102 pentaerythritol ester of rosin, modified (Hercules) | 21.7 | 0.9 | 8.5 | 15.8 | 16–21 | 17–22 | 21–24 |
| Versamid® 930 thermoplastic polyamide (General Mills) | 17.4 | −1.9 | 14.9 | 9.6 | — | — | 19–23 |
| Poly(methyl methacrylate) (Rohm and Haas) | 18.6 | 10.5 | 7.5 | 8.6 | 18–26 | 17–27 | — |

**TABLE III Preferred Hildebrand Parameter Values for Selected Polymers**

| Polymer | $\delta$/MPa$^{1/2}$ |
|---|---|
| Polyacrylonitrile | 26 |
| Polybutadiene | 17.0 |
| Poly(butyl acrylate) | 18.5 |
| Cellulose acetate | 24 |
| Cellulose nitrate | 21 |
| Polychloroprene | 18.5 |
| Poly(dimethylsiloxane) | 15.5 |
| Ethyl cellulose | 20 |
| Polyethylene | 17.0 |
| Poly(ethylene oxide) | 24 |
| Poly(ethyl methacrylate) | 18.5 |
| Polyisobutylene | 16.5 |
| Polyisoprene, natural rubber | 17.0 |
| Poly(methyl acrylate) | 20.5 |
| Poly(methyl methacrylate) | 19.0 |
| Polypropylene | 16.5 |
| Polystyrene | 18.5 |
| Poly(tetrafluoroethylene) | 13 |
| Poly(vinyl acetate) | 20 |
| Poly(vinyl chloride) | 19.5 |

stratospheric ozone depletion). On the basis of calculations such as those described in Section V, a Hildebrand parameter of 18 MPa$^{1/2}$ for cineole was deduced, which is within the range of values suggested by the polymer solubilities. This is close to the value for trichloroethane (17 MPa$^{1/2}$), successfully predicting the efficacy of cineole as a replacement solvent.

Further, in developing a new solvent or solvent blend, it is also necessary to determine what polymers are likely to be affected adversely if exposed to the liquid or vapor. Table V shows that those polymers having Hildebrand parameters within 1 MPa$^{1/2}$ unit of the cineole value are soluble, and within 3 MPa$^{1/2}$ units there can be significant swelling, which is useful as an initial guide. However, the Hildebrand parameters of polymers (such as polyethylene) showing good resistance to cineole, despite having similar Hildebrand parameter values, demonstrate the limitations of such predictions.

A demonstration of the great variety and extent of applications of cohesion parameters is provided by the results of an internet search using Alta Vista, <http://www.altavista.com/cgi-bin/query?pg=aq>, conducted in June 1999 for the expression ("solubility parameter*" or "cohesion parameter*" or "Hildebrand parameter*" or "Hansen parameter*"), yielding 460 hits.

The information resulting from this search has now been organized and collected at the two sites, <http://www.mallee.com/parameters.html> and <http://wwwchem.murdoch.edu.au/staff/barton/parameters.html>, with convenient hyperlinks to most of the sites found in the search.

Motivation for providing free internet information of the kind seen in these sites is determined by commercial considerations through the opportunity for attracting potential clients:

- Computer modeling and simulation software for chemical systems often include the ability to estimate cohesion parameters, a major growth area.
- Chemical manufactures incorporate cohesion parameter values (of either the Hildebrand or Hansen variety) in material safety data sheets.
- A few educational institutions provide cohesion parameter information as a component of chemical or polymer science, for example, University of Missouri—Rolla <http://www.umr.edu/~wlf/> and discussion lists such as <http://www.hmco.com/college/chemistry/resourcesite/digests/chemedl/cedjun96/ msg00149.htm>.
- Some publishers and conference organizers providing titles or abstracts of papers to be presented or published include cohesion parameter topics, but full texts are rarely available.

**TABLE IV Typical Hildebrand Parameter Values for Predominantly Covalent Crystalline Solids (Assumed to be Subcooled Liquids)**

| Solid | $\delta$/MPa$^{1/12}$ |
|---|---|
| Alcohol: 1-hexadecanol | 20 |
| Aliphatic acids | 18–22 |
| Amines, anilines, amides | 20–30 |
| Aromatic hydrocarbons | 20–22 |
| Barbituric acid derivatives | 23–28 |
| Benzoic acid, substituted benzoic acids | 23–29 |
| Cholesterol | 19 |
| Choleseryl esters | 15–19 |
| Cortisone and related compounds | 27–30 |
| Halogen compounds of Sn, As, Sb, Bi | 23–30 |
| Iodine | 29 |
| Lipids | 18–27 |
| Metal soaps | 18–19 |
| Methyl xanthines, including caffeine | 24–29 |
| Norethindrone and derivatives | 20–22 |
| Phenols, including antioxidants and nitrophenols | 19–22 |
| Phosphorus | 27 |
| Sulfonamides | 25–30 |
| Sulfur | 26 |
| Testosterone and derivatives | 19–20 |

**TABLE V Relative Resistance of Polymers to 1,8-Cineole ($\delta = 18$ MPa$^{1/2}$)**

| Effect of cineole on polymer (4-month continuous exposure) | Polymer | $\delta$ (MPa$^{1/2}$) |
|---|---|---|
| **Soluble** | Natural rubber | 17.0 |
| | Polystyrene | 18.5 |
| | Styrene-butadiene elastomer | — |
| **Strongly swollen (>100%)** | Neoprene | 18.5 |
| | Polyurethane rubber | 20–21 |
| | Silicone rubber | 15.5 |
| **Little swelling (<10%)** | Acrylic-styrene-acrylonitrile | ~22 |
| | Polyester urethane | ~20 |
| | Polyester urethane | ~20 |
| | Polycarbonate/acrylic-styrene-acrylonitrile | — |
| | Fluoro elastomer | — |
| **Resistant** | Acrylonitrile-butadiene-styrene | — |
| | Acetal | 21–22 |
| | Nylon-6 | 24–32 |
| | Nylon-6,6 | 28–32 |
| | Polyethylene | 17 |
| | Poly(butylene terephthalate) | 21–22 |
| | Poly(ethylene terephthalate) | 21–22 |
| | Poly(methyl methacrylate) | 20.5 |
| | Polypropylene | 16.5 |
| | Polycarbonate | 19–22 |
| | Styrene-acrylonitrile | — |
| | Poly(tetrafluoroethylene) | 13 |
| | Poly(vinyl chloride) | 19.5 |

- Scientific and technical consultancies use cohesion parameters to promote their services.
- Institutions and individuals include reference to their publications or research activities in resumes and bibliographies, such as the University of Geneva, <http://www.unige.ch/sciences/pharm/fagal/barra~4.html#pubi>, and Princeton University, <http://www.princeton.edu/~chemical/faculty/pubs.html>.

As with most other areas of the physical sciences, the amount of internet-accessible material describing the **theoretical background** is limited. Probably the most comprehensive source is that written by John Burke (Oakland Museum of California), <http://sul-server-2.stanford.edu/byauth/burke/solpar/>.

For values of individual **liquids**, by far the greatest amount of quantitative information is provided in the pages of Charles Tennant & Co, <http://www.ctennant.co.uk/tenn04.htm>, with smaller numbers of compounds being listed by companies such as Shell, DuPont, Monsanto, Aeropres, Lal, and Eastman.

A smaller but increasing number of sites deal with **compressed gases and supercritical fluids**, for example, *Polymerizations in Supercritical $CO_2$* from the University of Groningen, <http://polysg2.chem.rug.nl/>.

**Solvent applications** of fluids have always been an important use of cohesion parameters, such as *A Review of Supercritical Carbon Dioxide Extraction of Natural Products from Engineering World*, <http://www.exicom.org/cew/oct97/awasthi.htm>.

**Chemical properties**, as well as physical properties, are now being correlated by means of cohesion parameters, for example, *Solution Effects on Cesium Complexation with Calixarene—Crown Ethers from Liquid to Supercritical Fluids* (University of Idaho), <http://www.doe.gov/em52/65351.html>.

**Liquid crystal** applications are still poorly represented, and studies of **solids** are mostly limited to **pharmaceuticals**, such as *Partial-Solubility Parameters of Naproxen and Sodium Diclofenac* from the *Journal of Pharmacy and Pharmacology*, <http://dialspace.dial.pipex.com/town/avenue/ax60/j98020.htm>.

As in the printed literature, applications to **polymers** dominate internet cohesion parameter sites, for example, American Polywater Corporation on polycarbonate stress cracking, <http://www.polywater.com/cracking.html>; as well as Millipore, <http://millispider.millipore.com/micro/mieliq/MAL104.htm>; and *Evaluating Environmental Stress Cracking of Medical Plastics* from Medical Device Link, <http://www.devicelink.com/mpb/archive/98/05/001.html>.

**Polymer design** is represented by IF/Prolog, <http://www.ifcomputer.com/Products/IFProlog/Applications/PolymerDesign/home_en.html>.

**Surfaces** receive less attention than homogeneous fluid and solid phases, although coatings and adhesives are important, with typical sites being University of Missouri—Rolla, <http://www.umr.edu/~wlf/Adhesion/young.html>, and Shell Chemicals, <http://www2.shell-chemical.com/CMM/WEB/GlobChem.NSF/Searchv/SC:2023-94?OpenDocument>.

**Chromatographic** properties, such as *Relationship Between Retention Behavior of Substituted Benzene Derivatives and Properties of the Mobile Phase in RPLC* in the *Journal of Chromatographic Science*, <http://www.j-chrom-sci.com/353sun.htm>, provide additional possibilities for applications of cohesion parameters to complex systems.

**Nanoparticles** are beginning to be discussed in terms of cohesion parameters, for example, on the formation of gliadin nanoparticles: the influence of the solubility parameter of the protein solvent, <http://link.springerny

.com/link/service/journals/00396/bibs/8276004/82760321.htm>, but **dyes** are surprisingly underrepresented.

Among **environmental** applications, the area of materials substitution is the most potentially productive use of cohesion parameters, for example, *Toluene Replacement in Solvent Borne Pressure Sensitive Adhesive Formulations* (Shell Chemicals), <http://www2.shellchemical.com/CMM/WEB/GlobChem.NSF/Searchv/SC:2023-94? OpenDocument>; *Solvents—the Alternatives* from the Waste Reduction Resource Center, Raleigh, NC, <http://www.p2pays.org/ref/01/00023.htm/index.htm>; and *Solvent Substitution Data Systems* (U.S. EPA), <http://es.epa.gov/ssds/ssds.html>.

**Database and compilations** include ThermoDex, <http://thermodex.lib.utexas.edu/search.jsp>, and Infochem, <http://www.infochem.demon.co.uk/data.htm>.

Examples of **modeling and simulation sites** are *SciPolymer* from ScienceServe, <http://www.scienceserve.com/wwwscivision/scipolymer/prediciti.htm>; *Molecular Analysis Pro*™, <http://www.sge.com/software/soft/13051.htm>; *The Askadskii Approach* from MillionZillion Software, <http://www.millionzillion.com/cheops/metho1/askadskii.htm>; *Cerius2* from Molecular Simulations Inc., <http://www.man.poznan.pl/software/msi-doc/cerius38/TutMats/tut_synthia.doc.htm1#740585>; <http://www.msi.com/solutions/polymers/misc.html>; and *ProCAMD* from Capec, <http://www.capec.kt.dtu.dk/capec/docs/software/procamd/procamd.html>.

QSAR (quantitative structure activity relationships) and QSPR (quantitative structure property relationships), which study relationships between useful chemical and ideal candidates for this kind of approach, for example, Molecular Analysis Pro™, <http://www.chemsw.com/13051.htm>.

While internet sources do not provide the comprehensive and integrated information available in the *Handbook of Solubility Parameters and Other Cohesion Parameters* (Barton, 1992) or even the convenient summary in the *Encyclopedia of Physical Science and Technology*, they should not be overlooked for information on recent developments.

## SEE ALSO THE FOLLOWING ARTICLES

BONDING AND STRUCTURE IN SOLIDS • GAS CHROMATOGRAPHY • HYDROGEN BONDS • LIQUIDS, STRUCTURE AND DYNAMICS • MOLECULAR HYDRODYNAMICS • SURFACE CHEMISTRY

## BIBLIOGRAPHY

Barton, A. F. M. (1983). In "Polymer Yearbook" (H. G. Elias, and R. A. Pethrick, eds.), p. 149, Harwood, Chur, Switzerland.

Barton, A. F. M. (1992). "Handbook of Solubility Parameters and Other Cohesion Parameters," 2nd ed., CRC Press, Boca Raton, FL.

Barton, A. F. M. (1990). "Handbook of Polymer-Liquid Interaction Parameters and Solubility Parameters," CRC Press, Boca Raton, FL.

Barton, A. F. M., and Knight, A. R. (1996). *J. Chem. Soc. Faraday Trans.* **92,** 753.

Burrell, H. (1975). In "Polymer Yearbook" (J. Brandrup, and G. H. Immergut, eds.), 2nd ed., IV-337, Wiley (Interscience), New York.

Hansen, C. M. (1969). *Ind. Eng. Chem. Prod. Res. Dev*. **8,** 2.

Hansen, C. M., and Beerbower, A. (1971). In "Kirk-Othmer Encyclopedia of Chemical Technology," 2nd ed. (A. Standen, ed.), Suppl. Vol., p. 889, Wiley (Interscience), New York.

Hoy, K. L. (1970). *J. Paint Technol*. **42,** 76.

Karger, B. L., Snyder, L. R., and Eon, C. (1978). *Anal. Chem*. **50,** 2126.

Rowes, R. (1985). "Chemists propose universal solubility equation," *Chem. Eng. News* (March 18), 20.



# Crystallography

**Jeffrey R. Deschamps**
**Judith L. Flippen-Anderson**
*Laboratory for the Structure of Matter, Naval Research Laboratory*

I. Introduction
II. Evolution of Crystallography
III. Structure of a Crystal
IV. Steps in Crystal Structure Analysis
V. Comparison of X-Ray and Neutron Diffraction
VI. Results
Appendix I: Factors Affecting Intensities
Appendix II: Methods of Structure Solution
Appendix III: Methods of Refinement

## GLOSSARY

**Absorption edge** Sharp discontinuity in the variation of the linear absorption coefficient with the wavelength of the incident radiation. The discontinuity occurs when the energy of the incident radiation, $E = h\nu$, matches the excitation energy of an electron in an atom of the sample.

**Anomalous dispersion** A phenomenon that influences the intensities of X-ray reflections and causes a difference in the intensity of equivalent reflections. The effect is particularly important in studies of single crystals in polar space groups and is used in some experiments to determine phase information.

**Bragg reflection** When X-rays strike a crystal they are diffracted only when the Bragg equation, $n\lambda = 2d \sin\theta$ (where $n$ is an integer and $d$ is the spacing of a set of lattice planes), is satisfied. The diffracted beam is considered a *reflection*.

**Bravais lattice** One of the 14 possible arrays of points repeated periodically in three-dimensional space such that the arrangement of points about any one point is identical in every respect to the arrangement of points about any other point in the lattice.

**Centrosymmetric** A structure or space group containing an inversion center is centrosymmetric, if there is no inversion center it is noncentrosymmetric.

**Diffractometer** An instrument used to measure the position (i.e., Bragg angle) and relative intensity of the diffraction pattern produced by a crystalline material.

**Lattice** Any repetitive pattern can be described by noting the motif (the unit of pattern that repeats by translation) and the translation interval. In the case of a three-dimensional pattern such as a crystal, the lattice describes translations in three dimensions. It is an imaginary, mathematical construct characterized by three translations, **a**, **b**, **c**, and three angles, $\alpha$, $\beta$, $\gamma$.

**Miller indices** A set of integers with no common factors, inversely proportional to the intercepts with the crystal axes of a lattice plane.

**Orientation matrix** A matrix relating the crystal axes to the instrument axes such that one can predict the values of the instrument angles ($2\theta$, $\omega$, $\chi$, and $\Phi$) for a given reflection of the crystal.

**Patterson function** A Fourier summation that uses the squares of the structure factor magnitudes as coefficients. The peaks in this map correspond to vectors between atoms. The peak height is related to the scattering powers of the atoms at the two ends of the vector. The region around the origin gives information about bonded distances.

**Phase problem** A central problem of crystallography. The intensities of the different reflections allow derivation of the amplitude of the structure factors but not their phases. The phases are required in order to calculate the electron density, which is a "map" showing the position of atoms in the unit cell.

**Point group** A group of symmetry operations that leave unmoved at least one point within the object to which they apply.

**Polar space group** Space group in which the origin is not fixed by symmetry and hence must be defined (e.g., the space group $P2_1$).

**Reciprocal lattice** A set of imaginary points constructed in such a way that the direction of a vector from one point to another coincides with the direction of a normal to the real space planes within the crystal. The separation of those points (absolute value of the vector) is equal to the reciprocal of the real inter-planar distance.

**Space group** Identical atom groups are usually symmetrically arranged within the crystal lattice. The symmetry relating the groups may be due to rotations, inversions, mirror planes, or some other relational operation. The space group constitutes a mathematical shorthand description of the symmetry operations required to produce the unit cell.

**Special position** A point left invariant by at least two symmetry operations of the space group.

**Structure factor** $F_{hkl}$, complex quantity corresponding to the amplitude and phase of the diffraction maximum associated with the reciprocal lattice point *hkl*:

$$F_{hkl} = \sum_{j=1}^{N} f_j \exp[2\pi i(hx_j + ky_j + lz_j)]$$

where $N$ is the number of atoms in the unit cell and $x_j$, $y_j$, $z_j$ are the fractional coordinates of the $j$th atom.

**Torsion angle** If a group of four atoms (ABCD) is projected onto a plane normal to the bond between B and C, the angle between bonds connecting A and B, and C and D is the torsion angle.

**Unit cell** Parallelepiped bounded by three noncoplanar vectors **a**, **b**, **c** with angles $\alpha$, $\beta$, $\gamma$ that repeats by translation. If this unit is the smallest volume that meets these criteria it is referred to as the *primitive unit cell*.

**X-ray** Electromagnetic radiation with wavelengths in the range 0.01 to 1.0 nm (0.1 to 10 Å). The wavelength range most commonly used in diffraction experiments is between 0.71 and 1.54 Å. Shorter wavelengths require longer path-lengths (distance from crystal to detector) in order to resolve adjacent peaks in the diffraction pattern. Both the sample and the air along the beam path can significantly attenuate longer wavelengths.

**CRYSTALLOGRAPHY** is a broadly encompassing discipline that involves a variety of fields of study. The primary concern of modern crystallography is the three-dimensional arrangement of atoms in matter. Although the term most often refers to studies of crystalline solids (either single crystals or crystalline powders) using X-ray or neutron diffraction, it encompasses a much broader range of methodologies.

## I. INTRODUCTION

In the late 1660s, crystallography began as the study of the macroscopic geometry of crystals. Crystals were grouped into systems on the basis of the symmetry of their external shapes. Based on these observations, seven crystal systems were identified: triclinic, monoclinic, trigonal, tetragonal, hexagonal, orthorhombic, and cubic. It was theorized that the observed crystal shapes could be built up by stacking minute balls. This idea was refined, and in the late 18th century it was thought that crystals were composed of elementary building blocks (now referred to as the *unit cell*). This was supported by the orderly cleavage angles of calcite, which suggested a regular stacking of these elementary blocks. Further studies in the 1800s led to derivation of the 32 different *point groups*, the *Bravais lattices*, and the 230 *space groups*. All of these advances were made before any direct observations of the arrangement of atoms within a crystal were possible.

In 1912, Von Laue reasoned that the arrangement of atoms in crystals could help him measure the wavelength of X-rays. Based on experiments with copper sulfate, he demonstrated the ability of crystals to act as three-dimensional diffraction gratings. Crystallography had entered a new era: the analysis of the arrangement of atoms in a crystal by careful analysis of the diffraction pattern of that crystal. Crystal structures were now viewed as being built up from repeating units of an atomic pattern rather than the regular stacking of solid shapes.

As new types of scattering were discovered, they became incorporated into the discipline, and crystallography came to include structural studies of all classes of substances. Absorption, diffraction, or other scattering methods are used to study crystals, powders, amorphous materials, surfaces, liquids, and gases. The physics and methodologies associated with these techniques are also part of the science of crystallography.

Crystallographic studies play a vital role in materials science, chemistry, pharmacology, mineralogy, polymer science, and molecular biology. Accurate knowledge of molecular structures is a prerequisite for rational drug design and structure-based functional studies. Crystallography is the only method for determining the "absolute" configuration of a molecule. Absolute configuration is a critical property in biological systems, as changes in this may alter the response of the biologic system.

A requirement for the high accuracy of crystallographic structures is that a good crystal must be found, and this is often the rate-limiting step. Additionally, only limited information about the dynamic behavior of the molecule is available from a single diffraction experiment.

In the past three decades, new developments in detectors, increases in computer power, and powerful graphics capabilities have contributed to a dramatic increase in the number of materials characterized by crystallography. Synchrotron sources offer the possibility of time-resolved studies of physical, chemical, and biochemical processes in the millisecond to nanosecond range; the ability to study the nearest neighbors of cations present at parts per million concentrations; and the possibility of recording small-angle scattering data and powder data in seconds. Charge-density studies have been made on numerous light atom structures and are beginning to provide new insights into bonding of transition metals. Rietveld refinement is revolutionizing the study of powders and is being extended to fibers. Direct methods of structure solution are being applied successfully to structures of over 1000 atoms. The Human Genome Project has created many opportunities for crystallographic studies of biological macromolecules and resulted in intense activity in the areas of structural genomics and proteomics. Polarized neutrons are being used to determine the spin structure of magnetic materials and to probe the surface structure of such materials. The review that follows attempts to describe the fundamentals of crystallography and capture the excitement and diversity within the discipline.

## II. EVOLUTION OF CRYSTALLOGRAPHY

It was not very long ago that X-ray diffraction data were collected on photographic film, intensities of spots on the film (corresponding to data points) were "measured" by eye, and Fourier transforms were performed with Beevers–Lipson or Patterson–Tunell strips and summed by hand. Three-dimensional electron-density maps were plotted by hand, one section at a time, traced onto glass sheets, and stacked in frames for interpretation. As long as all calculations were done by hand, small, flat molecules were the ones most amenable to study. Early computers were difficult to program and had limited storage capacity, but they made it possible to solve and refine crystal structures of molecules of moderate size in 6 months to a year.

The rapid advances in crystallography owe much to the development of computer-controlled diffractometers for data collection, high-speed computers for data analyses, and, most recently, powerful graphics devices for displaying structures with the ability to perform real-time manipulations. Tasks that required months of effort can now be accomplished in minutes with the aid of a computer.

### A. The Early Years

The first experiments in X-ray diffraction were recorded on film. By 1913, W. H. Bragg had constructed the first "X-ray spectrometer" to allow a more careful study of X-rays. This instrument also proved useful in studies on crystals. Using measurements made with this X-ray spectrometer, Bragg's son determined the structures of fluorspar, cuprite, zincblende, iron pyrites, sodium nitrate, and the calcite group of minerals.

The first crystal structures reported were of substances crystallizing in cubic space groups. The structure of diamond was determined in 1913 by the Braggs, from symmetry considerations using the observed intensities to discriminate among possible structures. Their model established the carbon–carbon single bond distance of 1.52 Å and confirmed that bonds to carbon are directed tetrahedrally. The younger Bragg combined symmetry arguments with the notion that the scattering power of atoms is related to their atomic weight to explain the structures of the alkyl halides. These concepts were extended to ZnS, $CaF_2$, and $FeS_2$. The alkyl halide structures show anions surrounded by cations and cations surrounded by anions, demonstrating conclusively that discrete "molecules" of the type NaCl do not exist in crystals of ionic materials. Pauling published the first intermetallic structure, $Mg_2Sn$, in 1923. In that same year, Dickerson and Raymond showed that hexamethylene tetramine consisted of discrete molecules, each having the same structure, with C–N distance of 1.44 Å in a body-centered cubic lattice of edge 7.02 Å.

#### 1. Heavy-Atom Methods

The elder Bragg realized that the periodic pattern in the electron-density distribution could be represented by a

Fourier summation. The coefficients of this summation became known as the *structure factors*. This allowed the solution of structures where positions of atoms were not restricted to special positions. The determination of the structure of diopside, $CaMg(SiO_3)_2$, in 1928 was the first example of the use of this method.

In 1934, Patterson showed that a Fourier series with $|F|^2$ as coefficients could be summed without knowledge of the phases and would reveal interatomic vectors. Since the weight of a peak in the Patterson function is proportional to the product $Z_i Z_j$ of the atomic numbers of the atoms at the ends of the vector, vectors involving heavy atoms stand out among light atom–light atom vectors. In 1936, Harker showed that symmetry properties of the crystal caused vector density to accumulate on certain planes and lines (later known as the Harker section) in the Patterson function. These two papers were the foundation of the heavy-atom method for crystal structure solution. The method assumes that phases calculated from the heavy-atom positions will be sufficiently accurate that a Fourier synthesis (using $|F|$ as opposed to $|F|^2$) will reveal the positions of more atoms, thus allowing solution of the structure. Phase information from the new atoms could then be added to the Fourier synthesis to locate more new atoms and so on, until the full structure was revealed.

The method of isomorphous replacement was used first to solve the structure of the alums by Lipson and Beevers in 1935. If a centrosymmetric light-atom structure and its heavy-atom derivative differ only in the presence of the heavy atom, then the differences in the intensity of equivalent reflections can be used to determine the signs of the structure factors. Robertson applied this method to the phthalocyanines in 1936. These molecules form an isomorphous series and crystallize in $P2_1/a$. By comparing intensities of nickel phthalocyanine and the unsubstituted molecule, Robertson was able to assign phases to all but a few of the measured reflections from the *h0l* projection. Harker generalized the method to the noncentrosymmetric case in 1956. Isomorphous replacement can be combined with anomalous dispersion to obtain phase information for large molecules.

### 2. Absolute Configuration

Friedel's law states that the scattering from the front and back sides of a plane, *hkl*, are the same. This means that the measured intensities of the 111 and $\overline{111}$ reflections (and all other "Friedel" pairs) should be equivalent. However, in a 1930 study on ZnS by Coster, Knol, and Prins, it was noted that the 111 reflection was not equivalent to its Friedel mate. In the case of the ZnS crystal, the {111} faces are prominent but, even by visual examination, do not appear identical. One face is shiny and the other is dull. The crystal structure can be regarded as alternating layers of Zn and S atoms perpendicular to the {111} direction. Looking at one layer of Zn atoms, we find it lies closer to one of the two adjacent layers of S atoms. If we assume that the short Zn–S spacing is a "bonding" interaction and the long Zn . . . S gap is a van der Waals contact, we do not expect cleavage between bonded layers. This implies that one of the {111} faces corresponds to a layer of S atoms and the other to a layer of Zn atoms.

It was fortuitous that the intensities in the Coster, Knol, and Prins study were measured using AuL$\alpha$ radiation ($\lambda = 1.276$ Å). The K-absorption edge of Zn is 1.283 Å. Thus, the measured intensities of Friedel pairs were not the same. This was the first example of anomalous scattering. The difference due to anomalous scattering is greatest when data are collected near an absorption edge of a heavy atom in the structure. It was nearly 20 years after the ZnS experiment that Bijvoet realized this principle could be used to determine the absolute configuration of the sodium rubidium salt of (+)-tartaric acid.

## B. Modern Crystallography

The early years can be characterized as a period in which the size of a problem amenable to analysis was computationally and instrumentally limited. Data collection and analysis were manual processes. This began to change in the 1950s.

### 1. Data Collection

Progress in crystallographic data collection can be charted by examining early issues of *Acta Crystallographica*, published by the International Union of Crystallography (IUCr). Although a manual "diffractometer" was available in 1913, the primary method for collecting intensity data for crystallographic studies relied on X-ray cameras and film. In 1948, about two-thirds of the crystal structures reported used a camera and film to collect the intensity data, and less than one-quarter used a combination of film and a diffractometer. Despite the existence of automated single-crystal diffractometers, the situation was little changed by 1962.

Use of the Cambridge Structure Database (CSD) allows a more systematic study of the evolution of data collection methods (Fig. 1) with the caveat that this database references only organic and organometallic structures. Although automated diffractometers were first available in about 1955 it was not until the mid-1960s that this new technology made much of an impact. At that time, only a few hundred structures were being added to the CSD each year. By the mid-1970s, over 1000 structures were added to the CSD each year and data for over 80% were

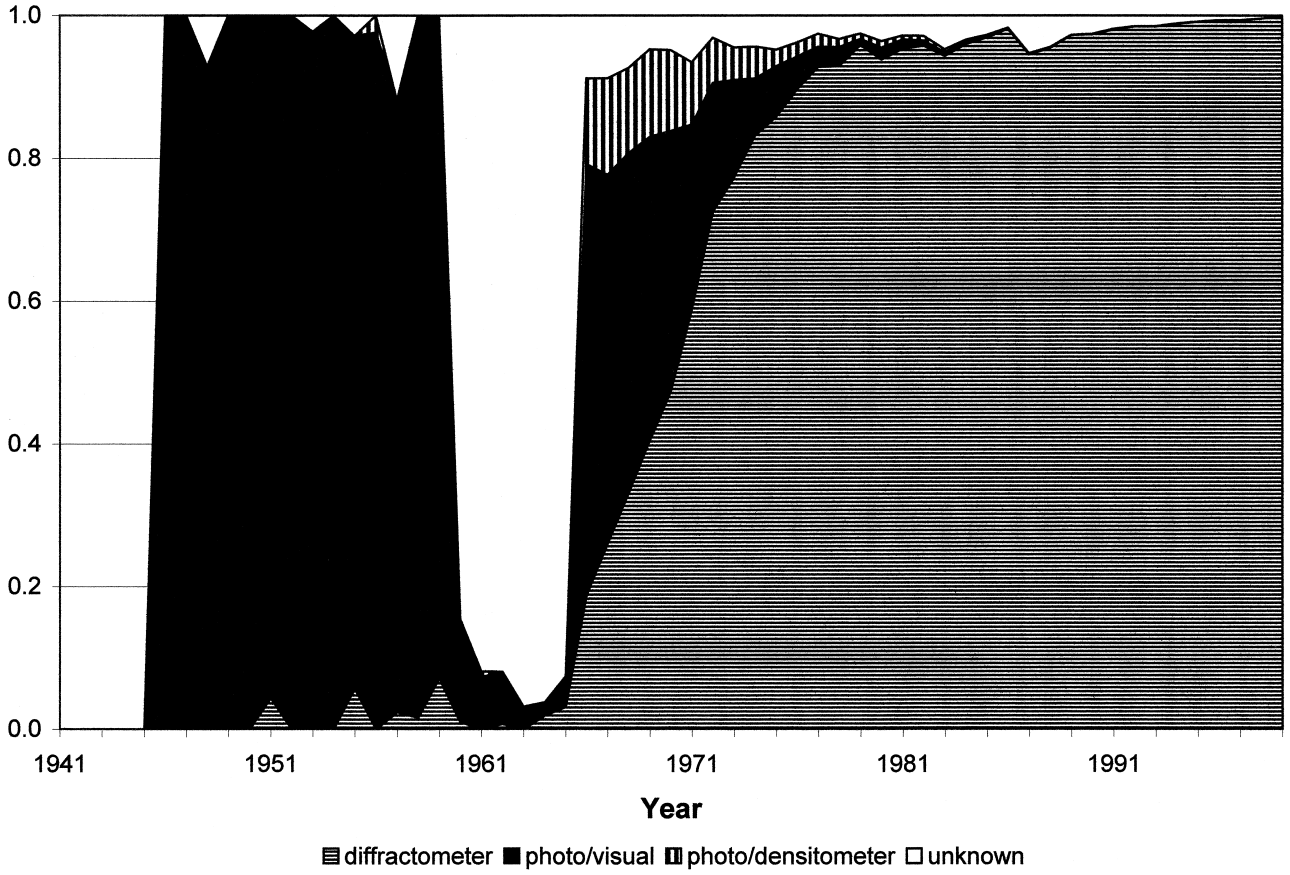



**FIGURE 1** Changes in data-collection methods with time. Using structures entered in the Cambridge Structure Database, changes in data collection strategies were determined by examining the fraction of structures reported in one year for various data collection methods (photograph with visual estimation of intensities, photograph with densitometer, diffractometer, or unknown). The large fraction listed as unknown between 1960 and 1965 is likely due to changes occurring at that time in how data were collected and how data-collection methods were reported.

collected on a diffractometer. In 1997, over 15,000 structures were added to the CSD and virtually all were collected on a diffractometer. The explosive growth of this database cannot be attributed solely to improvements in data collection, but certainly the routine use of automated instrumentation had a significant impact.

### 2. Structure Solution

Data collection was not the only beneficiary of the post-World War II progress in crystallography. No general method existed for solving unknown structures without heavy atoms until the advent of direct methods—a means of determining the values of phases from relationships among the structure factor magnitudes associated with those phases. The earliest structure solved by such methods was decaborane by use of some inequalities derived by Harker and Kasper. However, this technique was limited to centrosymmetric structures. The major effort in the 1950s concerned the development of the mathematical aspects of crystal structure analysis. The first general procedures for solving both centrosymmetric and noncentrosymmetric structures was developed in the early part of the 1960s. Use of this method grew with the power of computers and computer programs. It is now the most widely used method of solving crystal structures of moderate size. Efforts are currently being made to apply direct methods to very large structures such as proteins. For a more detail discussion of structure solution methods, see Section IV.H.

### 3. Charge-Density Distribution

The first papers to explore the difference between the results of X-ray and neutron diffraction experiments appeared in the early 1970s. Systematic differences between the positional and thermal parameters determined by the two techniques were reported. These differences were attributed to the difference in how neutrons and X-rays

interact with atoms. Neutrons are diffracted primarily from the nucleus; hence, neutron diffraction produces information about nuclear position. X-rays are diffracted by electrons and therefore yield information about the distribution of electrons in the molecule. It was logical to extend these ideas and attempt to map the redistribution of electron density that occurs on bonding.

A number of different kinds of mapping have been done. Subtracting core-electron densities from experimental electron densities (i.e., $p(\mathbf{r})$) should reveal details of the redistribution of valence electrons on bonding. The term "valence density" is used to describe the difference function:

$$p_{\text{valence}}(\mathbf{r}) = p(\mathbf{r}) - \sum_{\text{atoms}} p_{i,core}(\mathbf{r})$$

To define either of these functions the positions and thermal parameters must be known. One way to approach this is to use neutrons to determine the positional and thermal parameters and to use these parameters in conjunction with X-ray, spherical-atom scattering factors to calculate structure factors for the "promolecule" using:

$$F_{\text{calc},N} = \sum_{\text{atoms}} f_i \exp(2\pi i \mathbf{H}\cdot\mathbf{r}_i) T_i$$

$F_{\text{calc},N}$ is the X-ray structure factor calculated from neutron positions. The deformation density or $X$–$N$ map then corresponds to:

$$p^{X-N}_{\text{deformation}}(\mathbf{r}) = \frac{1}{V} \times \sum_H (F_{\text{obs},X} - F_{\text{calc},N}) \exp(-2\pi i \mathbf{H}\cdot\mathbf{r})$$

Since both $F_{\text{obs}}$ and $F_{\text{calc}}$ contain the effects of thermal motion, this deformation density map is thermally smeared. Its resolution is limited by the maximum $(\sin\theta)/\lambda$ value of the data.

Given that there are few neutron-diffraction facilities in the world and that it is difficult to correct adequately for systematic effects in the two experiments (i.e., absorption, extinction, thermal diffuse scattering, multiple reflections), it would be desirable to study bonding effects using exclusively X-ray data. There are several approaches to this problem.

The $X$–$X$ formalism is similar to the $X$–$N$ formalism described above, except that the calculated values for positional and thermal parameters are derived from refinement of high-angle X-ray data [$(\sin\theta)/\lambda > 0.70$ Å$^{-1}$]. For deformation density maps, neutral spherical atoms are subtracted from the observed density; in valence density maps, Hartree–Fock core-electron densities are used to evaluate $F_{\text{calc}}$. Comparison of $X$–$X$ and $X$–$N$ maps shows that they do indeed yield the same qualitative information. Bonding density shows in the middle of bonds, and lone-pair density is in the correct location. However, $X$–$X$ maps systematically underestimate lone-pair peak heights and place hydrogen atoms too close to the atoms to which they are bonded. The experiments must be conducted at low temperature (i.e., $-75^\circ$C or preferably less) for this method to succeed.

Other approaches include refinement of separate parameters for core and valence electrons or the direct refinement of a deformation model. The major advantage of the direct refinement methods is that they make no assumptions about the $(\sin\theta)/\lambda$ dependence of bonding features.

Errors in the deformation density maps arise from errors in both the model and the data. Demands on diffraction methodology and interpretation are many times more severe than those relevant to an average stereochemical investigation. The following considerations are extremely important:

1. The X-ray data set must be complete (all symmetry-related reflections measured) to a $(\sin\theta)/\lambda$ limit of about 1.3 Å$^{-1}$. This implies the use of short-wavelength radiation such as MoK$\alpha$ ($\lambda = 0.7107$ Å) or AgK$\alpha$ ($\lambda = 0.5612$ Å). The maximum value of $(\sin\theta)/\lambda$ is 0.65 for CuK$\alpha$ ($\lambda = 1.5418$ Å).
2. The data must be corrected for absorption, extinction, and thermal diffuse scattering. Multiple reflections must be either avoided or eliminated. If these conditions are met, internal agreements should be of the order of 2%; i.e.,

$$\left[\sum_H F^2(\mathbf{H}) - \langle F^2(\mathbf{H})\rangle\right] \Big/ \sum_H F^2(\mathbf{H}) = 0.02$$

## 4. Rietveld Analysis

There has been a renaissance in powder diffraction in recent years because Rietveld refinement allows determinations of positional and thermal parameters from powder data, even when the diffraction peaks are not well separated in the recorded pattern. Rietveld analysis is not a method of structure solution and can only be applied when cell dimensions and space group are known and when a reasonable model exists for the structure.

In a polycrystalline sample, information may be lost as a result of the random orientation of the crystallites. A more serious loss of information can result from the overlap of independent diffraction peaks in the powder pattern. Using the total integrated intensities of the separate groups of overlapping peaks in the least-squares refinement of a structure leads to the loss of all the information contained in the often-detailed profile of these composite peaks. Rietveld developed a refinement method that uses the profile intensities of the composite peaks instead of the integrated quantities. This is a pattern-fitting method

of structure refinement and allows extraction of the maximum amount of information contained in the powder pattern.

A powder pattern is recorded in a step-scan mode with a step width of 0.02 to 0.03° $2\theta$. No attempt is made to allocate observed intensity to individual reflections or to resolve overlapping reflections. Instead, the intensity of the powder diffraction pattern is calculated as a stepwise function of the angle, $2\theta$. Refinement allows calculation of the shifts in the parameters that will improve the fit of the calculated powder pattern to the observed one. The quantity minimized is

$$\sum_i w_i[y_i(\text{obs}) - y_i(\text{calc})]^2$$

where $y_i$(obs) is the observed intensity of the $i$th step of the profile, and $y_i$(calc) includes the usual structural parameters (i.e., positional parameters $x_i$, $y_i$, $z_i$; thermal parameter $B_{ij}$; and site-occupancy parameters $p_j$). However, the model must also include instrumental and sample parameters: $2\theta_0$ (overall scale), overall temperature factor, profile breadth ($\mathbf{H}^2 = \text{U} \tan 2\theta + \text{V} \tan\theta + \text{W}$), profile asymmetry, background, preferred orientation, lattice parameters, and wavelength. The agreement factors most often quoted are

$$R_{\text{weighted pattern}} = \left[\frac{\sum w_i[y_i(\text{obs}) - (1/c)y_i(\text{calc})]^2}{\sum w_i[y_i(\text{obs})]^2}\right]^{1/2}$$

and

$$R_{\text{Bragg}} = \frac{\sum |I_k(\text{“obs”}) - I_k(\text{calc})|}{\sum I_k(\text{“obs”})}$$

In $R_{\text{Bragg}}$, “obs” has quotation marks because $I_k$(“obs”) is calculated by partitioning the intensity.

To date, most of the papers published using this method have been neutron-diffraction studies from reactor sources. Advantages of neutron data include minimal preferred orientation, no polarization, and neutron absorption cross-sections smaller than X-ray values by a factor of $10^4$, scattering independent of $\theta$, and for fixed-wavelength experiments the peak shape is simple.

In X-ray experiments using radiation from conventional sources, peak shape is complicated by both $\alpha_1$–$\alpha_2$ splitting (at high angles) and by the fact that the peak shape is neither Lorentzian nor Gaussian, but is better described by a convolution of these two functions called the *Voigt function*. The pseudo-Voigt function used in many programs is an approximation of the Voigt function that can be evaluated much more quickly.

Spallation neutron sources, time-of-flight experiments with reactor sources, and synchrotron sources all have special problems in defining peak shapes, and this limits the precision of the resulting parameters. However, precision comparable to single-crystal X-ray diffraction can now be obtained from neutron diffraction at a fixed wavelength with Rietveld refinement. For X-rays from conventional sources, the precision of positional parameters is comparable to the single-crystal case, but thermal parameters are less reliable by a factor of two or three. Considerable effort is being expended to improve profile functions for the various X-ray and neutron sources.

## 5. Small-Angle Scattering

Scattering at small angles is derived from large structural units—units whose dimension $D$ is much larger than the wavelength of the radiation used in the experiment. The acronyms SAXS and SANS refer to small-angle X-ray scattering and neutron scattering, respectively.

Different sorts of small-angle experiments are typical of the kind of material studied and yield a characteristic type of pattern. Low-angle data from ordered or semi-ordered systems give Bragg peaks at specific values of scattering vector. Examples include aligned structures with long-range periodicity, such as two-dimensional biological structures or samples such as opal that present long-range order.

Scattering from polymers in dilute solution or from biological materials yields patterns that look rather like the Debye–Scherrer rings observed in wide-angle data from powders. When scattering arises from the spherical particles in a mono-disperse system, the pattern is a Bessel function consisting of a succession of peaks of diminishing magnitude that are broad relative to Bragg peaks. Analysis of the pattern yields the radius of the particles.

In some circumstances, multiple Bragg reflections give rise to scattering in the small-angle region. One of the advantages of the “tunable” sources of X-rays and neutrons is that multiple Bragg scattering can be avoided by choosing wavelengths larger than the lattice spacing.

Small-angle scattering can be applied to a wide variety of materials. In polymer science, it has been used to investigate chain conformation in amorphous polymers, the state of mixing in polymer blends, the compatibility ranges of polymer blends, and the measurement of domain structure and molecular conformation within those domains. In biological materials, examples include measurement of the radius of gyration of proteins in solution; aggregation of chlorophyll into micelles; diffraction patterns of semi-ordered materials such as muscle, collagen, etc.; and studies of the shapes and constitution of viruses. Separation processes, such as those used in refining of metals or extraction of tar sand, frequently involve micelle formation. Small-angle scattering can give information regarding the

shape, size, and degree of polymerization of the aggregates. Studies of materials such as cements, zeolites, and catalysts involve the measurement of size and distribution of pores and measurements of specific surface.

## 6. Extended X-Ray Absorption Fine Structure

Extended X-ray absorption fine structure (EXAFS) is a technique for studying the local environment of a specific atomic species in a complex matrix. Because the interaction of X-rays with the material under study is absorption rather than diffraction, the technique can be applied to gases, liquids, and amorphous solids as well as to crystals.

In the experiment, the X-ray absorption coefficient is measured from slightly below to about 1000 eV above the absorption edge for the atomic species whose environment is to be studied. Through analysis of the fine structure above the edge it is possible to determine the coordination of the atom. The actual position of the edge gives information about the oxidation state of the absorbing atom, while structure at energies at or just below the edge provides information about bound states associated with the absorbing atom and hence about the symmetry of the environment. The technique can be applied to any atom with $Z > 15$ to yield a determination of radial distance to a precision of $\pm 0.02$ to 0.01 Å. For elements with high values of $Z$, it may be preferable to use the L-absorption edge rather than the K edge because higher fluxes are available; however, theory for the L edge is not yet well developed.

Extended X-ray absorption fine structure has applications in many branches of science. In molecular biology, it has been used to study $Ca^{+2}$ transport in membranes, binding of oxygen in hemoglobin, and other coordination problems. It is an invaluable tool for the study of amorphous substances such as glass, since the manufacture of glasses with particular mechanical and thermal properties depends on structure. Catalysis has become extremely important in energy development, resource utilization, pollution abatement, refining of metals, etc. The chemical state and atomic environment of an atomic species in a catalyst *in situ* while reduced with hydrogen, chemisorbed with oxygen, heated, quenched, etc., can be determined with EXAFS. This allows the design of heterogeneous catalysts that are tailored from precise knowledge of electronic and structural parameters.

Synchrotron sources have revolutionized EXAFS studies. The intensity of the source and high collimation make it possible to collect the relevant data in about 20 minutes in a sample such as Cu metal with a resolution of 1 eV at 8.8 keV. To collect similar data with a rotating anode source would take about 2 weeks, and the precision would be reduced by a factor of about 100.

## 7. Implications of New X-Ray and Neutron Sources

Synchrotron sources present a unique combination of properties that are very attractive for X-ray scattering, absorption, and diffraction experiments. The radiation produced has extreme brightness over a broad spectral range. Synchrotron sources are five to six orders of magnitude brighter than the bremsstrahlung (that part of the X-ray spectrum caused by the slowing of electrons on impact with the target, also referred to as *white radiation*) available from a conventional rotating anode source. Monochromators currently available produce resolution of 0.1 eV at 8 keV. The radiation is naturally collimated with a divergence of the order of $2 \times 10^{-4}$ radians, is plane polarized, and has a precise time structure (subnanosecond pulses repeated every 0.5 to 1 $\mu$sec). These properties allow experiments that simply cannot be done with conventional sources: EXAFS on dilute samples (parts per million range), measurement of the magnitude and angular dependence of the real and imaginary components of anomalous dispersion, and determination of the structure of a protein using only one derivative and three wavelengths. (In this context, the use of anomalous dispersion is formally equivalent to multiple isomorphous replacement with the added feature that the isomorphism is exact.) Determinations of cation-site distribution in minerals and diffraction from monolayers on surfaces have many applications in such areas as catalysis or materials science. Perhaps the most exciting application is the ability to do time-resolved studies of physical, chemical, and biological processes using small-angle scattering, powder diffraction, and other scattering techniques.

One early example was the study by Larson of temperature and temperature gradients in silicon during pulsed-laser annealing. In this example, the duration of the laser pulse was 15 nsec and that of the synchrotron X-ray pulse was 0.15 nsec. The laser bursts were synchronized so that the probing X-rays arrived at 20, 55, and 155 nsec after the laser pulse. The experiments showed that the lattice temperature of silicon reaches the melting point during the 15-nsec pulse and remains at the melting point during the high reflectivity phase, after which time the temperature rapidly subsides. Temperature gradients at the liquid–solid interface were measured for the first time and were found to be in the range of $10^7$ °K/cm. Larson received the Warren Award for Diffraction Physics for this pioneering work in nanosecond, time-resolved X-ray diffraction.

The new pulsed spallation sources (such as the European Spallation Neutron Source at Forschungszentrum Jülich GmbH, or the Spallation Neutron Source at Oak Ridge National Laboratory) provide spectra substantially richer in short-wavelength neutrons than those available

from the reactor sources. Pulse duration and repetition are source parameters, but the time structure can be exploited by a variety of techniques. Essentially the same types of experiments are done at both the reactor and spallation sources. The higher neutron fluxes available with the new sources allow experiments to be done on smaller samples and/or in shorter times than was previously possible.

### 8. Contribution of Diffraction to Molecular Biology

Molecular biologists seek to unravel the mysteries of the cell by mapping gene location, function, and control. By understanding all of these we gain insight into how and when genes are turned on and how these might be used to perform useful tasks. One goal of such studies is to alter (i.e., reengineer) a natural biologic process to perform some other function. Examples of this include attempts to modify metal-binding proteins such that a protein that was originally selective for calcium is selective for zinc or copper. This reengineered metal-binding protein could then be used to construct a sensor for zinc or copper. Other examples include modifying enzymes for use in industrial processes rather than biologic processes. In order to effect these changes, the relationship between structure and function must be understood.

The structure of a macromolecule can be "known" at a number of levels. Primary structure is the linear sequence of building blocks (i.e., amino acids) from which the protein is built. For example, the $\beta$ chain of human hemoglobin contains 144 amino acids, of which the first five have the sequence valine, histamine, leucine, threonine, and proline. The term "secondary structure" refers to local interactions that determine the conformation of the polypeptide chain and the interchain hydrogen bonding scheme. "Extended chain," "$\alpha$-helix," and "$\beta$-sheet" are terms used to describe secondary structure. Tertiary structure is the three-dimensional arrangement of atoms within the macromolecule, while quaternary structure describes the arrangement and interaction of aggregates of the macromolecules themselves. Diffraction techniques are widely used to study the secondary, tertiary, and sometimes quaternary structures of macromolecules.

Single crystals of macromolecules may be studied by X-ray diffraction. A wide variety of techniques are now available for macromolecular structure solution. In the past, heavy-atom multiple isomorphous replacement (MIR) was the most common method of structure solution; however, anomalous dispersion is becoming more common as a method of structure solution. The relatively small protein, crambin, was an early example of using anomalous dispersion to effect structure solution of a macromolecule. In a few cases, neutron-diffraction studies have been carried out on single crystals of proteins. In myoglobin, for example, at 1.8 Å resolution, the negative density of the protons (which form about half of the scattering material in the cell) makes the polypeptide chain stand out clearly in the Fourier maps. Soaking the same crystal in $D_2O$ has allowed identification of exchangeable protons.

Neutron- and X-ray-scattering experiments using non-Bragg scattering can be used to study the size, shape, and aggregation of micelles. Contrast variation can be used to study the internal structure of viruses.

The average scattering densities from the protein coat and the RNA interior of viruses are different. Each can be "matched" by a different $H_2O/D_2O$ ratio. Thus, for virus particles in solution, the matched phase becomes "invisible" to the neutron beam and allows the radial distribution of scattering of the other component to be recorded. For spherical viruses, this allows measurement of the thickness of the protein coat and the degree of interpenetration of the two phases. Enzymes and other proteins bind some substrates so well that detailed, atomic-level analyses of the structure of the native protein can be compared with those of protein–substrate complexes, protein–inhibitor complexes, and proteins with catalytic groups bound to allosteric sites. Detailed comparisons between such structures have greatly enhanced our understanding of the mechanisms of biological catalysis.

Molecular graphics programs are now available that can display the full three-dimensional structure of a macromolecule and zoom in on any portion of it. It is possible to examine the active site and to attempt to fit known substrates and/or inhibitors into that site. The availability of coordinates for macromolecules in the Protein Data Bank (see Section VI.A.8) has allowed many fruitful applications of this type.

## III. STRUCTURE OF A CRYSTAL

### A. Choice of Unit Cell

A crystal is a multifaceted solid, similar in appearance to an unpolished gemstone. Internally it consists of a basic pattern, known as the *repeating unit*, of molecules that repeats itself by translation, in three dimensions, to the edges of the crystal. By choosing one corner of the repeating unit to be the origin, one can use three translational vectors, having both length and direction, to construct a parallelepiped that contains the entire basic pattern. This parallelepiped is defined as the *unit cell* and one need only determine the contents of the unit cell, the basic pattern, to know the structure of the entire crystal. The general symbols for the unit-cell vectors are **a**, **b**, **c** and for their magnitudes $a$, $b$, $c$. The coordinate axes, or directions of the sides of the unit cell, are referred to in general as the

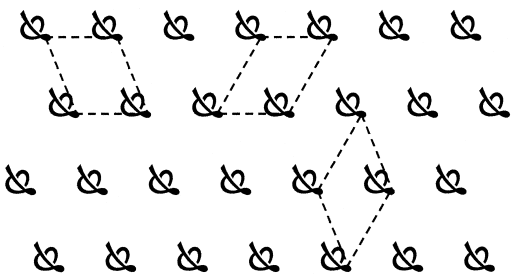


**FIGURE 2** Choice of unit cell. In the absence of symmetry, the unit cell may be chosen in a variety of ways. Each cell contains one unit of pattern. All such cells have the same volume.

$x$, $y$, and $z$ axes. The interaxial angles are denoted by $\alpha$, $\beta$, and $\gamma$. The unit cell can be defined in a variety of ways, but for a given pattern, the volumes of each cell will be equal (Fig. 2).

If the repeating unit itself has no internal symmetry, then the choice of unit cells is infinite. However, if the basic repeating unit contains additional symmetry, this influences the choice of the unit cell. If there are planes or axes of symmetry, the cell edges are generally chosen to be parallel or perpendicular to these directions. This places restrictions on how the unit cell may be defined and gives rise to seven crystal systems: triclinic, monoclinic, orthorhombic, tetragonal, hexagonal, trigonal, and cubic (see Table I).

## B. Diffraction Pattern

Friedrich, Knipping, and Laue first demonstrated the fact that crystals could act as three-dimensional diffraction gratings for X-rays in 1912. This work not only established the wave nature of X-rays but also established the relationship between a crystal and its diffraction pattern. The size and shape of the repeating unit in the crystal determine the position of the diffraction spots they recorded on film. It was some time later before it was realized that the intensity of the spots is related to the distribution of atoms within the unit cell.

**TABLE I Seven Crystal Systems and Their Unit Cell Constraints**

| Crystal System | Conditions imposed on cell geometry |
|---|---|
| Triclinic | $a \neq b \neq c$; $\alpha \neq \beta \neq \gamma$ |
| Monoclinic | $a \neq b \neq c$; $\alpha = \gamma = 90^\circ$ |
| Orthorhombic | $a \neq b \neq c$; $\alpha = \beta = \gamma = 90^\circ$ |
| Tetragonal | $a = b$; $\alpha = \beta = \gamma = 90^\circ$ |
| Trigonal | $a = b$; $\alpha = \beta = 90^\circ \gamma = 120^\circ$ (hexagonal axes) |
| | $a = b = c$; $\alpha = \beta = \gamma$ (trigonal axes) |
| Hexagonal | $a = b$; $\alpha = \beta = 90^\circ \gamma = 120^\circ$ |
| Cubic | $a = b = c$; $\alpha = \beta = \gamma = 90^\circ$ |

Bragg's contribution was to recognize the similarity between diffraction in a crystal and reflection in a mirror plane (Fig. 3). Consider a set of parallel planes with spacing $d$ and an incoming beam of monochromatic X-rays at a glancing angle, $\theta$. The condition for constructive interference is that the path difference between waves "reflected" from successive planes must be an integral number of wavelengths (i.e., $AB + BC = n\lambda$). However, $AB = BC = d \sin\theta$; thus, $n\lambda = 2d \sin\theta$, which is Bragg's law. Note that the angle between the incident beam direction and the reflected beam is $2\theta$.

The smaller the interplanar spacing $d$, the higher the angle at which the diffraction maximum or "reflection" is observed. This implies that large unit cells will give diffraction patterns with small spacing between the spots, while small unit cells will give patterns with wide spacing. Thus, lattices can describe both the crystal and its diffraction pattern. Since the lattice of the diffraction pattern is inversely proportional to the crystal lattice, it is defined as the reciprocal lattice.

In crystal space we can define a set of parallel planes of spacing $d$, and note that the first of these planes (for which the distance from the origin is $d$) has intercepts with the edges of the unit cell of $\mathbf{a}/h$, $\mathbf{b}/k$, $\mathbf{c}/l$. The Miller's index of that set of planes is then $hkl$, where $h$, $k$, and $l$ are small integers with no common factor. In Fig. 4 we have a plane with intercepts $\mathbf{a}/3$, $\mathbf{b}/4$, $\mathbf{c}/2$. The Miller's index is 342. The set of parallel planes of index 342 will give rise to a spot in the diffraction pattern with index 342. The reflection 684 can be regarded either as the second-order reflection from the planes 342 or as the first-order reflection from a set of parallel planes with spacing $d/2$. The number of diffraction planes possible for a given structure is directly related to the lengths of **a**, **b**, and **c**.

## C. Basic Formulas of Crystallography

The crystal structure is a pattern that repeats in three dimensions, and Fourier series can represent repetitive patterns.

Let the position of the $j$th atom in the unit have fractional coordinates $x_j, y_j, z_j$ (in this notation, $x_j$ means $x/\mathbf{a}$). Then, the vector from the origin to the $j$th atom would be $\mathbf{r}_j = \mathbf{a}x_j + \mathbf{b}y_j + \mathbf{c}z_j$. The vector representing the diffracted beam direction is $\mathbf{H} = h\mathbf{a}^* + k\mathbf{b}^* + l\mathbf{c}^*$ where $hkl$ are indices of the reflecting plane and $\mathbf{a}^*$, $\mathbf{b}^*$, $\mathbf{c}^*$ are the base vectors of the reciprocal lattice. The direction of the diffracted beam is given in terms of the indices $hkl$. The set of planes $hkl$ cuts **a** into $h$ divisions, **b** into $k$ divisions, and **c** into $l$ divisions. The phase difference for unit translation along **a** is $2\pi h$. Thus, if $h = 3$, a ray scattered by an electron at $\mathbf{a} = 1$ would be $2\pi 3$ or three wavelengths out of phase with one scattered by an electron at the origin.

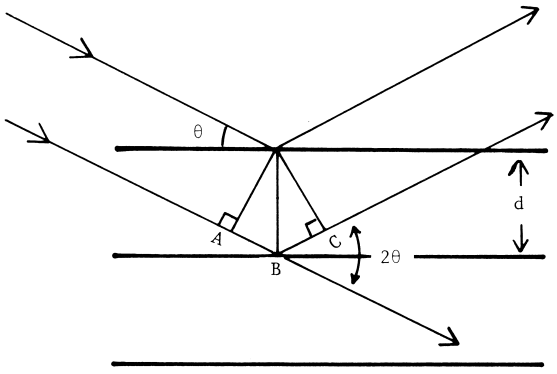



**FIGURE 3** Bragg's law. The distance AB = BC = $d \sin\theta$. Constructive interference occurs if the path difference is a whole number of wavelengths. Thus, $n\lambda = 2d\sin\theta$.

The amplitude of the wave scattered by the plane *hkl* is

$$F_{hkl} = V \int_0^1 \int_0^1 \int_0^1 \rho(xyz) \times \exp[2\pi i(hx + ky + lz)]\, dx\, dy\, dz$$

where $\rho(xyz)$ is the electron density at the point $x, y, z$ in the unit cell. The quantity $(hx + ky + lz)$ is the vector product $(\mathbf{H} \cdot \mathbf{r})$.

By the properties of Fourier series, $F_{hkl}$ is a Fourier coefficient of $\rho(xyz)$ so that:

$$\rho(xyz) = \frac{1}{V} \sum_h \sum_k \sum_l F_{hkl} \exp[-2\pi i(\mathbf{H} \cdot \mathbf{r})]$$

where the summations in $h, k$, and $l$ each run from $-\infty$ to $+\infty$. Note the change in sign of the exponent between the two expressions. $F_{hkl}$ is the Fourier transform of the electron density in the cell. The electron density is the inverse Fourier transform of structure factors.

If the electron density is a superposition of $N$ atomic densities, then the structure factor expression can be rewritten as:

$$F_{hkl} = \sum_{j=1}^{N} f_j \exp[2\pi i(hx_j + ky_j + lz_j)]$$

The summation is over all atoms in the cell and the scattering factor of the $j$th atom is

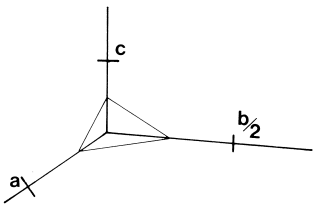



**FIGURE 4** Miller's index. The plane has intercepts **a**/3, **b**/4, **c**/2. The Miller's index is 342.

$$f_j = V \int_{-\infty}^{+\infty} \rho_j(uvw) \times \exp[2\pi i(hu + kv + lw)]\, du\, dv\, dw$$

where $\rho_j(uvw)$ is the electron density of the $j$th atom referred to $x_j, y_j, z_j$ as origin. Thus, if we know the position of an atom, then we can calculate the phases. Conversely, if we know the phases, then we can calculate the electron density and hence the positions of the atoms. The central problem in crystallography is that the phases are not observed in the diffraction experiment. Solving a structure consists of finding positions for the atoms or of finding phases for the structure amplitudes. Details of the major methods for structure solution are found in Appendix II.

## IV. STEPS IN CRYSTAL STRUCTURE ANALYSIS

The steps outlined here would be typical of those used for a moderately complex structure. Notes are also included on alternative techniques that may be more applicable to macromolecular crystallography.

### A. Growing Crystals

All strategies for the growth of crystals for diffraction experiments are aimed at bringing a concentrated solution of a homogeneous population of molecules very slowly toward a state of minimum solubility. The goal is to achieve a limited degree of supersaturation, from which the system can relax by formation of a crystalline precipitate. Many techniques developed for achieving these ends have been described.

Crystallization techniques used in routine synthetic methods tend not to produce crystals of the quality required for structural work. Suitable crystals must be grown slowly at near-equilibrium conditions. This implies low supersaturation ratios and small gradients. Generally, supersaturation is achieved by changing the composition of a solution containing the sample to be analyzed (and possibly other additives) or by altering the temperature. In either case, the concentration of the sample is driven beyond its saturation limit, whereby the sample is forced out of solution and crystal formation may result.

When crystals are grown from solution, changing the solvent may have a pronounced effect on their habit and size. Properties that may influence crystal growth, such as density, viscosity, dielectric constant, and solubility, may be varied over a wide range by mixing two or more solvents.

A small surface-to-volume ratio is useful for slow evaporation of solvent. For small samples, an NMR (nuclear

magnetic resonance) tube may be a suitable crystallization vessel.

Very insoluble compounds may be crystallized using a method known as *reactant diffusion*. In this method, reactants A and B are allowed to mix by diffusion; the very insoluble product C will crystallize in the zone of mixing. Sparingly soluble compounds can sometimes be crystallized from boiling solvent in a soxhlet extractor. Crude product is placed in the thimble and the reservoir is seeded.

Sublimation is effective for some classes of compounds. Vacuum sublimation reduces the temperature required and so increases the range of compounds for which it is suitable. If the temperature gradient is too large, only microcrystals will be formed. Large crystals grow at the expense of small ones only when the process is carried out slowly.

Vapor diffusion is a method that works well with milligram quantities. The solute is dissolved in a solvent in which it is relatively soluble. A small container of this solution is placed inside a closed beaker with a second solvent in which the solute is only sparingly soluble. The two solvents must be miscible in one another and the second solvent should be the more volatile. Suitable solvent pairs include ethanol/ether, benzene/ligroin, and water/ethanol. Diffusion of one liquid into another is also effective. The solute is dissolved in the solvent in which it is more soluble. Crystals form at the interface between the two solvents.

A form of vapor diffusion is commonly used for the growth of protein crystals. For proteins (and other macromolecules) the solution properties modified to achieve supersaturation include increasing the concentration of an additive (e.g., a precipitant), decreasing total solution volume, changing the solution pH, and/or changing the temperature. Although a wide variety of experimental setups have been used in protein crystallization, the most common technique is "hanging-drop microvapor diffusion" (HDMVD). In HDMVD, a droplet of 4 to 20 $\mu$l containing protein and precipitant is suspended from a glass coverslip which is sealed above a reservoir of a solution at a higher precipitant concentration. Since the droplet is at a lower precipitant concentration than the reservoir, the net migration of water vapor occurs from the droplet to the reservoir, resulting in a decrease in drop volume. The decrease in drop volume results in increased precipitant and protein concentrations, which should drive the protein out of solution. Crystals form if conditions are favorable. More often, protein precipitates as an amorphous solid along the bottom of the droplet.

Once initial conditions are found where crystals have formed, additional experiments often must be performed to perfect crystal growth in order to produce X-ray-diffraction-quality crystals. The crystals must be single, with no satellite growths or twinning, and on the order of 0.1 mm on each edge. Fine-tuning conditions, adding detergents or counterions to the precipitant, and seeding droplets with microcrystals are all techniques used to prepare large, single crystals from initial successful experiments.

## B. Microscopic Examination

The crystal chosen for analysis must have a uniform internal structure and be of an appropriate size and shape. The first criterion implies that the substance is pure—the crystal contains no voids or inclusions and is not bent, cracked, distorted, or composed of crystallites. It must be a single crystal, and ideally it should not be twinned (a twinned crystal has two or more different orientations of the lattice growing together).

The size of crystal required is determined by the conditions of the experiment. For X-ray diffraction, approximately 0.1 mm is preferred; for neutrons, an order of magnitude larger is appropriate. A crystal with roughly equal dimensions and well-defined edges is ideal; however, many crystals grow as plates or needles. Although a crystal with a highly asymmetric shape is far from ideal, useful structural information can often be obtained from such a crystal. In some cases, a more uniformly shaped fragment may be cut from a larger crystal with a sharp razor blade.

Examination with a binocular microscope allows a rapid screening of crystals. A few that appear suitable should then be examined more closely through crossed polarizers. As the crystals are rotated, they should either appear uniformly dark in all orientations or they should be bright and extinguish (appear uniformly dark) every 90° of rotation. An unsuitable crystal may show dark and light regions simultaneously, or regions that do not extinguish, or different regions that display different colors.

## C. Mounting a Crystal

The usual way of mounting a crystal for X-ray diffraction is to glue it to the end of a glass fiber that is mounted in a brass or aluminum pin. For photographic work, it is necessary to align a real lattice vector (Weissenberg technique) or a reciprocal lattice vector (precession technique) so that it is perpendicular to the X-ray beam. On a diffractometer, it is necessary for the crystal to be well centered in the X-ray beam, but alignment is neither necessary nor advisable since the aligned position is the one that often leads to overlapping reflections.

Materials that are air or moisture sensitive may be stabilized by covering the crystal with a light coating of oil or the mounting glue (provided that the crystal is not soluble

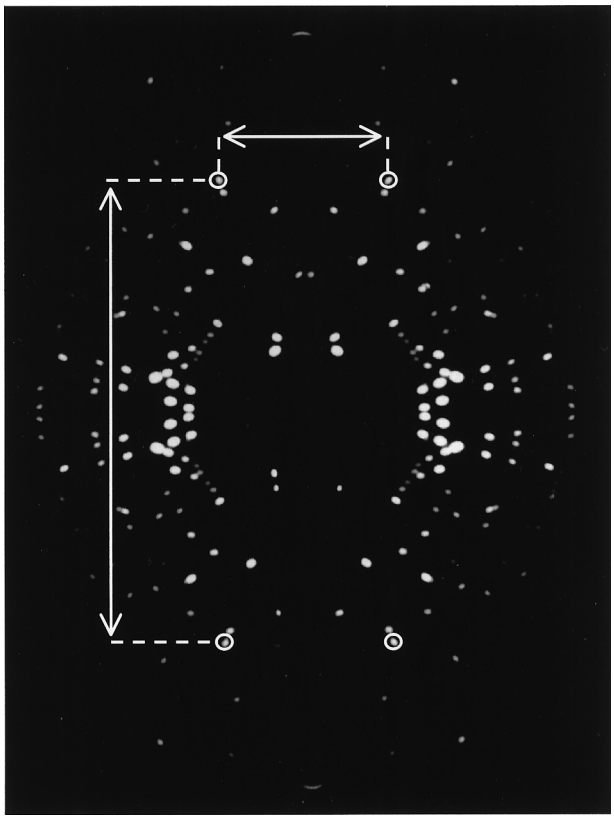


**FIGURE 5** Polaroid rotation photograph.

in the solvent of the glue). Very sensitive substances may have to be mounted in capillary tubes under an inert atmosphere, with all mounting operations carried out in a dry box.

### D. Preliminary Photographs

Most diffractometers are equipped with a Polaroid camera for a quick check of crystal quality. Alternatively a rotation photograph can be acquired with an area detector. The spots on the photograph should have similar shapes, without tails or streaks (see Fig. 5). A more sensitive check of crystal quality is provided by omega scans of a number of reflections measured with different orientations of the crystal. If the diffraction pattern falls off rapidly with angle, low-temperature data collection is advisable; however, some crystals crack when cooled. If suitable crystals are scarce, it may be advisable to collect a dataset at room temperature and then cool the sample to improve the resolution of the experiment

### E. Establishing the Orientation Matrix

The orientation matrix and the cell dimensions are established by determining the setting angles of about eight to ten reflections. These may be found using information available from the preliminary photograph or by allowing the computer-controlled diffractometer to search for them. The indices of these reflections are then determined (usually with a computer program). The cell dimensions are then refined by least squares. High-angle reflections are most sensitive, but fairly strong reflections are required. For this reason, it is common to use relatively low-angle data to establish a preliminary unit cell and matrix, and to recalculate the matrix when the intensities of the high-angle data have been established.

### F. Data Collection

It is always advisable to collect symmetry-equivalent data (reflection intensities). The degree of agreement between equivalent reflections allows assessment of crystal quality, absolute configuration, stability of the counting chain, suitability of absorption correction, and other systematic effects. For charge density studies, a complete dataset should be collected to a resolution of at least $\sin\theta/\lambda = 1.3$ Å$^{-1}$ with an internal agreement of about 2%. For determination of absolute configuration, effects of decomposition, absorption, and extinction errors are minimized if $hkl$ and $\overline{hkl}$ are measured consecutively at $+2\theta$ and $-2\theta$. This may be done for a few dozen of the most sensitive reflections or for the whole dataset. For large datasets, it is common practice to collect data in shells.

### G. Data Reduction

The process of deriving structure amplitudes $|\boldsymbol{F}_{hkl}|$ from the observed intensities $\boldsymbol{I}_{hkl}$ is known as *data reduction*. A number of geometrical factors influence the intensities observed in a diffraction experiment. The most important of these are Lorentz, polarization, absorption, and extinction corrections.

The first three corrections are normally applied to the observed intensities in the process of calculating the structure amplitudes, $|F_{hkl}|$. The Lorentz factor corrects for the relative speeds with which different reflections pass through the reflecting position. Since X-rays are polarized on reflection, and the degree of polarization depends on experimental conditions, a correction must be applied to account for how the polarization affects the observed intensities. Frequently, Lorentz and polarization corrections are combined. Absorption depends on the average path-length of both the incident and the reflected beam and hence may be very different for symmetry-related reflections.

Extinction is an interference process. An extinction correction is generally made during least-squares refinement, making it a correction applied to the model rather than to

the observations. A detailed discussion of the absorption, extinction, and other factors affecting diffraction intensity can be found in Appendix I.

## H. Solving the Phase Problem

The central problem in crystallography arises because the experimental data yield only the modulus of the structure factor, $|\boldsymbol{F}_{hkl}|$ and not the phase. The phase is required in order to evaluate the electron density in the unit cell, but it cannot be measured directly.

Several methods have been developed to determine the phases of the complex structure factors, $|\boldsymbol{F}_{hkl}|$, with no prior knowledge of atomic positions. These methods include multiple isomorphous replacement, single-isomorphous replacement with anomalous dispersion, multiple-wavelength and single-wavelength anomalous dispersion, heavy-atom, and direct methods. With the exception of direct methods, all of these methods take advantage of the scattering properties of "heavy" atoms (transition, actinides, and lanthanides). For very large molecules, such as proteins, neither direct nor heavy-atom methods are generally used. Molecular replacement (MR) also provides a powerful phasing method for the structure analysis. A model structure similar to the structure being analyzed is required for MR.

If the positions of the atoms are known, both the magnitude and phase of the structure factor can be calculated. The heavy-atom method of structure solution depends on this. If at least one atom in the structure is heavy enough to be located in a Patterson function, that position can be used to calculate phase angles. A Fourier summation using observed structure amplitudes and these calculated phases would reveal the heavy atom and some others. The additional atoms are included in the structure factor calculation, providing a better estimate of the phase angles. This process can be facilitated by use of the tangent formula to extend and refine the new phases and is repeated until all of the atoms are located.

Multiple isomorphous replacement (MIR) is a powerful method for determination of phases. MIR depends upon the phasing power of heavy-metal atoms bound to a compound in such a fashion that the positions of other atoms in the crystal are minimally perturbed (i.e., isomorphous derivatives). The MIR method requires that the crystallographer prepare more than one derivative of the parent crystal. These often turn out to be unstable, not isomorphous, or to have the metal bound with too low an occupancy to be useful. Even when successful, this method requires the collection and reduction of numerous datasets from multiple crystals. Anomalous dispersion methods can be combined with isomorphous replacement to circumvent some of the disadvantages.

Advances in the phasing of macromolecular data have been made by the use of a phenomenon called *anomalous dispersion* or *anomalous scattering*. Single isomorphous replacement with anomalous scattering (SIRAS) takes advantage of both the phasing power and anomalous scattering properties of certain heavy atoms. An advantage of the SIRAS technique is that data, in some cases, can be collected using a conventional Cu-K$\alpha$ radiation source. The SIRAS approach to phasing obviates the need for crystallization and data collection from multiple samples. The technique does require careful data collection but produces all of the information needed to determine phases from a single dataset collected from only one crystal.

A crystal containing an anomalous scattering atom may be used to collect data at multiple wavelengths and allow the phases to be determined using multiple-wavelength anomalous dispersion (MAD). A single crystal can be analyzed at multiple wavelengths, generating a variation in scattering factors that allows direct determination of crystal structures. With careful measurements, even weak signals from a single crystal can provide the necessary phasing information. Data must be collected using "tunable" X-ray radiation available only at synchrotron facilities.

A paper by J. Karle (see Bibliography) offers the possibility of nearly direct phasing for protein crystals. The potential for analysis of the data using this single-wavelength anomalous dispersion (SAD) technique is yet to be explored. Given the collection of sufficiently accurate anomalous dispersion data, one dataset at one wavelength may provide all the information required to determine the phases.

Direct methods of phase determination do not depend on any *a priori* structural information. Phases are determined from statistical relationships among the intensities. Powerful computer programs have facilitated the use of these methods for centrosymmetric and noncentrosymmetric structures of moderate size. Considerable effort has been made to extend these methods to protein structures, and structures containing over 1000 non-hydrogen atoms have been solved by direct methods.

In general, the conformationally rigid portions of the molecule are most easily located. Once the major features of the molecule have been recognized, the difference map provides a powerful means both for checking the accuracy of the partial model and for completing it. A Fourier summation using $\Delta F = |\boldsymbol{F}_{\text{obs}} - \boldsymbol{F}_{\text{calc}}|$ as coefficients and phases calculated from the known portion of the structure yields a map of the discrepancies between the crystal and the model. An atom present in the crystal but not included in the model will appear as a peak. Difference maps are also useful for locating light atoms in the presence of heavy ones, such as the location of hydrogen atoms

in compounds of first-row elements or carbon atoms in a tungsten compound.

### I. Refinement

Once approximate locations have been determined for all, or almost all, of the atoms in the structure, it must be refined or made more precise. The standard method for refinement of structures, known at *atomic resolution* (not for macromolecules), is full-matrix least squares. For structures that are not known at atomic resolution, or for very large structures, sparse-matrix or simulated annealing is used for refinement. Since the structure factors are not linear functions of the parameters, the process is an iterative one. Hydrogen atoms are not generally included in the early stages of refinement but are included in the final cycles of refinement.

The function minimized is $\sum \omega \Delta^2$, where $\omega$ is the weight assigned to a particular observation and $\Delta$ is the difference between the observed and calculated values of $\boldsymbol{F}_{hkl}$ (for a refinement based on $|\boldsymbol{F}|$) or $\boldsymbol{F}^2_{hkl}$ (for a refinement based on $|\boldsymbol{F}|^2$). A convenient parameter, referred to as the $R$-factor, for monitoring the progress of the refinement is

$$R = \sum(||F_o| - |F_c||) \Big/ \sum |F_c|$$

Correct structures generally have $R$-values under 0.10, and those that are well behaved are frequently under 0.05.

### J. Determination of Absolute Configuration

If a particular torsional angle has a positive sign in the right-handed enantiomer, it will have a negative sign in the left-handed molecule. Thus, determination of the absolute configuration in a chiral molecule can be regarded as the determination of the correct signs for the torsional angles. A torsional angle ABCD is positive if a clockwise rotation will cause the bond AB to eclipse the bond CD.

Determination of absolute configuration by X-ray crystallography requires a structural study in the presence of dispersive scatterers. Thus, if at least one atom in the structure is an anomalous (i.e., dispersive) scatterer, Friedel's law breaks down and reflections from two sides of the same plane are no longer equal. The differences in intensity are generally small, so careful measurement is required. Coster et al. demonstrated that a noncentrosymmetric crystal structure could be distinguished from its inverted image using these differences. Later, Bijvoet realized this principle was more general and used it to determine the absolute configuration of the sodium rubidium salt of (+)-tartaric acid. In 1983, Flack developed a method for distinguishing a noncentrosymmetric structure from its inverse. Using the method of Flack, any noncentrosymmetric crystal is treated as a twin by inversion and the contribution of the two components evaluated during refinement as the Flack parameter. In the case where the arrangement of atoms in the model and crystal are in agreement, the contribution of the Flack parameter is zero. If the model and crystal are inverted with respect to each other, the Flack parameter is one and the model needs to be inverted.

### K. Derived Parameters

The parameters produced directly by the least-squares refinement are the positions of the atoms and their thermal parameters. Bond lengths, bond angles, and torsional angles are derived from these positions. An examination of short inter- and intramolecular contacts may provide information about hydrogen bonding, van der Waals forces, packing forces, etc.

Room-temperature studies of organic compounds generally show appreciable thermal motion with $U_{ij}$ values of the order of 0.04 Å$^2$. This corresponds to a root-mean-square vibration amplitude of 0.2 Å and is a reminder that some caution is required in comparing bond lengths from diffraction experiments with those determined by spectroscopic and theoretical work. Bond lengths can sometimes be "corrected" for thermal motion, but it is generally preferable to reduce thermal motion by cooling the crystal. Many investigators routinely collect data at temperatures in the range of 100 to 200 °K.

A single crystal structure determination provides valuable information on chemical connectivity, relative conformation, and, under the proper experimental conditions, absolute configuration. However, an understanding of structure–function relationships requires correlating features from a number of different structures. The existence of computer-searchable databases of structural data greatly enhances the possibilities for such comparisons (see Section VI).

## V. COMPARISON OF X-RAY AND NEUTRON DIFFRACTION

For a structure with $N$ atoms, each with atomic scattering amplitude $f_i$ and position $\mathbf{r}_i$ in the unit cell, the structure factor for the Bragg reflection of index $\mathbf{h}$ is

$$F(\mathbf{h}) = \sum_{i=1}^{N} f_i(\mathbf{h}) T_i(\mathbf{h}) \exp(2\pi i \mathbf{h} \cdot \mathbf{r}_i)$$

where $T_i$ is the temperature factor of the $i$th atom. Differences between neutron diffraction and X-ray diffraction lie in the scattering amplitudes $f_i(\mathbf{h})$ and in the temperature

parameters $T_i(\mathbf{h})$. In X-ray diffraction, X-rays are scattered by the electrons of an atom, and the scattering factors $f_i(\mathbf{h})$ are strongly dependent on scattering angle. At a scattering angle of 0°, $f_i$ is proportional to $Z$, the atomic number of the scattering atom. As the scattering angle increases, the scattering factor $f_i(\mathbf{h})$ decreases. Thermal motion causes the scattering to fall off even more strongly. These factors limit the resolution available in an experiment and make it difficult to determine the positions of light atoms accurately in the presence of heavy ones or to distinguish among heavy atoms that have similar atomic numbers.

In neutron diffraction, the scattering is primarily from the nucleus. Since the diameter of the nucleus is small relative to the wavelength of thermal neutrons, the scattering factor is a constant, characteristic of the particular nucleus and independent of scattering angle. There is no simple relationship between the scattering amplitudes and the nuclear mass or charge. Nuclei with similar atomic number can have significantly different scattering amplitudes. Hence, neutron diffraction can distinguish among near neighbors in the periodic table and so is useful in the study of alloys.

Table II shows the relative scattering lengths of a number of atoms for X-rays and neutrons. If the average contribution to the intensity of a structure factor is

$$P(i) = f_i^2 \Big/ \sum_{i=1}^{n} f_i^2$$

Then, for a compound such as benzene with equal numbers of hydrogen and carbon atoms, carbon will contribute 97% and hydrogen will contribute 3%. Thus, in the X-ray experiment, positions of hydrogen atoms will be more poorly determined than positions of carbon atoms. In the neutron experiment on the same compound, hydrogen would contribute 24% and carbon 76%. (Because of the negative sign on the scattering length of hydrogen, a Fourier summation will show "holes" rather than peaks at hydrogen positions.) For deuterobenzene, the contributions from C and D (in the neutron-scattering experiment) are virtually identical, with each contributing 50% of the scattering. Thus, neutron diffraction can locate hydrogen and deuterium atoms with the same precision as carbon, nitrogen, and oxygen, while X-rays cannot. For this reason, studies of hydrogen bonding in biologically significant compounds such as amino acids and sugars were among the early experiments using neutron diffraction.

**TABLE II Scattering Amplitudes for Selected Elements ($10^{-12}$ cm)[a]**

| Element | Atomic number | Scattering amplitude | |
|---|---|---|---|
| | | X-ray | Neutron |
| $^{1}$H | 1 | 0.28 | −0.37 |
| $^{2}$D | 1 | 0.28 | 0.67 |
| $^{12}$C | 6 | 1.69 | 0.66 |
| $^{14}$N | 7 | 1.97 | 0.94 |
| $^{16}$O | 8 | 2.26 | 0.58 |
| $^{32}$S | 16 | 4.51 | 0.28 |
| Cl[b] | 17 | 4.79 | 0.96 |
| Br[b] | 35 | 9.87 | 0.68 |
| $^{127}$I | 53 | 14.95 | 0.52 |
| $^{238}$U | 92 | 25.92 | 0.84 |

[a] For X-rays, the scattering amplitude at 0° is given by $(e^2/mc^2)f_0$ or $((0.282 \times 10^{-12}$ cm) × atomic number). Values for neutrons are taken from "International Tables for X-Ray Crystallography," Vol. IV, 1974, pp. 270–271.

[b] These values are for the elements in their natural isotopic abundances.

As the precision of the experiments and the sophistication of the refinements improved, it became obvious that there were systematic differences in the positional and thermal parameters from the two experiments that were much larger than expected from the estimated standard deviations. The differences are very pronounced for hydrogen atoms. Even in the most precise, low-temperature studies, electron density maxima for hydrogen atoms were displaced by as much as 0.2 Å from the positions of protons determined from neutron diffraction. Since the C–H and O–H bond lengths in the neutron experiment agree with spectroscopic measurements, it was recognized that the apparent shortening of the bonds to hydrogen observed in X-ray experiments is a bonding effect. The position of maximum electron density does not coincide with the position of the nucleus because the formation of the covalent bond perturbs the electron-density distribution in the atom.

Similar, but smaller, effects are observed for first-row atoms, C, N, O, etc. Typically, the discrepancy is of the order of 0.01 Å, but it depends critically on the range of Bragg angles included in the X-ray refinement. If the refinement is based on very-high-order data [$(\sin\theta)/\lambda > 1.00$ Å$^{-1}$], the discrepancies will be much reduced; however, few organic compounds scatter to such high angles even at liquid-nitrogen temperature.

Systematic differences in temperature factors also reflect the very real differences in the scattering processes in the two experiments. In aromatic molecules, thermal vibration parameters are greater for X-rays than for neutrons in the plane of the ring and smaller perpendicular to the ring, implying that electron density is smeared in the plane of the ring by covalent bonding and contracted in the perpendicular direction.

Atoms with lone pairs of electrons show significant differences in both positional and thermal parameters in the two experiments because the lone-pair density is not centered on the nucleus. These very real differences suggest strongly that neutron diffraction is the preferred tool

for determining atomic parameters, whereas the X-rays measure the electron density in the solid. The combination of the two techniques provides a means of studying bonding effects (this topic was covered in greater detail in Section II.B.3).

## VI. RESULTS

The determination of a single crystal structure may answer a question about the connectivity of a molecule or some detail of its conformation. The direct results of structure analysis are the positional, thermal, and occupancy parameters of atoms in the asymmetric unit. Bond lengths and angles, torsion angles, and intermolecular associations (such as hydrogen bonding) are all derived from these basic structural parameters. While the final *R*-value that a structure refines to is often a good indicator of the quality of a structural determination, a plot showing the thermal ellipsoids (Fig. 6) can also give an indication of the quality of a structural determination. Errors in the structure determination and thermal motion in the molecule can distort these ellipsoids.

Understanding a complex process (such as the mechanism of a reaction, the biological activity of a class of drugs, the phenomenon of one-dimensional conduction, or any other structure–property relationship) requires the detailed analysis of a large number of structures and correlation with results from other disciplines. Databases provide the kind of information required for structure–function studies and are very valuable to the scientific community. The topics presented here are intended to be illustrative of the many ways that the databases have been used. There are many other uses not documented here.

### A. Crystallographic Databases

At present there are eight major databases for crystallographic results continuously maintained and updated in different laboratories in Europe and North America. In each case the data are available in machine-readable form, and considerable effort has been expended to develop efficient computational algorithms for searching the files and correlating the data. Perhaps the most important feature of the databases is that the data have been checked and errors are corrected when possible or flagged if uncorrected. Some of these databases have been incorporated into commercial packages. Some are available in both printed and computer-readable form. In the brief summary that follows, they are listed in alphabetical order.

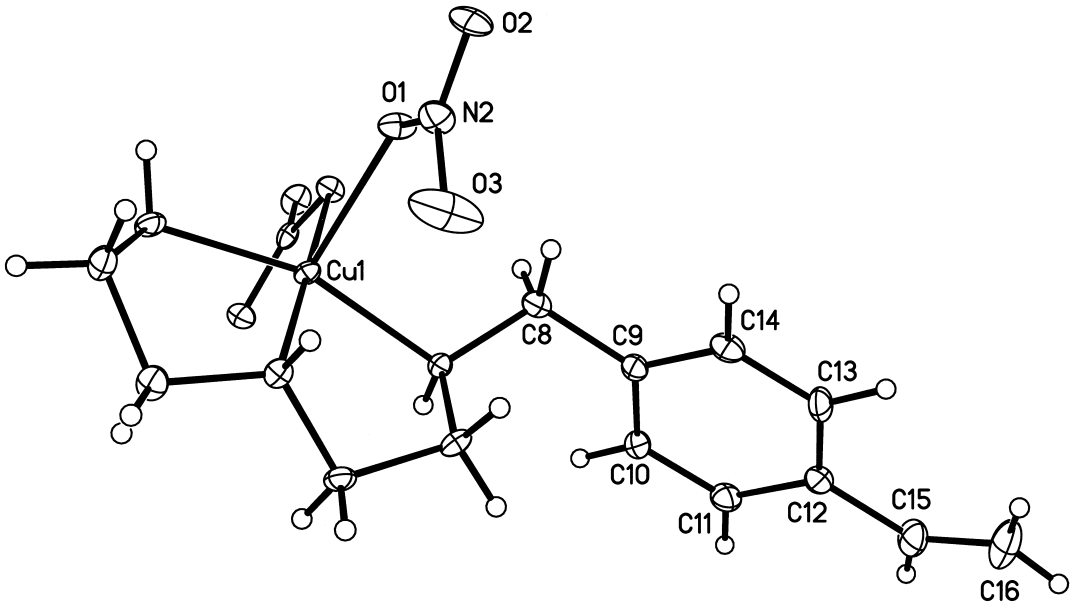



**FIGURE 6** Thermal ellipsoid plot from a crystallographic structure determination. Hydrogen atoms are plotted as small balls with an arbitrary radius; non-hydrogen atoms are plotted as ellipsoids, with the axes corresponding to the thermal parameters for the atom. Notice that the ellipsoids gradually increase in size and become more asymmetric moving out the chain from C8 to C16. The ellipsoids can become large and more asymmetric for noncovalently bounds groups such as the coordinated nitro group ($N_2$, $O_1$, $O_2$, and $O_3$). In both of these cases, the axes of the ellipsoids are aligned with expected vibrational modes and do not indicate any major problems with the structure. If the axes were poorly aligned and/or the ellipsoids more asymmetric, this could be an indication of problems with the structural determination.

### 1. Cambridge Structural Database

(http://www.ccdc.cam.ac.uk/prods/csd/csd.html; Cambridge Crystallographic Data Center (CCDC), University Chemical Laboratory, Lensfield Road, Cambridge, CB2 1EW, U.K.) The Cambridge Structure Database (CSD) is the largest database of experimentally determined organic and metallo-organic crystal structures in the world; inorganic carbon compounds (such as carbonates and cyanides) are excluded. The CCDC also provides a suite of graphical search, retrieval, data manipulation, and visualization software for use with the database. The CSD contains bibliographic data, tables of connectivity, atomic positions, cell dimensions, and quality indicators for virtually all three-dimensional structures of organic compounds published since 1935. The CSD currently contains over 210,000 entries and is growing at a rate of approximately 15,000 entries per year. The CSD is available to scientists throughout the world.

### 2. Crystal Data

(http://www.nist.gov/srd/nist3.htm; Crystallographic Section, National Bureau of Standards, Washington, D.C. 20234.) This database contains lattice parameters for all crystals whose dimensions have been reported by X-ray, neutron, or electron diffraction on single crystals or fully indexed powders. Data include name, formula, cell dimensions, space groups, number of molecules in the cell, density (measured and calculated), bibliographical data, crystal habit, melting point, etc. Crystal Data accepts data from the other databases.

### 3. Electron Density Data Base

(Prof. H. Burzlaff, Lehrstuhl fur Kristallographie, Institut fur Angewandte Physik der Universitat, Bismarckstrasse 10, D-91054 Erlangen, Germany.) This database contains accurate structure factors for crystal structures whose electron densities have been carefully determined. This is the type of data required for studying bonding effects, covalency in organometallics, and other details of electron distribution.

### 4. Inorganic Crystal Structure Database

(http://crystal.fiz-karlsruhe.de/portal/cryst/ab_icsd.html.) The Inorganic Crystal Structure Database (ICSD) was initiated in 1978 at the Institute for Inorganic Chemistry at the University of Bonn. Today the database is produced by FIZ Karlsruhe (P.O. Box 2465, D-76012 Karlsruhe, Germany) in cooperation with NIST. "Inorganic" is defined to exclude metals, alloys, and compounds with C–H and C–C bonds (with the exception of graphites). The data stored include chemical name, chemical formula, density, lattice parameters, space groups, atomic coordinates, oxidation state, temperature factors, remarks regarding conditions of measurement, $R$-values, and bibliographical references. Online access to the file is available. It is also possible to lease the entire database. The database now contains over 53,000 entries and is updated twice a year.

### 5. Metals Data File

(Manager, CAN/SND, Canada Institute for Scientific and Technical Information, National Research Council (NRC) of Canada, Ottawa, Canada K1A OS2.) The Metals file contains structural data for metal and alloy structures determined since 1913 based on either powder or single-crystal diffraction. Under an exclusive license from the NRC, Toth Information Systems has maintained and updated the database. If available, the following information is included: formula, cell dimensions, structure type, Pearson symbol, atomic coordinate, temperature parameters, occupancy factors, $R$-values, method of refinement, instrument used, radiation, and bibliographical information. New software for manipulating the file has been developed by Toth Information Systems.

### 6. Nucleic Acid Database

(http://ndb-mirror-2.rutgers.edu/NDB/ndb.html; Dr. H. M. Berman, Department of Chemistry, Rutgers University, 610 Taylor Road, Piscataway, NJ 08854-8087.) The goal of the Nucleic Acid Database (NDB) project is to assemble and distribute structural information about nucleic acids. Online access to the NDB is freely available. A variety of tools have been developed in conjunction with the NDB to provide a robust user interface.

### 7. Powder Diffraction File

(http://www.icdd.com/_productsold/pdf2.htm; International Center for Diffraction Data (ICDD), 1601 Park Lane, Swarthmore, PA 19081.) The Powder Diffraction File (PDF) is a compilation of powder diffraction patterns produced and published by the Joint Committee on Powder Diffraction Standards of the ICDD. The PDF is the world's largest and most complete collection of X-ray powder diffraction patterns. The 1999 release of the PDF contains about 115,000 patterns, of which 20,000 are organic and 95,000 inorganic. Included in the database are calculated powder patterns for almost 40,000 compounds. The PDF is used for identification of crystalline materials by matching d-spacings and diffraction intensity measurements. Each pattern includes a table of interplanar

d-spacings, relative intensities, and Miller indices, as well as additional helpful information such as chemical formula, compound name, mineral name, structural formula, crystal system, physical data, experimental parameters, and references.

### 8. Protein Data Bank

(http://www.rcsb.org/pdb/; Dr. H.M. Berman, Department of Chemistry, Rutgers University, 610 Taylor Road, Piscataway, NJ 08854-8087.) The Protein Data Bank (PDB) is the single international repository for the processing and distribution of three-dimensional macromolecular structure data (primarily determined experimentally by X-ray crystallography and NMR). In June of 1999, Brookhaven National Laboratory ceased its operation of the PDB when the Research Collaboratory for Structural Bioinformatics (RCSB) took over. The RCSB operates under a contract from the U.S. National Science Foundation with additional support from the Department of Energy and two units of the National Institutes of Health. Contents of the PDB are in the public domain, but the original work as well as the PDB should be properly cited whenever referred to. Affiliated centers in Australia, England, and Japan undertake distribution of data in their respective areas.

## B. Structure Correlation

The working hypothesis behind structure-correlation studies is that changes observed in a structural fragment or subunit in a number of different environments occur along a potential-energy valley in the parameter space of that fragment. Each observed structure is a sample point. An example of this type of study examines the role of iodine in the binding of thyroid hormone. Since iodine-containing thyroid hormones are protein bound during most of their metabolic lifetime, attractive interactions with nucleophiles could play an important role in that binding. Short contacts between iodine bound to carbon and nucleophiles (such as O, N, and S) were studied in an effort to better understand these interactions. The shortest contacts are essentially linear with the lone pair of the nucleophile directed towards the C–I vector. Such contacts are similar to hydrogen bonds and have been estimated to contribute attractive energy of about 3 kcal/iodine atom. The importance of such interactions is further supported by the observation of a short (2.96 Å) I–O contact in the crystal structure of pre-albumin with bound thyroxine.

## C. Reaction Coordinate

A chemical reaction can be represented by a plot of energy as a function of reaction coordinate. Figure 7 shows

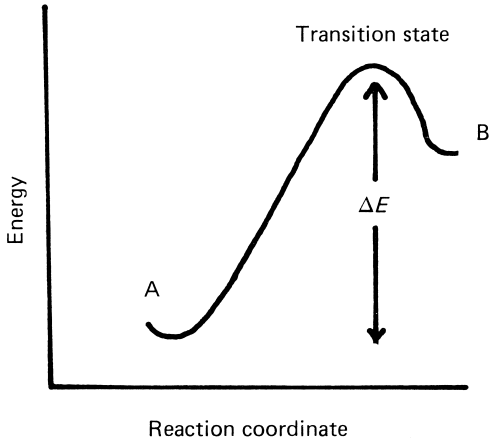



**FIGURE 7** Reaction coordinate. A plot of energy as a function of reaction coordinate for molecule A yielding product B.

the starting material A going to product B through a transition state with activation energy of $\Delta E$, which can be obtained from kinetic data. It is fairly obvious that we can determine the structures of the compounds A and B. Less obvious is that the databases provide a means of looking at the reaction path in some detail. Each individual structure provides a snapshot at one point along the reaction coordinate, but a whole family of structures can plot out a curve related to the potential-energy surface of the reaction.

An early example of this approach was provided by the study of the interactions between amino and carbonyl groups in nucleophilic addition reactions. By examining the data from six crystal structures, Burgi, Dunitz, and Schefter were able to show that interaction with the nucleophile causes the carbon of the carbonyl group to be displaced from the plane of its three substituents toward the approaching nucleophile. The direction of approach of the lone pair on the nucleophile is at an angle roughly 109° to the carbonyl bond (not perpendicular to the plane of the carbonyl). In addition, the displacement of the carbonyl carbon out of the plane of its substituents yields a smooth curve when plotted as a function of the observed C–N separation.

As further confirmation of the validity of this approach, the general conclusions have been reinforced by comparison with SCF-LCGO calculations on the system $CH_2O + H^- \rightarrow CH_3O^-$. The calculated reaction path for the nucleophilic attack of hydride ion on formaldehyde shows very close resemblance to the one predicted by the method of structure correlation.

A different type of correlation is provided in a set of papers published in 1984 by Kirby and colleagues on the length of the C–O bond. This investigation was prompted by the observation that C–O bond lengths for acetals showed an unusually broad range of values. To determine whether this was a general phenomenon or one peculiar to

acetals, the geometry of nearly 2400 ethers and esters was investigated. The results show clearly that there is substantial, systematic variation in C–O bond lengths. The fragment of interest is defined as $R_1$–O–$R_2$. Data are divided into four categories, depending on whether $R_1$ is methyl, primary, secondary, or tertiary, and each of those categories is then divided according to the effective electronegativity of the group $R_2$. The four categories of $R_2$ are alkyl, aryl, enol/ether, and acyl/carboxylic ester, giving 16 categories in all. The shortest C–O bonds (1.418 Å) are found in compounds where $R_1$ is methyl and $R_2$ is alkyl. Within each of the four categories of $R_1$, the length of the $R_1$–O bond increases with increasing electronegativity of the $R_2$ group. The effective electronegativity of a group can be estimated from the $pK_a$ of its conjugate acid. In the subset of 2-(aryloxy)-tetrahydropyrans, there is a linear relationship between the length of the exocyclic C–O bond and the $pK_a$ of the leaving group. This implies a linear relationship between the bond length and the free energy of activation for the hydrolysis reaction or any other reaction in which the C–O bond is cleaved. The consequences of these generalizations to the chemistry and reactivities of acetals and glycosides are fully explored in these papers. In the third paper in the series, the authors plot the reaction coordinate for six aryl tetrahydropyranyl acetals. Both relative free energy and $pK_a$ are plotted as a function of the bond length. The Morse function is also plotted, as is a reaction coordinate–energy contour diagram. This work by Kirby et al. is one of the most comprehensive attempts at deriving structure–reactivity relationships so far available.

## D. Drug Design

The coupling of the resources of databases with molecular graphics devices and QSAR (quantitative structure–activity relationships) techniques raises a tantalizing possibility that therapeutically useful new structures might be predictable. Detailed comparison of structures of known agonists and antagonists is the major strategy currently in use. Similarities in the three-dimensional structure of portions of the molecules allow identification of the features required for binding. Differences in other portions of the structure may account for the agonist/antagonist response after binding. Once the pharmacophore has been identified, binding studies can be combined with structural comparisons to map the receptor site.

A sum or superposition of active molecules can be used to define the available volume within a receptor. Inactive molecules then define excluded volumes—volumes occupied either by the receptor itself or by a cofactor. Any new drug must be designed to present the correct pharmacophore and to occupy only the available volume. Examples of this include mapping of the methionine binding site in the enzyme *S*-adenosyl ATP transferase by Marshall and colleagues, and the postulating of a common site of action for gamma-butyrolactone analogs and picrotoxinin. On the basis of this model, it has been proposed that the convulsant activity of these compounds is related to their ability to block the passage of chloride ions through channels. The model would appear to be applicable to a number of convulsant and anticonvulsant drugs.

## E. Crystallography and Molecular Mechanics

Molecular mechanics is an empirical method for calculation of properties of molecules such as molecular geometry, heat of formation, strain energy, dipole moment, and vibrational frequencies. Different programs use different parameter sets and reproduce these physical properties with different degrees of fidelity. Parameters are assumed to be transferable from one type of molecule to another.

The geometrical parameters used in molecular-mechanics programs are frequently derived from crystallographic data. However, the predictive value of the method is limited by the datasets used to derive the parameters. For instance, a parameter set derived from uncrowded hydrocarbons is not likely to predict structures of crowded hydrocarbons with satisfactory accuracy. Transition states, small rings, unusual states of hybridization, and other electronic effects may require special treatment. These caveats notwithstanding, the method has enjoyed considerable success.

It is not uncommon for molecular-mechanics calculations to be used to provide a starting point for the refinement of electron diffraction data. A recent example is provided by the study of bicyclo[3.2.0]heptane. Many other examples are available in the literature.

Molecular mechanics can also be useful in interpretation of crystal structures, particularly in differentiating electronic and steric effects or in estimating the effects of packing forces. In the structure of lepidoptrene, a C–C bond was observed to have a length of 1.64 Å in the X-ray experiment, while molecular mechanics predicted a length of 1.57 Å. Mislow and colleagues demonstrated that the additional lengthening of the bond beyond the amount expected from steric strain was caused by "through-bond" coupling of adjacent $\pi$ systems.

Although molecular-mechanics calculations can be useful in assessing which aspects of a structure control conformation, it should be pointed out that conformational parameters from molecular mechanics tend to be less reliable then bond lengths or bond angles. Also, conformational parameters in crystal structures are most influenced by packing forces. Molecular-mechanics calculations have been used to suggest alternative conformations

(having energy similar to that observed in the crystal) that might exist in the gas phase or in solution. Additionally, molecular-mechanics has been used to optimize the geometry of a pharmacophore in model studies of drug–receptor binding, to evaluate the interaction energies between dinucleoside monophosphates and cationic intercalators such as ethidium bromide, and to interpret conformational polymorphism. Although some caution is obviously necessary, molecular mechanics and crystallography can provide complementary information in a variety of cases.

## APPENDIX I: FACTORS AFFECTING INTENSITIES

A number of geometrical factors influence the intensities observed in a diffraction experiment. The most important of these are Lorentz, polarization, absorption, and extinction corrections. The first three corrections are normally applied to the observed intensities in the process of calculating the structure amplitudes $|F_{hkl}|$. Many researchers feel that the absence of this correction is the single largest source of systematic error in crystal structures in the current literature. Extinction is a correction made during least-squares refinement and so is a correction applied to the model rather than to the observations. However, the correct formulation for extinction depends on both the polarization and the path-length (calculated during absorption correction) and so is discussed with these other correction terms.

### A. Absorption

The absorption of X-rays by crystals obeys the relation $I = I_0 \exp(-\mu t)$, where $\mu$ is the linear absorption coefficient, in units of $(\text{length})^{-1}$, and $t$ is the thickness. The general effect is to reduce the intensity of reflections at low $\sin\theta$. If the crystal is centrosymmetric in cross-section, neglect of this correction will have little effect on positional parameters, although scale and thermal parameters may be very strongly affected. If the crystal cross-section is not centrosymmetric, all structural parameters will be systematically wrong. Neglect of absorption is probably the largest single source of systematic error in published structures.

Absorption corrections can be done by collecting $\varphi$-scans and calculating an empirical correction or by indexing and measuring crystal faces and calculating a face-indexed absorption correction. For a face-indexed correction, the size and shape of the crystal must be precisely determined. Once the crystal shape is established, a number of techniques are available for calculating the integrated path-length for incident and diffracted beams for all reflections.

For crystals with no reentrant angles, the most common methods are the analytical method and the method of Gaussian quadrature. In both cases, the objective is to evaluate the integral,

$$A = \frac{1}{V} \int_{\text{crystal}} \exp(-\mu T)\, d\upsilon$$

This can be done analytically if the crystal is divided into a number of polyhedra in each of which the path-length is a linear function of the coordinates. De Meulenaer and Tompa first programmed this method in 1965. Calculation time is independent of the severity of absorption. The factor-limiting accuracy in cases of severe absorption is the precision of measuring the crystal, particularly in its shortest direction.

Gaussian quadrature is a numerical integration method that evaluates an integral by summing an appropriate polynomial. It uses a nonisometric grid in which the interval is subdivided symmetrically about the midpoint, with large spacings near the middle and smaller ones toward the edges. This tends to put the maximum number of grid points near the surface of the crystal where the change of absorption with path-length is largest. The number of grid points determines the precision of the calculation. Thus, if a $4 \times 4 \times 8$ grid gives a precision of 2% in calculated transmission for $\mu = 2.5$ cm$^{-1}$, a grid of $8 \times 8 \times 16$ is required to produce the same precision if $\mu = 5.0$ cm$^{-1}$. By choosing sufficient points, the Gaussian method can reproduce the analytical result to any desired precision; however, for strongly absorbing crystals, the analytical method is the method of choice.

When a crystal has reentrant angles, neither of these methods can be used, and one possible resort is numerical integration with an isometric grid; however, the grid must be very fine in order to achieve reasonable precision. For crystals mounted in capillaries, crystals with irregular shapes, and other pathological cases, absorption may be dealt with by measurement of a transmission surface as proposed by Huber and Kopfmann in 1969. However, these problems can be dealt with by the program DIFABS, developed by Walker and Stuart in 1983, which models the absorption surface by a Fourier series in polar coordinates. The coefficients are obtained by minimizing the sum of squares of residuals between observed and absorption-modified values of the structure factors. The chief virtue of the method is that it can be used even when the crystal is no longer available

### B. Lorentz Factor

The Lorentz factor corrects for the relative speeds with which different reflections pass through the reflecting

position. The intensity of a reflection produced by a moving crystal depends on the time taken for the corresponding reciprocal lattice point to pass through the sphere of reflection. Using $1/(\omega S \cos\theta)$ as the angular velocity, the definition $S = 1/d$, and Bragg's law, $\lambda = 2d \sin\theta$, the correction takes the form:

$$d/(\omega \cos\theta) = \lambda/(2\omega \sin\theta \cos\theta) = \lambda/(\omega \sin 2\theta)$$

The factor $1/(\sin 2\theta)$ is the Lorentz factor.

### C. Polarization

X-rays are polarized on reflection, and the degree of polarization depends on experimental conditions. Neutrons are not polarized on reflection from ordinary crystals. X-rays produced from an X-ray tube or a rotating-anode generator are unpolarized; that is, all directions of the electric vector normal to the direction of propagation are equally represented. Thus, the beam can be regarded as composed of two components, one polarized parallel to the reflection plane and one perpendicular. The relative intensities of the two components are

$$I_{||} = I_{\perp} = \tfrac{1}{2} I_0$$

For an ideally imperfect crystal, the intensity of radiation scattered in a particular direction is proportional to $\sin^2\phi$, where $\phi$ is the angle between the electric vector and the direction of observation. It follows that the parallel component will be attenuated by reflection but the perpendicular component will not. Thus, the relative intensities of the two components after reflection at an angle $\theta$ will be $I_{\|} = \frac{1}{2} I_0 \sin^2(90 - 2\theta) = \frac{1}{2} I_0 \cos^2 2\theta$ and $I_{\perp} = \frac{1}{2} I_0 \sin^2 90^\circ = \frac{1}{2} I_0$. Thus, the beam is partially polarized at all angles and completely polarized at $2\theta = 90^\circ$.

For data monochromated by means of a $\beta$ filter, the polarization correction is

$$p = \tfrac{1}{2}(1 + \cos^2 2\theta)$$

Frequently, Lorentz and polarization corrections are combined to give:

$$Lp = (1 + \cos^2 2\theta)/2 \sin 2\theta$$

These corrections are applied to the observed intensity to derive structure amplitudes:

$$|F_{hkl}| = \sqrt{I_{hkl}/Lp}$$

In many modern diffractometers, monochromatization of the primary beam is achieved by Bragg reflection from a suitable crystal. Three of the most commonly used monochromator crystals are quartz ($d = 3.35$ Å), LiF ($d = 2.01$ Å), and highly oriented graphite ($d = 3.35$ Å). The monochromator may be installed in the incident beam or the diffracted beam and may be mounted with its axis parallel or perpendicular to the equatorial plane of the diffractometer. If the monochromator is in the incident beam and mounted so that its axis is perpendicular to the equatorial plane, the component that was attenuated by reflection from the monochromator is attenuated again by the sample, so that the polarization correction for the twice-reflected beam becomes:

$$p = \left(1 + \cos^2 2\theta_{\mathrm{m}} \cos^2 2\theta\right) \Big/ \left(1 + \cos^2 2\theta_{\mathrm{m}}\right)$$

For a monochromator mounted with its axis parallel to the equatorial plane, reflection from the monochromator attenuates the beam normal to the equatorial plane, while reflection from the sample attenuates the parallel component, so that

$$p = \left(\cos^2 2\theta_{\mathrm{m}} \cos^2 2\theta\right) \Big/ \left(1 + \cos^2 2\theta_{\mathrm{m}}\right)$$

These formulas assume that the monochromator crystal is an ideally mosaic crystal. For a perfect or non-mosaic crystal, the factor $\cos^2 2\theta_{\mathrm{m}}$ should be replaced by $|\cos\theta_{\mathrm{m}}|$.

In practice, the polarization ratio seldom corresponds to either of these ideal values and may not even lie between them. This has led some investigators to recast the equations in the form:

$$p = (1 + K \cos^2 2\theta)/(1 + K)$$

for the monochromator axis perpendicular to the equatorial plane, and

$$p = (K + \cos^2 2\theta)/(1 + K)$$

for the parallel orientation, where $K$ is the actual measured value of the polarization ratio for the monochromator in question. The value of $K$ is different for different wavelengths. For routine structural work with MoK$\alpha$ radiation, the error in assuming that the monochromator is an ideal mosaic will generally be small. However, maximum error occurs at $\theta = 45^\circ$ and so is important in the case of very precise studies that rely on high-angle data. Vincent and Flack have developed a method for determining $K$, the polarization ratio, without special equipment.

### D. Extinction

As early as 1922, Darwin realized that absorption is not the only effect that attenuates the X-ray beam as it passes through the crystal. He described two phenomena, which he designated as primary and secondary extinction, and showed how they could be treated mathematically.

Primary extinction is an interference process. If a set of planes is in a position to reflect, the reflected rays may also be reflected a second time. Since there is a phase change of $\pi/2$ on reflection, a beam that has been reflected $n$ times will be exactly out of phase with one that has reflected $(n–2)$ times. This causes the reflected intensity

to be proportional to |F| rather than $|F|^2$. A crystal for which this is true is called an *ideally perfect* crystal. Such crystals are rare. It is much more common to encounter crystals where $I \propto |F|^n$ where $I < n < 2$ but nearer to 2.

Darwin modeled the phenomena of extinction by assuming that crystals were made up of mosaic blocks, slightly misaligned with respect to one another. In perfect crystals, the blocks are assumed to be large and the misalignments small. Secondary extinction occurs because the planes first encountered by the incident beam reflect so strongly that deeper planes receive less radiation and so reflect with less power than they otherwise would have done. This effect is pronounced for strong reflections with $|F/V|$ of the order of $0.1 \times 10^{-24}$ cm$^{-3}$.

According to the mosaic model, the effect is expected in crystals where the mosaic blocks are small with respect to the size of the crystal. A mosaic crystal in which the blocks are sufficiently misaligned that secondary extinction is negligible is called an *ideally imperfect* crystal. Since such crystals are seldom encountered in experiments, it is convenient to correct for extinction in least-squares refinement. In most current programs, the correction is based on Zachariasen's 1967 formalism for isotropic extinction, in which the mosaic blocks are assumed spherical:

$$F_c^* = k|F_c|\left(1 + 2r^* Q_0 \bar{T} p_2 / p_1\right)^{-1/4}$$

where $k$ is the scale factor, $F_c$ is the calculated value of the structure factor, $r^* = \beta[1 + (\beta/g)^2]^{1/2}$ where $\beta = 2\bar{t}/3\lambda$, $\bar{t}$ is the mean path length in a single domain, and $g$ is related to the mosaic spread distribution and is frequently assumed to be Gaussian. For X-rays,

$$Q_0 = \left(\frac{e^2 FK}{mc^2 V}\right)\frac{\lambda^3}{\sin 2\theta}$$

where $e^2/mc^2$ is the classical radius of an electron, and $K$ is the polarization ratio. For neutrons,

$$Q_0 = \left(\frac{F}{V}\right)^2 \frac{\lambda^2}{\sin 2\theta}$$

The term $\bar{T}$ is the mean path-length in the crystal and represents an integration of incident and diffracted beams over all diffraction paths in the crystal. This is normally evaluated during the calculation of absorption correction. If absorption is small in the crystal, $\bar{T}$ may be arbitrarily set to some value such as 0.03 cm. Finally, the term $p_n$ is a polarization term generally assumed to have the form $1 + \cos^{2n} 2\theta$ appropriate to filtered radiation. If a monochromator is used, the appropriate form of the polarization factor should be incorporated into the extinction calculation. For very precise work, the actual polarization ratio of the monochromator should be determined experimentally. For extinction correction, the term refined is $r^*$. It contains both a breadth parameter and a misalignment parameter. If $\beta \gg g$, then the broadening of the diffraction peak is dominated by mosaic spread, and we have a type I extinction. Type II extinction, when $\beta \ll g$, is less commonly encountered and corresponds to the situation where the misalignment is small and the breadth of the diffraction peaks is controlled by small domain size. Separation of the terms is possible if two determinations are made on the same crystal with different wavelengths. This is rarely done.

Coppens and Hamilton first extended the treatment of extinction to an anisotropic model in 1970. The crystal is modeled as if it were composed of ellipsoidal particles whose misorientations follow a Gaussian probability distribution. Since there is no need for the distribution of mosaic blocks to obey any symmetry in the crystal, symmetry-equivalent data are not averaged in this treatment.

## E. Anomalous Dispersion

When the wavelength of the incident X-ray beam is close to the absorption edge of a scattering atom, the atomic scattering factor for that atom becomes complex:

$$f = f_0 + \Delta f' + i\Delta f''$$

where $f_0$ is the normal scattering factor for wavelengths far from the absorption edge, and $\Delta f'$ and $i\,\Delta f''$ are correction terms. The quantity $\Delta f'$ is usually negative, and $i\Delta f''$ is always $\pi/2$ radians ahead of the real part in phase. For structural work, the corrections are assumed to be independent of scattering angle. In addition, $\Delta f' = 0$ for wavelengths longer than the absorption edge.

Four aspects of anomalous dispersion important in normal structural work in crystallography are

1. Determination of absolute configuration
2. Solution of the phase problem (structure solution)
3. Distinguishing among atoms of similar scattering power
4. Avoiding systematic errors in structures with polar space groups

Knowledge of the absolute configuration is extremely important in physiologically active materials, since biological systems discriminate strongly between enantiomorphous forms of their substrates. Anomalous scattering with phase change causes a breakdown in Friedel's law, and $I_{hkl} \neq I_{\overline{hkl}}$. This effect was first exploited for determination of the absolute configuration of the sodium rubidium salt of (+)-tartaric acid using Zr radiation ($\lambda =$ 6.07 Å), whose wavelength is slightly shorter than that of the absorption edge of Rb ($\lambda = 6.86$ Å). The experiment

coupled with the known relationship between the stereochemistry of (+)-tartaric acid and (+)-glyceraldehyde showed that Fischer's arbitrary choice had been correct. With modern data collection techniques, the determination of absolute configuration is relatively uncomplicated. In favorable cases, the method can be applied to a compound with no atom heavier than oxygen when the incident radiation is CuK$\alpha$ ($\lambda = 1.54$ Å). Rabinovich and Hope have determined the absolute sign of the torsional angles in the achiral compound 4, 4′-dimethylchalcone, $C_{17}H_{16}O$. These authors believe that the determination of absolute configuration may be possible with hydrocarbons.

In structure solution, anomalous scattering without phase change is formally equivalent to isomorphous replacement. An anomalous-difference Patterson function is analogous to an isomorphous-difference Patterson and so contains peaks only for vectors between anomalously scattering atoms and vectors between anomalous scatterers and normal scatterers. Vectors between normal scatterers do not appear.

Anomalous dispersion with phase change ($\Delta f'' \neq 0$), can be used to determine the phase angles from noncentrosymmetric crystals. In the case where the position of the anomalous scatterer is known, the procedure requires that differences in intensity between be measurable for a significant number of intensities.

Anomalous dispersion provides an elegant means to distinguish among near neighbors in the periodic table that would otherwise have similar scattering power. Alloys such as $\beta$-brass (Cu–Zn) and $Cu_2MnAl$ were early examples of structures determined by this technique. The tunability of the synchrotron source allows choice of a wavelength close to the absorption edge of an element in the sample to maximize the anomalous component of scattering.

Neglecting to correct for anomalous dispersion will obviously introduce a small error in the magnitudes and in the phases of the calculated structure factors. It was long assumed that such errors would have little effect on atomic positions. However, Cruickshank and McDonald have pointed out that neglect of the correction will always cause errors in thermal parameters, and in the case of polar space groups very serious errors in coordinates can arise. The size of the error varies directly with $\Delta f''$ and inversely with the resolution of the data. For a moderately heavy atom such as Co ($Z = 27$), the error in coordinates can be as large as 0.06 Å in the experiment done with Cu radiation. Neglect of anomalous dispersion will cause errors of the order of 0.005 Å for structures with atoms no heavier than oxygen with Cu radiation or sulfur with Mo radiation. The error caused by including $\Delta f''$ and choosing the wrong enantiomer is twice as large.

## F. Scaling of Data

Virtually all methods of solving structures require a reasonable estimate of the relative scale factor between the observed and calculated structure factors. Most computer programs that calculate scale and temperature factors are based on Wilson's 1942 equation:

$$\langle I_{hkl} \rangle = \sum_i f_i^2$$

which says that the local average value of the intensity is equal to the sum of the squares of the scattering factors. It is assumed that the average is taken over a sufficiently narrow range of $(\sin\theta)/\lambda$ so that the $f$ values can be treated as constants:

$$I_{hkl} = k \sum_i f_i^2 \exp[-2B(\sin^2\theta/\lambda^2)]$$

$$I_{hkl} \Big/ \sum_i f_i^2 = k \exp[-2B(\sin^2\theta/\lambda^2)]$$

$$\log_e \left[ I_{hkl} \Big/ \sum_i f_i^2 \right] = \log_e k - 2B(\sin^2\theta/\lambda^2)$$

Thus, a plot of $\log_e[I_{hkl}/\sum_i f_i^2]$ versus $(\sin^2\theta)/\lambda^2$ will give a straight line of slope $2B$ and intercept $k$.

A number of conditions must hold so that the values of $B$ and $k$ are reasonable. The sampling interval should be small, so 40 to 50 intervals of $(\sin\theta)/\lambda$ are required for three-dimensional datasets. Weak reflections must be included. Elimination of all reflections with $I < 3\sigma(I)$ will cause the average intensity to be too high in the high ranges of $(\sin\theta)/\lambda$, and that will cause an underestimate of the temperature factor and a corresponding overestimate of the scale factor. Including weak reflections at half of the local minimum observed intensity is better than leaving them out, but a Baysian fill is probably a better strategy. Large excursions on both sides of the best-fit line are quite common and reflect the facts that Wilson's formula was derived for a random distribution of atoms and real structures contain many repetitions of certain bonded and nonbonded distances. The inflection points (the regions in the graph where the experimental points cross the straight line) are relatively constant for different types of structures. Hall and Subramanian recommend an "inflection point least squares" in which the least-squares line is fitted to 15 points: the five lowest angle points, five points nearest to $(\sin^2\theta)/\lambda^2 = 0.15$, and five points nearest $(\sin^2\theta)/\lambda^2 = 0.26$.

## G. Thermal Diffuse Scattering

Thermal diffuse scattering (TDS) arises mainly from low-frequency acoustic modes in the crystal. First- and

second-order TDS cause the scattering density to peak under Bragg peaks, with the degree of peaking related to the velocity of sound in the crystal. The effect is not removed in normal data-reduction techniques and is different in different directions in the crystal. In normal structure determination, TDS is ignored. The result is a systematic decrease in apparent thermal parameters. Since TDS increases with increased $(\sin\theta)/\lambda$, it enhances the apparent intensity of high-order diffraction data.

In very precise work, such as the determination of charge-density distribution, it is extremely important that the effect be eliminated or accounted for. It should be noted that the amount of TDS included in a diffraction profile will generally be different for X-ray and neutron experiments, since it depends on such experimental conditions as primary beam divergence, wavelength spread crystal dimensions, and counter aperture. Extensive calculations are required to correct for TDS, and most formulations demand that the elastic constants of the crystal be known. However, cooling the crystal can reduce the effect. If $\alpha$ is defined as $I(\text{TDS})/I(\text{Bragg})$, cooling from room temperature to liquid-nitrogen temperature will reduce $\alpha$ by a factor of 5. Cooling to liquid-helium temperature will reduce $\alpha$ by another factor of five. Facilities for X-ray diffraction experiments down to liquid-nitrogen temperature are fairly common. Helium cryostats are rare.

## APPENDIX II: METHODS OF STRUCTURE SOLUTION

### A. Trial-and-Error Methods

The earliest structures that were determined by X-ray diffraction were mineral structures with relatively high symmetry. Intensities were measured as strong, medium, and weak, and most of the atoms sat on special positions in the cell. From knowledge of the density of the material, its chemical formula, and the space group, one could postulate trial structures and see whether the pattern of intensities matched. The method cannot generally be applied to molecular structures in low-symmetry space groups.

### B. Transform Methods

A crystal can be regarded as convolution of the lattice with the unit-cell contents. By the convolution theorem, the transform of a convolution is the product of the transforms. The diffraction pattern is the transform of the crystal structure and so must be the product of the delta function representing the reciprocal lattice and the transform of the unit cell contents.

From the properties of a delta function we know that the product $f(x)\,\delta(x-x_0)$ has values only at $x_0$ since $\delta(x-x_0)=0$ if $x \neq x_0$. Thus, the diffraction pattern of a crystal can be regarded as the transform of the unit-cell contents sampled at the points of the reciprocal lattice. This implies that a direct plot of the weighted reciprocal lattice can give some information about the structure. The method is used frequently to solve the structures of polynuclear aromatic hydrocarbons. In these compounds the dominant features of the diffraction pattern are the benzene transform and the fringe function showing the separation of the molecules. The molecules generally crystallize in centrosymmetric space groups with one short axis roughly perpendicular to the plane of the molecule.

Figure 8a shows the calculated transform of a regular hexagon, the transform of a benzene ring. It is characterized by positive and negative regions with strong positive peaks at a distance 0.8 $\text{Å}^{-1}$ from the origin. If the hexagon is tilted along an axis, the distance perpendicular to the axis of tilt will be foreshortened in crystal space. The corresponding distances in the transform will be elongated. Simple geometry will allow calculation of the angle of tilt.

Determination of the separation is relatively simple. Consider two centrosymmetric molecules related by a center of symmetry with a separation of molecular centers

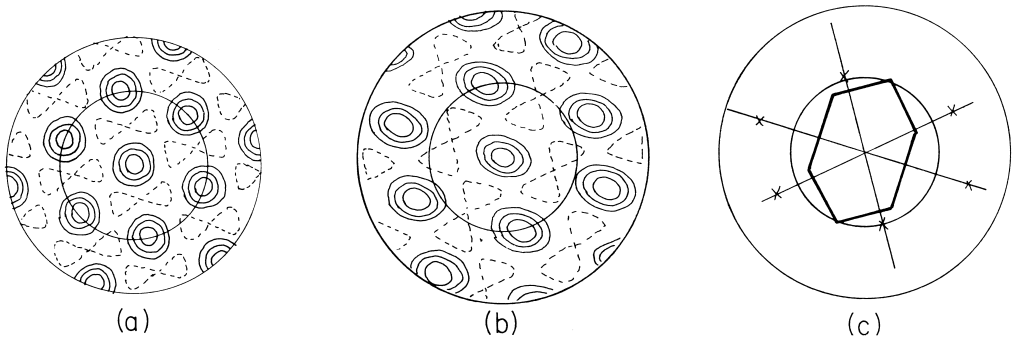



**FIGURE 8** Transform of a hexagon. The circle has a radius of 0.8 $\text{Å}^{-1}$. (a) Regular hexagon; (b) tilted hexagon; (c) geometric construction to determine tilt of hexagon in crystal space.

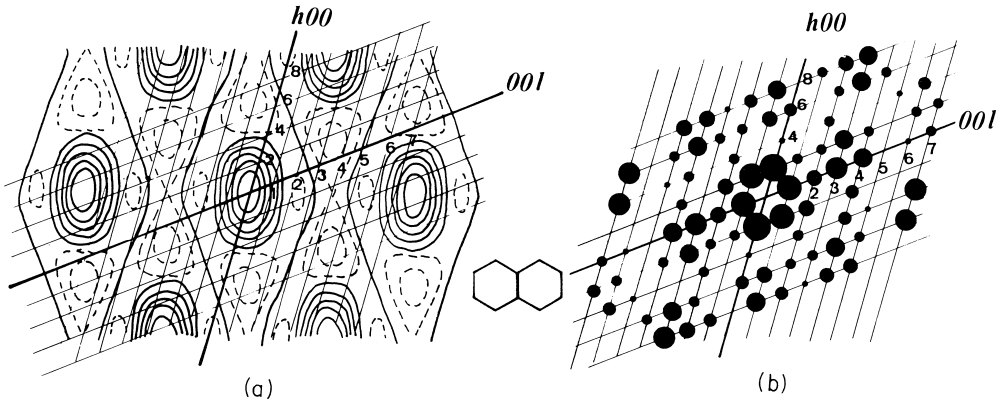



**FIGURE 9** (a) Transform of naphthalene with correctly oriented reciprocal lattice; inset shows naphthalene orientation used to calculate the transform. (b) Weighted reciprocal lattice of naphthalene for comparison with transform. Reflections such as 202 and 801, which lie on regions of the transform where density is changing, are much more sensitive to orientation than are those such as 203 that lie in the middle of strong areas.

of 5 Å. Since the molecules are identical in shape and orientation, the combined transform will be that of a single molecule crossed by straight fringes. Regions in the combined transform are strong only if the corresponding region in the single transform is strong, but weak regions arise either from weak regions in the single transform or from the zeros of the fringe system. A line perpendicular to the fringes is the direction of the line joining the centers of molecules, and the separation of the molecules is reciprocal to the spacing of the fringes.

If the transform and the reciprocal lattice are drawn on the same scale, the correct relative orientation of one to the other can be established by matching strong areas of the diffraction pattern with strong areas in the transform. For fine adjustment of the orientation, attention must be paid to those reflections most sensitive to orientation effects, those lying on rising or falling regions of the transform. See, for example, reflections 202 and 801 in Fig. 9. By contrast, the 203 reflection lies well within a strong area of the transform and its value will not be affected by even fairly large changes in orientation.

## C. Heavy-Atom Methods

In X-ray diffraction, the scattering power of an atom is proportional to the square of the atomic number, $Z^2$. If a molecule contains a heavy atom (high $Z$) and that atom can be located, then a set of phase angles can be calculated for the dataset that are approximations to the true phases. A Fourier synthesis calculated with observed structure amplitudes and phases appropriate to the heavy atom will give a map that contains the heavy atom, some light atoms, and some noise. Phases based on the known atom positions are better estimates of the true phase than the heavy atom alone. The iterative procedure is repeated until all atoms are located.

The location of the heavy atom can be determined from the Patterson function:

$$P(u\upsilon w) = \frac{2}{V}\sum_h\sum_k\sum_l (F_{hkl})^2 \times [\cos 2\pi(hu + k\upsilon + lw)]$$

$$= \int_{\text{volume of cell}} p(xyz) \times p(x+u, y+\upsilon, z+w)\, d\upsilon$$

The Patterson function can be calculated directly from the intensities with no previous knowledge of the phases. It is the self-convolution of the electron density. This means that a peak at $u\upsilon w$ represents a vector between two atoms whose separation is equal to the vector distance from the origin to the point $u\upsilon w$. The weight of that peak is proportional to the product of the atomic numbers of the atoms at each end of the vector. Vectors between heavy atoms tend to dominate these maps and hence allow the position of the heavy atom to be determined.

Figure 10 shows a hypothetical molecule with one heavy atom and its Patterson function. Note that the Patterson function is always centrosymmetric. If the heavy atom were iodine and the light atoms were carbons, I–I vectors would have weight 2809; I–C vectors, 318; and C–C vectors, 36.

If two molecules are related by a center of symmetry as in Fig. 11, then in addition to the intramolecular vectors near the origin of the vector maps, there are intermolecular vectors. The I–C and C–C vectors are double-weight vectors because both molecules give the same pattern of vectors. However, the I–I vectors, which represent vectors between iodine atoms across the center of symmetry, are single-weight peaks.

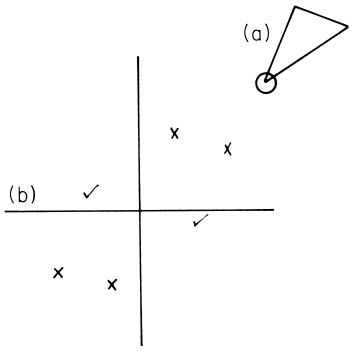



**FIGURE 10** Patterson function. (a) Three-atom "molecule" and (b) its vector set. If there are $N$ peaks in the Fourier, there are $N^2$ peaks in the Patterson function. Of these, $N$ are superimposed at the origin and $N(N-1)$ are distributed through the cell. The atom marked O is heavy; X indicates a heavy-atom–light-atom vector; ✔ indicates a light-atom–light-atom vector.

If the reader were to make a copy of Fig. 11 on transparent paper, place the origin of the transparent map on an I–I vector, and mark the places where the two maps overlap with a mark corresponding to the lower intensity in the overlapping functions, the original structure would be recovered. This is the basis for the "difference function," one of the methods for recovering the electron density from the Patterson function. If the heavy-atom peak had been a double-weight peak, two superpositions would be necessary to recover the original function.

There are constraints on the relative size of the scattering contribution of the heavy atoms. If the heavy atom

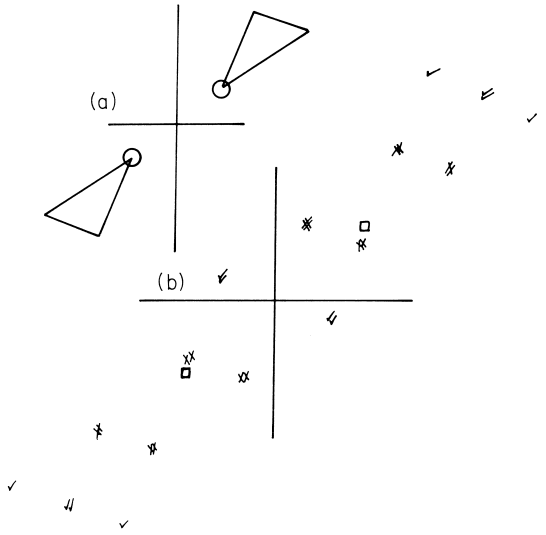



**FIGURE 11** Symmetry in the Patterson function. (a) Two molecules related by a center of symmetry, and (b) the corresponding vector set. The symbol □ indicates a heavy-atom–heavy-atom vector. Note that peaks corresponding to a vector distance between an atom and its symmetry mate are single-weight peaks. All other peaks are double weight.

is too light, phases calculated from its position are poor estimates of the true phases and it may be very difficult to find correct atom positions in a very noisy Fourier map. If it is too heavy, the scattering of the heavy atom will dominate to such an extent that the precision of the light-atom parameters may be seriously affected.

The rule of thumb is that the ratio $\sum Z^2_{\text{heavy}}/\sum Z^2_{\text{light}}$ should be approximately 1; however, the method will tolerate large deviations in either direction. For instance, the structure of vitamin $B_{12}$ ($C_{63}H_{88}N_{14}O_{14}PCo \cdot H_20$) was solved using the phases from the cobalt atom as a starting point. The $Z^2$ ratio is about 0.17!

A Sim-weighted Fourier is a Fourier series phased by the known portion of the structure with coefficients weighted according to the probability that the phase is correct. This is a very useful technique for improving the signal-to-noise ratio in poorly phased Fourier maps.

## D. Isomorphous Replacement

Two compounds are perfectly isomorphous if the only difference in their electron-density maps corresponds to the site of a replaceable atom. The method requires two isomorphous derivatives in the centrosymmetric case and three or more isomorphous derivatives in the noncentrosymmetric case. As direct methods have improved, the use of this method for general organic structures has declined; however, many all-protein structures have been solved by this method or by the combination of isomorphous replacement and anomalous scattering.

The centrosymmetric case is straightforward. Consider two derivatives, A and B, for which the light-atom portions are identical but the replaceable atoms are different. Then,

$$F_{\text{A}} = F_{\text{L}} + F_{\text{AR}}$$

$$F_{\text{B}} = F_{\text{L}} + F_{\text{BR}}$$

$$\Delta F_{\text{AB}} = F_{\text{A}} - F_{\text{B}} = F_{\text{AR}} - F_{\text{BR}}$$

The magnitudes $|F_{\text{A}}|$ and $|F_{\text{B}}|$ are available from the data collection; the magnitudes and signs of the replaceable components are available once the positions of the replaceable atoms are known. Thus, the four possible sign combinations corresponding to the left-hand side of the third equation are calculated, and the combination that gives best agreement with the right-hand side is accepted. Reflections whose phases are not well determined are omitted from the Fourier synthesis.

In the noncentrosymmetric case, the use of two isomorphous derivatives leads to a twofold ambiguity in phase and it is necessary to have a third derivative with a heavy atom in a different position in order to resolve the ambiguity. This is illustrated in Fig. 12. Again, let

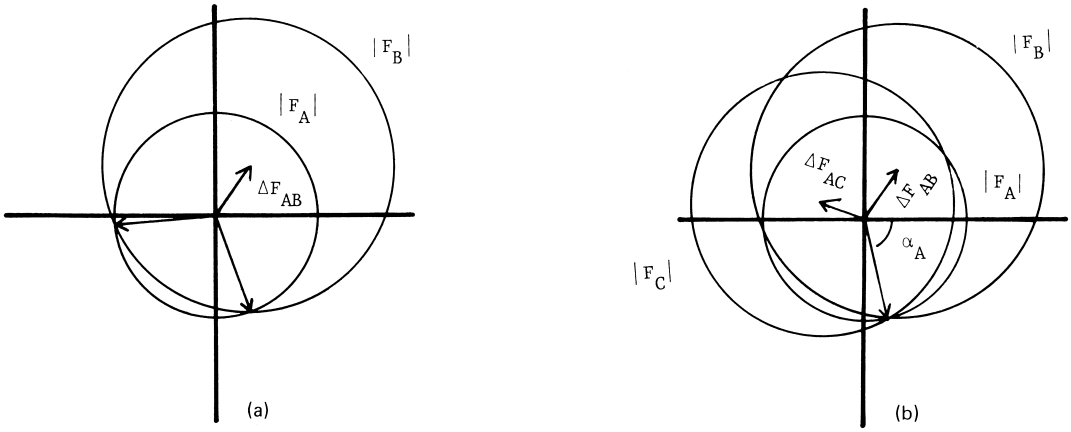



**FIGURE 12** Isomorphous replacement. (a) The difference $\Delta F_{AB}$ between the structure factors for the two isomorphs A and B provides the center for $|F_B|$. (b) The ambiguity in phase can be resolved by a third derivative.

$$F_A = F_L + F_{AR}$$
$$F_B = F_L + F_{BR}$$
$$\Delta F_{AB} = F_{AR} + F_{BR}$$

The circles of radius $|F_A|$ represents all possible values for the phase of $F_A$. The magnitude and phase of $\Delta F$ are indicated by the vector and a circle of radius $F_B$ that is drawn using the tip of $\Delta F$ as center. The two points of intersection correspond to the phase combinations that satisfy the third equation. A third derivative introduces a new equation,

$$\Delta F_{AC} = F_{AR} - F_{CR}$$

Drawing a third circle with radius $|F_C|$ and center at $\Delta F_{AC}$ resolves the ambiguity.

## E. Anomalous Scattering

When the energy of the radiation used in the experiment lies near an absorption edge for one (or more) of the atoms in the crystal, the scattering factor for that atom becomes complex:

$$f = f_0 + f' + if''$$

In this circumstance, Friedel's law breaks down and $|F|^2_{hkl} \neq |F|^2_{\overline{hkl}}$. For centrosymmetric crystals, the data are measured twice: once with a radiation for which the heavy atom scatters normally, and once for scattering anomalously. The two sets of data must be scaled carefully because the differences tend to be small. If the imaginary component is small, then anomalous scattering is exactly analogous to isomorphous replacement.

The results can be shown in a Harker diagram. In the centrosymmetric case with only one type of anomalous scatterer (Fig. 13a), the first circle (i.e., $F_N$) corresponds to the amplitude of the structure factor with no anomalous contribution. The second circle ($F_{NH}$) corresponds to the amplitude of the structure factor with an anomalous contribution. $F_{H'}$ and $F_{H''}$ represent the real and imaginary components of contribution of the anomalous scatterers. Since there is only one anomalous scatterer, the vectors $F_{H'}$ and $F_{H''}$ are perpendicular. The vector sum of $F_{H'}$ and $F_{H''}$ becomes the center for the $F_{NH}$ structure amplitude. The circles intersect in two places but there is no ambiguity, since the phase must lie on the real axis.

In the noncentrosymmetric case (Fig. 13b), the same procedure again leaves us with a phase ambiguity. In structures of moderate size, it may be sufficient to choose the solution near the phase of the heavy atom. This will not always be the correct choice but frequently leads to the correct solution.

In protein structures, the heavy atom is generally far too light to use to discriminate between the two choices. The ambiguity can be resolved by using the other member of the Bivoet pair (Fig. 13c) or by using data from an isomorphous derivative (not shown). In any case, three datasets are required.

An interesting variation on the method is possible using synchrotron radiation. If the protein contains one anomalous scatterer, the tunability of the source can be exploited to collect datasets at different wavelengths. Some protein structures contain a large number of atoms whose anomalous dispersion corrections are so weak they may be neglected. If they also contain only a few anomalously scattering atoms all of one type, a simple special case of the general system of equations results:

$$\begin{aligned}|F_{\lambda\mathbf{h}}|^2 = {} & \left|F^n_{1,\mathbf{h}}\right|^2 + [1 + Q(Q + 2\cos\delta_{\lambda 2})]\left|F^n_{2,\mathbf{h}}\right|^2 \\ & + 2(1 + Q\cos\delta_{\lambda 2})\left|F^n_{1,\mathbf{h}}\right|\left|F^n_{2,\mathbf{h}}\right|\cos\left(\phi^n_{1,\mathbf{h}} - \phi^n_{2,h}\right) \\ & + 2Q\sin\delta_{\lambda 2}\left|F^n_{1,\mathbf{h}}\right|\left|F^n_{2,\mathbf{h}}\right|\sin\left(\phi^n_{1,h} - \phi^n_{2,\mathbf{h}}\right)\end{aligned}$$

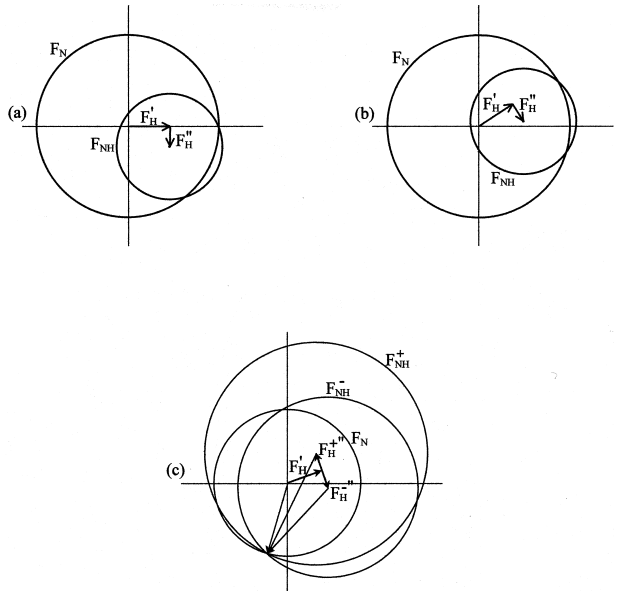



**FIGURE 13** Anomalous dispersion. The first circle is $F_N$. The vectors $F_{H'}$ and $F_{H''}$ are drawn and their vector sum becomes the center for $F_{NH}$. This gives a twofold ambiguity, which can be resolved in the (a) centrosymmetric case by noting that the phase must lie on the real axis. In the (b) noncentrosymmetric case, the ambiguity can be resolved by using the negative anomalous component as shown in (c) or an isomorphous derivative (not shown). $F^-_{NH}$ and $F^+_{NH}$ refer to $F_{hkl}$ and $F_{\overline{hkl}}$, respectively.

where $Q$ is the ratio $f^a_{\lambda 2}/f^n_{2,\mathbf{h}}$, $|F^n_{1,\mathbf{h}}|$ is the magnitude of the structure factor for the normally scattering atoms, $|F^n_{2,\mathbf{h}}|$ is the normal part of the structure factor for the anomalously scattering atoms, and $\phi^n_{1,\mathbf{h}}$ and $\phi^n_{2,\mathbf{h}}$ are the associated phase angles. There is a second equation for the Friedel mate, and it is the same as the previous one except for a minus sign before the last term. Thus, there will be two independent equations for each wavelength at which data are collected, plus a third equation resulting for the trigonometric identity $\sin^2 x + \cos^2 x = 1$. This set of equations can be solved if data are collected at two or more wavelengths. Generally, one dataset is collected at energies below the absorption edge so that $\Delta f'' = 0$; a second set is collected above the absorption edge, with $\Delta f'' \neq 0$ and $\Delta f'$ having the same value as for the first experiment. The actual values for the real and imaginary parts of the scattering factor are determined experimentally before the choice of wavelengths is made. The obvious advantage is that the isomorphism is exact. In principle, the data can be collected on one crystal; in practice, radiation damage will require the use of several different crystals. A generalization of the theory can take into account any number of types of anomalous scatterers and any number of anomalous scattereres within each type.

## F. Direct Methods

Methods of structure solution that attempt to evaluate the phases of the structure factors without recourse to structural information are known as direct methods. It is not necessary to determine all of the phases. In general, about $10\times$ the number of non-hydrogen atoms in the cell is sufficient.

Virtually all of the direct-method programs currently available make use of normalized structure factors, $E(\mathbf{h})$, which "correct" the structure factor $F(h)$ for fall-off with angle caused by the temperature factor and the scattering factors of the atoms,

$$E(\mathbf{h}) = F(\mathbf{h}) \Big/ \left(\varepsilon_{\mathbf{h}}^{-1/2} \sum f_i^2\right)^{1/2}$$

If the shapes of the scattering curves are similar, the values of $E(\mathbf{h})$ can be calculated from the relation:

$$E(\mathbf{h}) = \left[\varepsilon_{\mathbf{h}}^{-1/2} \sum_{j=1}^{N} Z_j \Big/ (\sigma_2)^{1/2}\right] \exp(2\pi i \mathbf{h} \cdot \mathbf{r}_j)$$

where $N$ is the number of atoms in the cell, and $\varepsilon_{\mathbf{h}}$ is the average intensity multiple of the **h**th reflection. The average value of $|E|^2(\mathbf{h}) = 1$. The most frequently used relationship in direct methods for centrosymmetric structures is the $\sum_2$ or triplet relationship:

$$S(E_{\mathbf{h}}) \approx S \sum_k E_{\mathbf{k}} E_{\mathbf{h}-\mathbf{k}}$$

where $S(E_{\mathbf{h}})$ is the sign of the reflection $hkl$, and $\approx$ means "probably equal to." The probability associated with this relationship is

$$P_+(E_{\mathbf{h}}) = \tfrac{1}{2} + \tfrac{1}{2}\tanh\left(\sigma_3(\sigma_2)^{-3/2}|E_{\mathbf{h}}|\sum_k E_{\mathbf{k}}E_{\mathbf{h}-\mathbf{k}}\right)$$

where $\sigma_n = \sum_{i=1}^{N} Z_i^n$ and $Z$ is the atomic number. For example,

$$\mathbf{h} = 63\bar{3} \quad \text{and} \quad \mathbf{k} = 790$$
$$\mathbf{h}\text{–}\mathbf{k} = \overline{163}$$

In this example from a centrosymmetric structure, if all of these reflections are strong, it is probable that the sign of $F(63\bar{3})$ is positive. The probability increases as the magnitude of the normalized structure factors increase. The steps in solving a centrosymmetric structure (where the phases can have values of 0 or $\pi$) are

1. Evaluate $E$ terms. This includes calculation of the normalized structure factors and sorting of the reflections among eight subgroups defined by the parity of the $h$, $k$, and $l$ indices.
2. Form $\sum_2$ relationships with the strongest reflections—those that have $E$ values greater than some arbitrary value such as 1.2 or 1.5.
3. Determine phase. Historically, phases were determined by the symbolic addition method. In this method, origin-determining reflections are given signs and a few others, chosen from the strongest $|E|$ values with the most $\sum_2$ interactions, are given symbols ($a$, $b$, $c$, etc.). Signs can be used because, in the case of a centrosymmetric structure, the phases can only be 0 or $\pi$. Thus, $S_{\mathbf{h}}$ means the sign of the structure factor for the **h** reflection. The value for $S_{\mathbf{h}}$ could arise in one of several ways. One way is its assignment as described above. Alternatively, $S_{\mathbf{h}}$ could acquire a value through the triplet relationship as follows: If $S_{\mathbf{k}}$ is know to be positive, and $S_{\mathbf{h}-\mathbf{k}}$ is known to be $a$, then:

$$S_{\mathbf{h}} = S_{\mathbf{k}}S_{\mathbf{h}-\mathbf{k}} = a$$

During the course of the analysis, relationships appear among the symbols, such as $ac = e$. Manipulation of these relationships usually allows the number of unknowns to be reduced at the end of phase determination.

4. Calculate the $E$ map—a Fourier summation using $E$ values as coefficients and phases determined in step 3.

In the noncentrosymmetric case, the solution is more difficult since the phase can take any value between 0 and $2\pi$. Hence, a different set of relationships was developed for this case:

$$\varphi_{\mathbf{h}} = \langle\varphi_{\mathbf{k}} + \varphi_{\mathbf{h}-\mathbf{k}}\rangle_{\mathbf{k}_r}$$
$$\varphi_{\mathbf{h}} = \frac{\sum_{\mathbf{k}_r}|E_{\mathbf{k}}E_{\mathbf{h}-\mathbf{k}}|(\varphi_{\mathbf{k}} + \varphi_{\mathbf{h}-\mathbf{k}})}{\sum_{\mathbf{k}_r}|E_{\mathbf{k}}E_{\mathbf{h}-\mathbf{k}}|}$$

The symbol $\mathbf{k}_r$ implies that **k** ranges only over those vectors associated with large $|E|$ values.

The process of choosing initial origin-determining phases is similar to the centrosymmetric case, but an additional enantiomorph-determining reflection must also be specified. The symbols are assigned in the same way as before and result in assignments such as $\phi_{\mathbf{h}} = 2a - b$. These are converted to numerical values, and each set is then expanded and refined using the tangent formula. The first computer programs for structure solution used this method. The symbolic addition procedure generates only a few alternatives for the values of the phases which must be considered because the number of resulting unknown symbols is usually no more than three for four.

In multisolution (i.e., multitrial) methods, a small number of phases are assigned arbitrarily to fix the origin and, in the case of noncentrosymmetric space groups, the enantiomorph. Additional reflections are each assigned many different starting values in the hope that one (or more) of these sets of starting conditions will lead to a solution. Some programs use random-number generators to set starting values for some 20 to 200 phases, which are then extended and refined by the tangent formula:

$$\tan\phi_{\mathbf{h}} = \frac{\sum_{\mathbf{k}} w_{\mathbf{k}}|E_{\mathbf{k}}E_{\mathbf{h}-\mathbf{k}}|\sin(\phi_{\mathbf{k}} + \phi_{\mathbf{h}-\mathbf{k}})}{\sum_{\mathbf{k}} w_{\mathbf{k}}|E_{\mathbf{k}}E_{\mathbf{h}-\mathbf{k}}|\cos(\phi_{\mathbf{k}} + \phi_{\mathbf{h}-\mathbf{k}})}$$

The weighting function in the tangent formula is useful in some approaches. A new phase $\phi_{\mathbf{h}}$ is assigned a weight that is the minimum of $0.2\alpha_{\mathbf{h}}$ and 1.0:

$$w_0 = \min(0.2\alpha_{\mathbf{h}}, 1.0)$$

Although this allows rapid development of a phase set for each trial, it tends to lead to an incorrect centrosymmetric solution in the case of polar space groups. For Hall–Irwin weights, $w = w_0 f(\alpha/\alpha_{\text{est}})$, where $f = 1.0$ for $\alpha < \alpha_{\text{est}}$ and decreases for $\alpha > \alpha_{\text{est}}$. This weighting tends to conserve the enantiomorph and has been incorporated into more recent versions of the multisolution programs. The major difference in treatments for centric and acentric datasets lies in the values assigned to the extra reflections in the starting set. In the centric case, the possible values are 0 and $\pi$; in the acentric case, general reflections have four possible values, $\pm\pi/2$ and $\pm 3\pi/2$.

A recent modification to the multisolution approach consists of phase refinement alternated with the imposition of constraints by peak picking in real space. This approach, referred to as a *dual-space method*, differs from other multisolution methods in that the phases are initially assigned values by computing structure factors for a randomly positioned set of atoms. The occurrence of two Fourier transformations per cycle results in an algorithm that is computationally intensive. However, the

new approach also extends the limits of these programs to much larger structures. It has proven to be capable of solving complete structures containing as many as 2000 independent non-hydrogen atoms (provided that accurate diffraction data have been measured to a resolution of 1.2 Å or better and some heavy atoms are present in the structure).

There has been progress in the experimental evaluation of triplet-phase invariants. The phenomenon of simultaneous diffraction has long been considered a nuisance in single-crystal work. If three reflections lie on the sphere of reflection simultaneously, there is a power transfer that tends to enhance such weak reflections at the expense of strong ones. In 1977, it was discovered that the shape of the simultaneous diffraction profile is sensitive to the phase of the triplet that gave rise to it. The sense of the asymmetry is opposite for positive and negative triplets. This effect has been observed with ordinary mosaic crystals of relatively heavy scatterers ($ZnWO_4$) using CuK$\alpha$ radiation from a fine-focus tube. This technique could be enhanced by the use of synchrotron radiation, since it is desirable that the beam be monochromatic, intense, and highly collimated. It is envisioned that experimental phase determination could be used to establish a starting set of 50 or so triplets and that the tangent formula or some similar technique could then expand these relationships.

## APPENDIX III: METHODS OF REFINEMENT

Once a trial structure has been proposed, improvements in the values of the parameters are sought so that the model corresponds as closely as possible to reality. Exact agreement between observed and calculated structure factors would yield a difference Fourier synthesis that was flat and an $R$-value of zero. The methods of refinement discussed here are in the context of small molecules where the ratio of reflections to parameters is commonly of the order of 10:1. Such a degree of overdetermination does not exist in protein structures. The modification of these techniques that would be required to handle proteins is not discussed in this review.

### A. Difference Fourier

A Fourier synthesis with coefficients $\Delta F = |F_0| - |F_c|$ reflects differences between the crystal and the model. The difference map is routinely used during structure solution to check the integrity of the model as it is being developed. Large peaks in the map ($Z/3$ to $Z/2$) correspond to atoms not yet included in the model. Smaller peaks may indicate a slightly misplaced atom or wrong scattering type or incorrect thermal parameters. Holes indicate electron density in the model where none exists in the crystal (see Fig. 14).

It is common to use difference maps to find hydrogen atoms once the positions of the heavier atoms have been refined isotropically. In centrosymmetric structures, the phases are either positive or negative and are usually correct for all but the smallest structure factors by this stage. Hydrogen atoms will have electron densities in the range 0.6 to 0.9 $\Delta e$ Å$^{-3}$. Phase errors in the noncentrosymmetric case cause difference maps to be less well defined. At the end of refinement, a difference map should show no significant features.

Since it is much faster to calculate a Fourier synthesis than to carry out least squares, refinement by difference synthesis was popular before the advent of modern

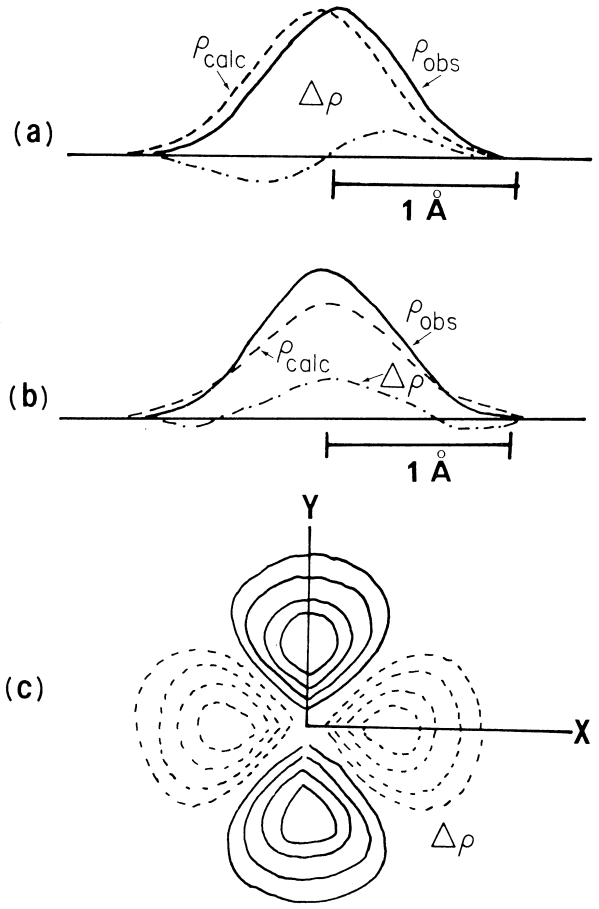



**FIGURE 14** Difference map. Comparison of observed, calculated, and difference electron density for (a) an atom displaced by 0.1 Å from its correct position, and (b) a temperature factor overestimated. (c) Difference density observed when an isotropic temperature factor is used in the model and the thermal vibration of the crystal is actually anisotropic. Dashed contours are negative.

computers. For small- and medium-sized molecules (<400 atoms), least squares is the preferred method of refinement, and difference maps are used mainly for error detection.

## B. Least-Squares Analysis

The set of parameters that minimizes the sum of the squares of the difference between the observed and calculated values of the structure factors is the most satisfactory set. The function minimized is $\sum_{hkl} \omega\Delta^2$ where $\Delta = |F_0| - |F_c|$, and $\omega$ is the weight. The minimization condition is

$$\sum_{hkl} \omega\Delta \frac{\partial F_c}{\Delta p_j} = 0$$

But, the relationship between $\Delta$ and the parameters is nonlinear. For a set of parameters close to the true values, $\Delta$ may be expanded as a function of the parameters by a Taylor series truncated to first order:

$$\Delta(p + \varepsilon) = \Delta(p) - \sum_{i=1}^{n} \varepsilon_i \frac{\partial |F_c|}{\partial p_i}$$

where $p$ stands for the whole set of $n$ parameters, $\varepsilon_i$ is a small change in the $i$th parameter, and the minus sign reminds us that $\Delta = |F_0| - |F_c|$, and it is $F_c$ that must be changed.

The *normal equations* have the form:

$$\sum \omega \frac{\partial |F|^2}{\partial p_i} \varepsilon_1 + \sum \omega \frac{\partial |F|}{\partial p_1} \frac{\partial |F|}{\partial p_2} \varepsilon_2 + \cdots = \sum \omega\Delta \frac{\partial |F|}{\partial p_1}$$

$$\sum \omega \frac{\partial |F|}{\partial p_1} \frac{\partial |F|}{\partial p_2} \varepsilon_1 + \sum \omega \frac{\partial |F|^2}{\partial p_2} \varepsilon_2 + \cdots = \sum \omega\Delta \frac{\partial |F|}{\partial p_2}$$

The left-hand side yields an $n \times n$ matrix, symmetrical about its diagonal. Computer programs store only the upper diagonal half of this matrix. For a 30-atom structure there are 90 positional parameters, 180 temperature parameters, and a scale factor. Thus, $n = 271$. Evaluation of the left-hand side involves evaluating $271 \times 270/2 = 36{,}585$ terms, with the sum over all reflections formed for each term. The matrix is then inverted and solved for the shifts in the parameters. These shifts are applied and the process continues in an iterative manner.

## C. Constrained Refinement

There are a number of reasons for wishing to add constraints to a refinement, such as when the resolution is limited, the structure is disordered, or light atoms are being refined in the presence of heavy ones. Examples of where this type of refinement has been used are inorganic coordination compounds with triphenylphosphine as a ligand. The benzene ring can be constrained to the shape of a regular hexagon with attached hydrogen atoms in their correct positions. Only six positional parameters need be refined: three for the center of mass and three for the orientation angles. The model for thermal motion may allow individual thermal parameters for each atom of the group, or a single temperature parameter for the whole group.

The advantage of using rigid-body constraints is that the number of parameters is reduced. In addition, calculation of the coefficients of the derivatives of the individual atom coordinates with respect to the rigid-body coordinates is performed only once (not for each reflection). Thus, the refinement will require less computer storage and less computer time. As an added bonus, the refinement tends to converge more quickly than an unconstrained refinement. However, it must be kept in mind that a group constrained to an incorrect geometry will result in a structure that is systematically wrong. If the number and quality of data will allow, the constraints may be released at the end of the refinement. Hamilton's $R$-value ratio test may help to determine whether the change in $R$-value on release of constraints is significant.

The "hard" constraints described above reduce the number of parameters needed to solve the structure and hence reduce the size of the matrix. "Soft" constraints do not change the size of the matrix but increase the number of observational equations. Bond length and angles are permitted to deviate from standard values by a fixed amount, or a ring may be constrained to be planar. This technique can be very useful in dealing with highly correlated parameters.

In general, the addition of constraints at the end of refinement will cause the $R$-value to rise slightly. Strict enforcement of an inappropriate constraint will increase the effect. It is unusual for constraints to be in place during the whole refinement; if they are, it is important that they be weighted appropriately.

## D. Thermal Motion Analysis

Most modern crystal-structure analyses refine values for the anisotropic vibrational parameters $U_{ij}$ for individual atoms in the structure. This model assumes that the vibrations are harmonic and that vibrations of neighboring

atoms are uncorrelated. If systematic effects such as absorption and extinction have been ignored, if the quality of the crystals is poor, or if the weighting scheme is inappropriate, these parameters may have little physical meaning.

Hirshfeld has proposed a convenient test for reasonableness of a set of vibration parameters. Since bond-stretching vibrations can be assumed to have much higher energy than other intramolecular modes such as torsion or angle bending, bonds can be considered "rigid" to a first approximation. If two atoms A and B are bonded, it is expected that the component of vibration along the bond will be the same for both atoms. For carbon atoms, agreement within 0.001 $Å^2$ is expected for low-temperature structure determinations where bonding effects have been accounted for. In the spherical atom approximation, the discrepancy is more likely to be of the order of 0.005 $Å^2$. In molecular crystals, vibrations of neighboring atoms are correlated to some degree. For relatively rigid molecules, translational and vibrational oscillations are the major contributors to the internal modes of vibration. If the rigid body model is a good approximation, it is possible to "correct" bond lengths for thermal vibration.

The reason for the correction can be seen in Fig. 15. If a molecule is undergoing librational motion, with a root-mean-square amplitude of libration $\omega$ about an axis through 0, then an atom at a distance $l$ from the center of libration sweeps out an arc. However, the X-ray analysis will place the center of the atom at C, the centroid of the electron distribution. Thus, bond distances are systematically shortened by an amount $PC = l - l\cos\omega \approx l\omega^2/2$. In three dimensions, the corrections are additive.

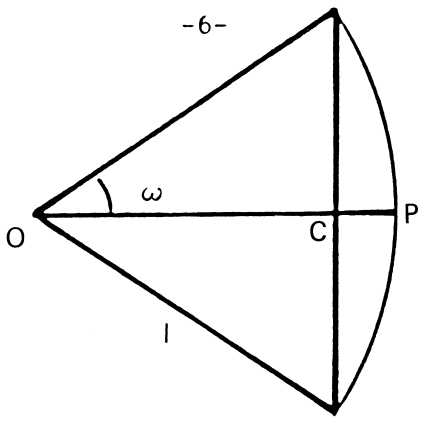



**FIGURE 15** Rigid-body motion. An atom at a distance *l* from the center of libration sweeps out in an arc. Least-squares analysis places the atom at *C* rather than *P*.

The rigid-body models require 12 parameters for centrosymmetric molecules and 21 for noncentrosymmetric. In the former, both the translational and librational axes pass through the center of gravity of the molecule. In the noncentrosymmetric case, the libration axes do not necessarily coincide, and an additional tenser, $S$, is required to account for the correlation between the librational and translational motions.

Two indicators are used to test the goodness of fit between the refined values of the $U_{ij}$ parameters and those calculated from the rigid-body model. The root-mean-square difference between observed and calculated values, given by $\langle(\Delta U_{ij})^2\rangle^{1/2}$, would be considered excellent in the region of $2 \times 10^{-4}$ $Å^2$ but only fair at $2 \times 10^{-3}$ $Å^2$. $R = \sum|\Delta U_{ij}|/\sum U_{ij}$ would be considered excellent at 0.04.

Obviously, the interatomic vectors in a crystal structure can be interpreted as bond distances only if thermal vibrations are negligible. To correct for the effects of thermal vibration, a model for that vibration is required. Often it is better to reduce the effects of thermal motion by working at the lowest available temperature than to attempt an *a posteriori* correction.

## SEE ALSO THE FOLLOWING ARTICLES

CRYSTAL GROWTH • CRYSTALLIZATION PROCESSES • FERROMAGNETISM • MICROSCOPY • PHASE TRANSFORMATIONS, CRYSTALLOGRAPHIC ASPECTS • X-RAY SMALL-ANGLE SCATTERING • X-RAY, SYNCHROTRON AND NEUTRON DIFFRACTION

## BIBLIOGRAPHY

Carter, C. W., Jr., and Sweet, R. M., Eds. (1997). "Methods in Enzymology," Vol. 277, "Macromolecular Crystallography." Academic Press, New York.

Dunitz, J. D. (1979). "X-Ray Analysis and Structure of Organic Molecules," Cornell University Press, Ithaca, NY.

Ewald, P. P., Ed. (1962). "Fifty Years of X-Ray Diffraction," N.V.A. Oosthoek's Uitgeversmaatschappij, Utrecht.

*Journal of Physics D, Applied Physics*. (1991). Special issue on structural aspects of crystal growth, **24**(2), February 14.

Karle, J. (1994). *J. Chem. Inf. Comput. Sci.* **34,** 381–390.

Ladd, M. C. F., and Palmer, R. A., Eds. (1980). "Theory and Practice of Direct Methods in Crystallography," Plenum, New York.

McPherson, A. (1982). "Preparation and Analysis of Protein Crystals," Wiley, New York.

Schmidt, P. W., Ed. (1983). "Proc. symp. small-angle scattering." *Trans. Am. Crystallogr. Assoc*. **19**.

Schoenborn, B. P., Ed. (1985). "Proc. symp. structure determination with synchrotron radiation." *Trans. Am. Crysrallogr. Assoc.* **21**.

Wyckoff, H. W., Hirs, C. H. W., and Timasheff, S. N., Eds. (1985). "Methods in Enzymology," Vol. 114, "Diffraction Methods for Biological Macromolecules." Academic Press, New York.



# Electrons in Solids

**Richard H. Bube**
*Stanford University*

I. A Bond Picture
II. Free-Electron Model of Metals
III. Energy Bands
IV. Optical Properties
V. Electrical Properties
VI. Galvanomagnetothermoelectric Effects
VII. Amorphous Semiconductors
VIII. Superconductors
IX. Conducting Polymers
X. Junctions
XI. Magnetic Properties

## GLOSSARY

**Amorphous material** Material without long-range order as in a crystalline material, but with many optical and electrical properties resembling those of a crystalline material.

**Effective mass** The apparent mass of an electron in a solid when it is treated as if it were a free electron.

**Electrical conductivity** Description of the transport of electrical charge in a metal, a semiconductor, or an insulator.

**Energy bands** The allowed energies for electrons in solids lying in a series of allowed bands separated by forbidden bands.

**Fermi-Dirac distribution** The appropriate statistics for occupancy of allowed electron states in a solid in view of the Pauli exclusion principle.

**Fermi energy** The electrochemical potential of electrons in a solid, an important parameter for describing electrical properties.

**Galvanomagnetoelectric effects** A variety of effects in solids resulting when any two of the following—electric field, magnetic field, and thermal gradient—interact.

**Imperfections** Impurities or defects in a solid that may play a wide variety of roles in affecting the optical and electrical properties.

**Junctions** Particular energy configurations occurring as a result of junctions between different kinds of materials, whether metals or semiconductors, which form the basis for many useful devices.

**Magnetic effects** Effects related either to the orbital motion of electrons in a solid or to the intrinsic angular momentum of an electron, its spin, in interaction with a magnetic field.

**Mobility** The proportionality constant between the drift velocity of a carrier in a conducting state in a solid and the applied electric field.

**Pauli exclusion principle** The recognition that no two electrons in the same system can have all of their quantum numbers the same; that is, only one electron can exist in a specific energy state.

**Periodic potential** The kind of potential characteristic of a crystalline material with atoms located in a regular periodic lattice.

**Scattering** Processes that interact with an electron moving in a solid under the influence of an electric field to counteract the effects of the field and return the carrier distribution toward the equilibrium condition.

**Schroedinger equation** The basic quantum-mechanical equation used to describe the energy levels allowed for electrons in a particular potential energy.

**SOLIDS** may be conveniently separated into those that are crystalline and those that are amorphous. A crystalline material consists of atoms arranged in a periodic array, so that the whole crystal can be constructed by repeating a basic unit called the unit cell. Usually this periodic array is in three dimensions, but special significant cases are known of two-dimensional periodic arrays (as in the plane of an interface). Because the atoms have a periodic arrangement, the electrical potential in the solid also is periodic. The allowed energies for electrons, and many of the optical and electrical properties of electrons in crystalline solids, can be associated with the effects of this periodic potential. An amorphous material does not have a periodic potential or the long-range order typical of a crystalline material, but it does exhibit some of the same basic electrical and optical properties as those of a crystalline material, insofar as these are based on the properties of chemical bonds rather than uniquely on long-range order.

## I. A BOND PICTURE

Some of the most basic optical and electrical properties of both crystalline and amorphous materials can be described qualitatively by considering the nature of the chemical bonds between atoms rather than the effects of a periodic potential.

In both kinds of materials, there is a minimum energy $\Delta E$ that must be supplied to electrons in the chemical bonds of the material to give them enough energy to be able to move freely through the material. Thus there is a minimum photon energy equal to $\Delta E$ for optical absorption and an electrical conductivity with a thermal activation energy of $\Delta E$. We consider the behavior specifically in different kinds of crystalline materials as follows.

In an ionic crystal an electron is transferred from the cation to the anion, and the electron is subsequently localized around the anion. So that the electron can be removed from the anion and can move freely in the ionic crystal, a certain basic energy $\Delta E$ must be supplied. If we think of the electron on the anion as being characterized by a range of energies characteristic of these electronic states, then the next allowed state where conductivity is possible lies at higher energies, separated from the allowed energies below by some finite energy gap $\Delta E$. If the electron is to make the transition from the lower states where conduction is not possible to the upper states where conduction is possible, energy must be supplied to overcome this energy gap. This energy may be supplied by photons, in which case the absorption spectrum consists of an absorption edge at the value of the photon energy equal to this energy gap, which rises to large values of absorption when the photon energy exceeds this energy gap.

Since the energy required to remove an electron from a covalent bond (in a covalent material) to allow it to be free is in general less than the energy required to accomplish this in an ionic material, the absorption edge for a covalent material occurs for lower values of photon energy. The largest energy gap for any material known is that for ionic lithium fluoride, for which $\Delta E = 11.5$ eV, whereas in the widely used semiconductor silicon, $\Delta E = 1.1$ eV.

In a material with metallic binding, on the other hand, there is a constantly available "sea" of free electrons capable of absorbing energy from photons to be excited to higher energy states. The metal is not characterized, therefore, by an absorption edge as the ionic and covalent insulators and semiconductors are, but rather has a continuous absorption due to free electrons over a wide photon-energy range.

This same kind of general picture can be connected with the basic features of the electrical properties of ionic, covalent, and metallic materials. In insulators and semiconductors, electrons must be excited thermally from the highest-lying filled states to the next higher-lying empty allowed states for conductivity to be possible. An activation energy for this process, with a value corresponding to that of the optical absorption edge $\Delta E$ described above, can be measured approximately from the slope in a plot of the logarithm of the electrical conductivity versus reciprocal temperature. Therefore ionic materials have a large activation energy and a small electrical conductivity. Covalent materials, in general, have a smaller activation energy and a larger electrical conductivity. Metals, having a ready supply of free electrons, have no activation energy and the largest value of electrical conductivity.

## II. FREE-ELECTRON MODEL OF METALS

It is often found that the outermost valence electrons of atoms making up a solid can be treated as if they were essentially free electrons. This is particularly true in the class of materials known as metals.

These valence electrons in metals are of course not completely free, since they still move in the presence of the positively charged ions that are located on the crystal lattice sites of the metal. But to an often striking degree it is sufficient to consider the valence electrons to be moving in a potential energy of zero, partially shielded from the positively charged ions by the other electrons, and with the repulsive interactions between electrons canceled out by the residual attractive effects of the positively charged ions. When we consider such electrons from a quantum-mechanical view, it is appropriate to consider them as contained in a box the size of the metal with zero potential energy inside the box and infinite potential energy outside the box. To escape from a real metal, an electron in the metal needs to acquire sufficient energy to overcome the finite potential barrier at the surface, known as the work function of the metal, which is much larger than the average electron energy and therefore approximately infinite.

From calculations based on the Schroedinger equation for free particles in a box, it is possible to derive what energy states are allowed. For a cube with side $L$, the allowed energies are given by

$$E(n_x, n_y, n_z) = (\hbar^2\pi^2/2mL^2)\left(n_x^2 + n_y^2 + n_z^2\right), \qquad (1)$$

where $n_x, n_y$, and $n_z$ are positive integers equal to one or greater. This quantization of the allowed energy levels arises from the necessity to have solutions with wavelength $\lambda$ such that $n\lambda/2 = L$ in each of the three dimensions, to satisfy the geometric boundary conditions that the solution must go to zero at the faces of the cube. If we examine the magnitude of $E(1, 1, 1)$ we see that it is of the order of $10^{-14}/L^2$ eV/cm$^{-2}$, so that for a cube of macroscopic dimensions (e.g., $L = 1$ cm), $E(1, 1, 1)$ can be set equal to zero. The same is true of the energy separation between different energy levels, so that the distribution of allowed energy levels can be treated effectively as a pseudocontinuous distribution.

The variation of the density of these states $N(E)$ per unit volume [$N(E)dE$ is the number of states per unit volume between $E$ and $E + dE$] can be calculated from Eq. (1) with the result that

$$N(E) = (1/2\pi^2)(2m/\hbar^2)^{3/2}E^{1/2}, \qquad (2)$$

which includes a factor of 2 to account for the two possible spin states (intrinsic angular momentum) of an electron.

When these results are coupled with the postulate consistent with the Pauli exclusion principle that only one electron can occupy a particular energy state, we are led to describe the occupation of a state with energy $E$ by the Fermi-Dirac distribution function

$$f(E, T) = 1/[\exp\{(E - E_F/kT\} + 1], \qquad (3)$$

where $E_F$ is a particular energy called the Fermi energy, $k$ is Boltzmann's constant, and $T$ is the temperature. At 0 K the allowed energy states are occupied for all energies up to the Fermi energy, so that if $n$ is the total number of free electrons, $n = \int N(E)\,dE$ from $E = 0$ to $E_F$, and as a result,

$$E_F = (\hbar^2/2m)(3\pi^2 n)^{2/3}. \qquad (4)$$

From the expression for $f(E, T)$, it can be seen that the Fermi energy is the energy for which $f(E, T) = \frac{1}{2}$. The Fermi energy can also be related to thermodynamic quantities by recognizing that it is the electrochemical potential. Using $f(E, T)$ makes it possible to calculate the density of occupied electronic states $n(E)$ as a function of energy by multiplying Eqs. (2) and (3), $n(E) = f(E, T)N(E)$, with the result shown in Fig. 1.

This simple model is able to give qualitative and even semiquantitative agreement with a number of phenomena related to electrons in metals, such as photoemission (emission of electrons from a metal as the result of absorption of photons), thermionic emission (emission of electrons from a metal because of high temperatures), field emission (emission of electrons from a metal because of high electric fields), and heat capacity (a measure of how rapidly the energy of a collection of electrons changes with temperature). If the concept of electron spin is included, this same simple model can be used to describe the magnetic properties of free electrons (free-electron or Pauli paramagnetism). A typical energy model for thermionic emission, for example, is given in Fig. 2.

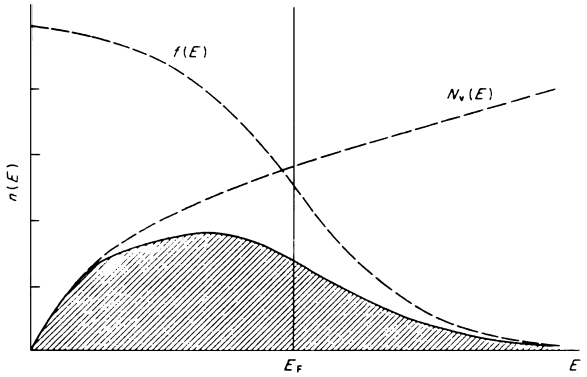



**FIGURE 1** The total density of allowed states per unit volume $N(E)$, the Fermi distribution function $f(E)$, and the density of occupied states per unit volume $n(E) = f(E)N(E)$, at a finite temperature in a metal with Fermi energy $E_F$.

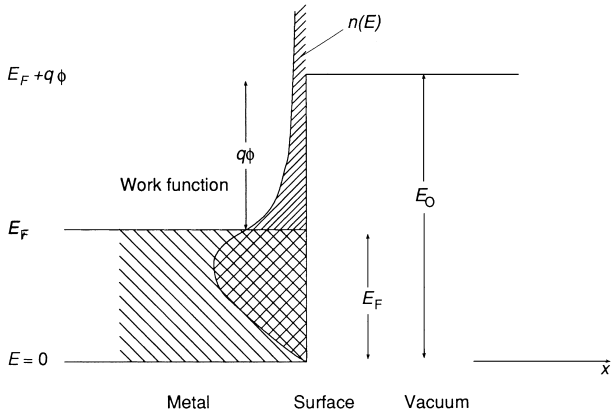



**FIGURE 2** Free-electron model of a metal surface suitable for the discussion of a variety of electronic interactions with a metal, such as photoemission, thermionic emission, or field emission. The specific diagram shown here is for the case of thermionic emission. Electrons with velocity $v_x > [2(E_F + q\phi)/m]^{1/2}$ in the higher-energy tail of the Fermi distribution can be emitted into vacuum. The result is a thermionic emission current that varies as $J_x \propto T^2 \exp(-q\phi/kT)$.

## III. ENERGY BANDS

In the free-electron model, the electrons occupy positive energy levels from $E = 0$ to higher values of energy. We could think of this as a "band" of energies with a lower limit at $E = 0$ but with no upper limit. Because the electrons involved in this band are the valence electrons, we call this the "valence band" of the material.

The free electrons no longer belong to isolated atoms but rather belong to the whole crystal. The allowed band of energies is the result of the fact that the electrons are describable by Fermi-Dirac statistics and obey the Pauli exclusion principle: no two electrons in the whole crystal are allowed to be in exactly the same energy state.

What is true of the valence electrons is also true of the other electrons present in the crystal, corresponding to the more tightly bound inner-shell electrons of the atoms. They also can be thought of as belonging to the whole crystal and therefore requiring that their energies be expressed by a band of energies rather than by the discrete energy-level scheme characteristic of isolated atoms.

There are several complementary ways of showing that the existence of a periodic potential in a crystal to the existence of energy bands:

1. It can be shown mathematically that such a periodic potential gives rise directly to the presence of a series of allowed energy bands separated by energy gaps in which electron states are not allowed.

2. We may pursue the consequences for the allowed energy levels of interactions between atoms as these are brought together in a periodic array to form a crystal (the *tight-binding approximation*). Because of the Pauli exclusion principle such interactions cause the discrete atomic levels to broaden into allowed bands in the crystal, separated by forbidden bands corresponding to the forbidden energies between the discrete levels of the isolated atoms.

3. We may start with effectively free electrons and inquire as to what happens to the allowed energies if we superpose a small periodic potential (the *weak-binding approximation*). We find that the effect of such a potential is to open up forbidden gaps in the previously continuous energy distribution, which once again produces a series of allowed bands and a series of forbidden bands.

### A. Representations of Energy Bands

The major ways of describing energy bands can be seen in Fig. 3. The plot of energy versus distance through the crystal, $E$ versus $x$, in Fig. 3a shows the allowed energy bands separated by the forbidden energy gaps and emphasizes the nonlocalized nature of the band states that extend throughout the whole crystal.

Within an energy band the relationship between the frequency $\omega$ of the electron waves (related to the energy through $E = \hbar\omega$) and the wavelength $\lambda$ of the electron waves (described by the wavenumber $k = 2\pi/\lambda$) is described by giving the specific dependence of $E$ on $k$. Typical variations are shown in Fig. 3b, illustrating the common result that electrons near the top or bottom of an energy band usually behave as if they were free-like electrons for which the relationship between $E$ and $k$ is parabolic (for free electrons, $E = \hbar^2 k^2/2m$, where $m$ is the electron mass). The upper band in Fig. 3b corresponds to a typical conduction band in a semiconductor, while the lower band corresponds to a typical valence band. For a simple one-dimensional crystal with lattice constant $a$, the limiting values of $k$ for the energy band are $-\pi/a$ and $+\pi/a$; these values are equivalent to the Bragg reflection conditions in one dimension. They mean that an electron wave cannot propagate with wavenumber $k = \pm\pi/a$. The range from $k = -\pi/a$ to $k = +\pi/a$ is called the first Brillouin zone. This dependence of $E$ on $k$ is illustrated further in Fig. 4 for the weak-binding model for the formation of energy bands as described above.

When the electrons in a band are free-like, then the corresponding density of states is free-like, with $N(E) \propto E^{1/2}$ as stated above. Figure 3c shows the variation of the density of states with energy throughout the bands. Near the top and bottom of the band $N(E) \propto E^{1/2}$ (as measured away from the minium or maximum point), but in the middle of the band the electrons do not exhibit free-like behavior, and the density of states passes through a maximum.

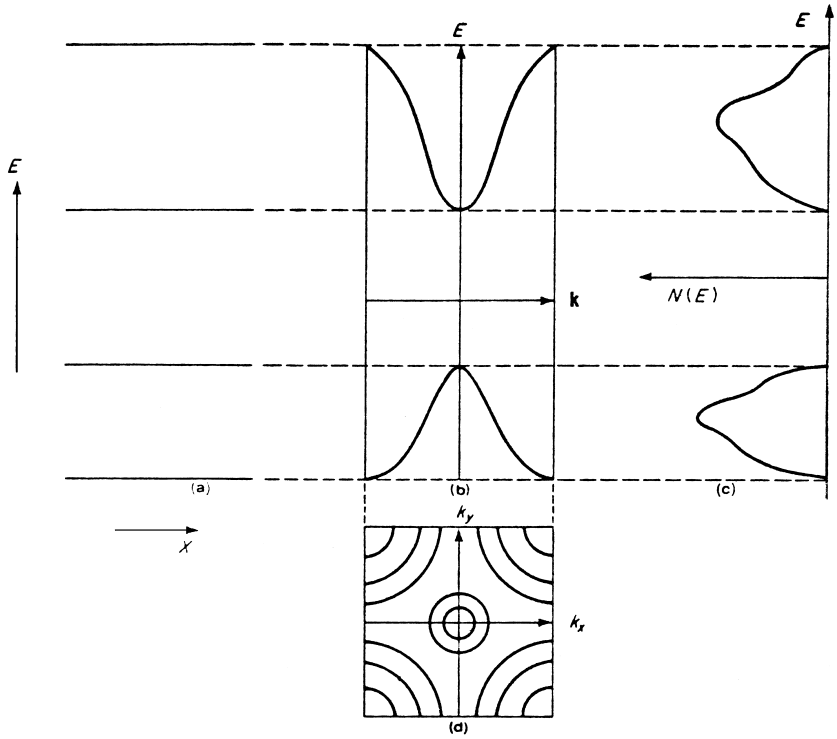



**FIGURE 3** Four ways of describing band properties: (a) flat band $E$ versus $x$, (b) electron wave dispersion relationship $E$ versus $k$, (c) density of states $N(E)$ versus $E$, and (d) equal-energy surfaces in $k$-space.

The basic electrial differences between solids—metals, semiconductors, and insulators—are readily understandable in terms of the band picture. Possibilities are illustrated in Fig. 5. If the highest-energy-allowed band occupied by electrons is only partially occupied, then there are available allowed states at very small energies above occupied states, and the drift of electrons in an electric field can be readily achieved. A partially filled valence band therefore corresponds to a metal.

Even if the highest-energy-allowed band occupied by electrons is totally filled, metallic properties can be found if the next higher-lying band overlaps the filled band (possible in a three-dimensional crystal) to again produce a continuum of allowed states separated by only small energies from occupied states. Overlapping bands also produce metallic behavior.

If the highest-energy-allowed band occupied by electrons is totally filled, and the next higher-lying band lies an appreciable energy above the top of the filled band, insulator-like properties are observed. Only electrons in the upper empty band can contribute to electrical conductivity; their density is very small, since thermal excitation across a large energy gap is required to raise them from the filled valence band to the empty conduction band.

Finally, if the gap between the top of the filled band and the higher-lying empty band is small, appreciable excitation of electrons into the conduction band may occur at normal temperatures, and intermediate conductivity is observed typical of a semiconductor.

If electrons exhibit free-like behavior, then a plot of equal-energy surface in $k$-space, as shown in Fig. 3d, will have spherical symmetry, since the magnitude of $E$ does not depend on the direction of $k$. Thus the equal-energy surfaces in many materials are frequently spherical with center at either the minimum or maximum value of $k$ for the energy band. A notable exception is the widely used semiconductor silicon, for which the minimum of the conduction band does not lie at $k = 0$ and for which the equal-energy surfaces are ellipsoidal rather than spherical.

## B. Interpretation of *E* versus *k* Curves

From a knowledge of the dependence of $E$ on $k$ in an energy band, two significant properties of electronic energy states can be deduced:

1. The velocity associated with an electronic state is the group velocity, given for the one-dimensional case by

$$v_g = \hbar^{-1} \partial E / \partial k. \tag{5}$$

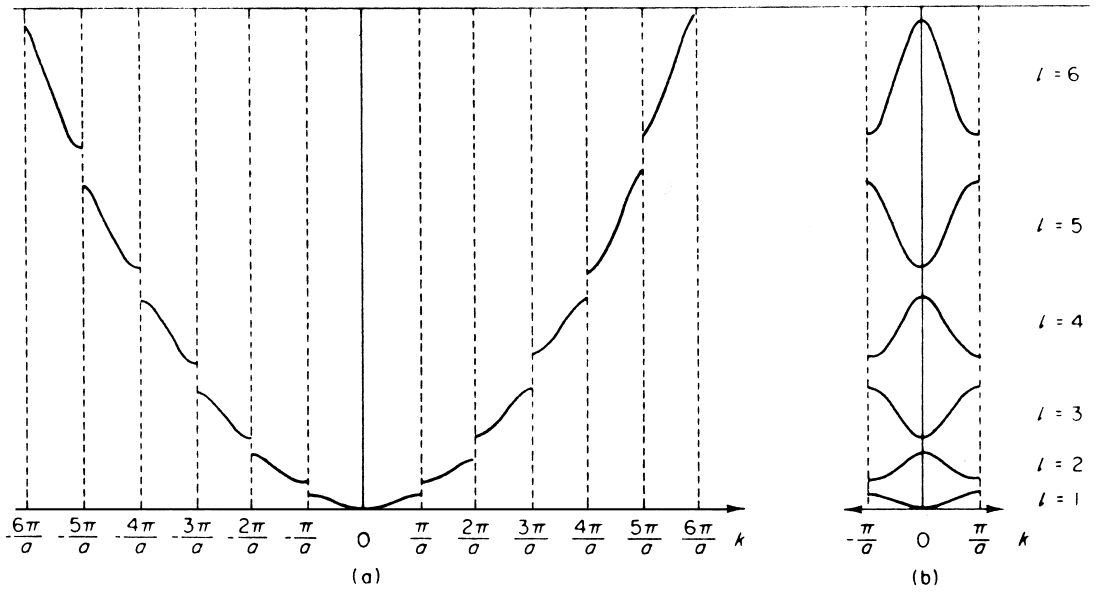



**FIGURE 4** (a) The free-electron model of electrons as perturbed by a small periodic potential (weak-binding approximation) causing Bragg reflection conditions for $k = n\pi/a$ in the extended zone representation. (b) The reduced representation in the first Brillouin zone achieved by translating band segments in (a) by $n2\pi/a$ to bring them into the basic zone between $k = -\pi/a$ and $k = +\pi/a$. A wavenumber $k' = (k + n2\pi/a)$ corresponds to the same solutions as does $k$. In the reduced representation, separate bands are designated by a band index $l$.

For free electrons this means that $v_g = \hbar k/m$, which illustrates the fact that the momentum $mv_g = \hbar k$, in agreement with the de Broglie relationship that the momentum $p = h/\lambda$. If this result is applied to the upper conduction band in Fig. 3a, for example, we see that $v_g = 0$ at the top and bottom of the band, is positive for $+k$ and negative for $-k$, and has a maximum some-where near the middle of the band. We see immediately why the application of an electric field to a totally filled band produces no electric current; the number of occupied states with positive $v_g$ is equal to that with negative $v_g$, and the application of an electric field cannot change this equality.

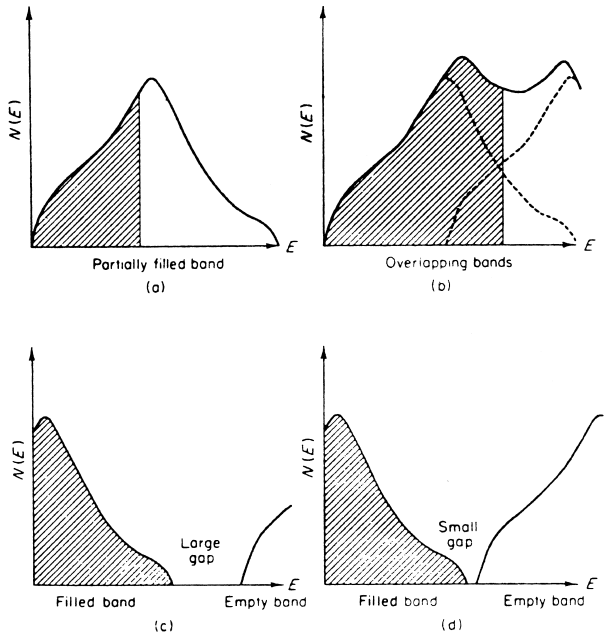



**FIGURE 5** Illustrations of the dependence of electrical properties of solids on energy-band filling and spacing: (a) and (b) metals, (c) insulators, (d) semiconductors.

2. Although electrons near the top or bottom of a band may be said to behave in a free-like manner, it is clear that they are moving in a periodic potential and are not really free. To describe them in a free-like model, we must usually define an effective mass $m^*$ that is different from the free mass $m$. From the result that the force is the time derivative of the momentum, we obtain $F = \hbar\, \partial k/\partial t$. If we also desire to write $F = m^* a$, after the classical form, then it follows that since $a = \partial v_g/\partial t = \hbar^{-1}\partial/\partial t(\partial E/\partial k) = \hbar^{-1}\partial/\partial k[(\partial k/\partial t)(\partial E/\partial k)]$,

$$m^* = \hbar^2/(\partial^2 E/\partial k^2) \tag{6}$$

for the one-dimensional case. If we apply this to the free-electron case, we find, as expected, that $m^* = m$. We see that the effective mass $m^*$ is inversely proportional to the curvature of the $E$ versus $k$ curve. Near the top or bottom of the band, $E$ versus $k$ may be parabolic and we can write $E = \hbar^2 k^2/2m^*$ with a constant $m^*$ over a range of $E$ and $k$. At the bottom of the upper band in Fig. 3b the effective mass for electrons is positive, whereas at the top

of the lower band in Fig. 3b the effective mass for electrons is negative. It is in this negative sign that we see most strongly the effects of the periodic potential. Physically, a negative effective mass corresponds to an acceleration that is in the opposite direction to the applied force; it corresponds to a condition like that encountered in Bragg reflection. Typical variations of $v_g$ and $m^*$ are shown in Fig. 6. Effective masses in most solids vary between about $10^{-2}\,m$ ($m_e^* = 0.014\,m$ in InSb, one of the smallest values) to values of the order of $m$ (for free electrons in the alkali metals or for holes in many semiconductors).

For the filled valence band of a semiconductor to contribute to the electrical conductivity, there must be some empty states corresponding, for example, to electrons that have been excited up to the conduction band. Instead of describing the electrical properties of the valence band in terms of all the electronic states with positive and negative velocities and negative and positive effective masses (depending on which state in the band is being examined), it is conceptually much simpler to describe the conductivity in a partially empty band in terms of the empty states themselves, called "holes." If we treat these holes as if they had a positive charge and a positive effective mass, a symmetric treatment of electrical properties involving them at the top of the valence band can be given, as is commonly given for electrons at the bottom of the conduction band.

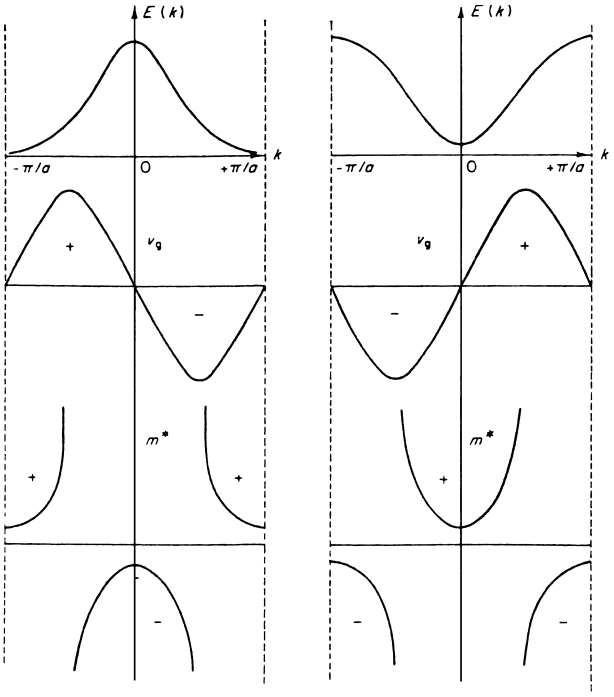



**FIGURE 6** Typical variations of group velocity $v_g$ and effective mass $m^*$ as a function of $k$.

## IV. OPTICAL PROPERTIES

Optical properties of solids include a wide range of phenomena involving either the interaction of light with crystals or the generation of light by crystals under suitable conditions.

The velocity of light $v$ in a material is reduced compared with its value in vacuum $c$ by a factor known as the index of refraction $r = c/v$. In general, the value of $r$ for near-visible light can be expressed by

$$r^2 = \varepsilon_r + c^2\alpha^2/4\omega^2, \tag{7}$$

where $\varepsilon_r$ is the relative dielectric constant associated with polarization of electrons in the material, $\alpha$ is the optical absorption constant (a sample with thickness of $1/\alpha$ reduces the light intensity passing through it by a factor of $e$), and $\omega$ is the light frequency.

### A. Optical Reflection

Reflection occurs for any wavelike phenomenon when the wave passes from a material with one set of properties to a material with another set of properties. Light is reflected on passing from a material with value $\varepsilon_{r1}$ and $\alpha_1$ to a second mateial with value $\varepsilon_{r2}$ and $\alpha_2$. For simple reflection at an air–material interface ($\varepsilon_{r1} = 1$ and $\alpha_1 = 0$ for air) with a material without absorption ($\alpha_2 = 0$), the reflection coefficient $R$ is given by

$$R = (r_2 - 1)^2/(r_2 + 1)^2. \tag{8}$$

If the material does have a finite $\alpha$, the result is that the higher the absorption index of the material, the more light it reflects, until in the extreme case $R$ approaches unity:

$$R = \left[(r_2 - 1)^2 + \Gamma_2^2\right]/\left[(r_2 + 1)^2 + \Gamma_2^2\right], \tag{9}$$

where $\Gamma_2$ is the absorption index expressed through the complex index of refraction $r^* = r + i\Gamma$. The absorption index is related to the absorption coefficient by the relationship

$$\alpha = 2\omega\Gamma/c. \tag{10}$$

This effect accounts for the very high reflectivity of metals and other materials with high values of $\alpha$ due to specific absorption processes.

One application of considerable practical importance that involves the reflectivity properties of thin films of one material on another is the antireflecting coating. As a way to reduce the reflection at an air–material A interface, a second material B with suitable index of refraction (ideally the square root of the index of refraction of A) is deposited in thin-film form on A, with a thickness such that light reflected from the air–B interface interferes destructively with light reflected from the B–A interface. This occurs

when the film B thickness is an odd multiple of quarter wavelengths of the light in the film, $\lambda_B = \lambda_{air}/r_B$.

## B. Absorption Processes

Optical absorption in solids may be categorized under six principal headings. One of these is the absorption of photons in the excitation of optical mode vibrations of the crystal lattice, known as reststrahlen absorption, which usually occurs in the infrared between 10 and 200 $\mu$m. This is the major absorption process that does not involve electronic transitions. Other kinds of optical absorption associated with the excitation of lattice vibrations may occur when an impurity is present in the material and local vibrational modes associated with the impurity become possible.

Figure 7 illustrates the other five major types of electronic absorption processes:

1. Excitation of an electron from the valence band to higher-lying conduction bands, characterized by high-absorption processes, showing only relatively small variations of absorption coefficient with photon energy depending on the density-of-states distributions in the bands involved. The optical absorption constant is usually in the range of $10^5$ to $10^6$ cm$^{-1}$.
2. Excitation of an electron from the valence band to the lowest-lying conduction band with the minimum required energy of the forbidden band gap. The magnitude and variation with energy of the absorption constant depends on whether the transition involves a photon only (direct transition) or whether it involves both a photon and a phonon (indirect transition). The absorption constant decreases by many orders of magnitude as the photon energy drops below the band-gap energy. Direct and indirect optical transitions are illustrated in Fig. 8.

If a direct optical transition is involved, a plot of $(\alpha\hbar\omega)^2$ versus $\hbar\omega$, where $\omega$ is the light frequency, yields a straight line with the energy intercept of the direct band gap. If an indirect optical transition is involved, a plot of $\alpha^{1/2}$ versus $\hbar\omega$ gives two straight-line segments with intercepts of $\hbar\omega_1$ and $\hbar\omega_2$, as shown in Fig. 8b. The upper segment corresponds to an indirect transition with emission of a phonon, whereas the lower segment corresponds to an indirect transition with absorption of a phonon. The indirect band gap is given by $(\hbar\omega_1 + \hbar\omega_2)/2$, and the energy of the participating phonon is given by $(\hbar\omega_2 - \hbar\omega_1)/2$.

3. Excitation of a bound electron–hole pair, known as an exciton, requiring less energy than that needed to produce a free electron–hole pair by excitation across the band gap. The exciton can be thought of as a hydrogenic system, capable of moving and transporting energy through the crystal without transporting net charge. The electron and hole making up an exciton may ultimately be thermally dissociated into free carriers or may recombine with the emission of photons or phonons.
4. Excitation of an electron from an imperfection level to the conduction band, as shown in Fig. 7, or from the

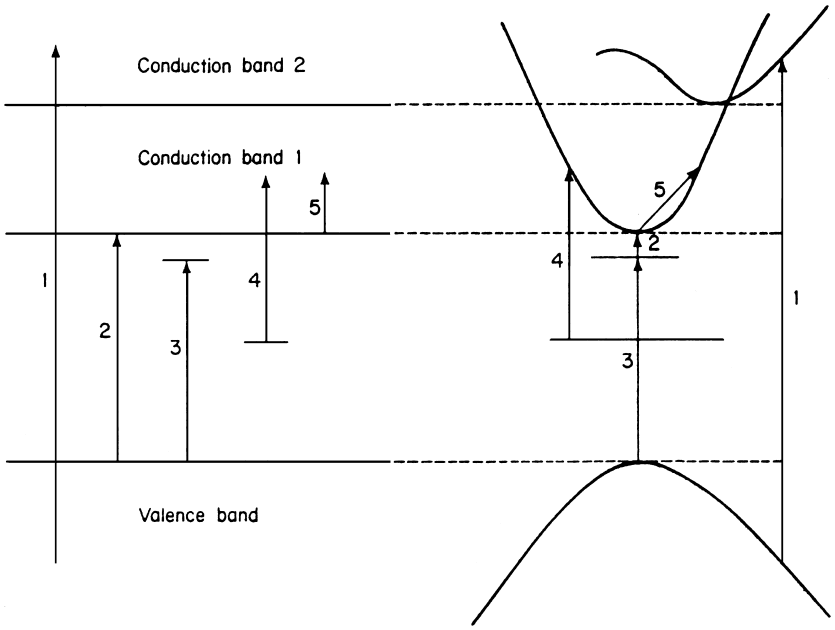



**FIGURE 7** Characteristic types of optical transitions shown both for the flat-band model and for the *E* versus *k* plot. (1) Excitation from the valence band to higher-lying conduction bands, (2) excitation across the band gap, (3) exciton formation, (4) excitation from imperfections, and (5) free-carrier excitation.

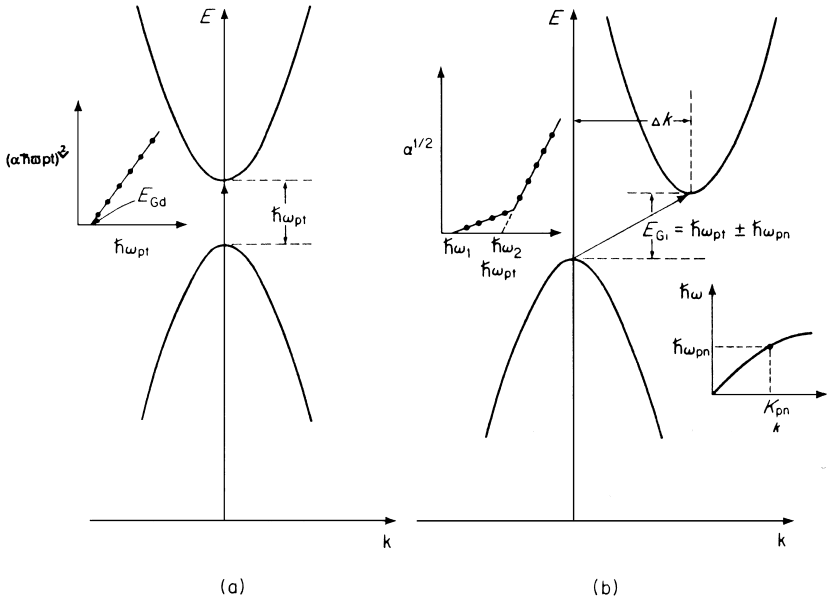



**FIGURE 8** (a) Band-to-band direct transitions, and (b) band-to-band indirect transitions. Inserts show the variation of absorption coefficient with photon energy expected for each type of transition, and a typical phonon dispersion curve for the indirect material.

valence band to an imperfection. If imperfections are present in the crystal, they create energy levels that lie in the forbidden gap. Therefore, at energies less than the band gap it is still possible to excite electrons to the conduction band from imperfection levels occupied by electrons, or to excite electrons from the valence band to unoccupied imperfection levels, each process giving rise to optical absorption. This absorption in turn ceases when the photon energy is less than the energy required to make a transition from the imperfection level to one of the bands. The absorption coefficient may be expressed as $\alpha = S_o N_I$ (where $N_I$ is the density of suitably occupied or unoccupied imperfections, and $S_o$ is the optical cross section, of the order of $10^{-16}$ cm$^2$). For very high imperfection densities, the corresponding absorption constant may have values as high as $10^3$ cm$^{-1}$, but in general it is considerably smaller. The absorption spectrum for this type of absorption consists of a threshold corresponding to the ionization energy of the imperfection and a relatively slow variation of absorption for higher photon energies.

Another kind of imperfection absorption may occur that is not illustrated in Fig. 7. It is most commonly encountered when an impurity with incomplete inner-shell atomic levels is present in a material. Absorption between two such atomic levels yields an absorption spectrum with a Gaussian shape, peaked at the energy separation between the two levels.

5. Excitation of a free electron (or free hole) to a higher energy state within the same band or to higher bands. This process can occur over a wide range of photon energies. It involves the absorption of photons and the absorption or emission of phonons, since both energy and $k$ must be changed in the transition, and is thus an indirect optical transition. The specific quantitative description for the effect depends on the magnitude of both the electrical conductivity and the frequency of the light; for optical frequency excitation of free carriers in a nondegenerate semiconductor $\alpha \propto \omega^{-n}$ where $n$ is between 2 and 3. The results of a classical calculation for the dependence of absorption coefficient on frequency for materials with different low-frequency conductivity are given in Fig. 9.

Another kind of optical absorption due to free electrons (or holes) corresponds to the motion of free carriers acting collectively as a kind of "electron gas," which has a characteristic frequency $\omega_p = (nq^2/\varepsilon_0 m^*)^{1/2}$, where $n$ is the density of the carriers, $q$ is the electronic charge, $\varepsilon_0$ is the permittivity of free space in SI units, and $m^*$ is the effective mass of the carriers. When photons with this frequency are incident on the material, resonant absorption of energy occurs. The effect is known as plasma resonance. The plasma frequency occurs in the ultraviolet for metals and in the infrared for most semiconductors.

## C. Photoelectronic Effects

When light is absorbed by a material so as to raise electrons to higher-energy states, several possibilities occur. If the excited electrons are in the conduction band, then the conductivity of the material is increased as a result of the absorption of light, and the effect is known as

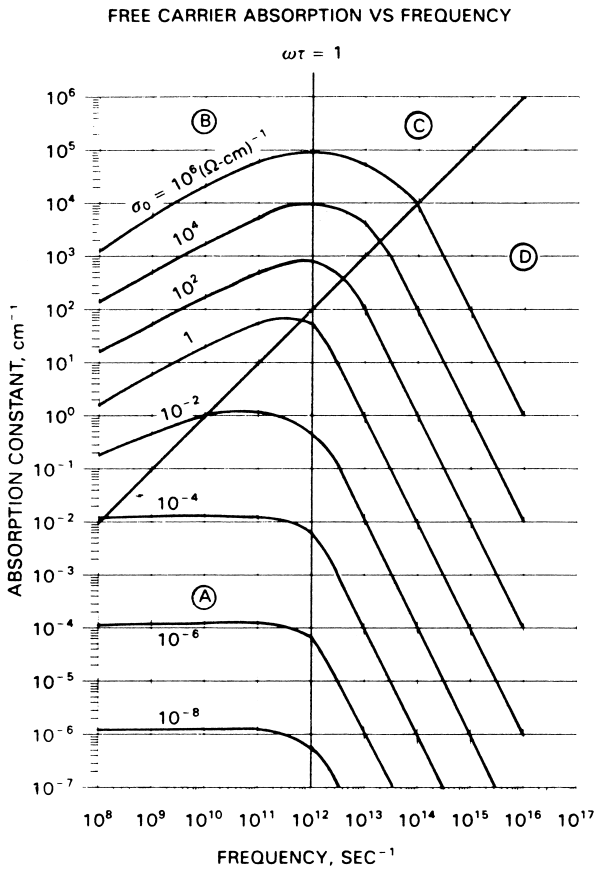



**FIGURE 9** Dependence of the optical absorption coefficient $\alpha$ on the frequency of the electric field $\omega$, for different valus of the low-frequency conductivity $\sigma_0$. The plot is divided into four regions, A through D, corresponding to low and high $\omega$ and low and high $\sigma/\omega$. (A) Low $\omega$, low $\sigma/\omega$: $\alpha \propto \sigma_0$. (B) Low $\omega$, high $\sigma/\omega$: $\alpha \propto (\sigma_0\omega)^{1/2}$. (C) High $\omega$, high $\sigma/\omega$: $\alpha \propto (\sigma_0/\omega)^{1/2}$. The plot assumes that $\tau = 10^{-12}$ sec and that $\varepsilon_r = 10$. (D) High $\omega$, low $\sigma/\omega$: $\alpha \propto \sigma_0\omega^{-2}$.

photoconductivity. If the excited electrons give up their excess energy when they return to their initial state in the form of photons, then the effect is known as luminescence; in particular, if the initial excitation is by light, the emitted radiation is called photoluminescence emission. Processes described here as being associated with the creation of free electrons by light absorption can, of course, also be associated with the creation of free holes.

Luminescence can also be excited by other high-energy sources: excitation by an electron beam produces cathodoluminescence [the conductivity analogue is electron-beam induced current (EBIC)], excitation by friction produces triboluminescence, and luminescence may also be excited by exposure to X-rays or high-energy particles. A material in thermal equilibrium emits radiation due to the recombination of thermally excited electrons and holes, which is commonly known as blackbody radiation and is described by Planck's radiation law. Luminescence is distinguished from this blackbody radiation, and the term is confined to radiation emitted over and above the blackbody radiation.

If photoexcitation produces a free electron–hole pair by excitation across the band gap of a semiconductor, then both the electron and the hole contribute to the increased conductivity of the material until (a) they are captured by localized imperfections, (b) they recombine with each other directly or at localized imperfections, or (c) they pass out of the material at one end without being replaced at the other, an effect that depends on the nature of the electrical contacts to the material. Most often recombination occurs not between a free electron and a free hole (although this process does occur and can be observed through intrinsic luminescence emission when the recombination occurs by photon emission), but between a free carrier of one type and a trapped carrier of the other type, or between trapped carriers of both types that have been trapped near each other in the crystal.

There are three processes by which the excess energy of an excited carrier can be released during recombination:

1. Radiative recombination with emission of photons such that $\Delta E = \hbar\omega_{pt}$, where $\Delta E$ is the energy released during recombination and $\hbar\omega_{pt}$ is the photon energy of the photons generated.
2. Nonradiative recombination with emission of phonons such that $\Delta E = n\hbar\omega_{pn}$, where $\hbar\omega_{pn}$ is the phonon energy of the phonons generated, and where normally it will require the simultaneous (or sequential, in the case of a Coulomb-attractive imperfection) release of $n$ such phonons to release the energy $\Delta E$.
3. Nonradiative (Auger) recombination in which the energy is given up to another free carrier such that $\Delta E = E_{carrier}$, which raises it to a higher energy state in the band, from which it can again drop back to its lower energy with the emission of phonons. This process increases in probability with the density of free carriers present.

Photoelectronic effects are frequently described in terms of optical spectra. These are illustrated in Fig. 10. The major types of such spectra are as follows:

1. Absorption spectra, plotted as $\alpha$ versus $\hbar\omega_{inc}$, where $\hbar\omega_{inc}$ is the photon energy of the incident photons
2. Photoconductivity excitation spectra, plotted as photoconductivity $\Delta\sigma(\sigma_{light} - \sigma_{dark})$ versus $\hbar\omega_{inc}$
3. Luminescence excitation spectra, plotted as luminescence emission intensity at a particular photon energy as a function of $\hbar\omega_{inc}$
4. Luminescence emission spectra, plotted as luminescence emission intensity as a function of $\hbar\omega_{emit}$ of the emitted light

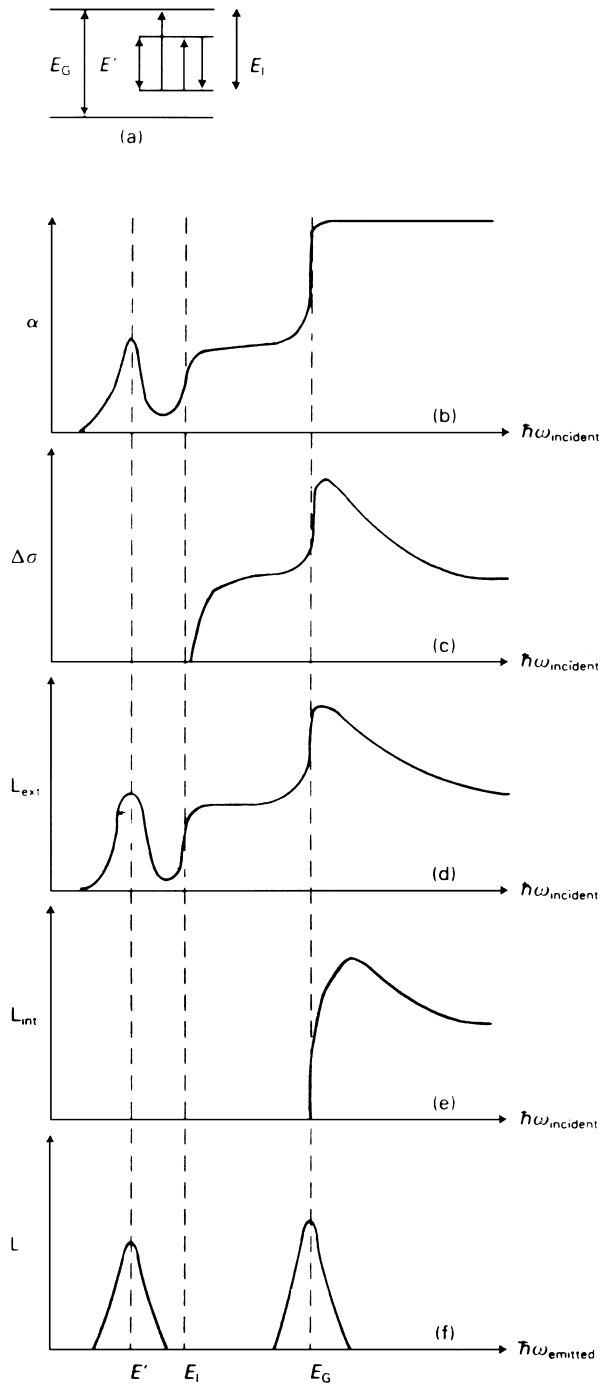



**FIGURE 10** Optical spectra for an illustrative situation. (a) Chosen band structure involving band-to-band transitions and an impurity with atomic ground state and excited state, in which it is assumed that only recombinations with energy $E_G$ or $E'$ are radiative; (b) optical absorption spectrum; (c) photoconductivity excitation spectrum; (d) extrinsic luminescence excitation spectrum, where $L_{ext}$ is the emission intensity of the extrinsic emission band; (e) intrinsic luminescence excitation spectrum, where $L_{int}$ is the emission intensity of the intrinsic emission band; and (f) luminescence emission spectrum showing both extrinsic and intrinsic emission bands.

## V. ELECTRICAL PROPERTIES

The electrical conductivity of different types of materials varies over a wide range, from values of the order of $10^8\ (\Omega\,\mathrm{m})^{-1}$ for metals to less than $10^{-14}\ (\Omega\,\mathrm{m})^{-1}$ for insulators. Semiconductors usually have a room-temperature conductivity of the order of $10^2\ (\Omega\ \mathrm{m})^{-1}$, although this value is strongly dependent on both the temperature and the purity of the semiconductor.

### A. Ohm's Law

The basic equation describing electrical properties in elementary discussions is Ohm's law, $I = V/R$, where $I$ is the electrical current measured in a circuit with resistance $R$ when a voltage difference $V$ is applied. This relation can be rewritten to explicitly indicate the role of electrical conductivity, $\sigma$:

$$\mathbf{J} = \sigma\mathbf{E}, \tag{11}$$

where $\mathbf{J}$ is the electrical current density (current per unit area), and $\mathbf{E}$ is the electric field. Alternatively, we can write $\mathbf{J} = nq\mathbf{v}_\mathrm{d}$, where $n$ is the density of free carriers, $q$ is the charge per carrier, and $\mathbf{v}_\mathrm{d}$ is the drift velocity of the electron caused by the electric field. We define a quantity called the carrier mobility $\mu$ so that

$$\mathbf{v}_\mathrm{d} = \mu\mathbf{E}; \tag{12}$$

then the conductivity can be expressed as

$$\sigma = nq\mu. \tag{13}$$

If both electrons and holes contribute to the conductivity, then the total conductivity can be expressed as

$$\sigma = q(n\mu_n + p\mu_p), \tag{14}$$

where $p$ is the density of holes and $\mu_n$ and $\mu_p$ are, respectively, the mobilities of electrons and holes.

For Ohm's law to hold, it is necessary that neither $n$ nor $\mu$ be a function of electric field. This requirement breaks down for high values of electric field for which $n$ may be increased by the electric field through mechanisms of impact ionization, field emission, or contact injection, or for which $\mu$ may become a function of electric field if the mobility is a function of carrier energy that in turn is increased by high electric fields.

The temperature dependence of electrical conductivity is determined by the temperature dependence of the free-carrier density and the temperature dependence of the free-carrier mobility. In a metal, the free-carrier density is independent of temperature, and therefore the temperature dependence of the conductivity for a metal arises totally from the temperature dependence of the mobility. The free-carrier density in a semiconductor or an insulator

is thermally activated over a wide temperature range and therefore increases exponentially with temperature over this range. The temperature dependence of the mobility depends on the specific scattering process that limits the drift of carriers in an electric field.

## B. Scattering and Mobility

The effect of an electric field on a distribution of free (or quasi-free, to be more exact) carriers in a material is to shift the distribution so that more carriers have velocity in one direction than in the opposite. The magnitude of this shift is determined by the interaction between the effects of the electric field and the effects of a variety of scattering processes that act to return the distribution to its equilibrium condition. If a force $q\mathbf{E}$ acts on an electron for an average time $\tau$, then the net shift in the distribution will be given by $qE\tau = \hbar\,\Delta k$ [see discussion preceding Eq. (6)], so that $\Delta k = qE\tau/\hbar$. We may consider $\tau$ to be the average time between scattering events, which acts in such a way that if $n_0$ carriers are unscattered at $t = 0$, then the density of unscattered carriers $n(t)$ at time $t > 0$ is given by $dn(t)/dt = -n/\tau$, or $n(t) = n_0 \exp(-t/\tau)$.

Using this approach, we may rewrite the above equation for $\Delta k$, as $qE\tau = m^* v_d$, from which it follows that

$$\mu = (q/m^*)\tau. \tag{15}$$

In general the scattering relaxation time is a function of energy $\tau(E)$, and the quantity that enters Eq. (15) as $\tau$ is a suitable average of $\tau(E)$ over electron energies.

The specific form of $\tau(E)$ depends on the scattering mechanism. These mechanisms include scattering by acoustic lattice waves, optical lattice waves, charged imperfections, neutral imperfections, piezoelectric effects, dislocations, grain boundaries, surfaces, and inhomogeneities. Each scattering mechanism is characterized by its own temperature dependence of mobility. As examples of these processes, we consider scattering by acoustic lattice waves and charged imperfections.

Acoustic lattice scattering corresponds to scattering of free carriers by interaction with lattice atoms as they move due to thermal energy. The probability for scattering is proportional to the average energy in the lattice waves, that is, to $kT$. The mean free path for scattering (the average distance traveled between scattering events, equal to the product of the thermal velocity of the carrier $v_{th}$ and $\tau$) by acoustic lattice waves is therefore proportional to $T^{-1}$. The relaxation time $\tau(E) \propto (Tv_{th})^{-1}$. Since in a semiconductor $v_{th} = (2kT/m^*)^{1/2}$, it follows that $\tau \propto \mu \propto T^{-3/2}$. In a metal, on the other hand, scattering events are experienced only by electrons near the Fermi energy, and so the average value of $\tau(E) = \tau(E_F)$; as a result, $\tau \propto \mu \propto T^{-1}$.

Actually, in a metal this is true only at sufficiently high temperatures where scattering by acoustic lattice waves is elastic; in other words, the change in energy on scattering is small compared with the thermal energy of an electron, $kT$. At low temperatures (temperatures less than the Debye temperature $\theta_D$, which is defined approximately by $k\theta_D = \hbar\omega_{max}$, where $\hbar\omega_{max}$ is the maximum vibrational phonon energy) this is no longer true in a metal, and acoustic lattice wave scattering in a metal becomes inelastic; an appropriate calculation shows that over this range $\mu \propto T^{-5}$.

An illustration of the temperature dependence of electrical resistivity $\rho = 1/\sigma$ as a function of temperature for a metal is given in Fig. 11, with data for silver with $\theta_D = 226$ K. Above this temperature the resistivity varies linearly with $T$, whereas below this temperature $\rho$ changes to a $T^5$ dependence, until at very low temperatures the resistivity is limited by impurity scattering. Since the total resistivity for a metal is equal to the sum of the resistivity due to lattice scattering and the resistivity due to impurities

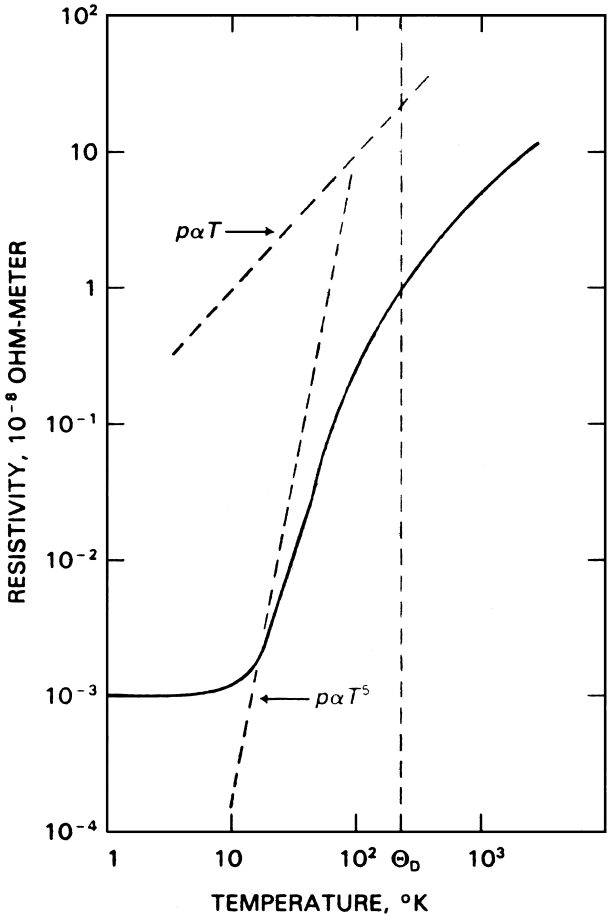



**FIGURE 11** Typical temperature dependence of electrical resistivity for silver with a Debye temperature of 226 K.

(Mathiessen's rule), a large value of $\rho_{300K}/\rho_{4K}$ indicates a pure metal.

Scattering of electrons or holes by charged impurities is illustrated in Fig. 12, which shows that scattering of both electrons and holes by a charged impurity, regardless of the sign of the charge on the impurity (i.e., whether the Coulomb interaction is attractive or repulsive), can be treated in the same mathematical way. In a simple model we consider the scattering effect to be large only if the Coulomb-interaction energy is comparable to the thermal energy of the carrier, that is, if

$$Zq^2/(4\pi\varepsilon_r\varepsilon_0 r_s) = kT, \tag{16}$$

where $Zq$ is the charge on the impurity, $\varepsilon_r$ is the relative dielectric constant of the material, and $r_s$ is a critical radius defined by Eq. (16). The effective scattering cross section $S_I = \pi r_s^2$ and is given by

$$S_I = Z^2q^4/\left(16\pi\varepsilon_r^2\varepsilon_0^2k^2T^2\right). \tag{17}$$

The physical meaning of $S_I$ is that if the carrier comes within the area $S_I$ of the scattering center, then scattering occurs; if not, then no scattering occurs. If there are $N_I$ charged impurities, then the rate of scattering is given by $1/\tau = N_I S_I \nu$, where $\nu$ is the thermal velocity of the carrier. Since $\nu \propto T^{1/2}$, it follows that $\tau \propto \mu \propto T^{3/2}$. At room temperature and for $\varepsilon_r = 10$, $S_I \approx 10^{-12}$ cm$^2$; for $N_I = 10^{17}$ cm$^{-3}$, $\tau = 10^{-13}$ sec.

If a semiconductor has scattering by both acoustic lattice waves and by charged-impurity scattering, then scattering rates add and

$$1/\mu \approx 1/\mu_L + 1/\mu_I, \tag{18}$$

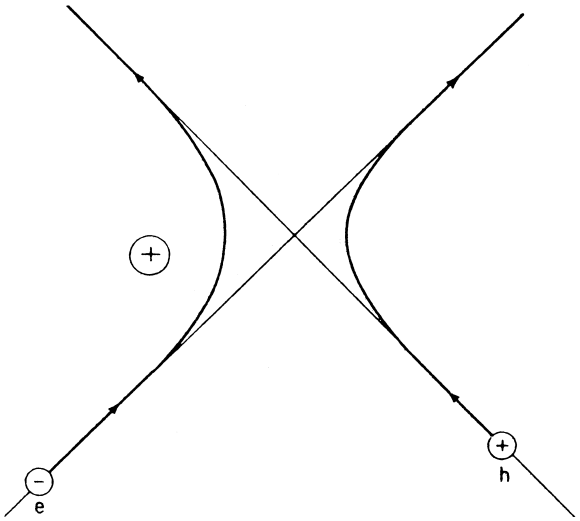



**FIGURE 12** Scattering of an electron by a positively charged impurity by Coulomb attraction, and scattering of a hole by a positively charged impurity by Coulomb repulsion. The hyperbolic paths of electrons and holes are mirror images of each other.

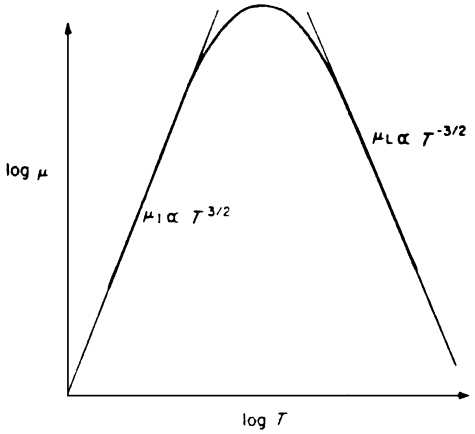



**FIGURE 13** Temperature dependence of mobility in a semiconductor with scattering by both acoustic lattice waves and charged imperfections, which results in a maximum mobility at a particular temperature.

where $\mu_L$ is the lattice-scattering determined mobility and $\mu_I$ is the charged-impurity-scattering determined mobility. The relationship of Eq. (18) is only approximate since the need for different averaging procedures in calculating the various mobilities introduces a correction factor of order unity. A typical situation is illustrated in Fig. 13. If $\mu_L = AT^{-3/2}$ and $\mu_I = BT^{3/2}$, a maximum mobility occurs for $T = (A/B)^{1/3}$.

Scattering by optical lattice waves is an inelastic process in semiconductors, since the optical phonon energy is larger in general than $kT$. Since the scattering probability is proportional to the density of optical phonons present, the temperature dependence for scattering by optical modes is approximately given by the Bose-Einstein distribution for optical phonons of energy $\hbar\omega_{pn}$ at a temperature $T$, $\mu_0 \propto [\exp(\hbar\omega_{pn}/kT) - 1]$.

Electron mobilities are usually larger than hole mobilities in most semiconductors because $m_e^* < m_h^*$. Electron mobilities at room temperature range from about $10^{-2}$ m$^2$/V-s (ZnS) to almost 10 m$^2$/V-s (InSb).

## C. Imperfections in Semiconductors

In most practical semiconductors, the electrical conductivity is controlled not by thermal excitation across the band gap of the material, but by thermal excitation from localized imperfections. The electrical behavior of imperfections depends on their location in the crystal and their effective number of valence electrons.

The simplest kind of imperfection is one that differs by one in valence from the atom for which it substitutes in the crystal lattice. Such imperfections give rise to localized energy levels in the forbidden gap of the semiconductor lying close to either the conduction or the valence band

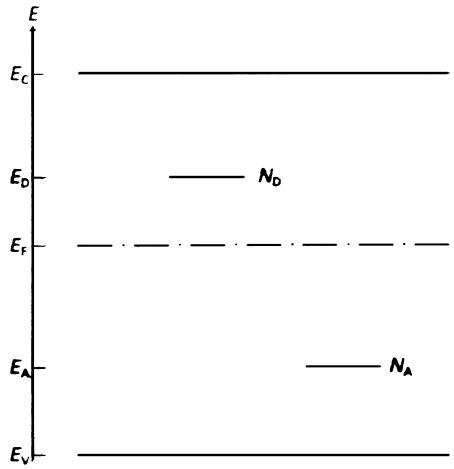



**FIGURE 14** Localized donor and acceptor levels in a semiconductor, showing the donor ionization energy ($E_c - E_D$) and the acceptor ionization energy ($E_A - E_v$).

edge. They can be treated approximately as if they were miniature hydrogenic systems, corrected for the effective mass and the dielectric constant of the semiconductor.

Such imperfections can be conveniently divided into donors and acceptors. A donor has one more valence electron than the atom for which it substitutes, is able to give up this electron to the conduction band if sufficient thermal excitation energy is available, and is neutral when electron-occupied and positive when ionized. Figure 14 shows typical donor and acceptor energy levels in a band diagram. The amount of energy required to free an electron from the donor is the ionization energy of the donor. The donor energy level is located at an energy $E_D$ so that the energy difference $(E_c - E_D)$ is the ionization energy of the donor. Ionization of a donor can be thought of as described by the equation $D^0 \rightarrow D^+ + e^-$.

An acceptor has one less valence electron than the atom for which it substitutes, is able to give up a hole to the valence band if sufficient thermal excitation energy is available, and is neutral when hole-occupied and negative when ionized (electron-occupied). The amount of energy required to free a hole from the acceptor is the ionization energy of the acceptor. The acceptor energy level is located at an energy $E_A$ so that the energy difference $(E_A - E_v)$ is the ionization energy of the acceptor. Ionization of an acceptor can be thought of as described by the equation $A^0 \rightarrow A^- + h^+$. Simple donors and acceptors of this type usually have ionization energies between about 10 and 30 meV.

When donors and acceptors are simultaneously present in the same semiconductor, the electrons donated by the donors may be accepted by the acceptors. By this process, ionized donors and acceptors are produced without the corresponding free electrons or holes. When this happens, the semiconductor is said to be compensated.

An imperfection differing in valence by more than one from the atom for which it substitutes usually gives rise to levels that lie near the middle of the forbidden gap. Different energy levels are needed to describe different states of ionization of the imperfection. For example, Zn (2 valence electrons) in Si (4 valence electrons) can accept either one electron (becoming a $Zn^-$ state) or two electrons (becoming a $Zn^{-2}$ state); the energy level corresponding to $Zn^{-2}$ lies 0.58 eV above the valence band in Si (with a band gap of 1.1 eV), whereas the level corresponding to $Zn^-$ lies 0.33 eV above the valence band in Si. In this kind of discussion the charges given on the Zn represent differences in charge between the site occupied by a Zn and the normal site in the Si lattice.

## D. Fermi Level in Semiconductors

In a nondegenerate semiconductor (one in which the Fermi level lies in the forbidden gap more than $kT$ away from a band edge), the location of the Fermi level can be calculated in a way similar to that used for a metal; that is, by calculating the total free-electron density $n$ in the conduction band by integrating $n = \int N(E) f(E)\, dE$ from the bottom of the conduction band at $E_c$ to the top of the conduction band, which can be taken effectively as infinity. In this case $f(E)$ can be expressed by a simple Boltzmann factor, since the energies of interest in the conduction band $E \gg E_F$. The result is that

$$n = N_c \exp[-(E_c - E_F)/kT], \tag{19}$$

where $N_c$ is called the effective density of states in the conduction band and is given by $N_c = 2(2\pi m_e^* kT/h^2)^{3/2}$. By performing a similar calculation for the free-hole density by integrating over the valence band, we obtain the result that

$$p = N_v \exp[-(E_F - E_v)/kT], \tag{20}$$

with $N_v = 2(2\pi m_h^* kT/h^2)^{3/2}$, and is called the effective density of states in the valence band. When $n$ becomes comparable to $N_c$ (or $p$ to $N_v$), the Fermi level lies in the conduction band (valence band) and the semiconductor is said to be degenerate. The consequence is that the full Fermi function, and not just the Boltzmann function tail, must be used to calculate the occupancy of band states.

Equations (19) and (20) show that if any two of the following three quantities are known—free-carrier density, $E_F$, or $T$—the third can be calculated. The product of Eqs. (19) and (20) gives a constant for the material at a given temperature:

$$np = N_c N_v \exp(-E_G/kT), \tag{21}$$

since $(E_c - E_v) = E_G$, the band gap of the semiconductor.

### E. Electrical Conductivity in Semiconductors

In an intrinsic semiconductor (a semiconductor with properties controlled by excitation and recombination across the band gap rather than through imperfections), the temperature dependence of the Fermi level (and hence the temperature dependence of the free-carrier densities) can be calculated simply by equating Eqs. (19) and (20), since in an intrinsic semiconductor $n = p = n_i$. The result is that

$$(E_c - E_F) = E_G/2 + (3kT/4)\ln\left(m_e^*/m_h^*\right). \qquad (22)$$

In an extrinsic semiconductor (a semiconductor with properties controlled by excitation and recombination through imperfections), the situation may be a little more complicated, but a simple rule may be stated that covers all cases. In the intrinsic case above, the requirement that $n = p$ may be recognized as a condition of charge neutrality. This condition of charge neutrality may be used in general to determine the temperature dependence of the Fermi level (and hence of the free-carrier density and the conductivity).

Consider the case shown in Fig. 14 with one kind of donor and one kind of acceptor imperfection present in the material. The general charge-neutrality condition that governs this case is given by

$$n + n_A = (N_D - n_D) + p, \qquad (23)$$

where the negative species on the left are the free-electron density and the density of ionized (electron-occupied) acceptors, and the positive species on the right are the density of ionized (electron-unoccupied) donors and the free-hole density. Each of these terms can be written in terms of the location of the Fermi level, and hence the Fermi-level position satisfying the equation can be readily determined by a computer even if there are many different donors and acceptors present.

The information needed to accomplish this lies in the appropriate expressions for electron-occupied donors and hole-occupied acceptors. These may be summarized as follows:

$$n_D = N_D\left\{\left(\tfrac{1}{2}\right)\exp[(E_D - E_F)/kT] + 1\right\}^{-1} \qquad (24a)$$

and

$$(N_A - n_A) = N_A\left\{\left(\tfrac{1}{2}\right)\exp[(E_F - E_A)/kT + 1\right\}^{-1}. \qquad (24b)$$

In these equations $N_D$ is the total density of donors with an ionization energy of $(E_c - E_D)$, and $N_A$ is the total density of acceptors with an ionization energy of $(E_A - E_v)$. Notice that the expressions are similar only if we compare the expression for electron-occupied donors with that for hole-occupied acceptors. Equations (24) are what would be expected from our previous discussion of the Fermi distribution, except for the additional degeneracy factor of $(\frac{1}{2})$ that appears in Eqs. (24). This factor arises because donor and acceptor states can accommodate an electron with either of two possible spin orientations; taking this into account leads to the insertion of the factors of $(\frac{1}{2})$ shown in Eqs. (24).

The general statement of Eq. (23) can be simplified in several special cases with physical significance. For an intrinsic material, $n_A = (N_D - n_D) = 0$ and $n = p$, as we have seen. If only donors are present, then $n_A = 0$ and $p$ can be neglected in the extrinsic conductivity range, so that $n = (N_D - n_D)$. Similarly, if only acceptors are present, then $(N_D - n_D) = 0$ and $n$ can be neglected in the extrinsic conductivity range, so that $p = n_A$. If both donors and acceptors are present in approximately equal densities, so that almost complete compensation occurs, $n_A = (N_D - n_D)$.

Analytical results can be obtained for simple cases. As an example, consider the case of donors only. Two ranges may be defined. In the high-temperature range where all the donors are ionized, $n = N_D$ and $(E_c - E_F) = kT\ln(N_c/N_D)$ from Eq. (19). In the low-temperature range, the donors are only partially ionized and

$$(E_c - E_F) = (E_c - E_D)/2 + (kT/2)\ln(2N_c/ND) \qquad (25a)$$

with

$$n = (N_c N_D/2)^{1/2}\exp[-(E_c - E_D)/2kT]. \qquad (25b)$$

Therefore a plot of $\ln(nT^{-3/4})$ versus $1/T$ yields a straight line at low temperatures with a slope proportional to one-half the donor ionization energy. These results are summarized in Fig. 15.

The case of one kind of donor with density $N_D$ and one kind of acceptor with density $N_A$ can also be solved analytically. Consider the case first where $N_D > N_A$. At low temperatures for which $n \ll N_A$ or $(N_D - N_A)$, a plot of $\ln(nT^{-3/2})$ versus $1/T$ has an activation energy of $(E_c - E_D)$. For intermediate temperatures for which $N_A \ll n \ll (N_D - N_A)$, a plot of $\ln(nT^{-3/4})$ versus $1/T$ has an activation energy of $(E_c - E_D)/2$. Finally, all of the uncompensated donors are ionized at high temperatures and $n = (N_D - N_A)$. If $N_D = N_A$, we have exact compensation, and the Fermi level must lie halfway between the donor and acceptor levels to achieve charge neutrality; the observed activation energy is therefore $(E_D + E_A)/2$.

## VI. GALVANOMAGNETOTHERMOELECTRIC EFFECTS

When any two of the following interact—an electric field, a magnetic field, and a thermal gradient—a number of electrical effects result. Some of these are small

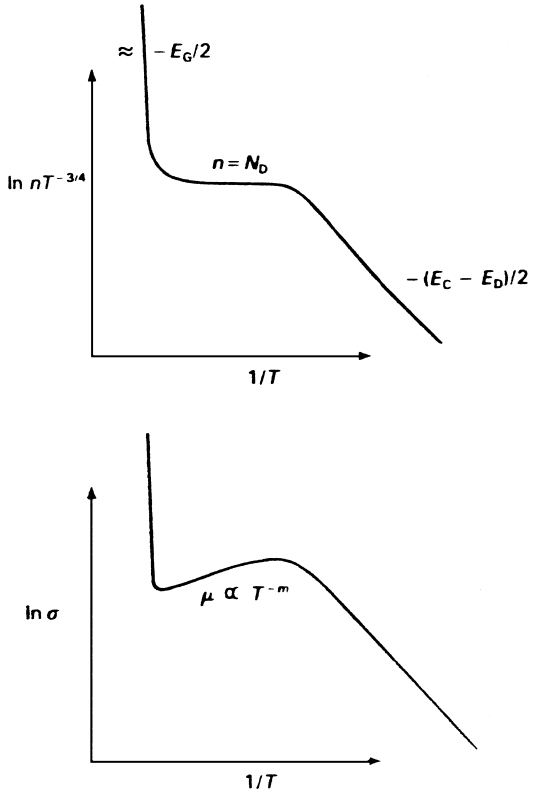



**FIGURE 15** Electron density and conductivity as a function of temperature for a semiconductor with donors only with indicated activation energies. The slopes for the ln $\sigma$ versus $1/T$ plot differ slightly from those of the ln $(nT^{-3/4})$ versus $1/T$ plot because of the temperature dependence of $N_c$ and $\mu_n$ in the extrinsic range. In the intrinsic range, none of the slopes seen will be exactly $-E_G/2k$ because in this range a plot should be made of ln $(nT^{-3/2})$ versus $1/T$.

second-order effects (Nernst, Ettingshausen, and Righi-Leduc effects). Here, we consider three of the major effects: the Hall effect, magnetoresistance, and thermoelectric power.

## A. Cyclotron Resonance Frequency

Classically, the effect of a magnetic field on free electrons can be readily described. The force exerted on a moving electric charge $q$ with velocity $\mathbf{v}$ by a magnetic field $\mathbf{B}$ is given by $q\mathbf{v} \times \mathbf{B}$. This force causes an electron to move in a circular orbit in the plane orthogonal to $\mathbf{B}$, that is, in the $x$–$y$ plane for $B_z$. The radius of the circular orbit can be determined by equating the magnetic force to the centrifugal force of the circular motion: $m_e^* v^2/r = qvB_z$. The angular frequency $\omega_c = v/r$ is called the cyclotron frequency and is given by

$$\omega_c = qB_z/m_e^*. \tag{26}$$

If a circularly polarized electric field with frequency $\omega = \omega_c$ is applied to such a free electron in a magnetic field, resonant absorption takes place. This effect provides a direct method for the measurement of the effective mass of carriers.

For this to be possible the product $\omega_c \tau \gg 1$, where $\tau$ is the relaxation time for scattering. This in turn implies that large values of $B_z$ must be used, which puts the problem out of the classical realm and into the quantum-mechanical realm where Schroedinger's equation must now be solved with an energy term involving the magnetic field. The net result is that the resonant absorption that occurs when $\hbar\omega = \hbar\omega_c$ corresponds in the quantum-mechanical picture to a transition between two of the quantized band states that result from the high magnetic field.

## B. Hall Effect

In the Hall effect, an electric field and a magnetic field applied at right angles to a material with either free electrons or free holes produce an electric field in the third orthogonal direction, in order to produce zero current in that direction. A magnetic field $B_z$ in the presence of an electric field $E_x$ producing a current density $J_x$ causes electrons to acquire velocity in the $y$-direction. Since there can be no net current flow in the $y$-direction, however, it follows that an electric field $E_y$ will form such that $qE_y = qv_xB_z$. Then the induced Hall field $E_y$ can be measured and the carrier density deduced from it:

$$E_y = \pm(1/nq)J_xB_z. \tag{27}$$

The polarity of $E_y$ indicated by the choice of sign tells whether the carriers are electrons (negative sign) or holes (positive sign). Two commonly defined parameters are the Hall coefficient $R_H = \pm 1/nq$ and the Hall mobility $\mu_H = \sigma R_H = \pm\mu$.

The Hall mobility, always defined as $\mu_H = \sigma R_H$, is a second kind of mobility, in addition to the conductivity mobility defined in Eq. (15). A third kind of mobility often referred to is the drift mobility $\mu_d = d/Et$, obtained from a direct measurement of the time $t$ required for carriers to travel a distance $d$ in the material under an electric field $E$. If there are localized trapping states in the material associated with imperfections, a carrier injected at $x = 0$ may spend a major portion of its time in the material in a trapped state rather than in a free state, and therefore may spend much longer in reaching the detection point at $x = d$ than simply $d/(\mu_{con}E)$. We may express the drift mobility in terms of the conductivity mobility, the density of free carriers $n$, and the density of trapped carriers $n_t$ by

$$\mu_d = \mu_{con}[n/(n + n_t)]. \tag{28}$$

If both electrons and holes are present at the same time, since the value of $E_y$ associated with the electrons is of opposite sign to that associated with the holes, the possibility exists for exact cancellation. By considering the total current due to both electrons and holes, we find that in general (for small values of $B_z$),

$$E_y = \left[q\left(p\mu_p^2 - n\mu_n^2\right)/\alpha^2\right]J_x B_z, \tag{29a}$$

$$R_H = E_y/J_x B_z = \left(p\mu_p^2 - n\mu_n^2\right)/\left[q(p\mu_p + n\mu_n)^2\right], \tag{29b}$$

and

$$\mu_H = \sigma R_H = \left(p\mu_p^2 - n\mu_n^2\right)/(p\mu_p + n\mu_n). \tag{29c}$$

The Hall coefficient and the Hall mobility both become zero if $p\mu_p^2 = n\mu_n^2$.

### C. Magnetoresistance

Magnetoresistance is the change in resistance corresponding to an applied electric field because of the effects of a simultaneously applied magnetic field. In the Hall-effect geometry, to be specific, it means a change in electrical resistance corresponding to an electric field $E_x$ because of $B_z$. To bring about $J_y = 0$ by the development of a specific value of $E_y$, as described above for the Hall effect, for all the carriers, essentially requires that all the carriers have identical properties. When all the carriers do not have identical properties (e.g., some are electrons and some holes, the scattering relaxation time varies with energy and carriers have a range of energies, or the energy-band structure is such that the effective mass is not a scalar), the development of $E_y$ cannot remove all of the $y$-component of velocity for all the carriers. Some carriers travel a relatively greater distance than others between end contacts, and an increase in resistance results. Although the effect is small, it has been of research interest because of the dependence of magnetoresistance effects on various kinds of energy-band structures.

### D. Thermoelectric (Seebeck) Effect

The application of a temperature gradient to a material causes the average energy of free carriers at the hot end to increase, which thus establishes a concentration gradient along the material. Diffusion associated with this concentration gradient is counteracted by the buildup of an electric field due to the displaced charge, to satisfy the condition that the total current be zero. The magnitude of the voltage per degree difference is called $\alpha$, the thermoelectric power in V/K. The effect is much smaller in a metal than in a semiconductor, being of the order of microvolts per degree in the former and of a millivolt per degree in the latter. Typical expressions for $\alpha$ are as follows:

$$\text{Metal } \alpha = \pi^2 k^2 T/qE_F, \tag{30a}$$

$n$-type Semiconductor $\alpha$

$$= -(k/q)[A + (E_c - E_F)/kT], \tag{30b}$$

and

$p$-type Semiconductor $\alpha$

$$= +(k/q)[A + (E_F - E_v)/kT]. \tag{30c}$$

Here $A$ is a constant depending on the specific scattering mechanism for the free carriers involved; for acoustic lattice scattering $A = 2$, and for charged-impurity scattering $A = 4$. Since the thermoelectric power in a semiconductor provides knowledge of the location of the Fermi level at a given temperature, it can also be used to provide direct information about the magnitude of the free-carrier density in the semiconductor.

## VII. AMORPHOUS SEMICONDUCTORS

Amorphous semiconductors are a class of semiconducting materials that do not show the long-range order typical of crystalline materials with a periodic potential that are discussed primarily in this article. Amorphous materials are generally made by one of three methods: (1) deposition from the vapor phase, (2) cooling from a liquid melt (called "glasses"), and (3) transformation of a crystalline solid by particle bombardment, oxidation, and so forth.

There are three general categories of amorphous materials: (1) covalent solids such as tetrahedral films of Group IV elements, or III–V materials; tetrahedral glasses formed from II–IV–V ternary materials (e.g., $CdGeAs_2$); or chalcogenide glasses formed from the Group VI elements, or IV–V–VI binary and ternary materials; (2) oxide glasses such as $V_2O_5$–$P_2O_5$ that have ionic bonds and show electrical conductivity between different valence states of the transition metal ion; and (3) dielectric films such as $SiO_x$ and $Al_2O_3$.

Because of the lack of long-range order, amorphous semiconductors do not show the same kind of density-of-states band gap as described above for crystalline semiconductors. A schematic density-of-states band diagram for an amorphous semiconductor is given in Fig. 16. The density of states continues through the points $E_v$ and $E_c$ that would mark the band gap in a typical crystalline material, but the mode of transport changes from that characteristic of extended states (somewhat like that in the allowed bands of a crystalline material) for electrons above $E_c$ (holes below $E_v$) to that characteristic of localized states

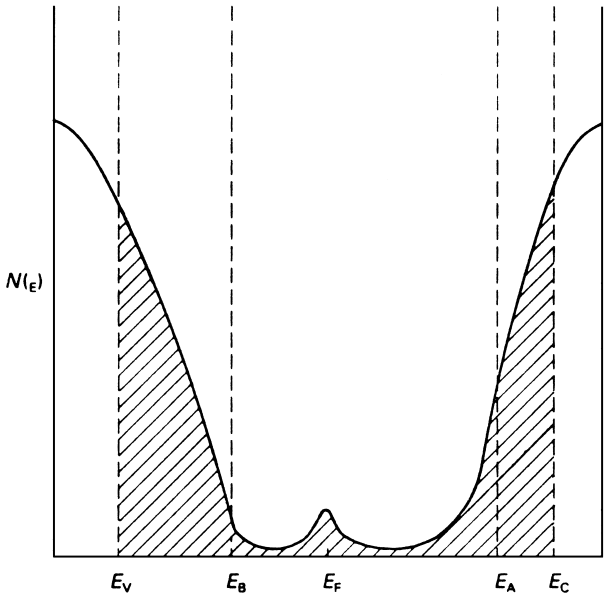



**FIGURE 16** Mott-Davis energy-band model for an amorphous semiconductor, with extended states ($E < E_v$) and $E > E_c$), localized tail states ($E_v < E < E_B$ and $E_A < E < E_c$), and localized defect states ($E_B < E < E_A$).

for electrons immediately below $E_c$ (holes immediately above $E_v$). Since the mobility for transport in extended states is appreciably larger than for that in localized states (where transport must be by some kind of localized hopping from one state to another), we may say that the crystalline density-of-states gap has become a mobility gap in an amorphous material. Furthermore, the density of localized states remains very high throughout the gap in a typical amorphous semiconductor, to the extent that doping of the semiconductor (changes in its free-carrier density by impurity incorporation) is essentially impossible.

In the early 1970s a major breakthrough occurred when W. E. Spear and P. G. LeComber showed that if hydrogen was incorporated into amorphous silicon, the density of the localized states in the gap could be greatly decreased, which would make possible the control of the conductivity in *a*-Si:H by incorporation of impurities similar to normal expectations. Such *a*-Si:H is found to have a tail of localized states decreasing exponentially in density with energy separation from either electron or hole mobility edge, and a residual density of defects in the center of the gap that may be as low as $10^{15}$ cm$^{-3}$ rather than the value of $10^{18}$ cm$^{-3}$ or higher typical of other amorphous semiconductors. The possibility of a thin-film material with electronic properties at least resembling those of crystalline silicon has led to a large research and development effort with applications of the *a*-Si:H material to solar cells, electrophotography, thin-film transistors, solid-state image sensors, optical recording, and a variety of amorphous junction devices.

Two additional properties of *a*-Si:H are of interest: (1) an increase in localized defect density caused by doping, and (2) an increase in localized defect density caused by illumination (the Staebler-Wronski effect), a reversible effect that can be removed by low-temperature annealing but constitutes a degradation mechanism for *a*-Si:H devices.

## VIII. SUPERCONDUCTORS

In 1911 H. K. Onnes showed that the electrical resistance of mercury suddenly vanished when the temperature was reduced below 4.15 K, and a new state of matter was discovered. It has been subsequently found that a number of metals and alloys show this zero resistance at sufficiently low temperatures, and a search has been under way since then to produce materials with higher critical temperature, the temperature below which superconductivity becomes possible. Superconductors are characterized not only by a critical temperature, but also by a critical magnetic field and a critical current density, none of which can be exceeded if the superconducting state is to be retained.

The accepted theory that describes superconductivity in these materials was developed in the 1950s by Bardeen, Cooper, and Schrieffer (the BCS theory). It proposes that in certain metals a new state of matter is possible at low temperatures, which results from an attraction between pairs of electrons through a phonon interaction that overcomes the Coulomb repulsion between them. This new state has an energy lower than that of the normal $E = 0$ state of free electrons and is separated from it by a superconducting energy gap that is larger than the energy of phonons available for scattering at this low temperature. Scattering therefore ceases, since there is no energy conserving final states for the scattering transition; the scattering relaxation time becomes infinite; and the resistance goes to zero.

A search for materials with higher critical temperature had extended the range up to 22.3 K for $Nb_3Ge$ by 1973, an increase of only 4 K for the previous 20 years of effort. Suddenly in 1987 new types of superconducting materials were discovered, not in the family of metals and alloys, but in a variety of materials based on copper oxide. These materials have critical temperatures much higher than any found earlier. The materials involved include such complex systems as $La_{2-x}Ba_xCuO_4$ and variations produced by replacing La with Y and replacing Ba by Sr or Ca. A representative material is $YBa_2Cu_3O_7$ which has a critical temperature of 95 K. Since the temperature of

liquid nitrogen is 77 K, this material has a critical temperature well into the range where practical applications become feasible, provided that a variety of materials problems can be solved.

A search continues for an adequate theoretical description of these materials that is able to account for the much higher values of critical temperature. The central need is for a mechanism that provides stronger attractive coupling of Cooper pairs than the conventional electron–phonon interaction.

## IX. CONDUCTING POLYMERS

Inorganic metals, semiconductors, and insulators are not the only materials that exhibit electrical conductivity effects. Polymers are a class of compounds that were thought to be insulators but have been shown to conduct electricity when treated or prepared in appropriate ways. Unlike crystalline three-dimensional lattices, polymers are more nearly characterized by a chain of repeating units in an approximately one-dimensional layout.

An example of such a polymer is 7,7,8,8-tetracyano-*p*-quinodimethane (TCNQ) and tetrathiafulvalene (TTF). In the compound (TTF) (TCNQ) a conduction band is formed by the overlapping of wave functions that allows a transfer of charge between the two parts of the compound, with the TTF playing the role of electron donor and the TCNQ playing the role of electron acceptor. After the charge transfer, both polymers have partially filled bands and can sustain a current in an essentially one-dimensional geometry. When this material is cooled to 60 K, its conductivity is about the same as that of copper.

Other examples of conducting polymers include poly(sulfer nitride), $[SN]_x$, which is an inorganic polymer and has a larger work function than that of any of the elemental metals; polyacetylene, $[CH_x]$, which can be made with conductivities over a wide range by incorporating impurities; poly(*p*-phenylene) (PPP) and poly(*p*-phenylene sulfide); polypyrrole; and phthalocyanine. Polymers have also shown superconductivity with low values of critical temperature.

Indirectly related to polymers are development involving large spherical molecules constructed of carbon atoms, of which $C_{60}$ is perhaps the most widely known and has been given the name "buckminsterfullerene" after the geodesic dome structures advanced by Buckminster Fuller. Other carbon clusters such as $C_{70}$ and $C_{84}$ have also been observed. Such structures have been nicknamed "buckyballs." They are structurally sturdy but able to accept electrons and hence react with many organic chemicals. Potassium-doped $K_3C_{60}$ has been shown to exhibit superconductivity with a critical temperature of 19.3 K.

## X. JUNCTIONS

Since the 1950s, perhaps no single property of materials has been developed as much with such far-reaching consequences as the cornucopia made available by junctions between different kinds of materials. The whole modern solid-state electronics world is made possible by the special properties of junctions in the semiconductor silicon between *p*-type (free-hole carriers) and *n*-type (free-electron carriers) portions of the material.

Since the subject of "Junctions" is treated in detail elsewhere in this Encyclopedia, our purpose in this section is simply to give a brief introduction and survey.

At least five types of junctions can be enumerated, with a variety of combinations of these also being of interest: (1) a material surface, representing a junction between a material and vacuum or a gaseous environment; (2) junctions between two different metals with different work functions; (3) junctions between a metal and a semiconductor (an MS junction) as commonly encountered in making ohmic (low-resistance) or blocking (rectifying) electrical contacts to a material; (4) junctions between two portions of the same material (homojunctions) with different electrical properties, most commonly one having *p* type and the other *n* type, to form a *p*–*n* junction; and (5) junctions between two different materials (heterojunctions) with different electrical properties. Another class of junctions consists of variations of (3)–(5) with a thin insulating layer (I) such as an oxide between the two junction materials to produce such structures as MIS junctions or SIS junctions.

Junctions between two materials occur when the work function (the energy separation between the vacuum level and the Fermi level) in the two materials is different. When junctions are formed, charge transfer takes place between the two different constituents so as to build up an internal electric field that allows the Fermi energy to be constant throughout the whole two joined materials. When two metals with different work functions are brought into contact, the potential difference between them is equal to the difference in their work functions and is known as the contact potential. When metals or semiconductors are brought into contact, the potential difference between them is known as the diffusion potential $\phi_D$. In every case, the diffusion potential of a junction is equal to the difference in the work functions of the two constituents forming the junction.

Typical energy-band diagrams for junctions are given in Fig. 17 for a rectifying contact between a metal and an *n*-type semiconductor (known as a Schottky barrier), in Fig. 18 for an ohmic contact between a metal and an *n*-type semiconductor, in Fig. 19 for a *p*–*n* homojunction, and in Fig. 20 for *p*–*n* heterojunctions.

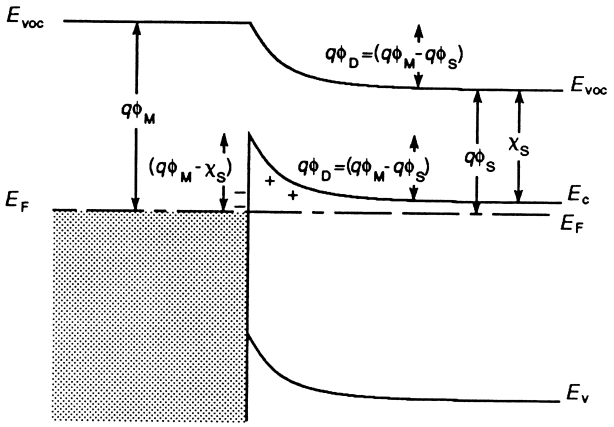



**FIGURE 17** A Schottky barrier between a metal and an *n*-type semiconductor, formed when the work function of the metal is larger than that of the semiconductor. Shown are the work functions of the metal $q\phi_M$ and the semiconductor $q\phi_S$, the diffusion potential $\phi_D$ of the junction, and $\chi_S$, the electron affinity of the semiconductor.

The energy-band diagrams in Figs. 17, 19, and 20 all show semiconductor regions that have been depleted of carriers in the formation of the internal fields required to equalize the Fermi levels. These depleted regions are known as depletion layers. Their width $w_d$ can be calculated from Poisson's equation, which for a depletion layer in an *n*-type material with a Schottky barrier is $\partial^2\phi/\partial x^2 = -qN_D^+/\varepsilon_r\varepsilon_0$. The solution for the appropriate

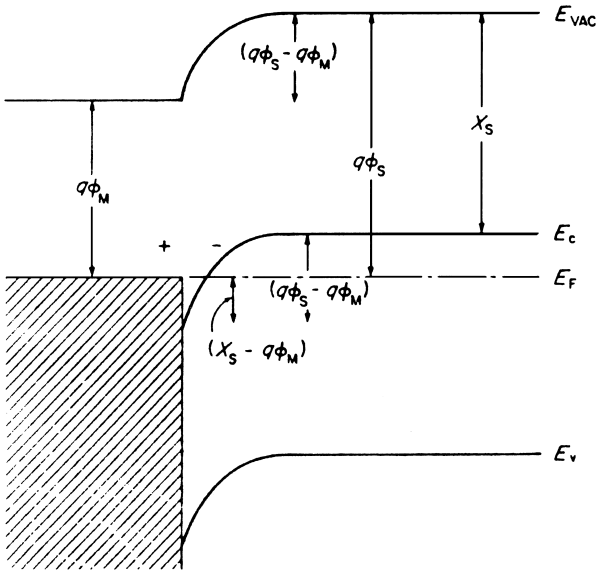



**FIGURE 18** An ohmic contact between a metal and an *n*-type semiconductor, formed when the work function of the metal is smaller than that of the semiconductor. Instead of the depletion layer in the semiconductor shown in Fig. 17, the ohmic contact has an accumulation layer in the semiconductor.

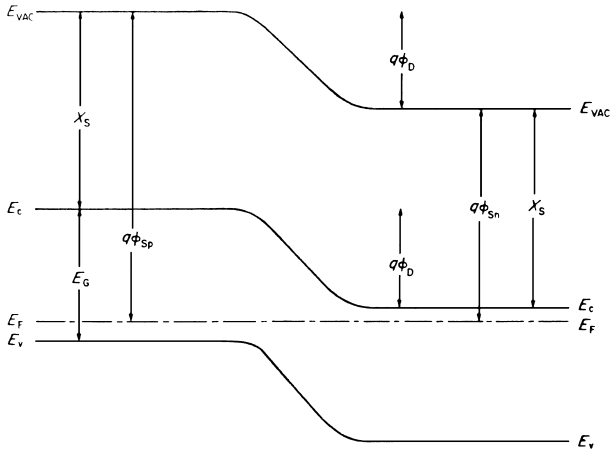



**FIGURE 19** Energy-band diagram of a *p*–*n* homojunction.

boundary conditions that $E_x = \partial\phi/\partial x = 0$ and $\phi = \phi_D$ at $x = w_d$, and $\phi = 0$ at $x = 0$, where $x = 0$ marks the junction interface and $\phi_D$ is the diffusion potential, is

$$w_d = \left(2\varepsilon_r\varepsilon_0\phi_D \big/ N_D^+ q\right)^{1/2}. \tag{31}$$

In cases involving homojunctions and heterojunctions, the depletion layer will be shared by the materials on both sides of the junction, depending on the respective charged-imperfection densities and the dielectric constants; in the event that one side of the junction is much more conducting than the other, the depletion layer will be essentially limited to the less-conducting material, as in the case of a metal–semiconductor junction.

An extremely useful situation results from the realization that, given normal plane geometry, a junction can be thought of as a parallel-plate condenser with the depletion region as the dielectric between the condenser plates.

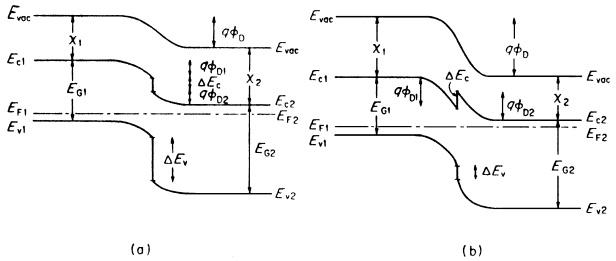



**FIGURE 20** Energy-band diagrams for *p*–*n* heterojunctions. The materials in (a) and (b) have the same band gaps, but in (a) the *p*-type material has a smaller electron affinity than that of the *n*-type material, whereas in (b) the situation is reversed. Heterojunctions are generally characterized by abrupt changes in the conduction band $\Delta E_c$ and valence band $\Delta E_v$ because of differences in the electron affinities and band gaps of the two materials making up the junction; the sign of these changes is of critical significance for the behavior of the heterojunction.

Since the capacitance $C = \varepsilon_r \varepsilon_0 A/w_d$ where $A$ is the area of the junction,

$$(C/A)^{-2} = \left[2(q\phi_D - q\phi_{app})/\varepsilon_r \varepsilon_0 q^2 N_D^+\right], \qquad (32)$$

where $\phi_{app}$ is the magnitude of the voltage applied to the junction from an external source. A plot of $(C/A)^{-2}$ versus $\phi_{app}$ according to Eq. (32) yields a straight line with intercept on the voltage axis of $\phi_D$ and with slope inversely proportional to $N_D^+$.

Figure 18 differs from the other three cases in that for an ohmic contact between a metal and a semiconductor, there is an excess of free carriers in the semiconductor such that an accumulation layer exists. The width of this accumulation layer is determined by the balance between drift and diffusion currents occurring in this region.

The dependence of current on applied voltage for the junctions of Figs. 17, 19, and 20 can be described by a relationship of general form

$$J = J_0[\exp(q\phi_{app}/AkT) - 1]. \qquad (33)$$

This equation describes a current that increases exponentially with $+\phi_{app}$ (forward bias) and that reaches a voltage-independent reverse current of $-J_0$ for $-\phi_{app}$ (reverse bias). Actual current transport mechanisms determine the specific dependences of $J_0$ and $A$. For a Schottky barrier with transport controlled by thermionic emission, or for an ideal homojunction with transport controlled by diffusion, $A = 1$.

Because of the extreme versatility of the standard $p$–$n$ homojunction, it is useful for a wide spectrum of applications, including rectifiers, amplifiers (transistors), photodetectors, photovoltaic energy converters, and radiation emitters, both noncoherent and lasers.

Properties and applications of single-crystal or thin-film $p$–$n$ junction devices are being greatly expanded by the ability acquired in recent years that makes it possible to deposit multilayer heterojunctions with atomically abrupt interfaces and controlled composition and doping in individual layers that are only a few tens of nanometers thick. These layers are so thin that the energy levels in them show the quantum effects of small potential well thickness, which causes the continuous allowed levels associated with macroscopic thicknesses to become the discrete allowed spectrum described by Eq. (1) for small values of $L$. It becomes possible to fine-tune a whole range of properties through choice of the appropriate structural parameters.

Structures called quantum wells are formed by sandwiching a very thin layer of a small-band-gap material between two layers of a wide-band-gap material. The actual values of the discrete levels in a quantum well are determined by the thickness and the depth of the well, which is the band discontinuity (electron band discontinuity $\Delta E_c$ for electrons and valence band discontinuity $\Delta E_v$ for holes). If many quantum wells are grown on top of one another, and the barriers are made so thin that tunneling between them is significant, the result is a superlattice, a concept first proposed by Esaki and Tsu in 1969. Properties of superlattices can be varied, not only by the choice of the materials used to make up the heterojunctions, but also by the spacing of the layers and their thickness. Thus, it becomes possible to fine-tune a whole range of properties through choice of the appropriate structural parameters.

## XI. MAGNETIC PROPERTIES

In various places in the previous paragraphs we have mentioned the fact that electrons have an intrinsic angular momentum known as spin. In our discussion thus far this property has played little role except in determining appropriate statistics. We have also discussed the interaction between moving free electrons and a magnetic field in such effects as cyclotron resonance, the Hall effect, and magnetoresistance, but these have been interactions between a magnetic field and a moving charge, not with the electron spin or with magnetic moments associated with electron motion. It is a consideration of these latter interactions that leads to the large field known as magnetic properties.

Since the subject of magnetic properties is treated in detail elsewhere in this Encyclopedia, our purpose in this section is simply to give a brief introduction and survey.

There are two sources of an electronic magnetic moment: (1) the orbital electronic motion of electrons in atoms, and (2) the intrinsic electron angular momentum, the spin.

A current $I$ flowing in a loop enclosing an area $A$ gives rise to a magnetic moment

$$\mu = \mu_0 IA, \qquad (34)$$

where $\mu_0$ is the permeability of free space, and the direction of the magnetic moment is perpendicular to the plane in which $A$ is defined. In the completed electronic shell of an atom (2 $s$ electrons, 6 $p$ electrons, 10 $d$ electrons, etc.), the net magnetic moment in the absence of a magnetic field is zero. If a magnetic field is applied to such a material, however, a magnetic field is set up according to Eq. (34) that opposes the applied field (Lenz's law). This phenomenon is known as diamagnetism and corresponds to a slight repulsion of a material by a magnetic field. The magnetization, the magnetic moment per unit volume, $\mathbf{M}$ is related to the magnetic field $\mathbf{H}$ by $\mathbf{M} = \kappa \mathbf{H}$, where $\kappa$ is the magnetic susceptibility; for a diamagnetic effect, $\kappa$ is small and negative. All materials exhibit some degree of diamagnetism. The classical Langevin equation for diamagnetic susceptibility is

$$\kappa_{\text{dia}} = -NZq^2\mu_0\langle r^2\rangle/6m, \tag{35}$$

where $N$ is the number of atoms per unit volume, $Z$ is the atomic number of the material, and $\langle r^2\rangle$ is the average value of $r$ corresponding to the average electronic orbit radius.

If a magnetic moment is present in the absence of a magnetic field, the magnetic moments line up with the magnetic field to decrease the total energy. This phenomenon is known as paramagnetism and correspond to a slight attraction of a material by a magnetic field and a small positive value of $\kappa$. Such magnetic moments associated with the spin of electrons can arise either from free electrons or from bound electrons. Paramagnetism due to free electrons is temperature independent and is given by the free-electron model as

$$\kappa_{\text{para(free)}} = \left(\mu_B^2 m/\pi^2\hbar^2\right)(3\pi^2 n)^{1/3}, \tag{36}$$

where $\mu_B$ is the Bohr magneton, $\mu_B = q\mu_0\hbar/2m$, the magnetic moment associated with the electron spin. Paramagnetism due to bound electrons varies inversely with temperature and is about 100 times larger than paramagnetism due to free electrons; the susceptibility is given by

$$\kappa_{\text{para(bound)}} = N\mu_B^2/kT. \tag{37}$$

The presence of a large magnetic moment that persists in the absence of a magnetic field due to spontaneous ordering of the moments by direct interaction at temperatures below some critical temperature, called the Curie temperature, is known as ferromagnetism (e.g., Fe, Co, and Ni). The energy that makes ferromagnetism possible is called the exchange energy and is a consequence of interaction between spins on neighboring atoms. The exchange energy between unpaired spins of neighboring atoms in suitable materials is lowest when these spins are parallel. In the ferromagnetic state, the susceptibility is positive, many orders of magnitude larger than in paramagnetism, and is field dependent. Above the Curie temperature, a ferromagnetic material shows only paramagnetism.

Antiferromagnetism is another magnetic effect that resembles ferromagnetism in that it results from internal interactions between magnetic moments, but in these materials an ordered situation is produced where the magnetic moments of nearby atoms are oppositely oriented (e.g., Cr, Mn, MnO, MnS, and NiO). The total moment over a finite volume is zero. The antiferromagnetic ordering is also destroyed above a critical temperature, the Neel temperature, above which the behavior is paramagnetic.

Ferrimagnetism involves an ordered structure not unlike that of antiferromagnetism but involves a case where the number of atoms with opposite spin are unequal, which therefore yields a net magnetic moment [e.g., $(MO)(Fe_2O_3)$ with M = Mn, Fe, Co, Ni, Cu]. Otherwise the behavior is much like that of ferromagnetism, yielding large and field-dependent susceptibilities below a Curie temperature.

All practical ferromagnetic materials consist of magnetic domains in the absence of an applied magnetic field, the orientation of the magnetic moment in a domain corresponding to overall minimization of various forms of magnetic energy. When a magnetic field is applied to such a material, the magnetic field needs to line up the various domains or so alter their extent that the saturation magnetization is achieved. Thus the variation of magnetization with magnetic field involves phenomena related to the motion of domain walls as well as the other magnetic phenomena described above.

The magnetic energies that must be minimized include the exchange energy (which by itself would lead to one domain); the magnetostatic energy, or the energy in the external magnetic field (which by itself would lead to many closure domains to restrict magnetic lines of force within the material); the magnetocrystalline anisotropy energy, or the energy required to align moments in crystalline directions not favored by a particular crystal structure (which by itself would limit the formation of closure domains); and the magnetoelastic or magnetostrictive energy, or an energy related to a change in physical dimensions accompanying magnetization in a particular direction (which by itself would make domains smaller).

Another consideration that limits the formation of domains is the energy required to form a wall between two differently oriented domains. Minimization of the exchange energy favors wide walls with small changes of moment orientation between neighboring atoms, whereas minimization of the magnetocrystalline anisotropy energy favors narrow walls to avoid spin alignment in difficult directions.

A number of these considerations are brought together in a consideration of the form of the variation of magnetization with magnetic field strength, as illustrated in Fig. 21. If the applied magnetic field to a ferromagnetic material is varied, the magnetization of the material exhibits hysteresis. The area of the hysteresis loop is the energy required to traverse one hysteresis cycle and is therefore an indication of the defect structure of the material and of the type of applications for which it would be best suited.

If a material contains many defects and inclusions, domain wall motion is difficult, and large-area hysteresis loops result, with high values of magnetization at zero applied field (the remanent magnetization) and large values of magnetic field required to reduce the remanent magnetization to zero (the coercive field). Such materials are useful as permanent magnets and are called hard magnetic materials. If the material contains few defects

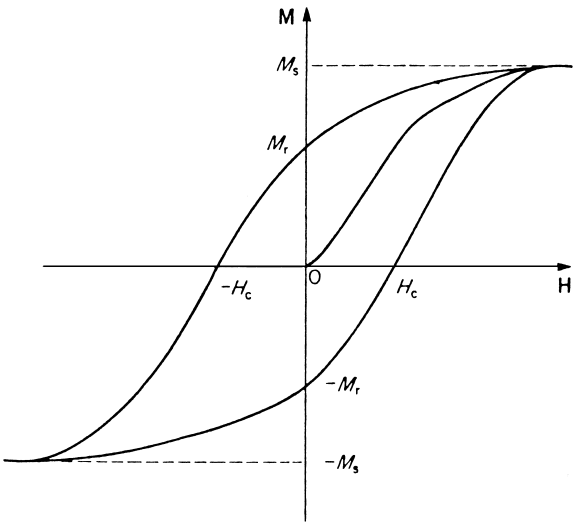



**FIGURE 21** Typical ferromagnetic hysteresis curve showing saturation magnetization $M_s$, remanent magnetization $M_r$, and coercive field $H_c$.

and inclusions such that domains can be easily aligned by an applied magnetic field, the area of the hysteresis loop is small. Such materials are useful for transformer cores or for electromagnets and are called soft magnetic materials.

Magnetic materials are playing an increasingly prominent role in power distribution, conversion between electrical and mechanical energy, microwave communications, and data storage.

One significant development in magnetic memory has been the magnetic bubble. A cylindrically shaped magnetic domain—a small region of oppositely magnetized material in a matrix of magnetic material with the opposite orientation—is stable for a wide range of magnetic fields. These configurations allow the retention of memory when the power is turned off and also allow a high packing density.

Increasingly stringent requirements are imposed for magnetic recording heads as the recording density increases, and of course magnetic materials are the basis for magnetic recording media themselves. These media are of two types: (1) particulate, in which the magnetic component consists of tiny, single-domain particles in a binder, and (2) thin-film ferromagnetic metals and alloys.

Techniques for growing thin-film magnetic materials have made possible not only high-quality thin films, but also multilayer films in which successive layers are only a few atomic layers thick. Such structures, for example, exhibit (a) "giant magnetoresistance," a change in resistance as large as 100% with a change in the magnetic state (relative orientation of successive layers) of the system, (b) strong magnetic anisotropy in-plane and out-of-plane, and (c) interesting magneto-optic properties.

## SEE ALSO THE FOLLOWING ARTICLES

BONDING AND STRUCTURE IN SOLIDS • POLYMERS, ELECTRONIC PROPERTIES • SOLID-STATE CHEMISTRY • SOLID-STATE ELECTROCHEMISTRY • SUPERCONDUCTIVITY

## BIBLIOGRAPHY

Bube, R. H. (1992). "Electrons in Solids," 3rd ed., Academic Press, New York.

Colclaser, R. A., and Diehl-Nagle, S. (1985). "Materials and Devices for Electrical Engineers and Physicists," McGraw-Hill, New York.

Gerlach, B., and Löwen, H. (1991). *Rev. Mod. Phys.* **63**(1), 63–90.

Hummel, R. E. (1985). "Electronic Properties of Materials," Springer-Verlag, New York.

Mayer, J. W., and Lau, S. S. (1990). "Electronic Materials Science: For Integrated Circuits in Si and GaAs," Macmillan Co., New York.

Solymar, L., and Walsh, D. (1979). "Lectures on the Electrical Properties of Materials," 2nd ed., Oxford Univ. Press, Oxford.

Wilson, J., and Hawkes, J. F. B. (1983). "Optoelectronics: An Introduction," Prentice-Hall, Englewood Cliffs, N.J.

Wolfe, C. M., Holonyak, N., Jr., and Stillman, G. E. (1989). "Physical Properties of Semiconductors," Prentice-Hall, Englewood Cliffs, N.J.



# Excitons, Semiconductor

**Donald C. Reynolds**
*Wright Patterson Air Force Base*

**Thomas C. Collins**
*University of Tennessee*

I. Intrinsic Exciton Characteristics
II. Extrinsic Exciton Characteristics
III. Interaction of Excitons with Other Systems
IV. Special Properties of Excitons

## GLOSSARY

**Band gap** Energy difference between two allowed bands of electron energy in a solid.
**Bound exciton** Exciton localized at an impurity or defect in the crystal.
**Brillouin zones** Those volumes in $K$ space bounded by intersecting surfaces defined by points in $K$ space where the energy is discontinuous and Bragg reflection occurs.
**Central-cell correction** Differences in the binding energies of different chemical impurity atoms (donors or acceptors) in a host lattice resulting from different core configurations.
**Degenerate semiconductor** Semiconductor in which a band has orbital as well as spin degeneracy.
**Direct semiconductor** Semiconductor in which the minimum in the conduction band and the maximum in the valence band occur at the same wave vector.
**Hamiltonian** Total energy operator of the system, kinetic plus potential energy operator.
**Indirect semiconductor** Semiconductor in which the minimum in the conduction band and the maximum in the valence band occur at different wave vectors.
**Nondegenerate semiconductor** Semiconductor in which the bands have only spin degeneracy.
**Oscillator strength** Measure of the intensity of a particular energy transition.
**Phonon** Expression for a quantized lattice vibration. Acoustic phonons refer to in-phase motion of neighboring ions in a lattice vibration; optical phonons refer to out-of-phase motion of neighboring ions.

**THE EXCITON** is a quantum of electronic excitation produced in a periodic structure such as an insulating or semiconducting solid. This quantum of energy has motion, and the motion is characterized by a wave vector. Frenkel was the first to treat the theory of optical absorption in a solid as a quantum process consisting of atomic excitations. The excitation process implies that the excited electron does not leave the cell from which it was excited. In his attempt to gain insight into the transformation of light into heat in solids, Frenkel was able to explain the transformation

by first-order perturbation theory of a system of $N$ atoms having one valence electron per atom with the following properties:

(a) The coupling between different atoms in a crystal is small compared with the forces holding the electron within the separate atoms.
(b) The Born–Oppenheimer approximation is valid.
(c) The total wave function is a product of one-electron functions.

Thus, the Frenkel exciton is a tight-binding description of an electron and a hole bound at a single site such that their separate identities are not lost. This model of the exciton emerges as a limiting case of the general theory of excitons and is applicable to insulating crystals. In the case of semiconductors, nonequilibrium electrons and holes are bound in excitonic states at low temperatures by Coulomb attraction. Semiconducting crystals are characterized by large dielectric constants and small effective masses; therefore the electrons and holes may be treated in a good approximation as completely independent particles, despite the Coulomb interaction. This results because the dielectric constant reduces the Coulomb interaction between the hole and electron to the extent that it produces a weakly bound pair of particles that still retain much of their free character. The exciton represents a state of slightly lower energy than the unbound hole–electron. The effective mass theory used to describe such weakly bound particles was developed by Wannier. These weakly bound excitons are appropriately described using the one-band electronic structure picture by adding the Coulomb interaction between the hole and electron. Semiconductor materials are the heart of most modern optical and electronic devices; as a result, the dominant technological interest is focused on these materials. In view of this interest, this article emphasizes the Wannier excitons, which are appropriate for these materials.

## I. INTRINSIC EXCITON CHARACTERISTICS

### A. Introduction

The intrinsic fundamental-gap exciton in semiconductors is a hydrogenically bound hole–electron pair, the hole being derived from the top valence band and the electron from the bottom conduction band. It is a normal mode of the crystal created by an optical excitation wave, and its wave functions are analogous to those of the Block wave states of free electrons and holes. When most semiconductors are optically excited at low temperatures, it is the intrinsic excitons that are excited. The energies of the ground and excited states of the exciton lie below the band-gap energy of the semiconductor. Hence, the exciton structure must first be determined in order to determine the band-gap energy. The exciton binding energy can be determined from spectral analysis of its hydrogenic ground and excited state transitions (this also gives central-cell corrections). Precise band-gap energies can be determined by adding the exciton binding energy to the experimentally measured photon energy of the ground-state transition.

Both direct and indirect exciton formation occurs in semiconductors, depending on the band structure. The former is characteristic of many of the II–VI and III–V compounds, and the latter is characteristic of germanium and silicon. For indirect optical transitions, momentum is conserved by the emission or absorption of phonons. The detailed nature of the valence-band structure of degenerate and nondegenerate semiconductors is elucidated by understanding the intrinsic-exciton structure of these semiconductors.

### B. Excitation of the *n*-Particle System

Excitons are excited states of the system in which the number of electrons does not change. Ordinarily in solid-state physics one thinks of calculations of one-body approximations as excitation of the system. However, this only works in metals near the Fermi surface or where one adds an electron or takes one away from the system. In the case of metals, the electrons and holes are very diffuse and the interaction between the excited "particle" and the other "particles" is very small. When an electron is added to the system, one obtains electron affinities or $N + 1$ solutions, and when an electron is removed one obtains ionization energies or $N - 1$ solutions. All solutions to one-body calculations such as the one-particle Green's function method, Hartree–Fock calculations, and similar approximations (unless specifically added) do not contain the interaction of the excited "particle" with the other "particles." It is necessary to calculate solutions to the two-body Green's function (or some approximation to the two-body problem) in order to have an exciton.

Another approach is to calculate the many-body excited states by construction of excited wave functions that are orthogonal to the approximate ground-state wave function. The excited-state approximate total energy is calculated, from which is subtracted the approximate total ground-state energy. The formalism gives an effective operator that looks like a scaled or screened hydrogen Hamiltonian. This approach was used by both Frenkel and Wannier. In the case of Frenkel, the excited states were required to be a linear combination of very local excitations. In particular, he required the atoms in his model to all be in the ground state except one. This one was excited, but the excitation

was localized to the one atomic site. The many-body solution using these trial wave functions was identical to the normal modes of vibration of the phonon spectra. The solution is also similar to spin waves where one starts with either an up or down spin on each lattice site.

Wannier, on the other hand, let the excitation trial wave functions be more diffuse than one lattice site. This representation he called "exciton waves," which could vary from very local excitations (Wannier-like functions) to very diffuse (Bloch-like functions). His effective excitation operator was a hydrogenic operator with multipole corrections. The many-body effects were approximated through screening of the Coulomb potential between electrons and the hole left by the excitation. This formalism also has the advantage that when small perturbing fields are turned on, the resulting effective operator adds the perturbing operator, similar to what one finds for the hydrogen atom or molecule case with screened potentials.

## C. Systems of Excitons in Various Crystal Symmetries

### 1. Direct Nondegenerate Semiconductors

Nondegenerate semiconductors are typified by those materials that belong to the wurtzite crystal structure. This is a uniaxial structure having sixfold rotational symmetry and belongs to the $C_{6V}$ crystal point group. In this structure the degeneracy in the valence band is removed by crystal-field interactions.

The tight-binding approximation in conjunction with group theory was first used to describe the irreducible representations, band symmetries, and selection rules for the wurtzite structure. If one considers the absorption (emission) of electromagnetic radiation by atoms, the probability of the occurrence of a transition between two unperturbed states $\psi_i$ and $\psi_f$ as caused by the interaction of an electromagnetic radiation field and a crystal is dependent on the matrix element

$$\int \psi_f^* H_{\text{int}} \psi_i \, dr \tag{1}$$

where

$$H_{\text{int}} = \frac{e\hbar}{imc} \mathbf{A} \cdot \boldsymbol{\nabla} \tag{2}$$

where $\mathbf{A}$ is the vector potential of the radiation field and has the form

$$\mathbf{A} = \mathbf{n} \, |A_0| e^{i(\mathbf{q}\cdot\mathbf{r} - \omega t)}$$

where $e$ is the electron charge, $m$ is the electron mass, $c$ is the velocity of light, $\mathbf{n}$ is a unit vector in the direction of polarization, and $\mathbf{q}$ is the wave vector. Expanding the spatial part of $\mathbf{A}$ in a series gives

$$H_{\text{int}} = \sum_{j=0}^{\infty} H_{\text{int}}^j$$

where

$$H_{\text{int}}^j \approx (\mathbf{q} \cdot \mathbf{r})^j (\mathbf{n} \cdot \boldsymbol{\nabla})$$

and the dipole term is then the first term ($j = 0$). The matrix element in Eq. (1) is now expressed as a series, and for an electric dipole transition to be allowed, the matrix element between the initial and final states must be nonzero.

In the case of transitions between two states of an atom (that is in a crystalline field), the initial and final states of the atom are characterized by irreducible representations of the point group of the crystal field. Also, the dipole moment operation must transform like one of the irreducible representations of the group. If one denotes the representations that correspond to the initial and final states of the transition and to the multipole radiation of order $j$ ($j = 0$ for electric dipole radiation) by $\Gamma_i$, $\Gamma_f$, and $\Gamma_r^{(j)}$, respectively, at $\mathbf{k} = 0$, then the matrix element in Eq. (1) transforms under rotations like the triple direct product

$$\Gamma_f \times \Gamma_r^{(j)} \times \Gamma_i \tag{3}$$

The selection rules are then determined by which of the triple-direct-product matrix elements in question do not vanish.

The dipole moment operator for electric dipole radiation transforms like $x$, $y$, or $z$, depending on the polarization. When the electric vector $\mathbf{E}$ of the incident light is parallel to the crystal axis, the operator corresponds to the $\Gamma_1$ representation. When it is perpendicular to the crystal axis, the operator corresponds to the $\Gamma_5$ representation.

Since the crystal has a principal axis, the crystal field removes part of the degeneracy of the $p$ levels. Thus, disregarding spin–orbit coupling, the following decomposition at the center of the Brillouin zone is obtained:

$$\begin{aligned} &\text{conduction band} && S \to \Gamma_1 \\ & && P_x, P_y \to \Gamma_5 \\ &\text{valence band} && P_z \to \Gamma_1 \end{aligned}$$

Introducing the spin doubles the number of levels. The splitting caused by the presence of spin is represented by the inner products

$$\Gamma_5 \times D_{1/2} \to \Gamma_7 + \Gamma_9$$

$$\Gamma_1 \times D_{1/2} \to \Gamma_7$$

and the band structure at $\mathbf{k} = 0$ along with the band symmetries is shown in Fig. 1.

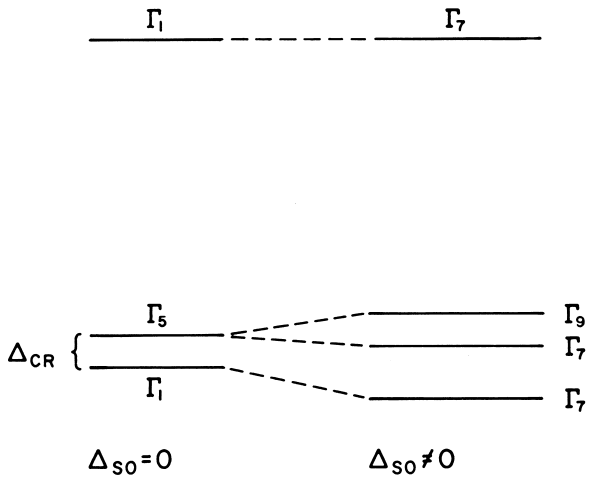



**FIGURE 1** Band structure and band symmetries for the wurtzite structure.

## 2. Direct Degenerate Semiconductors

Materials that crystallize in the diamond or the zinc-blende structures are representative of degenerate semiconductors. Two materials that have been extensively investigated and that are characteristic of direct degenerate semiconductors are GaAs and InP. These materials crystallize in the zinc-blende structure, which has $T_d$ point-group symmetry.

The dipole momement operator for electric dipole radiation in zinc-blende structures transforms like $\Gamma_5$. The conduction band is s-like, while the valence band is p-like. This structure does not have a principal axis; therefore, the crystal-field energy is zero and the full degeneracy of the p levels is retained. Thus, disregarding spin–orbit coupling, the following decomposition at the center of the Brillouin zone is obtained:

$$\text{conduction band} \qquad S \rightarrow \Gamma_1$$

$$\text{valence band} \qquad P \rightarrow \Gamma_4$$

Introducing the spin doubles the number of levels. Consider the $\Gamma_1$ s-like conduction band and the triply degenerate p-like valence band. The states at the center of the Brillouin zone, which belong to $\Gamma_1$ and $\Gamma_4$ representations of the single group, are shown in Fig. 2. The splitting caused by the presence of spin is represented by the inner product as follows:

$$\Gamma_1 \times D_{1/2} \rightarrow \Gamma_6$$

$$\Gamma_4 \times D_{3/2} \rightarrow \Gamma_7 + \Gamma_8$$

Physically this result means that the six valence-band states, consisting of the three p-like states, each associated with one or the other of the two spin states, and that are degenerate in the absence of spin–orbit interaction, now split into two levels, one having $\Gamma_7$ symmetry and the other having $\Gamma_8$ symmetry. The $\Gamma_8$ level is fourfold degenerate, while the $\Gamma_7$ level is twofold degenerate.

The energy of the exciton is described by the Hamiltonian

$$H_{ex} = H_e + H_h + H_{eh} \tag{4}$$

where $H_e$ and $H_h$ are the Hamiltonians for the electron and the hole and $H_{eh}$ is the interaction Hamiltonian between the electron and the hole.

## 3. Indirect Transitions

Two of the most extensively studied indirect materials are the elemental semiconductors Si and Ge. Both of these materials have indirect band gaps, and therefore the lowest energy electronic state is an indirect exciton. For this lowest energy state to be optically excited, momentum must be conserved; thus, additional momentum must be supplied by the creation or annihilation of an appropriate phonon. These materials crystallize in the diamond structure and belong to the $O_h$ point-group symmetry.

The band structures of Si and Ge are similar; as a result, Si will be used as the example in this discussion to describe the indirect exciton. The band structure of Si is shown in Fig. 3. The conduction-band minimum occurs as a $\Delta_1$ symmetry approximately 85% of the way to the zone boundary in the $\langle 100 \rangle$ direction. The corresponding valence band symmetry is $\Delta_5$. Using the group character tables at $\Gamma = (0, 0, 0)$ and $\Delta = (k, 0, 0)$,

$$\Delta_1 \times \Delta_5 = \Gamma_{15}^{+} + \Gamma_{25}^{+} + \Gamma_{15}^{-} + \Gamma_{25}^{-} \tag{5}$$

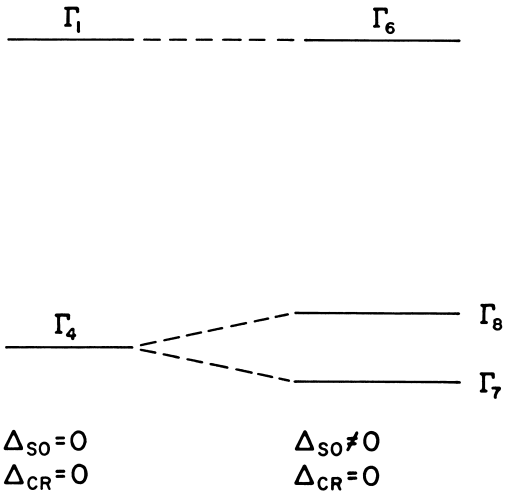



**FIGURE 2** Band structure and band symmetries for the zinc-blende structure.

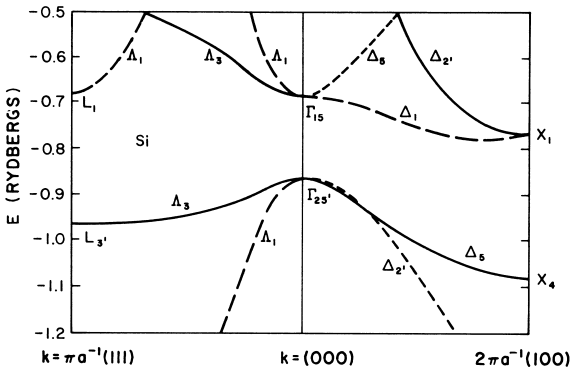



**FIGURE 3** The band structure of silicon near the energy gap. [As computed by Phillips (1958). *Phys. Rev.* **112,** 685.]

The presence of the $\Gamma_{15}^-$ symmetry means that this is an allowed transition. The maximum in the valence band occurs at $\mathbf{k} = 0$ having $\Gamma_{25}^+$ symmetry. The $\Gamma_{25}^+ \rightarrow \Gamma_{15}^-$ transition is allowed as

$$\Gamma_{25}^+ \times \Gamma_{15}^- = \Gamma_2^- + \Gamma_{12}^- + \Gamma_{25}^- + \Gamma_{15}^- \tag{6}$$

The momentum-conserving phonon in the valence band from $\Delta_5$ to $\Gamma_{25}^+$ is

$$\begin{aligned}\Gamma_{25}^+ \times \Delta_5 &= (\Delta_5) + (\Delta_1 + \Delta_2') + \Delta_1' + \Delta_2 \\ &= (\mathrm{TO} + \mathrm{TA}) + (\mathrm{LA} + \mathrm{LO})\end{aligned} \tag{7}$$

where TO is transverse optical, TA is transverse acoustic, and L indicates longitudinal. All phonons are allowed, as seen from Fig. 4. In this case, the energy denominator in

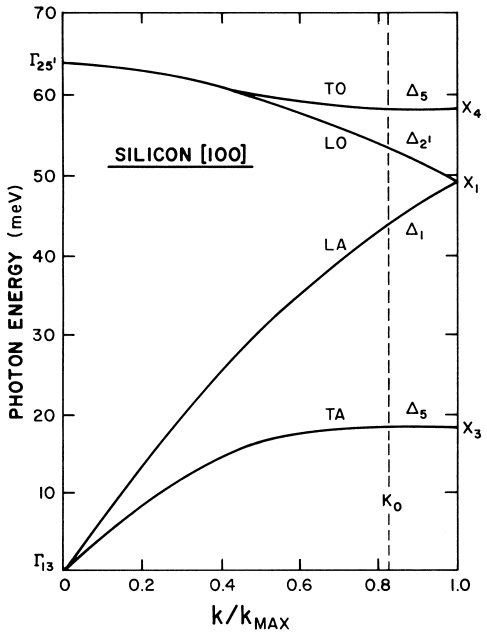



**FIGURE 4** The vibration spectrum of silicon. [As determined by Brockhouse (1959). *Phys. Rev. Lett.* **2,** 256.]

the intermediate state is fairly large where the radiative transition occurs first; therefore the transitions may be weak.

The more important phonon for conserving momentum is in the conduction band from $\Delta_1$ to $\Gamma_{15}^-$,

$$\Gamma_{15}^- \times \Delta_1 = \Delta_1 + \Delta_5 = \mathrm{LA} + (\mathrm{TO} + \mathrm{TA}) \tag{8}$$

Here only the LO phonon is forbidden.

The same momentum considerations apply to the exciton as apply to the bands, since an exciton that is constructed from bands, whose extrema differ by an amount $\mathbf{k}\Delta$ will have a momentum $\mathbf{k}\Delta$, which must be supplied by the phonon field during an optical transition.

## D. Perturbations

### 1. Magnetic Fields

When the crystal is placed in a uniform magnetic field, there are several new terms in the Hamiltonian, which will be described below. In this description, the band-gap extrema are at $\mathbf{k} = 0$ with their shape parabolic at least to second order in $\mathbf{k}$ and with only double spin degeneracy. One may write the exciton equation as a simple hydrogen Schrodinger equation including mass and dielectric anisotropy. In general one finds the mass anisotropy is small, allowing first-order perturbation calculations to be made for the energy states as well as for the magnetic-field effects.

Since in this model the valence and conduction band extrema are at $\mathbf{k} = 0$, the wave vector of the light that creates the exciton $\mathbf{k}$ will also represent the position of the exciton in $\mathbf{k}$ space. If one divides the momentum and space coordinates into the center-of-mass coordinates and the internal coordinates, the exciton Hamiltonian can be divided into seven terms as follows:

$$H = H_1 + H_2 + H_3 + H_4 + H_{k1} + H_{k2} + H_{k3}; \tag{9}$$

$$\begin{aligned}H_1 = {} & \frac{-\hbar}{2m}\left[\frac{1}{\mu_x}\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right) + \frac{1}{\mu_z}\left(\frac{\partial^2}{\partial z^2}\right)\right] \\ & - \frac{e^2}{\varepsilon\eta^{1/2}}(x^2 + y^2 + \eta^{-1}z^2)^{-1/2}\end{aligned} \tag{10}$$

$$H_2 = -2i\zeta\left[\frac{A_x}{\Delta_x}\left(\frac{\partial}{\partial x}\right) + \frac{A_y}{\Delta_y}\left(\frac{\partial}{\partial y}\right) + \frac{A_z}{\Delta_z}\frac{\partial}{\partial z}\right] \tag{11}$$

$$H_3 = \frac{e^2}{2mc^2}\left\{\frac{A_x^2}{\mu_x} + \frac{A_y^2}{\mu_y} + \frac{A_z^2}{\mu_z}\right\} \tag{12}$$

$$H_4 = \frac{\zeta}{2}\sum_{\gamma=x,y,z}(g_{e\gamma}S_{e\gamma} + g_{h\gamma}S_{h\gamma})H_\gamma \tag{13}$$

$$H_{k1} = \frac{i\hbar}{2m}\left[\frac{K_x}{\Delta_x}\left(\frac{\partial}{\partial x}\right) + \frac{K_y}{\Delta_y}\left(\frac{\partial}{\partial y}\right) + \frac{K_z}{\Delta_z}\left(\frac{\partial}{\partial z}\right)\right] \tag{14}$$

$$H_{k2} = \zeta\left(\frac{K_x}{\mu_x}A_x + \frac{K_y}{\mu_y}A_y + \frac{K_z}{\mu_z}A_z\right) \tag{15}$$

$$H_{k3} = \frac{\hbar^2}{8m}\left\{\frac{K_x^2}{\mu_x} + \frac{K_y^2}{\mu_y} + \frac{K_z^2}{\mu_z}\right\}, \tag{16}$$

where $m$ is the free-electron mass and $\mu_\gamma$ is the reduced effective mass of the exciton in the direction $\gamma$. Also,

$$\zeta = \frac{e\hbar}{2mc}, \qquad \eta = \frac{\varepsilon_z}{\varepsilon_x} \tag{17}$$

$$A = \tfrac{1}{2}(\mathbf{H} \times \mathbf{r}) \tag{18}$$

$$\frac{1}{\Delta_\gamma} = \left(\frac{m}{m^*_{e\gamma}} - \frac{m}{m^*_{h\gamma}}\right) \tag{19}$$

in the wurtzite structure $\mu_x = \mu_y$.

The first term is the Hamiltonian for a hydrogenic system in the absence of external fields. This term has the possibility of including the mass and dielectric anisotropies. The second term is an $\mathbf{A} \cdot \mathbf{p}$ term, which leads to the linear (Zeeman) magnetic field term. In this term, the momentum operator $\mathbf{p}_i$ becomes $\mathbf{p}_i - (2e/c)\mathbf{A}_i$, where $\mathbf{A}_i = \frac{1}{2}\mathbf{H} \times \mathbf{r}_i$ is the vector potential, $\mathbf{H}$ is the magnetic field, and $\mathbf{r}_i$ is the coordinate of the $i$th electron. The $A^2$ term is the diamagnetic field term proportional to $|\mathbf{H}|^2$. The fourth term is the linear interaction of the magnetic field with the spin of the electron and hole. If one has small effective reduced mass for the electron and a large dielectric constant, the radii of the exciton states are much larger than the corresponding hydrogen-state radii. Hence, since the spin–orbit coupling is proportional to $r^{-3}$ and thus quite small, it is legitimate to write the magnetic-field perturbations in the Paschen–Back limit as done above.

The last three terms are the $\mathbf{K} \cdot \mathbf{P}$, $\mathbf{K} \cdot \mathbf{A}$, and $K^2$ terms; $\mathbf{K}$ is the center-of-mass momentum. Treating the $\mathbf{K} \cdot \mathbf{P}$ to second order and adding the $K^2$ term, one can obtain an energy term that appears like the center-of-mass kinetic energy. The $\mathbf{K} \cdot \mathbf{A}$ term has little effect upon the energy; however, it has very interesting properties. This term represents the quasi-electric field that an observer riding with the center-of-mass of the exciton would experience because of the magnetic field in the laboratory. This quasi-field would produce a stark effect linear in $\mathbf{H}$, and this would give rise to a maximum splitting interpretable as a "$g$ value."

## 2. Strain Field

A detailed study has been made of the stress-induced splitting of the exciton states in both wurtzite and zinc-blende structures. The band symmetries for the structures are shown in Figs. 1 and 2, respectively. In these materials the conduction band is s-like and the valence bands are p-like. The theory was developed around the effective Hamiltonian

$$H = H_{v-0} + H_{v-p} + H_c + H_{ex}, \tag{20}$$

where $H_{v-0}$ describes the zero pressure mixing of the three p-like valence bands, $H_{v-p}$ describes the strain-dependent mixing of the valence bands, $H_c$ describes the conduction-band energy, and $H_{ex}$ describes the valence-hole–conduction-electron interaction. The final term in the Hamiltonian is where the spin-exchange interaction is introduced. This term is written as

$$H_{ex} = \mathrm{BE} + J\boldsymbol{\sigma}_h \cdot \boldsymbol{\sigma}_e, \tag{21}$$

where BE is the exciton binding energy and the last term is the exchange interaction term. The exchange constant $J$ can be calculated from known band properties. The Hamiltonians have been formalized for both zinc-blende and wurtzite structures. Solutions of the Hamiltonians lead to the matrix elements for the allowed optical transitions $\langle 0|\boldsymbol{\epsilon} \cdot \boldsymbol{\nabla}|\psi\rangle$, where $\boldsymbol{\epsilon}$ gives the polarization of the light and $|0\rangle$ corresponds to a filled valence band and empty conduction band.

Experimentally observed splittings of exciton lines under uniaxial stress were observed in wurtzite-type II-VI compounds. In wurtzite structures, all of the orbital degeneracies of the valence band are removed by the spin–orbit interaction and by the trigonal crystal field. It was evident that this splitting could not be accounted for by the usual deformation-potential theory based on one-electron energy bands. It was determined that the observed splitting could be attributed to the decomposition of the degenerate $\Gamma_5$ exciton state by the deformation of the wurtzite lattice. It is the combined effect of stress and exchange coupling that gives rise to the splitting. In the wurtzite structure the conduction band has $\Gamma_7$ symmetry while the top valence band has $\Gamma_9$ symmetry, and the next two lower valence bands have $\Gamma_7$ symmetry as shown in Fig. 1. Exciton symmetries associated with optical transitions between these bands are as follows:

$$\Gamma_7 \times \Gamma_9 \rightarrow \Gamma_5 + \Gamma_6$$

$$\Gamma_7 \times \Gamma_7 \rightarrow \Gamma_5 + \Gamma_1 + \Gamma_2$$

In the above analysis of strain splitting only the splitting of the $\Gamma_5$ exciton was treated. The $\Gamma_6$ and $\Gamma_2$ excitons were not considered since they are forbidden. In a later study of ZnO the excitons that make up the top valence band were observed in emission, and both the $\Gamma_5$ and $\Gamma_6$ exciton were resolved. In the absence of a magnetic field the $\Gamma_6$ exciton was observed in samples containing in-grown strain. It would be expected that strain would relax selection rules since it changes the symmetry of the sample. Not only was the $\Gamma_6$ exciton observed in the presence of strain but it also

showed a splitting, resulting from combined strain and electron-hole spin exchange. The inclusion of the spin-exchange interaction also allows one to bridge the gap between the description of excitons in the $J$–$J$ coupling scheme and the $L$–$S$ coupling scheme.

## II. EXTRINSIC EXCITON CHARACTERISTICS

### A. Introduction

The intrinsic exciton may bind to various impurities, defects, and complexes, and the subsequent decay from the bound state yields information concerning the center to which it was bound. Bound-exciton complexes are extrinsic properties of materials. These complexes are observed as sharp-line optical transitions in both photoluminescence and absorption. The binding energy of the exciton to the impurity or defect is generally weak compared to the free-exciton binding energy. The resulting complex is molecular-like (analogous to the hydrogen molecule or molecule–ion) and has spectral properties that are analogous to those of simple diatomic molecules. The emission or absorption energies of these bound-exciton transitions are always below those of the corresponding free-exciton transitions, due to the molecular binding energy.

Bound excitons were first reported in the indirect semiconductor silicon. Here it was found that when group V elements were added to silicon, sharp photoluminescent lines were produced, and these lines were displaced in energy in a regular way. The binding energies of exciton complexes produced by adding different group V donors were described by the linear relation

$$E = 0.1E_i \tag{22}$$

where $E$ is the binding energy of the exciton and $E_i$ is the ionization energy of the donor. The small differences in ionization energies for different-effective-mass chemical donors result from central cell corrections. A similar relationship was found when the group III acceptors were added to silicon. A modified linear relationship has been found for donors and acceptors in compound semiconductors.

The sharp spectral lines of bound exciton complexes can be very intense (large oscillator strength). The line intensities will, in general, depend on the concentrations of impurities and/or defects present in the sample.

If the absorption transition occurs at $\mathbf{k} = 0$ and if the discrete level associated with the impurity approaches the conduction band, the intensity of the absorption line increases. The explanation offered for this intensity behavior is that the optical excitation is not localized in the impurity but encompasses a number of neighboring lattice points of the host crystal. Hence, in the absorption process, light is absorbed by the entire region of the crystal consisting of the impurity and its surroundings.

The oscillator strength of the bound exciton, $F_\mathrm{d}$, relative to that of the free exciton $f_\mathrm{ex}$ can be expressed as

$$F_\mathrm{d} = (E_0/|E|)^{3/2} f_\mathrm{ex} \tag{23}$$

where $E_0 = (2\hbar^2/m)(\pi/\Omega_0)^{2/3}$, $E$ is the binding energy of the exciton to the impurity, $m$ is the effective mass of the intrinsic exciton, and $\Omega_0$ is the volume of the unit cell.

It has been shown in some materials that $F_\mathrm{d}$ exceeds $f_\mathrm{ex}$ by more than four orders of magnitude. An inspection of Eq. (23) reveals that, as the intrinsic exciton becomes more tightly bound to the associated center, the oscillator strength, and hence the intensity of the excitoncomplex line, should decrease as $(I/E)^{3/2}$.

In magnetic fields, bound excitons have unique Zeeman spectral characteristics, from which it is possible to identify the types of centers to which the free excitons are bound. Bound-exciton spectroscopy is a very powerful analytical tool for the study and identification of impurities and defects in semiconductor materials.

### B. Bound-Exciton Complexes in Different Symmetries

#### 1. Degenerate Semiconductors

The model of the donor–bound-exciton complex is used to describe bound-exciton complexes in zinc-blende structures. The Hamiltonian of the system may be written

$$H = H_\mathrm{ex} + H_\mathrm{d} + H_\mathrm{dex} \tag{24}$$

where $H_\mathrm{ex}$ and $H_\mathrm{d}$ are the exciton and donor Hamiltonians and $H_\mathrm{dex}$ describes the interaction between the exciton and donor. The Hamiltonian for the exciton is given in Eq. (4). In this equation, $H_\mathrm{eh}$ is the interaction Hamiltonian between the electron and hole,

$$H_\mathrm{eh} = -e^2/\varepsilon|\mathbf{r}_\mathrm{e} - \mathbf{r}_\mathrm{h}| + H_\mathrm{exch} \tag{25}$$

Here $\varepsilon$ is the dielectric constant and $H_\mathrm{exch}$ is the electron–hole exchange Hamiltonian. The exchange Hamiltonian $H_\mathrm{exch}$ is

$$H_\mathrm{exch} = A_1\boldsymbol{\sigma}\cdot\mathbf{J} + A_2\left(\sigma_x J_x^3 + \sigma_y J_y^3 + \sigma_z J_z^3\right) \tag{26}$$

where $\boldsymbol{\sigma}$ and $\mathbf{J}$ are the operators for electron spin and effective hole spin, respectively, and $A_1$ and $A_2$ are parameters describing the exchange energy.

The model of the exciton bound to a neutral donor is shown in Fig. 5. In the initial state the two electrons pair to form a bonding state, leaving an unpaired hole. When the

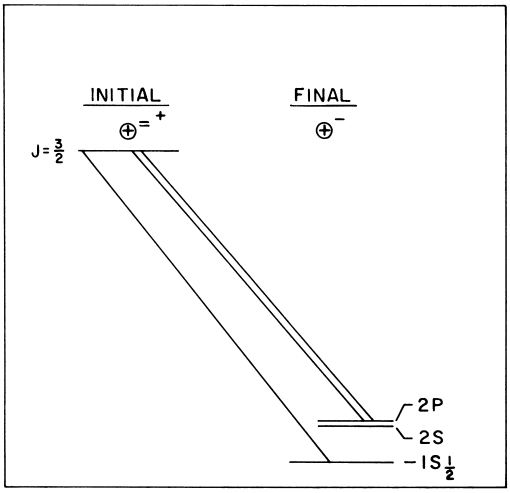



**FIGURE 5** Schematic representation of radiative recombination of an exciton bound to a neutral donor, where the final state is the donor in the ground or in the excited configuration.

exciton collapses from this state, the final state may consist of the donor in the ground state or in an excited state. The donor may pick up energy from the exciton recombination, thus leaving the donor in an excited state. The model of the exciton bound to the acceptor is more complicated than the donor. The initial state of the neutral acceptor-bound exciton consists of two $J = \frac{3}{2}$ holes and one $J = \frac{1}{2}$ electron as shown in Fig. 6. The two $J = \frac{3}{2}$ holes combine

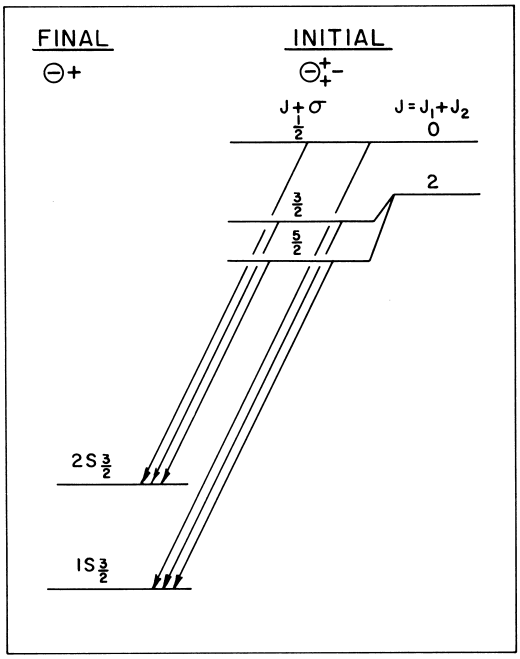



**FIGURE 6** Schematic representation of radiative recombination of an exciton bound to a neutral acceptor, where the final state is the acceptor in the ground or in the excited configuration.

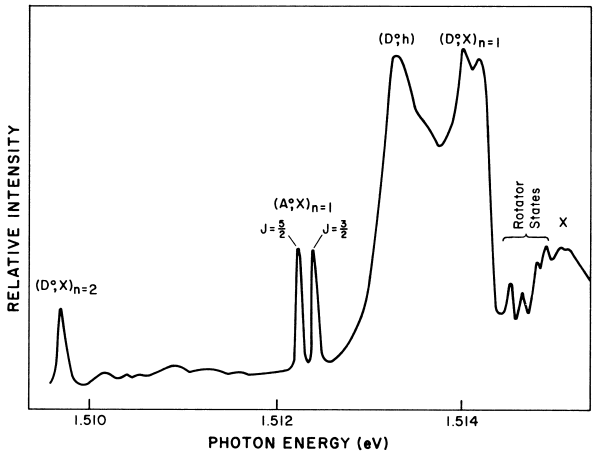



**FIGURE 7** Photoluminescence spectrum of GaAs in the near bandgap region.

to give a $J = 0$ and a $J = 2$ state. The interaction of the electron spin operator $\boldsymbol{\sigma}$ with the effective hole spin operator $\mathbf{J}$ results in three $\mathbf{J} + \boldsymbol{\sigma}$ states, $\frac{1}{2}$, $\frac{3}{2}$, and $\frac{5}{2}$. As in the case of the donor, when the exciton collapses, the final state will consist of the neutral acceptor in the ground or an excited state.

The photoluminescent spectrum for GaAs is shown in Fig. 7, where X is the free exciton transition; the D°X lines are associated with the neutral donor-bound excitons, and the A°X lines are associated with the neutral acceptor-bound excitons. The $J = \frac{5}{2}$ and $J = \frac{3}{2}$ neutral acceptor-bound exciton transitions are clearly observed. The (D°X) $n = 2$ lines are associated with the collapse of the exciton from the neutral donor-bound exciton state, leaving the donor in an excited state. The energy of the transition is expressed as

$$E_{\mathrm{T}} = E_{\mathrm{ex}} - E_{\mathrm{Dex}} - E_{\mathrm{D}}^{*} \tag{27}$$

where $E_{\mathrm{ex}}$ is the free exciton energy, $E_{\mathrm{Dex}}$ is the binding energy of the exciton to the donor, and $E_{\mathrm{D}}^{*}$ is the energy required to place the neutral donor in an excited state. The analogous neutral acceptor-bound exciton transitions in which the final state of the acceptor is left in an excited state are not shown. These transitions occur at appreciably lower energies, due to the larger binding energies of the acceptors.

### 2. Nondegenerate Semiconductors

The theory for the nondegenerate case is based on the wurtzite structure, with the salient factors of the band structure such as band symmetries and selection rules being derived from group theory.

Consider any simple optical transition in which an electron bound to an impurity is taken from one band to another. Suppose the initial and final states can be

approximately assigned effective mass wave functions. The initial-state wave function can then be written as

$$f_i(r)U_{i0}(r) \tag{28}$$

where $f_i(r)$ is a slowly varying function of $r$ and $U_{i0}$ is the periodic part of the Bloch function of band $i$ for wave vector zero. Similarly, the final-state wave function can be written

$$f_f(r)U_{f0}(r) \tag{29}$$

Since $f$ is a slowly varying function, the optical matrix element

$$\int f_i(r)U_{i0}(r)\mathbf{P}f_f^*(r)U_{f0}^*(r)\,d^3r \tag{30}$$

can be approximately written as

$$\left[\int f_i(r)f_f^*(r)\,d^3r\right]\frac{1}{\Omega}\left[\int U_{i0}(r)\mathbf{P}U_{f0}^*(r)\,d\tau\right] \tag{31}$$

where the second integration is carried out over the unit cell, whose volume is $\Omega$. In this approximation, the only large optical-matrix elements will arise when the analogous band-to-band transition is allowed.

A similar argument can be made to show that in this effective-mass approximation, large $g$ values can be expected only when the parent energy-band wave functions exhibit large $g$ values.

In the case of weakly bound states at substitutional impurities and energy bands at $\mathbf{k} = 0$ in the wurzite structure, it is reasonable to describe the states as though they belonged to the point group of the crystal rather than to the group of the impurity. Such a description gives the degeneracy of the states correctly. This description neglects certain optical transitions that are technically allowed, but that are weak in the effective mass approximation and will set equal to zero certain $g$ values that should be much smaller than usual $g$ values. The advantage of the description is that it neglects these small effects and thus permits the full use of group theory without the clutter of what should be small perturbations.

The electron $g$ value, $g_e$, should be very nearly isotropic, since the conduction band is simple and the $g$ shift of the free electron is small, only weakly dependent on the state of binding of the electron. The hole $g$ value, $g_h$, should be completely anisotropic with $g_h$ equal to zero (for the top $\Gamma_9$) for magnetic fields perpendicular to the hexagonal axis. It is to be expected that the hole $g$ value will be sensitive to its state of binding, since the different valence bands will be strongly mixed in bound-hole states. The model of the exciton bound to a neutral donor for the nondegenerate case is very similar to that for the degenerate case, shown in Fig. 5, for zero applied magnetic field. The unpaired hole for the nondegenerate semiconductor

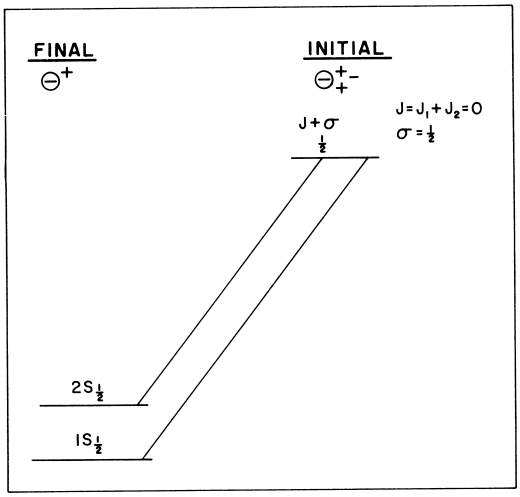



**FIGURE 8** Model of the exciton bound to the neutral acceptor for a nondegenerate semiconductor.

is twofold degenerate, as compared to the fourfold degeneracy for the degenerate semiconductor. The model of the exciton bound to the neutral acceptor for the nondegenerate case is less complicated than for the degenerate case. The initial state of the complex consists of paired doubly degenerate holes, leaving an unpaired electron as shown in Fig. 8. The final state consists of the acceptor, either in the ground state or an excited state. In the absence of an applied magnetic field, a single optical transition is observed.

### 3. Bound-Exciton Excited States

In many materials, on the high energy side of the neutral donor-bound-exciton complex lines is a similar set of lines, which are excited states of the lower energy complex structure. A rigid rotation model was proposed to explain these excied states in CdTe. In this model the hole is excited to rotate around the fixed donor, analogous to rotation of diatomic molecules. A non-rigid-rotator model was subsequently proposed, which was successful in predicting the excited state energies in InP and GaAs. A more sophisticated model followed, which was applied to the D,$^0$X ground and excited states, this model was successful in predicting the energy ordering of the excited states. A final model was proposed to explain the high-magnetic-field results in InP. In this model D,$^0$X is considered to be a free exciton orbiting a neutral donor; one electron was considered to be strongly correlated with the hole and the other with the donor. This model was capable of explaining the relative intensities of the photoluminescence transition in the ground- and excited-state regions of InP.

Excited states associated with the D,$^0$X ground-state transitions were later observed in ZnO. The transitions

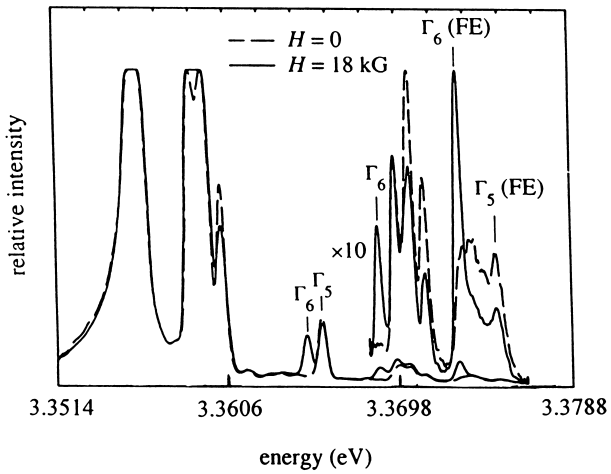



**FIGURE 9**

3.3662 eV ($\Gamma_6$) and 3.3670 eV ($\Gamma_5$) in Fig. 9 are excited states analogous to rotational states of the $H_2$ molecule. These states are rotational states associated with the 3.3564 eV ground state, and are not electronic excited states. As observed in Fig. 9, these transitions are on the low energy side of the 3.3772 eV ($\Gamma_5$) and 3.3750 eV ($\Gamma_6$) free exciton transitions. The solid curve in the figure represents spectra with an applied magnetic field of 18 KG. The $\Gamma_6$ exciton is an unallowed transition that becomes allowed in the presence of an applied magnetic field. The dashed curve shows the same transition in zero magnetic field. Note that the rotator state associated with the $\Gamma_6$ exciton was observed. The two lowest energy rotator states are associated with the lowest energy 3.3564 eV D,$^0$X transition. The next two lowest energy rotator states, 3.3714 eV ($\Gamma_5$) and 3.3702 eV ($\Gamma_6$), are associated with the next lowest energy 3.3594 eV D,$^0$X transition. It is noted that again one of the rotator states is associated with the $\Gamma_6$ exciton. Other rotator states associated with the $\Gamma_6$ exciton are most likely not resolved since they would come in the energy region where they would not be resolved from other $\Gamma_5$ rotator states. This was the first observation of rotator states associated with the $\Gamma_6$ unallowed exciton, and lends support to the model that the exciton itself rather than the hole is rotating.

## C. Perturbations

### 1. Magnetic Field

When a magnetic field is applied to the donor-bound exciton complexes in a degenerate semiconductor, the line splitting due to the presence of the magnetic field can be predicted from Fig. 5. In the initial state, the $J = \frac{3}{2}$ unpaired hole will split into a quartet while the 1s final state will split into a doublet. This splitting will result in six allowed transitions. When the final state consists of the excited $n = 2$ states, the splitting is much more complicated. The 2s state is doubly degenerate, while the 2p state is sixfold degenerate. In addition to the increased multiplicity of lines, rather large diamagnetic shifts are also observed. The energies of these transitions in a magnetic field have been calculated. In general it is not easy to solve the Hamiltonian for the donor-bound exciton complex in a magnetic field. In the low-field regime, the exciton Hamiltonian of Eq. (4) can be separated into two parts: the spherical symmetric (s-wave-like) and asymmetric (d-wave-like) parts. For the perturbation calculation, in the low-field regime one can treat the s-wave-like part as an unperturbed Hamiltonian and the d-wave-like part as a perturbed Hamiltonian.

In the high-field regime—i.e., when the magnetic energy is much greater than the Coulomb energy—the solution of Eq. (4) may be obtained by an adiabatic method which can be written as

$$E_{ij} = L_{ij}\gamma \tag{32}$$

where the $L_{ij}$ are the linear coefficients for the Landau-type solutions. They turn out to be of the order of $L_{ij} = 0.01$. These energies shift much more rapidly with magnetic field than is experimentally observed in the intermediate-field region.

In the intermediate-field regime where the magnetic energy is of the order of the Coulomb energy, the solution of Eq. (4) is not easily obtained. A phenomenological scheme for solution in this region was used to bridge the gap between the solutions in the low- and high-field regimes. In the intermediate-field region, a variety of functional forms for eigenvalues can be constructed. In the framework of infinite-order perturbation calculations, one may conclude that the dominant correction term will be an even function of an applied magnetic field, provided the linear Zeeman energy term is absorbed in the unperturbed Hamiltonian. For simplicity, the following form for the eigenvalues for all fields was chosen:

$$E_{ij} = E_B + \left(G_{ij}\gamma + D_{ij}\gamma^2 + \beta_{ij}L_{ij}\gamma^3\right)\Big/\left(1 + \beta_{ij}\gamma^2\right) \tag{33}$$

In the low-field regime, the above reduces to the perturbation scheme, and in the high-field regime, it reduces to the adiabatic scheme, that is, the Landau-level-type solutions.

The magnetic field splitting of the acceptor-bound exciton is quite complicated. It can be seen from Fig. 6 that the 3 initial states will split into a total of 12 states, and the final 1s state will split into a quartet. These transitions have been observed experimentally; however, the energies of the transitions have not been calculated due to the complexity of the problem, which involves the degenerate valence band.

The theory of bound excitons in nondegenerate semiconductors is based on the wurtzite structure. In considering transitions involving bound excitons formed from holes in the top valence band, the $g$ value of the electron is isotropic. The $g$ value of the hole has the form $g_h = g_{h||} \cos\theta$, where $\theta$ is the angle between the $c$ axis of the crystal and the magnetic field direction. The symbols $\oplus$ and $\ominus$ refer to ionized donors and acceptors, respectively, and $+$ and $-$ refer to electrons and holes. The neutral donor-bound exciton is very similar to that for the degenerate case in Fig. 5; however, in the initial state, the unpaired hole is only doubly degenerate. Therefore, a total of four transitions is observed in the presence of a magnetic field. For the orientation $\mathbf{C} \perp \mathbf{H}$, only two transitions will be observed, since the hole $g$ value goes to zero for this orientation. The neutral acceptor-bound exciton will also exhibit a four-line transition in the presence of a magnetic field. In the initial state, the unpaired electron is doubly degenerate, while the final state consisting of the unpaired hole is also doubly degenerate. In this case also, the hole splitting goes to zero for the orientation $\mathbf{C} \perp \mathbf{H}$, resulting in a two-line transition.

## 2. Stress Field

In zinc-blende-type semiconductors, the uniaxial strain patterns and electric dipole selection rules have been derived for lines arising from weakly bound exciton complexes. The effect of stress on excitons bound to shallow neutral acceptors in zinc-blende structures has been rather thoroughly investigated.

In the unstrained crystal, a hole from the $J = \frac{3}{2}$ ($\Gamma_8$) valence band in combination with an electron from the $J = \frac{1}{2}$ ($\Gamma_6$) conduction band gives rise to the ground-state exciton. Uniaxial stress splits the $J = \frac{3}{2}$ degenerate valence band into two bands, one with $M_j = \pm\frac{1}{2}$, the other with $M_j = \pm\frac{3}{2}$. This splitting is reflected in optical transitions involving holes from the valence band. The shallow acceptor removes an energy state from the valence band and establishes it as a quantum state of lower energy in the gap region. This state is made up of valence-band wave functions and therefore will also reflect valence-band splittings.

When an excition is captured by the shallow acceptor, an acceptor-bound exciton complex ($A^0X$) is formed that consists of two holes and one electron weakly bound to a negative acceptor ion. In the absence of stress, three transitions are observed, as shown in Fig. 6.

When a uniaxial stress is applied to the ($A^0X$) complex described above, the degeneracy of the states is lifted due to the splitting of the $\Gamma_8$ hole states. A schematic plot of the resulting energies is shown in Fig. 10. The lower part of the figure shows the splitting of the final (one-hole) state after the collapse of the exciton. The upper part shows

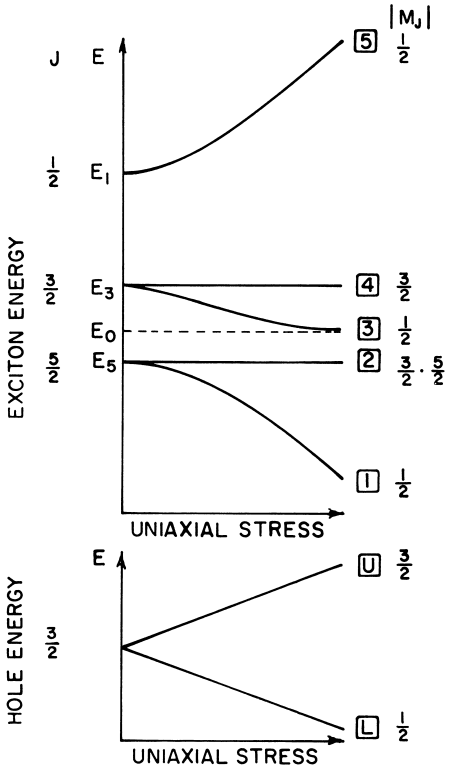



**FIGURE 10** Schematic plot of the splitting of the ($A^0X$) initial and ground states under uniaxial stress. [From Schmidt, M., Morgan, T. N., and Shairer, W. (1975). *Phys. Rev.* **B11,** 5002.]

the splitting and shifts of the initial energy states prior to exciton decay. The energies of these states have been calculated and compared with experimental observations, as shown in Fig. 11. The lines in Fig. 11. show the predicted energy levels (in the absence of a crystal field) for

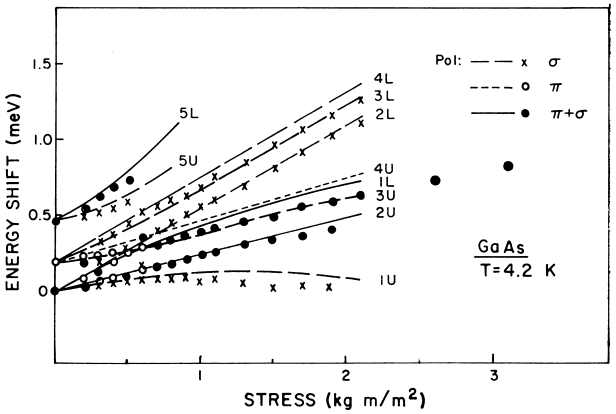



**FIGURE 11** Experimental (points) and calculated (lines) line shifts under increasing uniaxial stress. The stress is applied in a [100] direction. [From Schmidt, M., Morgan, T. N., and Shairer, W. (1975). *Phys. Rev.* **B11,** 5002.]

transitions between the initial and final states, with the center of gravity shift included. The $\sigma$ lines are for transitions polarized perpendicular to the applied stress, while the $\pi$ lines show transitions polarized parallel to the applied stress. The agreement between theory and experiment is very good.

## D. Multibound Excitons

Sharp photoluminescent lines have been observed at energies less than the energy of the line associated with an exciton bound to a neutral donor in silicon, germanium, and silicon carbide. Similar lines have also been observed that are associated with acceptors in silicon and gallium arsenide. The energies and widths of these lines were such that they could not be explained in terms of any recombination mechanism involving just a single exciton bound to a neutral shallow impurity center. A model involving a multiexciton complex bound to a donor (acceptor) was invoked in which each line was associated with radiative recombinations of an exciton in the bound multiexciton complex.

A series of emission lines was observed in silicon crystals also lightly doped with boron or phosphorus. The series began with the bound-exciton line and converged toward the energetic position of the maximum of emission of the condensed electron–hole state. The emission series is shown for both boron and phosphorus dopants in Fig. 12. The impurities can bind a series of intermediate "multiple-exciton states" containing the single bound exciton and electron–hole droplet state.

A model was proposed in which the multiple-exciton complex is built up by successive capture of free excitons at neutral impurity centers. A multiple-exciton complex having index $m$ can capture another free exciton and then have the index $m + 1$; the decay of an exciton would decrease the index to $m - 1$. The observed photon energy $h\nu_m$ is the difference between the energies of the initial and final states,

$$h\nu_m = E_g - E_{FE} - E_m = h\nu_{FE} - E_m \qquad (34)$$

where $E_{FE}$ is the binding energy of the free exciton and $E_m$ that of an exciton in the $m$ complex. The energy difference between the $m$th line $h\nu_m$ and the free-exciton line is a measure of the binding energy $E_m$.

The model was successful in obtaining an empirical fit to the series of emission lines with the series formulas

$$h\nu_m^* = -18.5[1 - \exp(-0.21m)]\ \text{meV} \qquad (35)$$

for Si : B (except for the bound exciton line) and

$$h\nu_m^* = -18.5[1 - \exp(-0.32m)]\ \text{meV} \qquad (36)$$

for Si : P. The calculated line positions are shown in Fig. 12, $h\nu_0^* = 0$ corresponds to the free-exciton line.

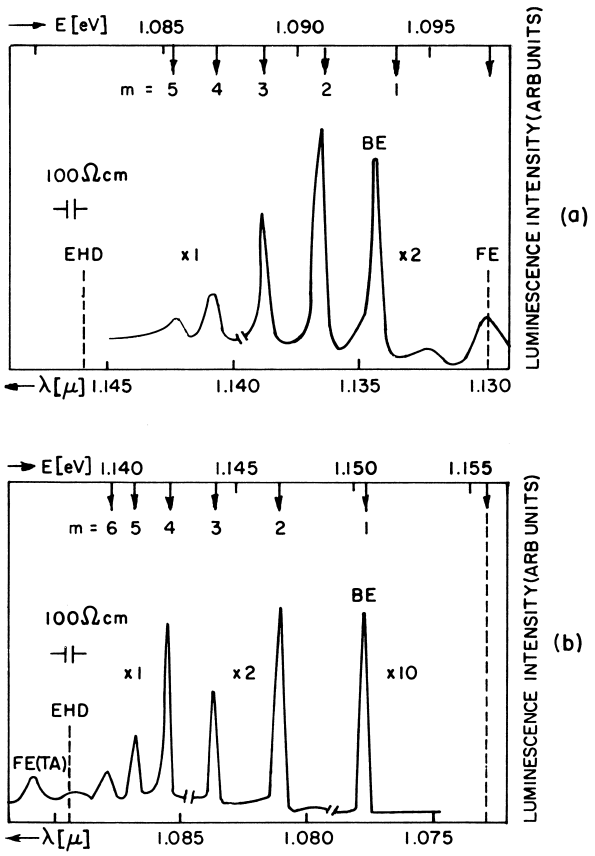



**FIGURE 12** Emission spectra of (a) Si : B with TO phonon and (b) Si : P without phonon assistance. Excitation intensity 7.5 W cm$^{-2}$, $T = 2$ K. The dashed lines indicate the positions of the FEs and the maxima of EHD emission. In Si : P(NP), the FE does not really appear; however, its position is known from the phonon-assisted FE spectrum. The EHD emission in the NP spectrum only appears at higher doping levels and under high excitation. The arrows mark the calculated values $h\nu_m^*$ given by $h\nu_m^* = -18.5[1 - \exp(-0.21m)]$ meV for Si : B and $h\nu_m^* = -18.5[1 - \exp(-0.32\ m)]$ meV for Si : P. [From Sauer, R. (1973). *Phys. Rev. Lett.* **31,** 376.]

Other models have been proposed, including a shell model in which all of the electrons and all of the holes in the bound multiexciton complex are assumed to be equivalent and therefore must conform to the Pauli principle. The complex is then built up along the lines of a shell model similar to what has been used to study nuclei and many-electron atoms.

The behavior of these new lines in the presence of magnetic and stress fields helped to establish the viability of the bound multiexciton complex model.

## E. Donor–Acceptor Pairs

Donor–acceptor pairs introduce transitions in the bound-exciton region whose behavior is quite different from

excitons bound to foreign impurities or defects. The pairs can produce bound states distributed in energy. The range of energies results both from the possible impurities or defects interacting as pairs and from a dependence on pair separation. Discrete pair spectra were first observed as a very complicated spectra in GaP consisting of very many sharp lines. The donors and acceptors will occupy substitutional or interstitial sites. In the case of substitutional sites, both the donor and acceptor can occupy sites on the same sublattice for a compound material such as GaP, or they may be on opposite sublattice sites. Another arrangement is with one impurity at an interstitial site and the other at a particular lattice site. All of these arrangements have been observed.

The energy required to bring a hole and an electron from infinity to an ionized donor–acceptor pair separated by a distance $R$ may be written as

$$E(R) = E_g - E_A - E_D + e^2/\varepsilon R \qquad (37)$$

In this expression $E(R)$ is the energy of the pair recombination line, $E_g$ the band gap of the semiconductor, $E_A$ and $E_D$ the acceptor- and donor-binding energies, respectively, $R$ the donor–acceptor separation, and $\varepsilon$ the low-frequency dielectric constant. When the donor–acceptor distances become small [$R < R_0 =$ (donor–acceptor concentration)$^{1/3}$], a van der Waals attractive term may become important, and Eq. (37) becomes

$$E(R) = E_g - E_A - E_D + e^2/\varepsilon R - (e/K)(a/R)^6 \quad (38)$$

In the case of random pair distribution, it would be expected that over a small range of $R$, the line intensity would reflect the statistical probability of a specific pair occurring. In considering GaP, which has the zinc-blende structure, and assuming that both the donors and acceptors result from substitutional impurities and that both occupy sites on the same sublattice, it is possible to relate $R$ to a given observed line. For the preceding case, $R_m = a_0(\frac{1}{2}m)^{1/2}$, where $a_0$ is the GaP lattice constant and $R_m$ the distance to the $m$th nearest neighbor on the radius of the $m$th shell. The donors and acceptors occupy face-centered cubic sites, and the number of pairs for a given $m$ can be tabulated. The variation in number of pairs allows a correlation with observed spectra. For the case when the donors and acceptors occupy opposite sublattice sites, $R_m = a_0(\frac{1}{2}m - \frac{5}{16})^{1/2}$ and $N(R) > 0$ for all $m$.

The value of $E(R)$ and $R$ are determined from experiment; therefore Eqs. (37) and (38) can be helpful in identifying the donors and acceptors involved in donor–acceptor pair recombination.

## III. INTERACTION OF EXCITONS WITH OTHER SYSTEMS

### A. Phonons

Emission from bound exciton complexes has been observed in many materials. These are very sharp transitions which in many cases are replicated by emission lines that are separated in energy from the parent transition by an optical phonon energy for the particular lattice in question. In crystals having the wurtzite symmetry there is a $\Gamma_1$ and a $\Gamma_5$ longitudinal optical–transverse optical (LO–TO) splitting due to long-range electrostatic forces as well as a $\Gamma_1$–$\Gamma_5$ LO–LO and TO–TO splitting due to anisotropic short-range interatomic forces. The $\Gamma_1$–$\Gamma_5$ LO–LO splitting has been observed on the phonon sidebands, due to the interaction with the macroscopic longitudinal optical phonon electric field, in both CdS and ZnO. In CdS, the phonon-assisted the results from the collapse of the exciton bound to a neutral acceptor with the creation of an LO phonon. Both the $\Gamma_1$ and the $\Gamma_5$ LO phonons are created in this process. These two phonons differ in energy by 2.4 cm$^{-1}$. This small difference in energy is clearly resolved, showing that the phonon-assisted transitions are not appreciably broadened by the phonon interaction.

In the case of CdS, the exciton is rather weakly bound (17 meV) to the acceptor. This results in a localized state in **K** space. The phonon energies show that it is localized near $\mathbf{K} = 0$. The phonon dispersion curves were calculated for CdS using a mixed binding model. In this model the potential contains a short-range part corresponding to covalent bonding and a long-range part due to Coulomb interactions between point ion charges. The calculations show that the LO phonon dispersion curves are quite flat in the vicinity of $\mathbf{K} = 0$. This would account for the very small line broadening observed in the phonon interaction.

A similar interaction between the LO phonon and an exciton bound to a neutral donor was observed in ZnO. In this material, the energy separation between the $\Gamma_1$ LO phonon and the $\Gamma_5$ LO phonon is larger (11 cm$^{-1}$).

#### 1. Exciton–Bound-Phonon Quasi-Particle

Optical transitions have been observed in a number of ionic crystals in which the energy separating the parent transition and its LO-phonon sideband is less than the LO-phonon energy $\hbar\omega_0$ by approximately 10%. In absorption spectra of AgBr : I, transitions associated with the bound exciton occur at energy separations approximately 30% less than $\hbar\omega_0$. These results can be explained in terms of a bound-phonon quasi-particle model. The

calculated binding energies and oscillator strengths for this new quasi-particle can account for phonon interactions whose energies are less than that of the LO phonon $\hbar\omega_0$. LO phonons bound to neutral donors have been observed in both Raman-scattered and luminescence spectra of GaP crystals that were doped with S, Te, Si, and Sn. These results were interpreted as impurity modes associated with dielectric effects of the neutral donors, rather than as local modes associated with mass defects of the substituents.

The spectra due to Raman scattering from the neutral donor LO-phonon bound states are shown in Fig. 13. The binding energies associated with the exciton–bound-phonon quasi-particle for several different donors are given in Table I.

The virtual process

$$\text{donor} + \text{LO phonon } \mathbf{k} \rightarrow \text{excited donor} \rightarrow \text{donor} + \text{LO phonon } \mathbf{k}'$$

produces the interaction of an LO phonon with a donor site. The effective scattering matrix element $H_{\mathbf{k}\mathbf{k}'}$ is proportional to

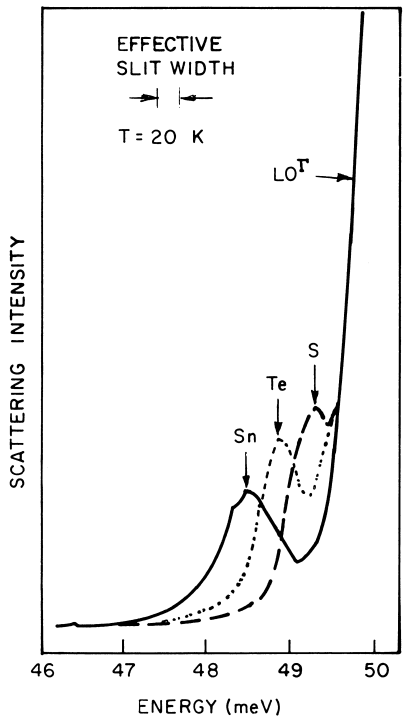



**FIGURE 13** Raman scattering of 5145- Å $Ar^+$ laser light from GaP containing $\sim 10^{15}$ cm$^{-3}$ neutral Sn, Te, or S donors, recorded just below the $k = 0$, LO phonon at 50.2 meV, showing the new local modes. These modes can be seen easily at donor concentrations as low as $10^{17}$ cm$^{-3}$, although their strengths relative to the LO $\Gamma$ lattice normal mode decrease in proportion to the neutral donor concentrations. [From Dean P. J., Manchon, Jr., D. D., and Hopfield, J. J. (1970). *Phys. Rev. Lett.* **25,** 1027.]

$$\sum_j \frac{(E_j - E_0)\langle 0|e^{i\mathbf{k}\cdot\mathbf{r}}|j\rangle\langle j|e^{i\mathbf{k}'\cdot\mathbf{r}}|0\rangle}{\left[(E_j - E_0)^2 - (\hbar\omega)^2\right]K K'} \tag{39}$$

where $|0\rangle$ is the donor ground-state wave function and the sum is over excited donor states $j$. The interaction is attractive when $\hbar\omega$ is less than the excitation energy of the donor. For a spherical approximation, the interaction will produce bound states for each angular momentum of the LO phonon around the donor. When the first excited state of the donor is comparable with the phonon frequency, an approximation in which only the lowest donor excited state is kept will give a reasonable lower bound to the binding energy for $H_{\mathbf{K}\mathbf{K}'}$.

Using this approximation, one obtains the following wave functions and energies for the s and p states. For s states,

$$E_B = \frac{32}{729}\hbar\omega_0\left(\frac{\varepsilon_0}{\varepsilon_\infty} - 1\right)\frac{e^2}{2a\varepsilon_0}\frac{E_{2s} - E_{1s}}{(E_{2s} - E_{1s})^2 - (\hbar\omega_0)^2} \tag{40}$$

$$\Psi_{\mathbf{K}} \alpha \frac{K}{\left[\left(\frac{3}{2}a\right)^2 + K^2\right]^3} \tag{41}$$

For p states,

$$E_B = \frac{224}{6561}\hbar\omega_0\left(\frac{\varepsilon_0}{\varepsilon_\infty} - 1\right)\frac{e^2}{2a\varepsilon_0}\frac{E_{2p} - E_{1s}}{(E_{2p} - E_{1s})^2 - (\hbar\omega_0)^2} \tag{42}$$

$$\Psi_{\mathbf{K}} \alpha \frac{\cos\theta_b}{\left[\left(\frac{3}{2}a\right)^2 K^2\right]^3} \tag{43}$$

In these expressions, $\varepsilon_0$ and $\varepsilon_\infty$ are the static and high-frequency dielectric constants of the Frohlich electron–phonon interaction. The donor Bohr radius is $a$, and the donor binding energy is $e^2/2a\varepsilon_0$ for a hydrogenic donor. The theoretical binding energies from Eqs. (40) and (42) are included in Table I.

## B. Photons

### 1. Spatial Resonance Dispersion

For those states in a crystal where the photon wave vector and the exciton wave vector are essentially equal, the energy denominator for exciton–photon mixing is small and the mixing becomes large. These states are not to be considered as pure photon states or pure exciton states, but rather mixed states. Such a mixed state has been called a polariton. When there is a dispersion of the dielectric constant, spatial dispersion has been invoked to explain certain optical effects of crystals. It was originally thought that it would introduce only small corrections to such things as the index of refraction, until it was

**TABLE I Experimental and Calculated Binding Energies $E_B$ of LO Phonons Localized at Neutral Donors in GaP Observed through Raman Scattering $(E_B)^p$, and as Sidebands in the Luminescence of Excitons Bound to Neutral Donors $(E_B)^{sa}$**

| Donor | (meV) | $E_{1s}(A_1) \rightarrow E_{2s}(A_1)$ (meV) | $E_{1s}(A_1) \rightarrow E^b_{2p}$ (meV) | $(E_B)^s_{calc}$ (meV) | $(E_B)^2_{exp}$ (meV) | $(E_B)^p_{calc}$ (meV) | $(E_B)^p_{exp}$ (meV) |
|---|---|---|---|---|---|---|---|
| S | 104.2[c] | 82.6[c] | 94.0[g] | 0.98 | $1.2 \pm 0.2$ | 0.58 | $0.8 \pm 0.2$ |
| Te | 89.8[d] | 68.2[d] | 79.9[d] | 1.40 | $1.9 \pm 0.2$ | 0.70 | $1.2 \pm 0.2$ |
| Si | 82.5[c] | 70.2[f] | 72.3[c] | 1.16 | ? | 0.82 | $1.3 \pm 0.2$ |
| Sn | 65.5[e] | 53.2[e] | 58.5[e] | [3.6][h] | ? | [1.35][h] | $1.6 \pm 0.2$ |

[a] From Dean, P. J. *et al*. (1970). *Phys. Rev. Lett*. **25,** 1027.
[b] Calculated as weighted mean of $E_{1s}(A_1) \rightarrow E_{2p0}$ and $E_{1s}(A_1) \rightarrow E_{2p\pm}$.
[c] From Onton, A. (1969). *Phys. Rev*. **186,** 786.
[d] From Onton, A., and Raylor, R. C. (1970). *Phys. Rev. B* **1,** 2587.
[e] From Dean, P. J., Faulkner, R. A., Schonherr, E. G. (1970). *Proc. Int. Conf. Phys. Semicond., 10th*, Cambridge, Mass., p. 286.
[f] Assuming the $E_{2s}(A_1)$ is the same as for Sn.
[g] Assuming the $E_{2po}$ and $E_{2p\pm}$ are the same for S and Te.
[h] Using degenerate perturbation theory.

demonstrated that if there was more than one energy transport mechanism, as in the case of excitons, this was not true. Spatial dispersion addresses the possibility that two different kinds of waves of the same energy and same polarization can exist in a crystal differing only in wave vector. The one with an anomalously large wave vector is an anomalous wave. In the treatment of dispersion by exciton theory, it was shown that if the normal modes of the system were allowed to depend on the wave vector, a much higher order equation for the index of refraction would result. The new solutions occur whenever there is any curvature of the ordinary exciton band in the region of large exciton–photon coupling. These results apply to the Lorentz model as well as to quantum-mechanical models whenever there is a dependence of frequency on wave vector.

It was pointed out early in the investigation of spatial dispersion that the specific dipole moment of polarization of a crystal and the electric field intensity are not in direct proportion. It was found that the two were related by a differential equation that resulted in giving Maxwell equations of higher order. This led to the existence of several waves of the same frequency, polarization, and direction but with different indices of refraction. Subsequent studies of the reflectivity of CdS demonstrated the effects of spatial dispersion. Extensive calculations resulted in the following expression for the index of refraction:

$$n^2 = \frac{c^2k^2}{\omega^2} = \varepsilon + \sum_j \frac{4\pi\left(\alpha_{0j} + \alpha_{2j}k^2\right)\omega_j^2}{\omega_{0j}^2 + \left(\hbar\omega_{0j}k^2/m_j^*\right) - \omega^2 - i\omega\Gamma_j} \quad (44)$$

In this equation the sum over $j$ is to include the excitons in the frequency region of interest, and the contributions from other oscillators are in included in a background dielectric constant $\varepsilon$. In Eq. (44) one has expanded both the numerator and denominator in powers of $k$, keeping terms to order of $k^2$, $m^*$ is the sum of the effective masses of the hole and electron that comprise the exciton, and $\omega_{0j}$ is the frequency of the $j$th oscillator at $\mathbf{k} = 0$. Eliminating $k^2$ from Eq. (44) and neglecting the line width of the oscillators, the working equation is

$$n^2 = \varepsilon + \sum_j \frac{4\pi\left(\alpha_{0j} + \alpha_{2j}\omega^2n^2/c^2\right)\omega_{0j}^2}{\omega_{0j}^2 + \left(\hbar\omega_{0j}^2\omega^2n^2/c^2m_j^*\right) - \omega^2} \quad (45)$$

The sum is over excitons from the top two valence bands, where the "allowed" excitons have been included with $\alpha_{0j} \neq 0$, $\alpha_{2j} = 0$ while the "forbidden" (which are seen only because $\mathbf{k} \neq 0$) excitons are included with $\alpha_{0j} = 0$, $\alpha_{2j} \neq 0$. The above equation reduces to a polynomial in $n^2$ whose roots give the wavelengths of the various "normal modes" for transfer of energy within the crystal.

In the classical case, $\alpha_{2j} = \hbar\omega_{0j}/m_j^* = 0$. For a given frequency, the two roots of $n^2$ for Eq. (45) are $-n$ and $+n$. Thus in the classical case, for a given principle polarization, frequency, and direction of propagation, only one transverse mode exists.

## 2. Two-Photon Processes

The concept of two-photon processes dates back more than 40 years, and was first treated theoretically. The observation of two-photon transitions occurred approximately 30 years later. The two-photon transition is a nonlinear process, and as such its full potential as a tool for investigating material parameters was not realized until the advent of laser sources.

The process is one in which two quanta are simultaneously absorbed in an electronic transition. The energy

sum of the two quanta must be equal to the energy of the electronic transition. In one-photon absorption spectra, the absorption coefficient as a function of photon energy is obtained. In two-photon transitions, the absorption is dependent on two frequencies; therefore, instead of a plane curve, the spectrum is a two-dimensional surface. The two processes are complementary; however, they yield different information. The single-photon process is allowed between states of different parity, while the two-photon transitions are allowed between states of the same parity. The polarization dependence of two-photon absorption is more complicated than it is for one-photon processes. Even for isotropic materials, the mutual orientation of the electric vectors of the absorbed fields is important. Thus, the two-photon spectrum contains more information than the one-photon spectrum. All two-quantum processes require an intermediate state in which one photon is absorbed or emitted and the atom is left in an excited state.

The fine structure of the 2P states of the exciton from the top valence level in CdS has been studied. In this experiment, a visible dye laser and a $CO_2$ laser were used. From the polarization selection rules, it was shown that the visible photon created the virtual 1s exciton, and the absorption of the infrared photon brings it to the final p state. Using these results, the two-photon absorption coefficient was calculated and compared with the experimentally measured results. The agreement between theory and experiment is very good.

## IV. SPECIAL PROPERTIES OF EXCITONS

### A. Introduction

The excitonic properties of the semiconductor are key to the development of a large number of devices. In this section five of the properties that have direct application will be outlined. The first is the exciton states in quantum wells. The binding energy of the exciton can be adjusted due to the thickness of the layers of the different semiconductor materials. This leads to having different energy responses, as well as many other electrical and optical properties.

The second area is hole–electron droplets, which are extremely useful in the study of many-body effects and the phase transition from a gas into a liquid. The role of excitons in developing a high-temperature superconductor is given in Section IV.D. When this property is exploited, it could have a major impact on modern electronic devices. The last two areas outlined are lasing transitions and optical bistability. These two areas lead to totally optical switching devices, which again may lead to many useful devices such as the optical computer.

### B. Excitons in Quantum Wells and Quantum Dots

Superlattice structures have generated considerable interest for more than a decade because of the novel transport phenomena predicted for such structures. The superlattice is a multilayered periodic structure having dimensions varying from a few-angstroms to hundreds of angstroms. Carriers in semiconductor superlattices may be confined to certain layers by the superlattice potential variations, resulting in new conductivity properties. The allowed carrier energy levels are determined by quantization effects when the confined regions are sufficiently small. The confined regions produce new optical effects as well as electrical effects. Superlattices emerged as practical structures when the metal organic chemical vapor deposition (MOCVD) and molecular beam epitaxy (MBE) crystal growth techniques evolved, making high-quality structures feasible. Very thin layers with smooth surface morphology can be grown by these techniques. One of the very common heterostructures produced by these techniques is GaAs/$Al_xGa_{1-x}As$. By cladding the GaAs layer with GaAIAs barriers, the electrons and holes are confined within the GaAs well, resulting in a modification of their energy levels in the well. Repeating the growth of these layers results in a multiquantum well (MQW) structure, the number of wells being equal to the number of repeated cycles. When the layer thicknesses are small, the electrons are confined as quantized electron waves. If the barrier-layer thicknesses are large enough, tunneling between wells does not occur. The confinement of carriers within the GaAs well results in an effective increase in the bandgap. The low-temperature bandgap of bulk GaAs is 1.5196 eV. $Al_xGa_{1-x}As$ has a direct bandgap for $x < 0.45$, with a bulk bandgap which is 1.25 $x$ eV greater than GaAs. In the quantum well structure, the difference in bandgap is divided between the conduction band and the valence band. The percentage contribution to each band is a measure of the confining barrier for that band. Well-size quantization results in a shift of the allowed energies for electrons. If infinite confining barriers are assumed, the allowed minimum energies for electrons are given by

$$E_n = h^2n^2/8m^*L^2 \tag{46}$$

where $h$ is Planck's constant, $n$ is an integer marking the number of half-wavelengths of the confined electron, $m^*$ is the effective mass, and $L$ is the well thickness. These energy shifts for electrons in the conduction band are shown in Fig. 14.

The energy levels in the valence band of the quantum well are also modified. In bulk GaAs, the light and heavy hole valence bands are degenerate at $K = 0$. From Eq. (46) it is seen that the energy of the confined particles is

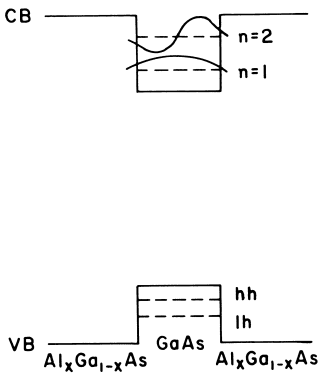



**FIGURE 14** Quantum-state energy levels in GaAs/$Al_xGa_{1-x}As$ quantum well. CB and VB refer to conduction and valence bands, respectively.

different for different masses. The layered structure has reduced the cubic symmetry of bulk GaAs to uniaxial symmetry. The optical transitions thus become nondegenerate. This feature is shown in the valence band of Fig. 14.

The line shape of the absorption and photoluminescence bands of GaAs (MQW) is excitonic. Verification of these bands as being assigned to light and heavy-hole free-exciton transitions has been made from polarization measurements. These measurements include optical spin-orientation measurements and linear polarization measurements and linear polarization measurements of emission emanating from a cleaved edge of the MQW.

The lowest level conduction and valence subband levels are shown in Fig. 14. In absorption and photoluminescence, it is the excitons associated with these subband levels that are observed. A summary of a calculation of the energy levels of heavy- and light-hole excitons associated with the lowest electron and hole subbands for finite values of the potential barrier heights is presented.

The Hamiltonian of an exciton associated with either the heavy- or the light-hole band, in a quantum well structure as shown in Fig. 14, within the framework of an effective mass approximation is given by:

$$H = \frac{-h^2}{2\mu_\pm}\left[\frac{1}{\rho}\frac{\partial}{\partial\rho}\rho\frac{\partial}{\partial\rho} + \frac{1}{\rho^2}\frac{\partial^2}{\partial\phi^2}\right] - \frac{\hbar^2}{2m_e}\frac{\partial^2}{\partial z_e^2} - \frac{\hbar^2}{2m_\pm}\frac{\partial^2}{\partial z_h^2} - \frac{e^2}{\varepsilon_0|\mathbf{r}_e - \mathbf{r}_h|} + V_{ew}(Z_e) + V_{hw}(Z_h) \tag{47}$$

where $m_e$ is the effective mass of the conduction electron, $\varepsilon_0$ is the static dielectric constant, $m_\pm$ is the heavy (+) or light (−) hole mass along the $Z$ direction, and $\mu_\pm$ is the reduced mass corresponding to heavy (+) or light (−) hole bands in the plane perpendicular to the $Z$ axis. Both $\mu_\pm$ and $m_\pm$ can be expressed in terms of the Kohn-Luttinger band parameters $\gamma_1$ and $\gamma_2$ as

$$\frac{1}{\mu_\pm} = \frac{1}{m_e} + \frac{1}{m_0}(\gamma_1 \pm \gamma_2) \tag{48}$$

and

$$\frac{1}{m_\pm} = \frac{1}{m_0}(\gamma_1 \mp 2\gamma_2) \tag{49}$$

where $m_0$ is the free electron mass. In these equations the upper sign refers to the $J_Z = \pm\frac{3}{2}$ (heavy-hole) band and the lower sign to the $J_Z = \pm\frac{1}{2}$ (light-hole) band. The positions of the electron and hole are designated by $\mathbf{r}_e$ and $\mathbf{r}_h$, respectively, $\rho, \phi$, and $Z$ are the cylindrical coordinates. The potential wells for the conduction electron, $V_{ew}(Z_e)$ and the holes $V_{hw}(Z_h)$ are assumed to be square wells of width $L$

$$V_{ew}(Z_e) = \begin{cases} 0, & |Z_e| < L/2 \\ V_e, & |Z_e| > L/2 \end{cases} \tag{49a}$$

and

$$V_{hw}(Z_h) = \begin{cases} 0, & |Z_h| < L/2 \\ V_h, & |Z_h| > L/2 \end{cases} \tag{49b}$$

The values of $V_e$ and $V_h$ are determined from the Al concentration in the $Al_xGa_{1-x}As$ barrier.

An exact solution of the Schrödinger equation corresponding to the exciton Hamiltonian Eq. (47) is not possible. A variational approach was used to calculate the ground-state energy $E_1$ of the Hamiltonian. The binding energy of the ground state of an exciton $E_{1s}$ is then obtained by subtracting $E_1$ from the sum of the lowest electron and hole subband energies ($E_e + E_h$). These subband energies are determined by solving the transcendental equations for finite square-wells.

The binding energy of the ground state of a heavy-hole exciton $E_{1s}(h)$ (solid lines) and a light-hole exciton $E_{1s}(\ell)$ (dashed lines) as a function of well-width $L$ for different values of the potential-barrier heights are shown in Fig. 15. For a given value of $X$, the value of $E_{1s}(h)$ increases as $L$ is reduced until it reaches a maximum and then drops quite rapidly. Similar behavior is exhibited by $E_{1s}(\ell)$. The explanation of this behavior is that as $L$ is reduced the exciton wave function is compressed in the quantum well, leading to increased binding. However, beyond a certain value of $L$ the spread of the exciton wave function into the barrier becomes important. This causes the binding energy to approach the value in bulk $Al_xGa_{1-x}As$ as $L$ becomes diminishingly small.

It is seen that for a given value of $X$, $E_{1s}(\ell)$ is larger than $E_{1s}(h)$ for $L$ greater than a certain critical value $L_c$, at which they become equal. For values below $L_c$, $E_{1s}(\ell)$ is smaller than $E_{1s}(h)$. The value of $L_c$ depends on $X$; the larger the value of $X$ the smaller the value of $L_c$. For $X = 0.3$, $L_c = 50$ Å. This behavior can be understood as follows: the value of $E_{1s}(\ell)$ is greater than that of $E_{1s}(h)$

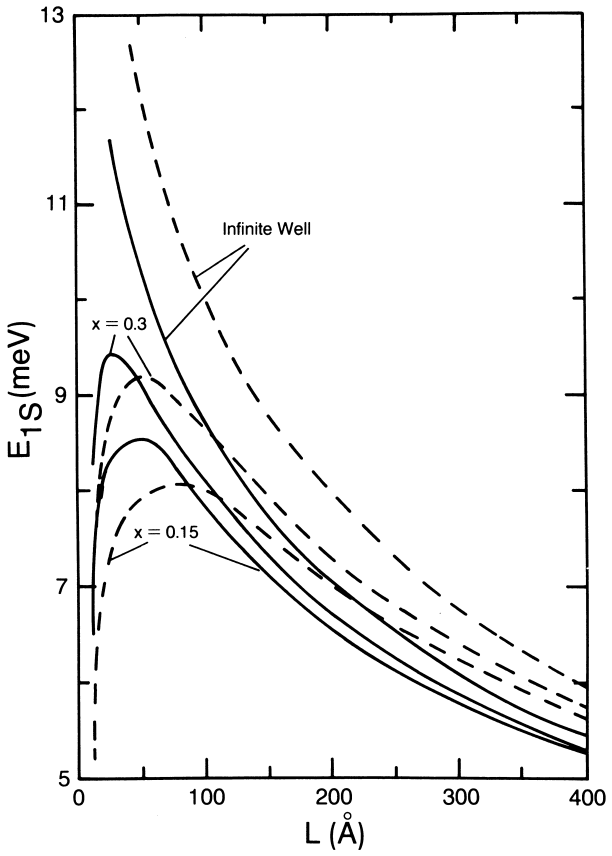



**FIGURE 15** Variation of the binding energy of the ground state, $E_{1s}$, of a heavy-hole exciton (solid lines) and a light-hole exciton (dashed lines) as a function of the GaAs quantum-well size ($\ell$) for A$\ell$ concentration $X = 0.15$ and 0.3, and for an infinite potential well.

for large $L$. Both increase as $L$ is reduced, $E_{1s}(\ell)$ less rapidly than $E_{1s}(h)$ as proportionately more of the light-hole exciton wave function penetrates into the barrier, thus reducing the increase in $E_{1s}(\ell)$. At a certain value of $L$, which depends on $X$, the two values become equal, and then $E_{1s}(\ell)$ becomes smaller as $L$ is reduced further. This contrasts with the behavior of $E_{1s}(h)$ and $E_{1s}(\ell)$ for infinite potential barriers where $E_{1s}(\ell)$ is always larger than $E_{1s}(h)$.

In the case of quantum dots, two general configurations are considered. The energy levels in the first case, type I quantum dots, are similar to those in Fig. 14. Here the conduction band of the quantum dot is lower in energy than the bulk conduction band and the valence-band energy is higher that that of the bulk valence band. If the combination dot size and potential barrier are large enough, an exciton may be bound to the quantum dot. The overlap of the electron and hole wave functions of the dot exciton will be larger than that of the bulk exciton leading to very interesting studies of the exchange and correlation energy terms of the exciton. One can also create excitons and other multiexcitons, which are also found within the quantum dot, leading to detailed studies of these many-body quasi-states as well.

In the second case, type II quantum dots, either the valence or conduction band energy of the quantum dot is lower than the corresponding bulk energy bands. This leads to one of the carriers of the exciton being found within the quantum dot, while the other carrier is made up of bulk band states and, in general, insensitive to the energy states of the dot. As in the case of quantum wells, the superlattice dots of both type I and type II quantum dots will lead to more and novel device applications.

## C. Hole-Electron Droplets

Nonequilibrium electrons and holes in semiconductors are bound in excitons at low temperatures by Coulomb attraction. The exciton forms because it represents a state of slightly lower energy than the unbound hole–electron. At high exciting intensities the density of hole–electron pairs is increased and excitons are formed at a higher rate. At high concentrations the interaction among electrons becomes very important, and when a certain threshold density is reached liquid droplets are formed. These collective droplets consist of nonequilibrium electrons and holes; therefore, when electron–hole recombination occurs, specific radiation is emitted.

This intense radiation was first observed in Si. The spectra contained, in addition to the well-known peaks due to the annihilation of free excitons with appropriate phonons, some broad bands of radiation shifted toward lower energies.

The formation of a condensed phase of nonequilibrium carriers was considered, to account for the new radiation peaks. If the collective interaction predominates in the condensed phase, survival of excitons as quasi-particles is doubtful. It was shown that the condensed phase consists of a degenerate electron–hole plasma characterized by metallic properties.

The most convincing evidence that the new substance is in a separate phase was obtained from light-scattering experiments. The effect is analogous to the scattering of light by drops dispersed in a fog. The range of angles through which the radiation is scattered depends on the size of the particles, the larger scattering angles resulting from smaller particles.

The far-infrared absorption of the condensed phase in Ge was investigated. The results were explained by assuming that the absorption is caused by the excitation of plasmons in the drops of the condensed phase. From these experiments the drop radius was estimated to be $R \approx 10^{-3}$ to $10^{-4}$ cm.

When the excitons enter the liquid state (electron–hole drops), the electron and hole give up their exclusive association and enter a sea of particles in which they are bound equally to all of the other charge carriers in the droplet. The droplet therefore is made up of independent electrons and holes. Since the density in the droplet is greater than in the exciton gas, it can be qualitatively understood why droplets form, by considering the relation of energy to charge carrier distance. Free electrons and holes which have the greatest separation recombine with the highest energy. Excitons form only when the electron and hole are coupled, resulting in an appropriate Bohr radius for the particular material being considered (in the case of Ge it is approximately 115 Å). The energy of the resulting recombination radiation from the excitons is less than the energy of the recombination radiation of the free electrons and holes, reflecting the exciton binding energy. In the case of the liquid droplet, the electron–hole distance is still further reduced (100 Å for Ge). Therefore, the liquid state forms because it is a reduced energy state of the system.

The liquid is made up of independent electrons and holes, which gives a metallic character to the liquid, whereas the exciton gas is an insulator. In the exciton gas the particles are far enough apart to behave as a classical gas, they move independently and their velocities are determined by random processes. The probability of finding a particle with a given energy falls off exponentially as the energy is increased. The exciton gas obeys Maxwell–Boltzmann statistics. The combination of the statistical distribution in the gas with the density of states gives the shape of the luminescent line as shown in Fig. 16. In the liquid droplet the electrons and holes are close enough together so that the droplet must be considered as a single system. The availability of states for occupation in such a system is determined by the Pauli exclusion principle. The probability of the charge carriers occupying the available states is determined by Fermi–Dirac statistics. The uppermost filled state at the absolute zero of temperature can be considered the Fermi level. Here again, the statistical distribution coupled with the density of states determines the shape of the luminescent peak. Here it is assumed that any electron is equally likely to recombine with any hole. Recombination in which both particles have very high or very low energies is unlikely. The maximum intensity will occur when the difference in energy of the two particles is near the medium value as shown in Fig. 16. The luminescence spectrum of free excitons and electron–hole drops in Ge is also shown in Fig. 16. The line shapes agree reasonably well with theory.

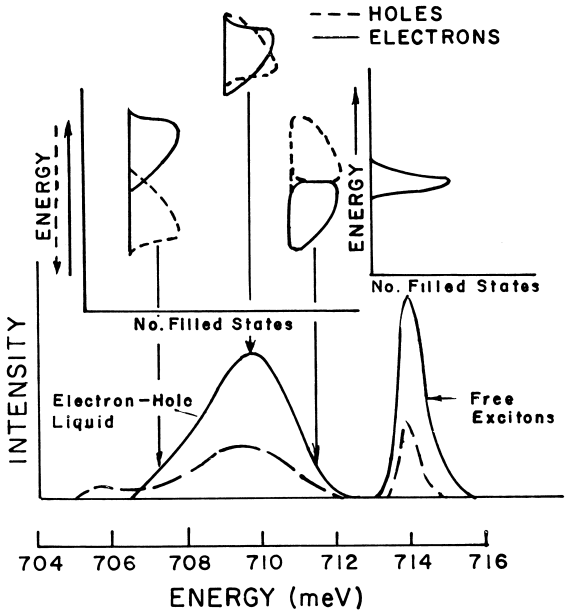



**FIGURE 16** Line shapes for free excitons and electron–hole drops in Ge(___), the combination of the statistical distribution with the density states(– – –), and experimental spectra. [From Lo, T. K. (1974). *Solid State Comm.* **15,** 1231.]

## D. Exciton Mechanism in Superconductivity

The experimental discovery of superconductivity in $La_{2-x}Sr_xCuO_4$ with an upper limit of the critical temperature $T_c$ in the 30–40 K range in 1986, and in $YBa_2Cu_3O_{7-x}$ with $T_c$ in the 90–100 K range in 1987 surpasses the $T_c$'s thought possible for normal phonon processes. Soon afterward $T_c$'s up to 107 and 125 K were discovered in similar compounds containing Bi and T1, respectively. Analyzing a large number of experiments, one can conclude that the superconductivity in these ceramic materials is due to weakly coupled, quasi-particle pairs. This indicates that the Bardeen, Cooper, Schrieffer (BCS) theory can be applied. This model would explain the superconductivity in these materials as resulting from the replacement of the phonon by the exciton as the mediating particle of the coupling field.

It is generally considered that one must have two different electronic systems closely located in space but separated by energy and character for one to have an excitonic mechanism for superconductivity. Such a situation exhibits itself in the compounds discussed above. For example, the $YBa_2Cu_3O_{7-x}$ system has some localized electrons, some electrons that are covalently bonded, and some electrons which form a free, quasi-particle structure near the Fermi surface. $YBa_2Cu_3O_{7-x}$ is made up of layers or layers or layer structures which are predominantly bonded together by ionic forces. In the first layer, Y gives up three electrons, then comes a Cu-O plane, where the oxygen can be assigned nearly two extra electrons and the Cu has given up two electrons and is almost doubly positively charged. Then, comes a Ba-O layer followed by a Cu-O chainlike structure. This is followed by a Ba-O

plane, another Cu-O plane, and $Y^{3+}$ plane. Focusing on the Cu-O plane and chain, it is found that the plane valence and conduction-like states are pushed to lower energy by the close proximity of the $Y^{3+}$ compared to the chainlike electronic states. One finds that some of the chain has holelike quasi-particle states at the Fermi surface. These states are located at the Brillouin zone boundary. These holelike states are intinerant and are the current carrying states.

There is another state found both in the chain and the Cu-O plane which appears to be caused by approximations that have been made in the electronic band structure calculations which do not account fully for correlation effects. To understand this more fully, one has to resort to atomiclike calculations and focus on the Cu-like states embedded both in the plane and the chain including correlation. One finds a ground state for the Cu to be $Cu^{2+}$ with one of the d electronic states unoccupied and at a much higher energy. Namely, one of the d $(x^2 - y^2)$ states is not converged and $Cu^{2+}$ does not exist in the ground state. Returning now to the approximations that are made in one electron band structure, one finds this artificial state arising very quickly from the center of the Brillouin zone to 2 to 3 eV above the Fermi surface. In this way the band calculation tries to account for these correlation effects and form $Cu^{+2}$ ($d^9$). There is no experimental verification for this state, as it is only an artifact of the approximation. Also, using the atomiclike calculations one finds the first excited state (the primary exciton) in the 4–6 eV range for the Cu electronic system. These states are seen in the optical spectra experimentally.

The above discussion demonstrates the two electron-like systems with which to develop an excitonic theory for superconductivity. The holelike states are predominantly related to the $O^{2-}$ electronic states of the chain. As this quasi-particle travels along the chain, it approaches the localized Cu-like states and polarizes the Cu valence states. The polarization states are represented by the mixing in of the 4 s and p conduction bands of the Cu electronic system. This is represented in the model by the exciton discussed above.

One would ordinarily expect that the holelike states of $O^{2-}$ would be able to screen out this interaction much more readily than the polarization of the d-like electrons. However, as stated before, the planelike states of Cu-O are suppressed to lower energy than those of the chain. Thus, when the holelike states attempt to readjust they tend to mix in the conductionlike states from the plane, causing their polarization charge to be distributed into other layers away from the chain. This leaves the d-like states polarized without the total dampening out of the effect by the quasi-particles at the Fermi surface.

Another important feature is that the polarization disturbance cannot keep up with the quasi-particle, but remains behind at the $Cu^{2+}$ site and cannot dissipate before the second quasi-particle arrives to absorb the energy from the field. The two quasi-particles thus form a Cooper pair mediated by the excitonic field of the Cu ion.

Starting with the energy band results and the correlation effects related to the $Cu^{2+}(d^9)$ one can add in the two quasi-particle pair states (Cooper pair states) and the polarization field to derive an effective interaction between the two quasi-particle states given below. These two quasi-particles are two hole states from $O^{2-}$ -like states that are at the Fermi surface. This effective interaction is given by

$$V_{\text{eff}}(q) = \frac{4\pi\varepsilon^2}{|\mathbf{q}|^2\kappa(q)}\left[\frac{\left(\omega_{\text{exc}}^2/\kappa(q)\right)(1 - 1/\varepsilon_\infty)}{q_0^2 - \omega_{\text{exc}}^2/\kappa(q) + i\delta} + 1\right] \quad (50)$$

In the above expression $\omega_{\text{exc}}$ is the Cu $3d - 4s$ and p exciton frequency and $\kappa(q)$ is the screening function caused by the $O^{2-}$ quasi-particles. $q_0$ are the calculated band energies of the hole states and $\varepsilon_\infty$ is the dielectric constant in infinite frequency of the crystal. $(4\pi e^2)/(|q|^2\kappa(q))$ represents the Coulomb interaction. One finds if the first term in the bracket on the right-hand side is negative and has an absolute value larger than one, superconductivity will exist. It is interesting to note that the first term on the right-hand side will have a different sign as long as $\omega_{\text{exc}}^2 \backslash \kappa(q)$ is larger than $q^2$. As more and more hole states are added $\kappa(q)$ becomes larger; thus, the ratio of $\omega_{\text{exc}}^2 \backslash \kappa(q)$ becomes smaller and the superconductivity will disappear. This happens experimentally in $YBa_2Cu_3O_{7-x}$ as $x$ increases, and superconductivity is totally gone in these compounds by the time $x$ has reached 0.5. As we remove oxygen from this compound, the vacancy appears in the chain and has the effect of lowering the Fermi energy. This creates the number of holes near the Fermi surface and, in turn, increases $\kappa$ and diminishes the ratio.

There are also four other well-defined experiments that to date are inexplicable without invoking a superconducting state in which the phonon-mediated electron–electron interaction is replaced by excitons.

The first experiment is one in which an attempt was made to measure the effects of gravity on the position of an electron in a copper tube. A large temperature-dependent transition in the magnitude of the ambient axial electric field inside the vertical copper tube was found. Above a temperature of 4.5 K the ambient field was $3 \times 10^{-7}$ V/m or greater. Below 4.5 K the magnitude of the ambient field drops very rapidly, to $5 \times 10^{-11}$ V/m at 4.2 K.

These measurements were made using time-of-flight spectra of an electron traveling in the center of the copper tube, and the field effects on the tube were screened out. The equation for the time of flight spectra is

$$T = \left(\frac{m}{2}\right)^{1/2} \int_0^h [W - ezE_{(\mathrm{amb})(z)} - ezE_{\mathrm{app}} - mgz]^{-1/2} dz \qquad (51)$$

where $W$ is the related kinetic-energy term of the electron, $h$ is the length of the tube, $mgz$ is the gravitational term, $E_{\mathrm{amb}}$ is the effective ambient electric field [$E_{\mathrm{amb}}$ is assumed to consist of a constant term due to gravitationally induced distortions of the tube and a term due to the patch (roughness of the surface) effect with a complicated $z$ dependence], and $E_{\mathrm{app}}$ is the uniform applied field. It is believed that this ambient field is screened out by superconducting electrons in the oxide layer that forms on the copper tube.

The next two experiments were measurements of CuCl that revealed large changes in electrical conductivity and magnetic susceptibility. It was found that when polycrystalline samples of CuCl under hydrostatic pressure of approximately 5 kbar were rapidly cooled (20 K/min), they went through repeated transitions from a state of weak diamagnetism to a state of strong diamagnetism. The diamagnetic susceptibility ($X$) varied between $10^{-5}$ and approximately $-1$. When $X \approx -1$ the Meissner effect, in which the magnet field is totally excluded from the sample, was observed. This phenomenon is characteristic of superconducting materials. The strong diamagnetic state was accompanied by a sharp increase in electrical conductivity.

Similar experiments were performed on the same material in carefully controlled environments. In these experiments, the diamagnetic anomaly was observed above 90 K over a temperature range of 10–20 K, accompanied by a sharp increase in electrical conductivity.

The above experiment can be explained using the exciton model to describe events within the calculated band structure of CuCl. In zero applied pressure, CuCl is a direct-gap material with minimum energy at the $\Gamma$ point. When pressure is introduced into the band calculation by reducing the lattice constant, the conduction band at the $X$ point moves down in energy relative to the conduction band at the $\Gamma$ point. The conduction band at $X$ becomes degenerate with the conduction band at $\Gamma$ for a lattice constant reduction of 0.2%, which is consistent with the pressures used in the experiments.

Another important point is that oxygen was present in all of the CuCl samples. The calculations reveal that the energy of the oxygen-bound electron is less than the binding energy of the exciton, and oxygen will give up some of its electrons to the conduction band at the experimental temperature.

An upper limit on the critical temperature $T_c$ of an electron–exciton coupled superconductor can be obtained assuming the static limit of the coupling of the electron and exciton and assuming the $T_c$ can be approximated by

$$T_c = 1.14 \frac{\hbar\omega_{\mathrm{exc}}}{K_B} \exp[-1/uD(E_F)] \qquad (52)$$

Here $\hbar\omega_{\mathrm{exc}}$ is the energy of the exciton, $u$ is the coupling coefficient of the exciton, $K_B$ is Boltzmann's constant, and $D(E_F)$ is the density of electrons at the Fermi surface. When the conduction bands at $\Gamma$ and $X$ are degenerate, $D(E_F)$ is a maximum. It was found that for carrier densities $N(E) = 10^{-1}$ $e$/unit cell, $T_c$ for a nondegenerate conduction band was 38 K, while it was 1745 K for the degenerate case. For $N(E) = 10^{-2}$ $e$/unit cell, the corresponding numbers are $10^{-2}$ K and 55 K, respectively. Therefore it is possible that with $10^{-2}$ $e$/unit cell, superconductivity close to or above liquid-nitrogen temperatures can be achieved.

In the final experiment, studies of CdS samples that had been pressure-quenched at 77 K showed strong diamagnetic effects. A super paramagnetic effect was also detected. In this experiment one must explain the pressure effects, the super paramagnetic effect, the diamagnetic anomaly, and the fact that these are only observed in selected samples. As in the case of CuCl, these effects are interpreted in terms of a superconducting state induced by the interaction of the band structure with applied pressure and specific impurity effects.

### E. Lasing Transitions

Semiconductor lasers were first reported in 1962, the first being the GaAs injection laser. Since that time many semiconductor lasers have been produced from III–V compounds, and they cover an appreciable portion of the spectrum from 0.65 to 8.5 $\mu$m.

Shortly thereafter, rapid developments were made in the area of II–VI compound lasers. The first report of high-efficiency laser action was in electron-beam-pumped CdS. Even higher efficiencies were later achieved in electron-beam-pumped CdSe. The spontaneous line, centered at 6800 Å for CdSe, corresponds to an emission line that has been observed in photoluminescence experiments and has been attributed to an exciton bound to an acceptor. The spontaneous line in CdS at 4.2 K is the 4888-Å line and was also associated with an exciton bound to a neutral acceptor site.

The recombination radiation from highly excited CdS crystals was investigated. The experimental technique allowed the determination of the spectral dependence of the optical gain. From these investigations, it was concluded that at least three different processes can contribute to laser action. A low-gain process results from the annihilation of a free exciton and the emission of a photon and an

LO phonon. A medium-gain process is due to an exciton–exciton interaction, and a high-gain process involves an exciton–electron interaction.

The interpretation of the excitation dependence of the spontaneous emission and of the gain for the free exciton-related processes are as follows:

1. For the excitation intensity $J < 1$ A/cm$^2$, only the $E_x$–LO process yielded some gain.
2. For 1 A/cm$^2$ $< J <$ 3 A/cm$^2$, the low-energy tail resulting from electron–exciton interaction line dominates.
3. For $J > 3$ A/cm$^2$, the low-energy tail resulting from electron–exciton interaction is the dominant gain process. CdS, CdS: Se, and CdSe lasers provide a tunability from 0.5 to 0.7 m. In CdS, mode-locked pulses shorter than 4 psec have been obtained.

### F. Optical Bistability

Optical bistability can be defined as any optical system possessing two different steady-state transmissions for the same input intensity. To achieve optical bistability, the optical device must have feedback. This implies that the transmission intensity must have some dependence on the output intensity. Many of the bistable devices have been Fabry–Perot etalons containing materials having nonlinear indices of refraction at high input light intensities. In this type of device, the cavity is tuned so that a transmission maximum lies close to the laser frequency, but still having low transmission at low input intensities. As the input intensity is increased, the light penetrating the cavity will be sufficient to cause the nonlinear index material to tune the cavity toward the laser frequency. This has been termed intrinsic dispersive optical bistability—intrinsic because the nonlinear index material provides the feedback, and dispersive because the reflective or real part of the nonlinear susceptibility is more important than the imaginary or absorption part.

For practical optical bistable devices, attention has been focused on semiconductor materials. Semiconductors have high absorption coefficients, particularly at resonant excitonic transitions. These materials will produce absorptive and dispersive bistable devices. Useful absorption in these materials is achieved in very short transversing paths, making very fast switching devices achievable. Both the free- and bound-exciton transitions in semiconductors have shown promise for high-speed, low-power switching devices. These characteristics make possible fast, all-optical, signal-processing devices.

## SEE ALSO THE FOLLOWING ARTICLES

BONDING AND STRUCTURE IN SOLIDS • CRYSTALLOGRAPHY • GROUP THEORY, APPLIED • LASERS, SEMICONDUCTOR • METALORGANIC CHEMICAL VAPOR DEPOSITION (MOCVD) • MOLECULAR BEAM EPITAXY, SEMICONDUCTORS • QUANTUM MECHANICS • SUPERCONDUCTIVITY

## BIBLIOGRAPHY

Baldereschi, A., and Lipari, N. O., (1971). *Phys. Rev*. **B3,** 439.
Cardona, M. (1969). "Modulation Spectroscopy," by F. Seitz and D. Turnbull and E. Ehrenreich. *Solid State Physics, Suppl. 11*. Academic Press, New York.
Craig, D. P., and Walmsley, (1968). "Excitons in Molecular Crystals," Benjamin, Elmsford, New York.
Davydov, A. S. (1962). "The Theory of Molecular Excitons," McGraw-Hill, New York.
Dexter, D. L., and Knox, R. S. (1965). "Excitons," Interscience Publishers, New York.
Dimmock, J. O., (1967). "Theory of Excton States," *Semicond. and Semimetals* **3,** Chap. 7, Academic Press, New York.
Green, R. L., Bajaj, K. K., and Phelps, D. E., (1984). *Phys. Rev*. **B29,** 1807.
Knox, R. S., (1963). "Theory of Excitons," *Solid State Phys. Suppl*. 5, Academic Press, New York.
Rashba, E. I., and Sturge, M. D., ed. (1982). "Excitons," North Holland Publishing Co., Amsterdam.
Reynolds, D. C., and Collins, T. C., (1981). "Excitons, Their Properties and Uses," Academic Press, New York.
Thomas, D. G., and Hopfield, J. J., (1959). *Phys. Rev*. **116,** 573.
Thomas, D. G., and Hopfield, J. J., (1962). *Phys. Rev*. **128,** 2135.
Wheeler, R. G., and Dimmock, J. O., (1962). *Phys. Rev*. **125,** 1805.

# Ferromagnetism

**H. R. Khan**

*Forschungsinstitut für Edelmetalle und Metallchemie and University of Tennessee at Knoxville*

I. Basic Concept of Magnetism
II. Origin of Magnetism
III. Magnetization Curves
IV. The Hysteresis Loop
V. Anisotropic Magnetization
VI. Magnetic Order
VII. Ferromagnetic Domains
VIII. Magnetostriction
IX. Magnons
X. Ferromagnetism and Superconductivity
XI. Magnetoresistance and Giant Magnetoresistance
XII. Ferromagnetic Materials and Their Applications

## GLOSSARY

**Magnetic field** A magnet attracts a piece of iron at a distance and this is caused by the magnetic field or the field of force of the magnet.

**Magnetic poles** A magnet has north and south poles. Like poles repel and unlike poles attract each other with a force that varies inversely as the square of the distance between them. A unit pole is defined in such a way that two like unit poles placed one centimeter apart in vacuum would repel each other with a force of one dyne.

**Magnetic field strength** The magnetic field strength may be defined in terms of magnetic poles, for example, one centimeter from a unit pole the field strength is one Oersted. In the MKS system, the unit of field strength is one ampere-turn/meter.

**Magnetic moment** The magnetic moment of a small plane coil is a product of the current $I$ flowing in the coil and the area of the coil $A$, $IA$ ($Am^2$). The magnetic moment of a small magnet is equal to the magnetic moment of a small coil that would experience the same torque when placed in the same orientation at the same location in the same magnetic field.

**Magnetic flux density** In the international system of units (SI), the magnetic field intensity $H$ (A/m), and the magnetization $M$ (A/m) are related to the magnetic

induction or the magnetic flux density $B$ (Wb/m$^2$) through the relation

$$B = \mu_0(H + M)$$

where $\mu_0$ is the permeability of the free space and has the value $12.57 \times 10^{-7}$ Wb/Am.

**Magnetization ($M$)** Magnetic moment per unit volume.

**Coercive force ($H_c$)** Negative value of the magnetizing field $H$ that makes the magnetic induction $B$ of a ferromagnetic material zero.

**Remanence ($M_r$)** If the magnetic flux density does not drop to zero upon reducing the magnetic field on the specimen to zero, then this remained magnetic flux density in the specimen is called the remanence.

**Permeability ($\mu$)** The ratio $B/H$ is the permeability of a material.

**Magnetic susceptibility ($\chi$)** The ratio $M/H$ is called the magnetic susceptibility of a material and may be expressed in mass, volume, or molar units.

**Curie temperature ($T_c$)** Temperature above which the spontaneous magnetization of a ferromagnetic material vanishes.

**Neel temperature ($T_N$)** Temperature below which the interaction between the atomic moments affecting antiparallel orientation surmounts the thermal agitation. At Neel temperature $T_N$, the susceptibility of a material has its maximum.

**A SPECIAL ARRANGEMENT** of electrons in the atoms causes a material to become a ferromagnetic material. For example, the incompletely filled M shells of iron, cobalt, and nickel atoms are responsible for the ferromagnetism in these metals. Atoms behave as small magnets ordered in parallel arrangement in ferromagnetic materials. The magnetization curve and the hysteresis loop determine whether it is a hard or soft ferromagnetic. The parameters determined from the hysteresis loop are the permeability, coercivity, remanence, and the area of the loop itself. The area of the loop gives the energy loss per unit volume of the specimen per cycle and is dissipated as heat energy called as hysteresis loss. The hysteresis loop and the parameter derived from it determine the suitability of a material in a particular application.

## I. BASIC CONCEPT OF MAGNETISM

All materials occurring in nature are magnetic. They may be paramagnetic, diamagnetic, ferromagnetic, antiferromagnetic, or ferrimagnetic. The magnetic behavior of a material depends on its electronic structure. For example, the ferromagnetism of iron, cobalt, and nickel in the periodic system is due to the incompletely filled M shells of their atoms. Due to these incompletely filled shells, the atoms behave as magnets ordered in parallel arrangement in ferromagnetic materials. In antiferromagnetic materials, the atomic magnets are ordered in antiparallel arrangement. Ferrimagnetic materials are a special case of ferromagnetic materials. The neighboring atoms interact with each other in a material, and this interaction force is dependent on the distance between neighboring atoms and the diameter of the atomic shell responsible for the atomic magnetic moment. The sign and magnitude of this interaction force cause a material to show different magnetic behavior. The usefulness of a ferromagnetic material is shown by its magnetization curve and hysteresis loop. The hysteresis loop provides information about the "permeability" and "coercivity" of a ferromagnetic material. For example, a soft magnetic material to be used as a transformer core should have a high value of permeability, whereas a hard or permanent magnetic material should have a high value of coercivity. The physical condition, purity, and composition of a material control the useful magnetic properties like permeability and coercivity, and they can be modified by controlling different parameters of a material. By rapid solidification of materials from the melt, very soft magnetic and hard magnetic materials can be produced. Both the hard magnetic and soft magnetic materials find applications in the electrical and electronic industry. Some examples of their uses are transformers, motors, generators, relays, telephone cables, audio and video recording and replaying, and data memory systems and computers.

A magnetic field gradient is generated by a magnet consisting of a flat north pole and a pointed south pole, as shown in Fig. 1. The magnetic field is stronger near the pointed pole. lf a piece of material in the cylindrical form is suspended with a string between the poles, then a magnetic force is generated on this material. The kind of magnetism on the material is determined from the direction of the

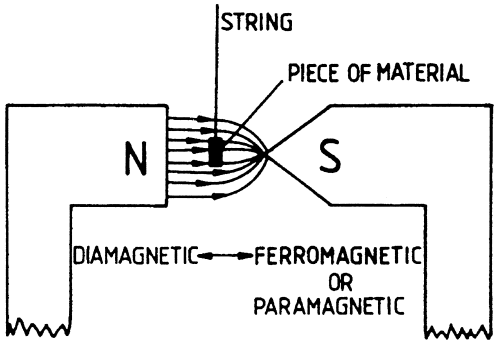



**FIGURE 1** An experimental setup to check the paramagnetism, diamagnetism, and ferromagnetism of a material.

magnetic force on this material. If the material is strongly attracted toward the pointed pole, then it is ferromagnetic. A paramagnetic material is weakly attracted toward the pointed pole, and a diamagnetic material is repelled by the pointed pole. For example, iron is ferromagnetic, aluminum is paramagnetic, and bismuth is diamagnetic. Materials can be synthesized that show ferromagnetism, although the constituents may be paramagnetic or diamagnetic. Ferrimagnetism and antiferromagnetism are closely related to ferromagnetism. Many compounds are ferrimagnetic. Ferrimagnetic materials in general are oxides of the ferromagnetic metals.

## II. ORIGIN OF MAGNETISM

The earliest human experience with magnetism involved a mineral magnetite (also known as lodestone). This is the only material that occurs naturally in a magnetic state. In the classical picture, an atom consists of a nucleus surrounded by a number of electrons that depends on the element and its position in the periodic system. The electrons are distributed in different shells, named K, L, M, and N, outward from the center. Each shell can accommodate only a certain number of electrons, the maximum number being $2n^2$, where $n$ is the number of the shell. The innermost shell K is complete with $2 \times 1^2 = 2$ electrons, the second shell L is complete with $2 \times 2^2 = 8$, the third shell M with $2 \times 3^2 = 18$ electrons, etc. For example, the ferromagnetic element iron has its M shell incompletely filled and contains only 14 electrons. An electron carries a negative charge, and its motion in an orbit gives rise to an electric current. The orbital motion of the electron is equivalent to a thin magnet and produces a magnetic field. Besides the orbital motion, the electron also spins around its own axis. The spinning negative charge also gives rise to an electric current and behaves like a small magnet. The completely filled shells are magnetically neutral, because an equal number of electrons spin in clockwise and anticlockwise directions. As mentioned earlier, the M shell of iron contains only 14 electrons instead of 18, and nine electrons spin in clockwise and five in anticlockwise directions. The uncompensated four electrons produce a magnetic field at a distance equal to four electrons.

Therefore iron has a magnetic moment of 4 units. The other ferromagnetic element, cobalt, has only 15 electrons in the M shell and carries a magnetic moment of 3 units. Nickel is also ferromagnetic because its atom in the M shell has only 16 electrons and carries a magnetic moment of 2 units. Thus we have seen that ferromagnetism originates from the unfilled M shells in iron, cobalt, and nickel atoms.

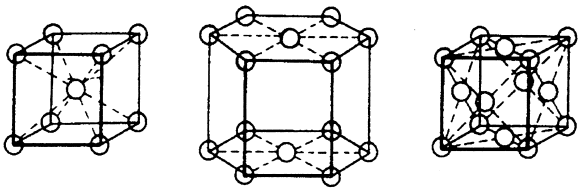



**FIGURE 2** The crystal structures of the ferromagnetic metals iron, cobalt, and nickel.

The free isolated atoms of iron, cobalt, and nickel have magnetic moments equivalent to 4, 3, and 2 units. In metals, the atoms are not isolated but are packed together. This packing of atoms influences the distribution of electrons in the M shell, because this is outside of an atom. Therefore the experimentally measured values of the magnetic moments of iron, cobalt, and nickel are 2.22, 1.71, and 0.606 units, and are lower compared to the isolated free atoms.

Each element in the periodic system has a definite crystal structure. In iron and nickel atoms, the atoms occupy the body-centered cubic (bcc) and face-centered cubic (fcc) lattice sites, respectively, whereas in cobalt the atoms occupy the hexagonal lattice sites as shown in Fig. 2. In ferromagnetic elements, each atom carries a magnetic moment and a magnetic axis, and even in the absence of an externally applied magnetic field these atomic magnets point in one direction, as shown in Fig. 3. The internal magnetic field required to order the atomic magnets in one direction is called a Weiss molecular field. For example, in the case of iron its intensity is $5.5 \times 10^6$ Å/cm.

The interaction of neighboring atoms with permanent dipole moments causes the alignment of atomic magnets in the same direction (ferromagnets; for example, iron, nickel, rare earths with $64 < Z < 69$; alloys like $Cu_2MnAl$, etc.; magnitude of $\chi$ is large below $T_c$), in opposite directions (antiferromagnets; for example, MnO, CoO, NiO, $Cr_2O_3$, $CuCl_2$ etc., magnitude of $\chi$ is like paramagnetic materials), or in alignment in opposite directions but moments are unequal [ferrimagnetic; $Fe_3O_4$ (magnetite), $\gamma Fe_2O_3$ (maghemite) etc., magnitude of $\chi$ is like

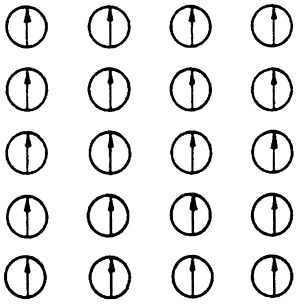


**FIGURE 3** Arrangement of the atomic magnets in a ferromagnetic material.

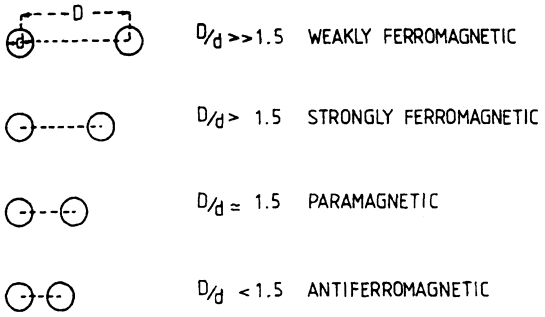



**FIGURE 4** Ratios of the interatomic distance $D$ to the unfilled atomic shell diameter $d$ causing different kinds of magnetism.

ferromagnetic materials]. The force of interaction is a function of the ratio of the distance $D$ between the neighboring atoms and the diameter $d$ of the atomic shell responsible for the atomic magnetic moment. The force of interaction changing from positive to negative value causes a material to become ferromagnetic, weakly ferromagnetic, paramagnetic, or antiferromagnetic as shown in Figs. 4 and 5. The iron, cobalt, and nickel are strongly ferromagnetic because the ratio $D/d$ is larger than 1.5, whereas gadolinium is weakly ferromagnetic because the value of $D/d$ is about 3.1.

When a ferromagnetic material with its atomic magnets pointed in one direction is heated, the thermal vibrations of the atoms become stronger with the increase in temperature. When the temperature reaches a value called a magnetic change point or Curie temperature, the atomic magnets orient themselves randomly and the ferromagnetic material transforms to a paramagnetic state. The Curie temperature of iron is 770°C, 1130°C for cobalt, and 360°C for nickel.

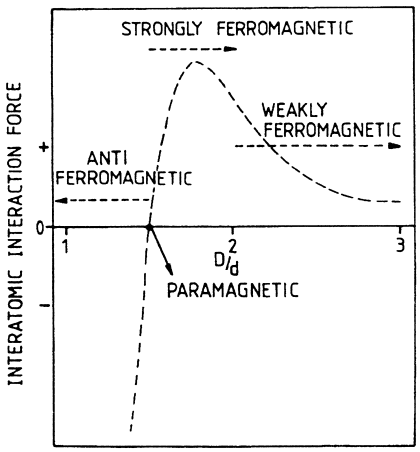



**FIGURE 5** Variation of the interatomic interaction force with the ratio $D/d$ of the interatomic distance $D$ to the diameter $d$ of the unfilled atomic shell.

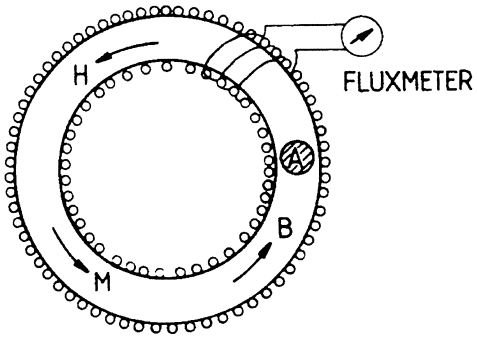



**FIGURE 6** Experimental setup for plotting the magnetization curve.

## III. MAGNETIZATION CURVES

The magnetization $M$ of a material is defined as the magnetic moment per unit volume. The practical usefulness of a ferromagnetic material is determined from its magnetization curve. The experimental setup for plotting the magnetization curve is shown in Fig. 6. A thin torroidal ring of the ferromagnetic material of cross section $A$ is wound with $N$ turns per meter. A current of $I$ amperes flowing through the winding generates a magnetic flux $\Phi = BA$ in the ring. A flux meter connected with a secondary coil of few turns measures this flux $\Phi$. The flux density $B = \Phi/A$ is composed of two parts: one arising from the external current flowing in the winding, and the second arising from the internal current associated with the motion of electrons in the ferromagnetic material.

The flux density arising from the external current $NI$ per meter of the winding is $\mu_0 NI$, where $\mu_0$ is a constant. The magnetization $M$ arising from the internal currents in the ferromagnetic material may be considered constant across the cross section. The magnetic field set up by this magnetization is $\mu_0 M$. The magnetic intensity in the core is the sum of these two contributions:

$$B_{\text{core}} = \mu_0 NI + \mu_0 M \tag{1}$$

In the absence of a ferromagnetic core

$$B_{\text{no core}} = \mu_0 NI \tag{2}$$

The ratio $B_{\text{core}}/B_{\text{no core}}$ is defined as the permeability of a ferromagnetic material. The permeability is the ratio of the magnetic intensity in a torroidal core to the intensity that the same current in the same winding would produce in the absence of a ferromagnetic core, and it is dimensionless. The permeability of a material depends on its history and is very high (~1000) for the soft magnetic materials like iron.

The value of the magnetizing field $H = NI$ is increased by increasing the current $I$ in the winding, which increases

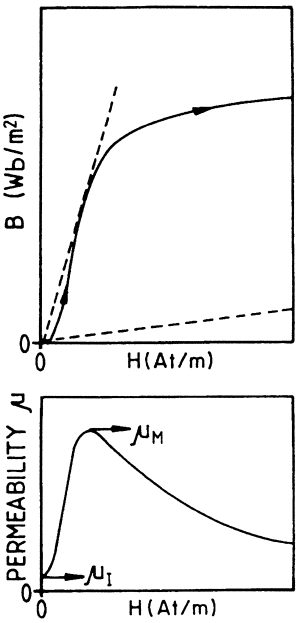



**FIGURE 7** Plot of the magnetic induction $B$ as a function of the magnetizing field $H$ of a ferromagnetic material, and plot of the permeability $\mu$ as a function of the magnetizing field $H$ (ampere turns per meter).

the value of $B$ measured by the flux meter. A plot of $B$ versus $H$ is called a magnetization curve ($B$–$H$ curve) and is shown in Fig. 7. The term $\mu_I$ is the initial permeability of the material obtained from initial slope. When the magnetizing field $H$ is increased, the ratio $B/H$ also increases until a maximum value is reached. The slope at this point, $\mu_m = B/H$, is called the maximum permeability. The term $B$ also achieves its maximum value, called the saturation magnetic intensity $B_s$.

The variation of the initial and maximum permeability as a function of the magnetizing field of a ferromagnetic material is shown in Fig. 7. All ferromagnetic materials show this kind of $B$–$H$ and $\mu$–$H$ behavior, but the magnitudes of the permeability and the scales of the $B$ and $H$ are different for different materials. Many practical uses of ferromagnetic materials require them to possess high values of magnetizing fields. Typical examples are the core of low-current transformers, low-current relays, inductive loading of telephone cables, and the sensitive detectors of small field changes. The best ferromagnetic materials for these applications are those with highest $\mu_I$ and $\mu_m$ values.

## IV. THE HYSTERESIS LOOP

The initial magnetization as shown in Fig. 7 is not reversible. When $H$ slowly increases, the value of $B$ also increases until it reaches its maximum value at point A as shown in Fig. 8.

Thus 0A represents the initial magnetization curve. When the magnetic field is slowly decreased, the flux density $B$ follows the curve AC. The magnetic field is zero at C, but at this point a flux density equal to 0C remains in the ferromagnetic material. If the magnetic field is reversed to 0D, the flux density is completely removed from the material, and further reducing the field in the negative direction brings the flux density to point E. The flux density follows the curve EFA if the direction of the magnetic field is changed and slowly increased. lf the cycle is repeated a few times, it brings the material in the cyclic state. The loop ACEFA is called the *hysteresis loop*. The flux density at the point C is called *remanence*, and the reverse field at point D is called *coercive force*.

The total energy required to magnetize a unit volume of the specimen from 0 to A on the initial curve is given by

$$W = \int_0^A H\, dB$$

lf the magnetic field is reversed to zero, then the returned path on the hysteresis loop is AC and the total energy taken from the magnetizing field $H$ is the area 0AC. If the curve 0A were traced back to the original path, then the energy taken from the magnetizing field $H$ would have been returned to it and there would have been no loss of energy. But in the case of the hysteresis loop shown in Fig. 8, there is a loss of energy. The total energy per unit volume of the specimen taken from the magnetizing field $H$ for one complete cycle of the hysteresis loop is the area of the hysteresis loop, which is

$$\int_0^A H\, dB$$

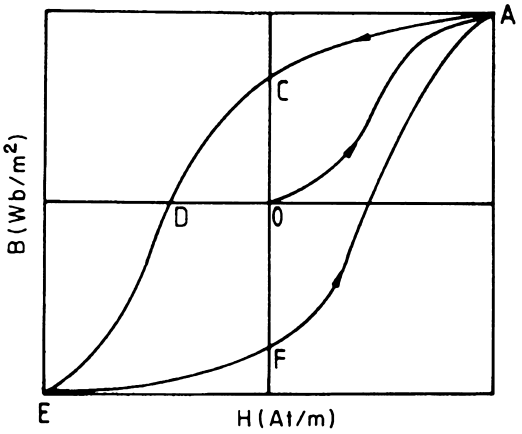



**FIGURE 8** Plot of the magnetic induction $B$ as a function of magnetizing field $H$, hysteresis loop.

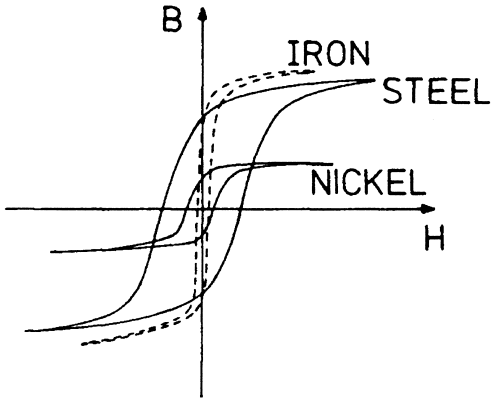



**FIGURE 9** Hysteresis loops of some ferromagnetic materials.

This energy is dissipated as heat energy and is called the hysteresis loss.

Hysteresis loops of small areas are observed for the soft ferromagnetic materials, whereas the hard ferromagnetic materials show large areas. The hysteresis loop of an ideal soft ferromagnet should show just a line, but in practice such materials do not exist. Some amorphous iron-based materials show hysteresis loops of very small area. A typical example of a hard ferromagnetic with a relatively large hysteresis loop area is carbon steel. The hysteresis loop area is also an important parameter in determining the application of a ferromagnetic material. The hysteresis loops of some ferromagnetic materials are shown in Fig. 9.

Good permanent magnets possess high values of residual flux and coercive force. These ferromagnetic materials with high residual flux and coercive force cannot be used in motors or transformers, because the flux changes continuously and there is an energy loss in each cycle. The energy loss is proportional to the area of the hysteresis loop. Therefore a ferromagnetic material subjected to cyclic magnetization, as in motors and transformers, should have as narrow a hysteresis loop as possible.

## V. ANISOTROPIC MAGNETIZATION

The energy in single crystals of ferromagnetic material that governs the magnetization along the crystallographic axes is called the magnetocrystalline or anisotropy energy. It is easy to magnetize an iron crystal along the cubic-edge directions rather than along directions of other crystal axes. However, a nickel crystal is easily magnetized along the long diagonal axis compared to the cubic-edge directions. Cobalt with its hexagonal crystal structure can be easily magnetized in the direction of the hexagonal axis. These crystallographic axes of a ferromagnetic crystal along which they are easily magnetized are called directions of easy magnetization.

## VI. MAGNETIC ORDER

The magnetic susceptibility $\chi$ per unit volume of a magnetic material is defined as the ratio of the magnetization $M$ to the macroscopic magnetizing field intensity $B$: $\chi = M/B$. The magnetic susceptibility may be defined in terms of the unit mass, "mass susceptibility," or mole (molar susceptibility) of a material, or unit volume, called volume susceptibility. The variation of $\chi$ with temperature of a paramagnetic material is shown in Fig. 9 and is related to the temperature as $\chi = C/T$, called the Curie law. Here the constant $C$ is called the Curie constant.

### A. Ferromagnetic Order

In a ferromagnetic material, the individual magnetic moments are ordered in parallel arangement, as shown in Fig. 3. A ferromagnet possesses a magnetic moment even in the absence of an externally applied magnetic field, and this spontaneous magnetic moment is also called the *saturation moment*. When a ferromagnet is heated, the parallel arrangement disappears above the Curie temperature. The magnetic susceptibility of a ferromagnetic material at temperatures close to the Curie temperature is related to the temperature by

$$\chi = \frac{C}{(T - T_c)}$$

which is also shown in Fig. 9.

### B. Antiferromagnetic Order

The atomic magnetic moments order in an antiparallel arrangement in an antiferromagnetic material is shown in Fig. 10. The resultant moment is zero below the ordering or *Neel*

$$\chi = \frac{C}{(T + \theta)}$$

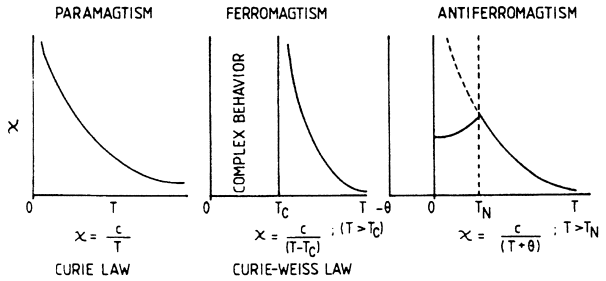



**FIGURE 10** The arrangement of the atomic magnets in an antiferromagnetic material.

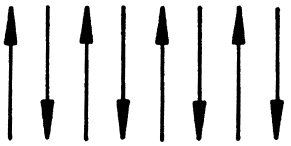


**FIGURE 11** Temperature-dependent magnetic susceptibility $\chi$ behavior of a paramagnetic, ferromagnetic material.

*tempetature* $T_N$. The variation of the magnetic susceptibility $\chi$ of an antiferromagnetic material with temperature is shown in Fig. 11 and is given by the relation

$$\chi = \frac{C}{(T + \theta)}$$

where $\theta$ is the Neel temperature $T_N$.

### C. Ferrimagnetic Order

In some ferromagnetic materials, the saturation magnetization does not correspond to the parallel alignment of the individual atomic magnetic moments. These materials are the magnetic oxides with the general chemical formula $MO \cdot Fe_2O_3$, where M may be the metals, for example, zinc, cadmium, iron, nickel, copper, cobalt, or magnesium. These ferrites have the spinel crystal structure. There are eight occupied tetrahedral sites and 16 occupied octahedral sites in this crystal structure. In these ferisites iron has two different ionic states, ferrous ($Fe^{2+}$) and ferric ($Fe^{3+}$). The eight tetrahedral sites in the cubic spinel structure are occupied by $Fe^{3+}$, whereas half of the 16 octahedral sites are occupied by $Fe^{3+}$ and the rest by $Fe^{2+}$ ions. The magnetic moments of eight $Fe^{3+}$ ions on the tetrahedral and octahedral sites cancel each other, leaving only the magnetic moments of the eight $Fe^{2+}$ ions, as shown in Fig. 12.

## VII. FERROMAGNETIC DOMAINS

At temperatures below the Curie point, the magnetic moment may be much less than the saturation moment of a ferromagnetic material.

The polycrystalline as well as the single-crystal specimen consist of small regions called *domains*, within each

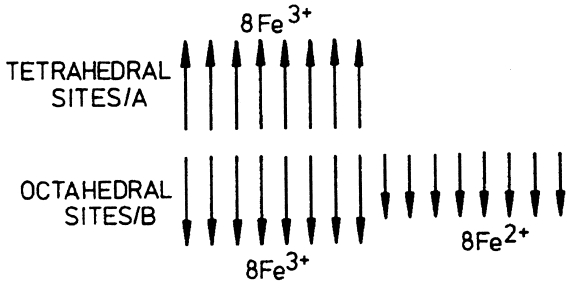



**FIGURE 12** The arrangement of the atomic magnets in a ferrite of cubic spinel structure.

of which the local magnetization is saturated. The magnetic axes of these domains may point in different directions, and it is possible that for a certain arrangement it might give a zero resultant magnetic moment of the specimen. The application of an external magnetic field saturates the specimen, because the external field causes the orientation of the domain magnetization in the direction of the applied magnetic field. Small cylindrical magnetic domains may be stabilized in a thin crystal of uniaxial material by applying a bias magnetic field. The bubble diameter is on the order of 10 $\mu$m. These magnetic bubbles are of interest in high-density memory-storage devices.

As discussed in Section III, the permeability and coercivity are important parameters that control the practical application of a ferromagnetic material. The domain structure of a ferromagnetic material affects both of these parameters. A pure, well-oriented, and homogeneous material facilitates the domain boundary displacement and possesses high permeability. On the other hand, an inhomogeneous material consisting of multiple phases suppresses the boundary displacement and possesses high "coercivity."

F. Bitter developed a simple method to observe the domain boundaries. A drop of a colloidal suspension of a finely divided ferromagnetic material such as magnetite is placed on the surface of a ferromagnetic material. The colloidal particles in the suspension concentrate strongly on the boundaries between the domains where the strong local magnetic fields exist that attract the magnetic particles. A simple domain structure in a silicon-iron single crystal is shown in Fig. 13.

## VIII. MAGNETOSTRICTION

When a ferromagnetic material is magnetized, small changes in the physical dimensions of the specimen take place, and this effect is called *magnetostriction*. This magnetostriction of a material is defined as the increase in length per unit length in the direction of the magnetization. Magnetostriction is different for different axes of a ferromagnetic single crystal. The useful parameter "permeability" of a ferromagnetic material is related to the magnetostriction.

## IX. MAGNONS

A ferromagnetic material in the ground state has all its spins arranged parallel in one direction, as shown in Fig. 14a. The excited state is obtained if the spins are reversed. Figure 14b shows the excited state where one spin is antiparallel. The elementare excitations are the spin

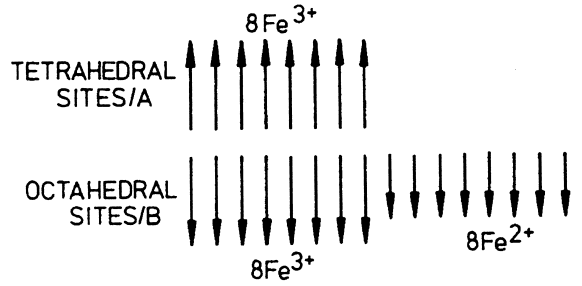



**FIGURE 13** Simple domain structure in silicon-iron single crystal. [After Williams, Bozorth, and Shockley (1949). *Phys. Rev.* **75,** 155.1

waves, and one wavelength is shown in Fig. 14c. These elementare excitations are called *magnons* and are analogous to the lattice vibrations or phonons. Spin waves are the relative orientation of the spins on a lattice, whereas the lattice vibrations are the oscillations in the relative position of the atoms on a lattice. Spin waves have been observed by neutron scattering experiments near the Curie temperature or even above the Curie temperature.

## X. FERROMAGNETISM AND SUPERCONDUCTIVITY

Both ferromagnetism and superconductivity involve spin ordering. The difference is that in a ferromagnet the spins order parallel, whereas in a superconductor they order antiparallel below the superconducting transition temperature and form the "Cooper pairs."

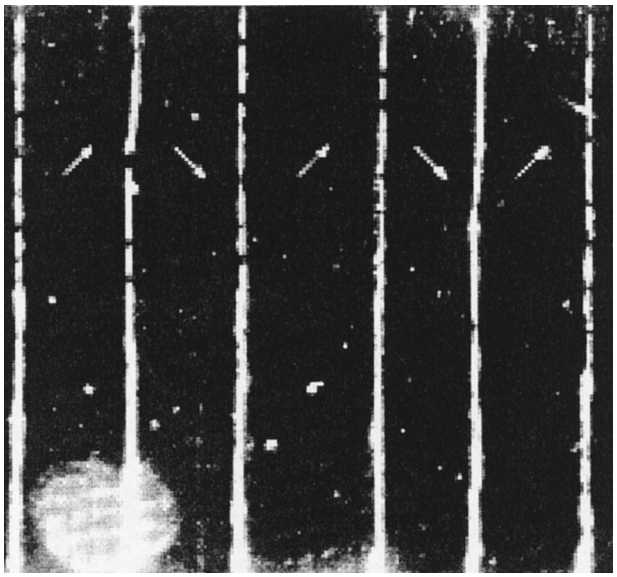


**FIGURE 14** (a) The arrangement of spins in a ferromagnetic material. (b) The elementare excitation occurs when one spin is antiparallel. (c) One wavelength of the spin wave.

A possibility of coexistence of superconductivity and ferromagnetism in the same material was proposed. To observe this coexistence, the ferromagnetic impurities were dissolved in superconducting materials, for example, gadolinium (ferromagnetic) in lanthanum (superconducting). The lanthanum-gadolinium compounds were superconducting up to 1 at.% gadolinium and became ferromagnetic for the concentrations of gadolinium above 2.5 at.%.

Recently some temary compounds of the formula $MRh_4B_4$, with B = as thorium, yttrium, neodymium, samarium, gadolinium, terbium, dysprosium, holmium, erbium, or lutetium, having $CeCo_4B_4$ structure, have been discovered. A typical example of a compound showing both ferromagnetism and superconductivity is $Er_4Rh_4B_4$, which is superconducting at 8.7 K and ferromagnetic at 0.9 K.

## XI. MAGNETORESISTANCE AND GIANT MAGNETORESISTANCE

William Thompson in 1857 showed that the electrical resistance of a ferromagnetic material, for example, iron, changes under the influence of a magnetic field. This phenomen is called magnetoresistance. In most magnetic materials this magnetoresistance increases with magnetization when current and magnetization are parallel and decreases when they are at right angles to each other. The magnitude of the change in resistivity caused by magnetization to the saturation magnetization is usually a few percent and rarely exeeds 5% at room temperature.

This change in resistance may be used for reading magnetically recorded information. Magnetoresistive read heads use permalloy as the magnetoresistive material. The advantages of magnetoresistive read head over the inductive read head are that it can be made very small to read the high-density recorded information and the head does not have to move relative to the medium. In the conventional inductive case, the read head has to move relative to the medium, because electromagnetic induction can only be produced by a changing magnetic flux. The advanced read head is based on the multilayered thin films system. The change in resistance of some of the multilayer system in a magnetic field can be as large as 60–70% and this effect is called Giant Magnetoresistace (Baibich *et al.* and Binasch *et al.*). These multilayered systems can be used to make very sensitive and small read heads. Typical examples of multilayer systems are the iron-chromium system with alternate layers of iron and chromium or the cobalt-copper system with alternate layers of cobalt and copper. The thickness of the magnetic and nonmagnetic layers is

few nanometers. In these systems the magnetic layers are magnetized antiparallel via exchange interaction across the nonmagnetic layers, for example, copper or chromium. When a sufficiently large magnetic field is applied, it can align the magnetization of the magnetic layers in a parallel direction. In a magnetic metal, the free electrons which carry the electric curent have their spins aligned either parallel or antiparallel to the magnetization of the metal. Electrons experience different resistance to their motion depending on the direction of their spin relative to the magnetization. When current flows in the plane of the multilayers with antiparallel magnetization, the electrons experience a high resistance; but when a magnetic field is applied and the magnetization of the magnetic layers changes to a parallel state, the electrons experience lower resistance. In the case of conventional magnetoresistance, the resistance increases with the application of magnetic field; whereas in the case of Giant Magnetoresistance, the resistance decreases with the application of magnetic field because of the change from antiparallel to parellel magnetization of the magnetic layers.

## XII. FERROMAGNETIC MATERIALS AND THEIR APPLICATIONS

The hysteresis loop of a ferromagnetic material provides information about its usefulness in technical applications, as discussed in Section IV. The hysteresis loop depends on the physical condition, composition, and purity of a specimen. Depending on the application of a ferromagnetic material, the important properties are the "permeability" and "coercivity."

When a strain-free material is cold-worked, the permeability of the material is reduced and the hysteresis loss is increased. The strain-relieving heat treatment of the cold-worked specimen again brings the original magnetic properties back, for example, the permeability is increased and the hysteresis loss reduced. In general, the strain-free crystals show the minimum hysteresis loss.

The presence of the impurities carbon, oxygen, nitrogen, sulfur, etc., affects the permeability and hysteresis loss of a ferromagnetic material. In general, the materials with high permeability and low hysteresis loss are pure materials.

The composition of a ferromagnetic material also influences its magnetic properties. The addition of silicon to iron increases the permeability and reduces the hysteresis loss. However, high concentrations of silicon decrease the saturation magnetization. Therefore, the iron-silicon alloys with low concentrations of silicon are desirable in applications like the cores of transformers and in electric motors and generators.

The iron-nickel alloys possess high values of initial and maximum permeability and very low hysteresis loss compared to the iron-silicon alloys. An alloy of composition with 78.5% nickel and 21.5% iron is called *permalloy* and has an initial relative permeability of $\sim$10,000 compared to 250 for the pure iron. These alloys are in general used for the magnetic screening of the electronic equipment. Small additions of the metals chromium or molybdenum further modify the magnetic properties of these materials to be used as cores in transformers or inductors working at the audio or higher frequencies. For example, the magnetic cores of inductors and transformers working at radiofrequencies ($\sim$100 Mc/sec) show high eddy current losses. Used here are the ferrites, which have high resistivity ($\sim 10^6$ times that of metals) and high permeability. In other applications, the ferromagnetic materials with a high value of "coercivity" and large area of hysteresis loop are required. These materials possess hard magnetism compared to the already discussed soft magnetism. The addition of carbon to iron increases the hysteresis loss. The carbon steel was used as a material for permanent magnets in earlier days. However, aging degenerates the magnetic properties of carbon-steel magnets. Addition of metals such as cobalt, chromium, or tungsten improves the magnetic properties, and these materials are less susceptible to aging. A large number of alloys composed of iron, nickel, cobalt, aluminum, copper, platinum, manganese, and oxides of iron and rare-earth metals have been developed that show high values of coercivity and are suitable for permanent magnets.

Some oxides like $\gamma$-$Fe_2O_3$ and $CrO_2$ possess high coercivity and are used as recording tapes in the form of this layers of fine powders. Permalloys are used to construct the inductive magnetic heads to write signals as residual magnetization on the tapes or to reproduce the electrical signals from the magnetized tapes. Magnetoresistive read heads based on permalloy are also used. Significant progress has been made in the development of very sensitive and small read heads based on magnetic-nonmagnetic multilayer systems. Magnetic discs or drums are made for the memory systems in computers. Development of modifying the magnetic properties by rapid solidification of alloys from the melt has created a new field. By rapid solidification, the microstructure can be affected—in some cases the phases may be finely dispersed, and in other the alloys may become noncrystalline or amorphous. The amorphous alloys have no crystal lattice and no magnetic anisotropy.

There are no extended defects that would otherwise interact strongly with the domain walls in these noncrystalline materials. In certain cobalt-based noncrystalline alloys, the magnetostriction can be adjusted to zero such that the internal and applied stresses have minimal effect

on the magnetic properties. Amorphous magnetic alloys have high hardness and yield strength and are magnetically soft. In particular, cobalt-containing alloys have vanishingly small magnetostriction. Combining the good mechanical and magnetic properties, they are very useful materials. They can be strained elastically over wide limits, are insensitive to irreversible magnetic damage, and are very suitable materials where the elastic deformability is desired. They are also useful materials for making recording heads in audio, video, and data recording systems, due to their good high-frequency response and wear resistance.

Soft magnetic and highly elastic mechanical behavior has led to the development of flexible magnetic shielding.

The amorphous magnetic materials consist of two main groups. In one group, the materials are composed of transition metals and metalloids; in the second group, they are composed of only different metals. Some amorphous materials used as soft magnetic materials are $Fe_{81}(Si, B, C)_{19}$; $(FeNi)_{78}(Mo, Si, B)_{22}$; $(Co, Fe)_{70 \ldots 76}(Mo, Si, B)_{30 \ldots 24}$; and $(Co, Mn)_{70 \ldots 76}(Mo, Si, B)_{30 \ldots} 24$.

Hard magnetic materials can also be produced by rapid solidification techniques. For example, a magnetic material of composition $Fe_{14}Nd_2B$ produced by rapid solidification is superior to the Co-Sm material. Other hard magnetic materials (some transition and rare-earth metals, and boron), have also been produced by rapid cooling. The rapid cooling technique is also less expensive compared to the conventional methods.

## SEE ALSO THE FOLLOWING ARTICLES

CRYSTALLOGRAPHY • ELECTROMAGNETISM • GEOMAGNETISM • MAGNETIC MATERIALS • SUPERCONDUCTIVITY • TRANSFORMERS, ELECTRICAL

## BIBLIOGRAPHY

Baibich, M. N., *et al.* (1988). *Phys. Rev. Lett.* **61,** 2472.
Binasch, G., Grünberg, P., Saurenbach, F., and Zinn, W. (1989). *Phys. Rev.* **B39,** 4282.
Brailsford, F. (1968). "An Introduction to the Magnetic Properties of Materials," Longmans Green and Co., London.
Chikazumi, S. (1964). "Physics of Magnetism," John Wiley and Sons, New York.
Craik, D. J., and Tebble, R. S. (1966). "Ferromagnetism and Ferromagnetic Domains," North-Holland, Amsterdam.
Della Torre, E., and Bobeck, A. H. (1974). "Magnetic Bubbles," North-Holland, Amsterdam.
Kittel, C. (1979). "Introduction to Solid State Physics," John Wiley and Sons, New York.
Morrish, A. H. (1965). "Physical Principles of Magnetism," John Wiley and Sons, New York.
Standley, K. J. (1972). "Oxide Magnetic Materials," 2nd ed., Oxford University Press, London.
Steeb, S., and Warlimont, H., eds. (1985). "Rapidly Quenched Metals," Vol. 11, North-Holland, Amsterdam.
Vonsovski, S. V. (1975). "Magnetism," Halsted Press, New York.
Wohlfarth, E. P. (1980–1982). "Ferromagnetic Materials," Vol. I (1980), Vol. II (1980), Vol. III (1982), North-Holland, Amsterdam.
Zeiger, H. J. (1973). "Magnetic Interaction in Solids," Oxford University Press, London.



# Gamma-Ray Spectroscopy

**R. F. Casten**
**C. W. Beausang**
*WNSL, Yale University*

I. Introduction
II. Gamma-Ray Detection
III. Gamma-Ray Spectroscopy and Nuclear Structure
IV. Conclusions

## GLOSSARY

**Detector efficiency** Loosely defined as the probability of detecting the full energy of a gamma-ray.
**Detector resolution** Refers to the width of the full energy gamma-ray peak measured in the detector. Typically the resolution is ∼2 keV for semiconductor and 10–20 keV for scintillator detectors for a 1000-keV gamma-ray.
**Doppler shift** The shift in frequency or energy of waves emitted from a moving source.
**Nuclear level scheme** A graph of the excited energy levels of a nucleus and their connecting gamma-ray transitions. The levels are usually labeled by their angular momentum and parity quantum numbers.
**Nucleons** The protons and neutrons that make up the nucleus.
**Pauli exclusion principle** Fundamental principle of quantum mechanics. It states that for certain types of elementary particles, including electrons, protons, and neutrons, no two identical particles can be in the same quantum state.
**Potential energy surface** A contour plot of the potential energy of the nucleus as a function of deformation. Stable deformations correspond to minima in the potential energy.
**Scintillator detector** A material, liquid or solid, that converts the energy lost by a gamma-ray into pulses of light.
**Semiconductor detector** Essentially a large diode usually constructed out of either silicon or germanium.
**Spin** Angular momentum.

## I. INTRODUCTION

**THE NUCLEUS** is a unique, strongly interacting, quantum mechanical system. Consisting of a few to a few hundred protons and neutrons, its structure combines the macroscopic features expected of bulk nuclear matter (shape, size, etc.) with the microscopic properties associated with the motion of a finite number of nucleons in a potential.

Atomic nuclei studied in the laboratory (whether they are produced in reactions or populated via radioactive decay) are often found in excited states. Since any physical

system seeks its lowest possible energy level, such excited nuclear configurations are unstable. They generally de-excite to the nuclear ground state in time scales of $10^{-15}$ to $10^{-6}$ sec by the emission of one or more gamma-rays. Hence, the study of the gamma-rays emitted from excited nuclei provides a means of studying the levels, de-excitation rates, and structure of these objects.

The study of the decay properties of the atomic nucleus has provided an enormous quantity of information on the behavior of such systems when stressed by the application of high temperature, high angular momentum, large deformation or by large isospin values (isospin is a quantum number which basically counts the difference in the numbers of protons and neutrons in a nucleus).

In this article we will touch upon some of these topics while attempting to give a flavor of the field of nuclear gamma-ray spectroscopy and charting some possible future directions. We begin by introducing some of the detector types used to detect gamma-rays and briefly discuss some of the design criteria for modern gamma-ray spectrometers. This is followed by a discussion of some features found in excited nuclear states, broadly separated into low-spin and high-spin properties, and chosen to illustrate the variety of macroscopic and microscopic features of the nuclear system. To discuss or even list the enormous number of practical applications of gamma-ray spectroscopy in medicine, in industry (e.g., the oil industry), in other sciences such as archeology and astronomy, and in the areas of security and defense is far beyond the possible scope of this article.

## II. GAMMA-RAY DETECTION

In this section we discuss the mechanisms by which gamma-rays interact with matter (i.e., detectors), the different types of detectors and detector systems, and the criteria that go into the design and choice of particular systems. To facilitate this discussion, a simplified example of a nuclear level scheme is shown in Fig. 1.

### A. Interaction Mechanisms

For energies ranging from a few kilo-electron volts to a few mega-electron volts, gamma-rays interact with matter via one of three principal mechanisms: the photoelectric effect, Compton scattering, or for energies above ~1 MeV, the electron-positron pair production. Most gamma-ray detectors exploit one or more of these effects both to detect the gamma-ray and to measure its energy. Of course, gamma-rays are electromagnetic waves and sometimes their wave properties are also used in their measurement, for example, with diffraction techniques. Before we discuss the design of gamma-ray detectors and spectrometers, we first briefly describe these interaction mechanisms. The relative probability for each mechanism is shown schematically in Fig. 2 as a function of gamma-ray energy.

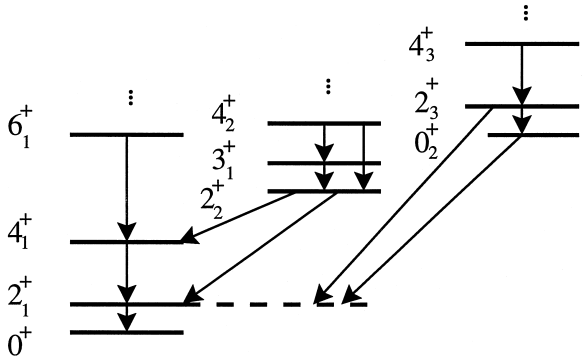



**FIGURE 1** A simplified nuclear level scheme showing some of the levels and gamma-ray transitions that might be observed in a typical heavy-ion fusion-evaporation reaction. The levels are labeled by their angular momentum and parity quantum numbers.

The photoelectric effect is the dominant interaction mechanism for low gamma-ray energies, below a few hundred kilo-electron volts. In this case, the gamma-ray interacts with an atomic electron somewhere in the bulk of the detector material. The gamma-ray energy is transferred to the electron, which is ejected from the atom with energy $E_e = E_\gamma - E_{BE}$, where $E_{BE}$ is the electron binding energy. The probability for the photoelectric effect interaction increases very rapidly with the atomic number ($Z$) of the material. This is why high-$Z$ materials are favored both for gamma-ray detectors and for absorbers and shields.

The Compton scattering mechanism is similar to the photoelectric effect in that the gamma-ray also interacts with an atomic electron in the detector material. In this case, however, the initial gamma-ray energy is shared between the electron and a scattered (lower energy)

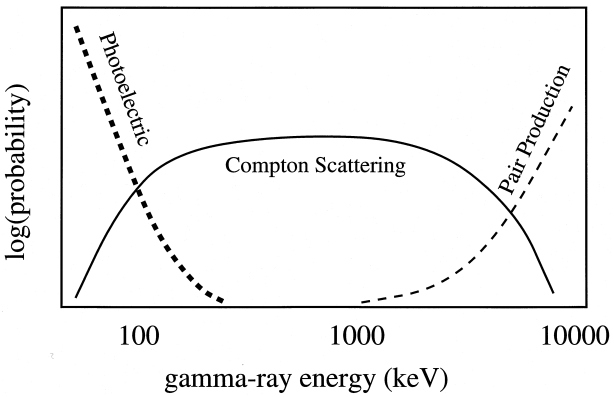



**FIGURE 2** Schematic diagram showing the relative probabilities for photoelectric, Compton scattering, and pair production as a function of gamma-ray energy.

gamma-ray. Compton scattering is the dominant interaction mechanism for gamma-ray energies in the range from a few hundred kilo-electron volts up to a few mega-electron volts. It is important to realize that this is the energy range of most gamma-rays produced in typical nuclear structure experiments.

For gamma-ray energies greater than $2m_ec^2$ ($m_ec^2$ being the rest mass of the electron $\sim$511 keV), electron-positron pair production is possible, and, for energies significantly higher than this threshold, pair production begins to dominate the interaction cross section. In this case, some of the incident gamma-ray energy is used to create the electron-positron pair while the remainder (in excess of $2m_ec^2$) is shared as kinetic energy between the electron and positron. Eventually, the positron annihilates with another electron in the detector medium, producing two photons each of energy 511 keV emitted back to back.

As one can see, all three of these interaction mechanisms result in the production of a single energetic electron (or, in the case of pair production, of an electron-positron pair) with a kinetic energy less than or equal to that of the incident gamma-ray energy. These energetic electrons recoil through the bulk material of the detector, their range being typically less than a millimeter or two. They rapidly slow down, losing energy through many collisions with other atomic electrons. In an ideal detector all of the incident gamma-ray energy is eventually absorbed in the detector material by a combination of photoelectric, Compton scattering, and (for high enough gamma-ray energies) pair-production processes.

It is intuitively obvious that the number of collisions, and hence the number of secondary electronic excitations produced, is proportional to the primary electron energy and, hence, is directly related to the incident gamma-ray energy. For example, in semiconductor detectors, which are essentially diodes made out of germanium (Ge) or silicon (Si), the electron-hole pairs produced following electron–electron collisions are extracted by a high voltage placed across the detector and produce a current pulse which is proportional to the deposited gamma-ray energy. For scintillation detectors, such as sodium iodide (NaI(Tl)) detectors, the collisions of the primary electron produce excited atomic or molecular states. The subsequent decay of these states produces scintillation photons (typically in the UV range). These photons are converted into a current pulse using a photocathode and photomultiplier tube. The size of the current pulse is again proportional to the deposited gamma-ray energy.

## B. Gamma-Ray Detectors

Different types of detectors have quite different efficiencies and energy resolutions. Indeed, generally speaking, gamma-ray spectroscopy is a constant trade-off between these two properties. For example, scintillation detectors, such as NaI(Tl) detectors, typically have high efficiencies but poor energy resolution compared to Ge detectors. Detectors based on the technique of crystal diffraction (see below) have superb energy resolution but very small efficiency.

### 1. Scintillation Detectors

The detection of gamma-rays (or other types of ionizing radiation) by the scintillation light produced in certain materials is one of the oldest techniques on record, and it is still one of the most useful and common techniques today.

A scintillator, either a solid or a liquid, is a material which converts the energy lost by the gamma-ray into pulses of light. The scintillation light is detected in turn by a light-sensitive material which usually forms the cathode of a photomultiplier tube. The light pulses are converted into electrons in the photocathode. These electrons are then accelerated and their number vastly (and linearly) amplified in the photomultiplier. The resulting current pulse is proportional to the energy of the absorbed gamma-ray.

The energy required to produce a light pulse is fairly large, on the order of 30–50 eV. Thus, the average number of light pulses produced when, say, a 500-keV gamma-ray is absorbed is on the order of 10,000. Fluctuations in this number and in the light collection process limit the resolution obtainable with scintillation detectors.

### 2. Semiconductor Detectors

A semiconductor detector is essentially a large diode constructed out of either Si or Ge. For gamma-ray spectroscopy, Ge detectors are preferred, as Ge has a larger stopping power. The diode is operated under a reverse bias and is normally fully depleted (i.e., with no free charge carriers). The gamma-ray interaction produces electron-hole pairs in the depletion region, which are collected because of the detector bias voltage and which produce a current pulse proportional to the absorbed gamma-ray energy.

In contrast to scintillator detectors, the average energy required to produce a single electron-hole pair is only about 2–3 eV. Therefore, a 500-keV gamma-ray can produce around 250,000 primary charge carriers, much larger than the corresponding number for scintillation detectors with a corresponding decrease in the statistical fluctuations and improvement in detector resolution. Figure 3a shows a typical spectrum of a $^{60}$Co source obtained using a modern Ge detector. The energy resolution obtained is about 2 keV for gamma-ray energies of about 1000 keV. Using a NaI(Tl) scintillation detector, the two peaks in

Fig. 3a (having energies of 1174 and 1332 keV) would be barely resolved. Following a brief discussion of crystal diffraction detectors, we will focus most of the remainder of this article on gamma-ray spectroscopy using Ge detectors.

### 3. Ultra-High Energy Resolution Spectroscopy: Crystal Diffraction

Gamma-rays are, after all, electromagnetic photons, and under certain circumstances their wave properties may be used in their detection and measurement. Indeed, the ultimate in current gamma-ray energy resolution and measurement precision is obtained by the use of crystal diffraction. With such techniques it is routine to measure a 1-MeV gamma-ray with an energy resolution of ~3 eV and an energy precision of better than 1 eV. The cost, however, is low efficiency and the need to scan the energy spectrum one small energy bite at a time.

The technique uses Bragg diffraction from a nearly perfect crystal, usually of Si. As for optical and X-ray transitions, the gamma-ray wavelength $\lambda$ and diffraction angle $\theta$ are related by the Bragg law: $n\lambda = 2d \sin\theta$, where the lattice spacing $d$ is known to an accuracy of 1 part in $10^{10}$ and $n$ is the order of diffraction. Clearly, the resolution scales with $n$. Higher order diffraction gives greater dispersion and, hence, energy precision, although the efficiency generally falls off with $n$. The accuracy depends on the precision of the angle measurement. In the realization of this technique at the Institut Laue Langevin in Grenoble, in the GAMS (GAMma-ray Spectrometer) family of instruments, accuracies of the latter are typically in the milli-arc seconds range (Koch *et al.*, 1980). Given the nature of the technique, the gamma-ray energy spectrum is stored energy interval by interval rather than sampled fully at each point in time. An example of a crystal diffraction spectrum compared to the corresponding Ge detector spectrum is shown in Fig. 3b. Generally speaking, crystal spectrometers offer the greatest advantages over Ge detectors for gamma-ray energies below ~1 MeV. At higher energies their efficiency drops quickly, and hence, lower orders of diffraction are used with poorer energy resolution.

## C. Level Scheme Construction

One might ask why the extraordinary energy precision of crystal diffraction techniques can be useful since nuclear models seldom predict nuclear states to accuracies better

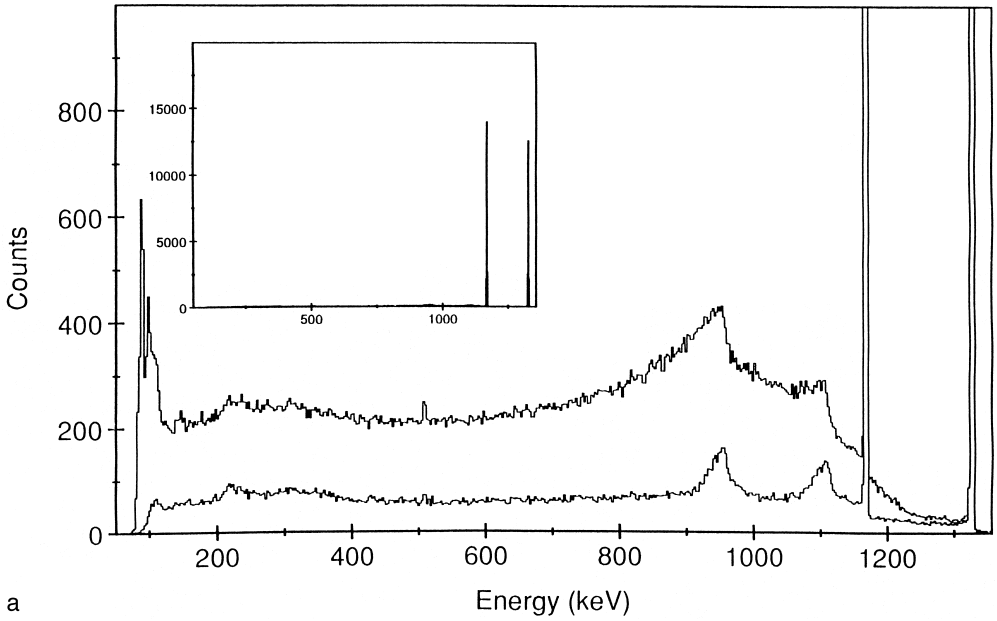
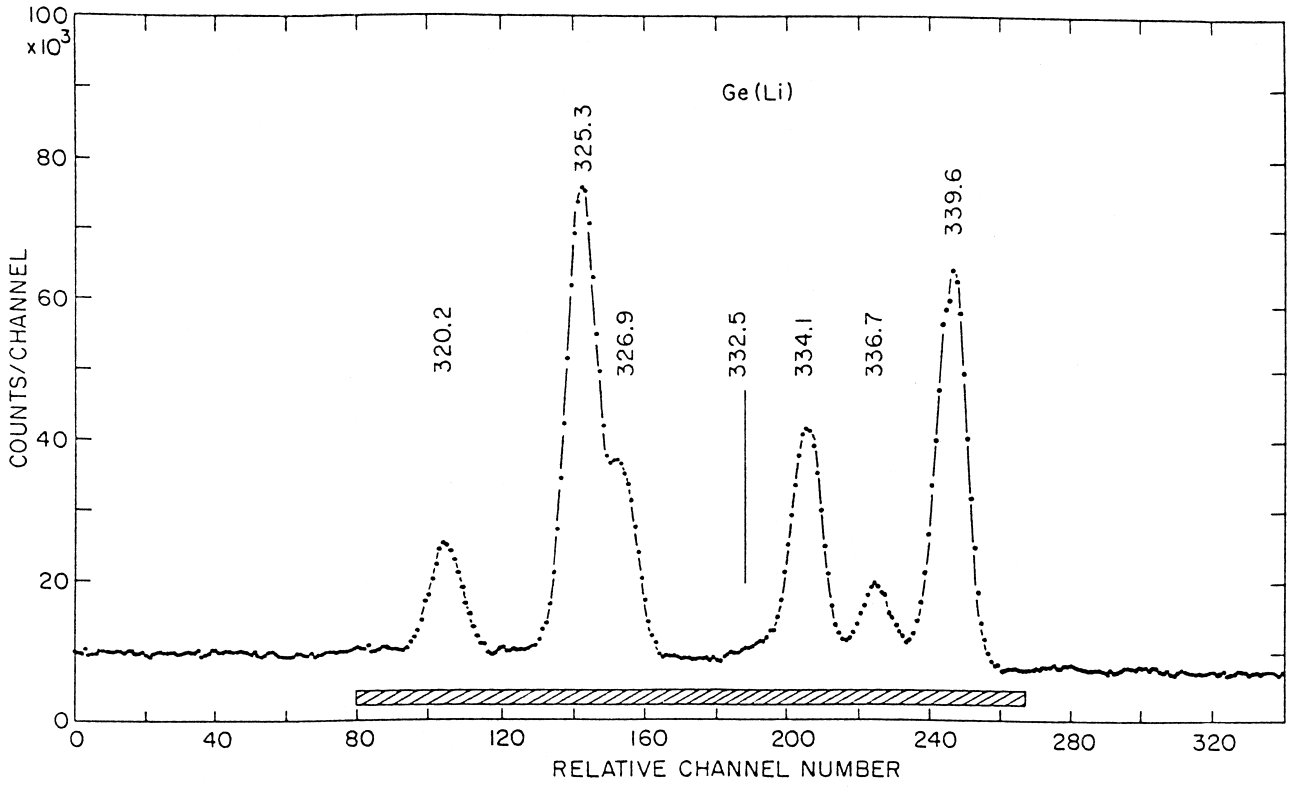
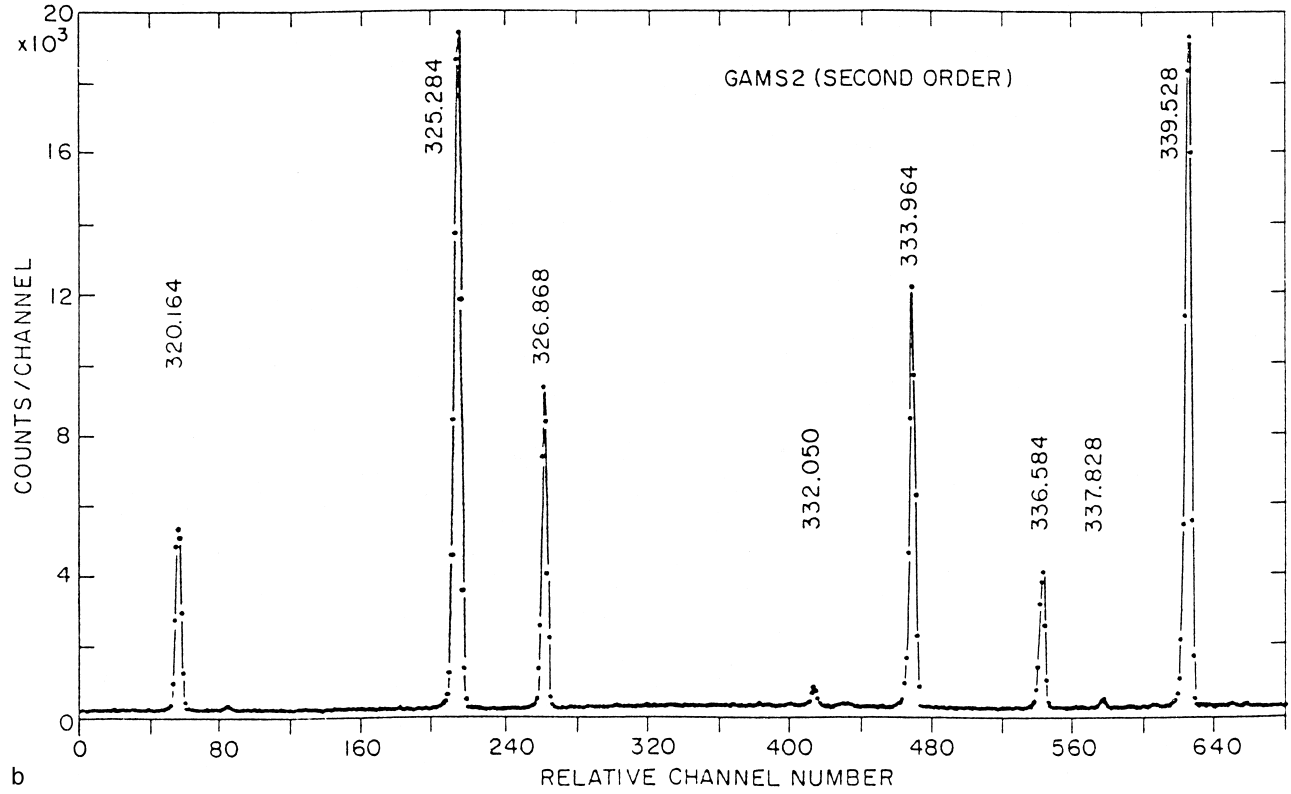



**FIGURE 3** (a) Typical spectrum of a $^{60}$Co source obtained using a modern Ge detector with and without escape suppression. The vertical scale has been greatly expanded in order to show the Compton background. The dramatic reduction in the height of the background when using a Compton suppression shield is obvious. The insert shows the same spectra but now with the full vertical scale to illustrate the height of the photopeaks compared to the background. (b) An example of the very high energy resolution obtainable using a crystal diffraction system. The top spectrum, obtained using a Ge detector, shows seven peaks, some of which are not resolved. The lower spectrum, obtained using the GAMS crystal diffraction system, shows the same portion of the spectrum, but with a dramatic improvement in resolution. In both spectra, the gamma-rays are labeled with their energies in kilo-electron volts. *Continued.*




b

**FIGURE 3** *Continued.*

than many kilo-electron volts? The principal reason relates to the construction of reliable level schemes. One method of constructing level schemes utilizes the Ritz Combination Principle. Here, nuclear level energies are determined by demanding that their energy differences be equal to the energy of the gamma-ray transition connecting them (see Fig. 1). A brief discussion of this technique enables us to see both the role of the ultra-high energy resolution in level scheme construction and the complementary and much more commonly used technique of coincidence spectroscopy with Ge detectors.

To see the point, imagine constructing a level scheme which consists of 50 levels spanning the excitation energy range from zero up to 2.0 MeV simply by using the Ritz Combination Principle in a case where one has detected, say, 500 $\gamma$-rays with energies below 1.2 MeV, with Ge detector energy accuracy of $\pm 0.1$ keV. This is a typical situation encountered in the spectroscopy of low-spin states of heavy nuclei. In such a case the probability of an accidental Ritz Combination, that is, a level energy difference that inadvertently coincides within uncertainties with a gamma-ray energy, is about 10%. Even with other experimental input, such as information on the angular momenta of the levels to rule out certain transition placements from conservation of angular momentum, it is clear that a large number of incorrect placements and, hence, incorrect physics will result. There are two ways of resolving this situation: either using time coincidence relations between successive gamma-rays to place them correctly in the level scheme of a nucleus or improving the energy resolution significantly. For the latter, the crystal diffraction approach is ideal. With an energy precision of, say, $\pm 5$eV, the probability of an accidental sum drops to negligible levels. Indeed, data from the GAMS spectrometers at the ILL have often shown that existing Ge detector results (usually data taken without coincidences) are in error.

However, the usual solution to this problem is the use of coincidence spectroscopy. This technique exploits the fact that nuclear levels are generally short lived, with typical half-lives in the pico- to nanosecond range, so that successive gamma-ray de-excitations effectively occur simultaneously, on the time scale of standard pulse analysis electronics. Therefore, if two or more gamma-rays are observed in separate detectors, in time coincidence (within say a few nenoseconds) then they must occur in a cascade in the nuclear level scheme. For example, in Fig. 1 the $6_1^+ \rightarrow 4_1^+$ and $4_1^+ \rightarrow 2_1^+$ transition would be in coincidence, as would the $3_1^+ \rightarrow 2_2^+$ and $2_2^+ \rightarrow 4_1^+$ or $2_2^+ \rightarrow 2_1^+$ transitions. However, the $3_1^+ \rightarrow 2_2^+$ transition is not in coincidence with the $6_1^+ \rightarrow 4_1^+$ transition. Such coincidence relations are of inestimable help in constructing complex nuclear level schemes.

While coincidence spectroscopy is a powerful technique, it also has limitations. For example, it does not help place ground state transitions or very weak transitions, and, in some cases, experimental constraints preclude its use. Nevertheless, it is, by far, the most common approach to sorting out the plethora of gamma-rays observed in nuclear de-excitation. We will further discuss coincidence spectroscopy below when we introduce advanced multidetector arrays of Ge detectors.

We briefly mention that another application of ultra-high resolution crystal spectroscopy is in the determination of fundamental constants such as the accepted standard for length measurements (the definition of the meter) through the precise measurement of gamma-ray wavelengths. These applications fall outside the scope of this article.

## D. The Evolution of Detector Arrays

To illustrate the increasing power and sophistication of gamma-ray detectors, particularly of Ge detector arrays, it is useful to consider the nuclear reactions by which the nuclei to be studied are formed. One of the most common reaction mechanisms used to populate high-spin states in atomic nuclei is the heavy-ion fusion-evaporation reaction. This type of reaction has the advantage of bringing large quantities of angular momentum into the product nucleus (often up to the limit allowed by fission), while at the same time populating only a few product nuclei with significant probability.

In such reactions, a heavy-ion beam is incident on a target at an energy just above the Coulomb barrier. A typical reaction might involve a $^{48}$Ca ($Z = 20$, $N = 28$) beam incident on a $^{108}$Pd ($Z = 46$, $N = 62$) target at a beam energy of 200 MeV. Following the collision, the beam and target nuclei fuse to form a compound nucleus, in this case $^{156}$Dy ($Z = 66 = 20 + 46$, $N = 90 = 28 + 62$). The compound nuclear system will be produced in a highly excited and rapidly rotating state, with typically 60 MeV of excitation energy and about $70\hbar$ of angular momentum.

The initial decay of the compound system is via the emission of a few (3–5) particles, usually neutrons and less frequently protons or alpha particles. This first stage of the decay process typically removes about 40 MeV of excitation energy and about $10\hbar$ of angular momentum. The remainder of the excitation energy and most of the angular momentum is subsequently removed by gamma-ray emission. Each gamma-ray photon removes either one or two units of angular momentum. Thus, we can expect the emission of cascades of up to 30 gamma-rays following each reaction. Because of this multiplicity of gamma-rays, the study of transitions following production of the compound nucleus imposes stringent requirements on detector

systems. Other reactions used, such as Coulomb excitation, $(n, \gamma)$, or $\beta$-$\gamma$ decay, generally present simpler experimental challenges. The driving force in the impressive developments in gamma-ray detector systems in the last 40 years has been the requirements imposed by the fusion evaporation reaction studies.

Information about the changes in nuclear structure during the decay, as the nucleus loses energy and angular momentum, is obtained by measuring the properties of the gamma-rays in these cascades, such as the gamma-ray energy, angular distribution, linear polarization, emission sequence, etc.

The evolution of gamma-ray spectroscopy with time over the past 40 years or so is illustrated in Fig. 4, which plots the population intensity of various nuclear states as a function of angular momentum or spin of the state. Pioneering experiments in the early 1960s were carried out with one or a few NaI(Tl) scintillation detectors (Morinaga and Gugelot, 1963). The sensitivity of these experiments was limited, both by the poor energy resolution of NaI(Tl) detectors (about 80 keV at 1000 keV) and by the small number and size of the detectors, to spins up to about spin 8–10 $\hbar$ and to states populated with about 10% the intensity of the strongest transition.

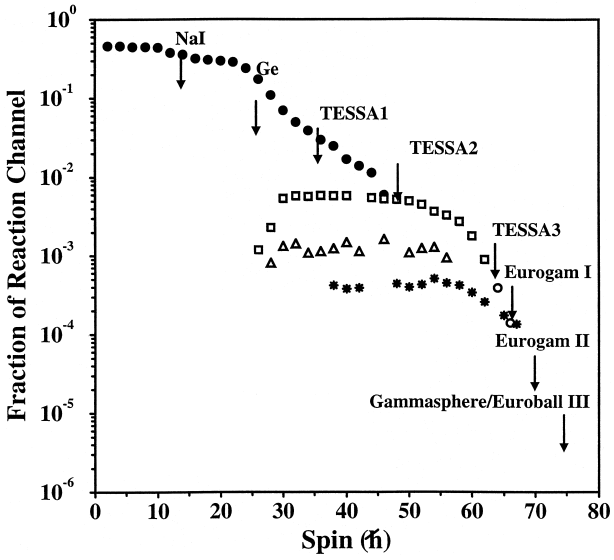



**FIGURE 4** A schematic diagram illustrating the evolution of gamma-ray spectroscopy. The various symbols plot the measured intensity of various nuclear states vs angular momentum, giving an indication of the sensitivity of various detector systems. Early experiments using NaI(Tl) and a few Ge detectors were sensitive to excited states which were populated with intensities down to about one-tenth of the reaction channel (solid symbols). As time went on, more sensitive arrays were developed. The current generation of arrays, the Gammasphere and Euroball arrays, are capable of observing excited states populated with a fraction as small as $10^{-6}$ of the reaction channel. The open symbols and stars plot the intensity of various superdeformed bands as a function of angular momentum.

The introduction of reversed bias, lithium drifted, Ge detectors in the mid 1960s led to a major increase in sensitivity and major breakthroughs in our physics knowledge. Germanium detectors have very good energy resolution, about 1 keV for $E_\gamma \sim 100$ keV and 2 keV for $E_\gamma \sim$ 1000 keV. On the other hand, the detection efficiency of early Ge detectors was often much lower than NaI(Tl) detectors. To compensate for the lower efficiency, and also to measure the time coincidence relationships of successive gamma-rays in a cascade, experiments with more than one Ge detector were soon commonplace. The phenomenon of backbending, at spin $\sim$15 $\hbar$ (see below), was discovered by Johnson, Ryde, and Sztarkier (1971) using just two Ge(Li) detectors, while the structure of $^{160,161}$Yb was investigated by Riedinger *et al.* (1980). up to spin $\sim$30 $\hbar$ using only four Ge detectors. In the last three decades, the study of the properties of the atomic nucleus through gamma-ray spectroscopy has evolved through the development of larger and more efficient Ge detector arrays and, indeed, has driven the development of these arrays.

Starting in the 1980s and continuing to today, large arrays of Ge detectors such as TESSA, GASP, Eurogam, Gammasphere (Lee, 1990) and Euroball (Gerl and Lieder, 1992) further revolutionized gamma-ray spectroscopy. Future arrays such as the proposed GRETA (Gamma-Ray Energy Tracing Array) spectrometer (Deleplanque *et al.*, 1999), which promise very large increases in sensitivity resulting from modern manufacturing techniques, electronics, and digital data processing, are in the planning stages.

## E. Germanium Detector Performance

### 1. Peak-to-Total Ratio and Escape Suppression

A major problem encountered with early Ge detectors, and still a problem today, is the poor peak-to-background ratio in the spectrum. The background, clearly seen in Fig. 3a, is caused by incomplete energy collection in the Ge detector occurring when a Compton scattered gamma-ray leaves the active bulk of the detector before being absorbed.

Even with today's large volume Ge detectors, irradiation with a standard $^{60}$Co source (which emits two gamma-rays with energies of 1174 and 1332 keV) yields a spectrum where only $\sim$25% of the events lie in the full energy photopeaks. This number, the ratio of the number of counts in the photopeak(s) to the total number of counts in the spectrum, is termed the peak-to-total ratio (PT). For PT = 0.25, the remaining 75% of the events in the detector form a continuous background extending to lower energies.

The preferred solution is to detect these scattered photons in a second, surrounding detector, termed an escape suppression or an anti-Compton shield, and to reject, using fast electronics, coincidence events between the Ge detector and the shield detector. The combination of Ge detector and suppression shield is termed an escape suppressed spectrometer (ESS). The material commonly used in the anti-Compton shield is bismuth germanate (BGO), a dense, high-efficiency scintillator material.

After suppression, typically about 65% of the remaining events are in the photopeaks (PT = 0.65). A typical ESS configuration showing a Ge detector and shield is shown in Fig. 5, while the improvement in the background, and hence spectrum quality, is illustrated in Fig. 3a.

The PT ratio is of prime importance for coincidence spectroscopy. For example, when requiring a coincidence between two Ge detectors, a PT ratio of 0.25 implies that only $(0.25)^2$ or ~6% of the events will be photopeak-photopeak coincidences. The remaining 94% will be background events. Using an ESS, however, the photopeak-photopeak coincidence fraction increases to $(0.65)^2$ or 42%, an improvement of a factor of 7. Even larger improvements are obtained when three- or higher fold coincidence events are recorded. For example, the improvement is a factor of 17 for triples coincidences, 45 for quadruples, and 120 for quintuples. Today's largest gamma-ray spectrometers, the Gammasphere array in the United States and the Euroball array in Europe, regularly record even higher fold coincidence events (Lee, 1990; Gerl and Lieder, 1992).

The efficiency and sensitivity of ESS arrays improved rapidly, so that by the mid 1980s arrays with more than 20 ESS having total absolute peak efficiencies of up to 1% were constructed. By convention, the total photopeak efficiency is defined as the probability of measuring the full energy of the 1332-keV $^{60}$Co gamma-ray when the source is placed at the center of the array. These ESS arrays enabled nuclear phenomena that occur at an intensity of about 1% of the total intensity of the nucleus to be studied. Worldwide there were about a dozen arrays with this level of sensitivity. One of the earliest of these, the TESSA3 array (Nolan, Gifford, and Twin, 1985), located at Daresbury Laboratory in the United Kingdom, was used in the discovery of the classic discrete line superdeformed band in $^{152}$Dy. Superdeformation will be discussed further below.

In the mid 1990s the latest generation of gamma-ray spectrometers with total photopeak efficiencies of up to ~10% came on line. These spectrometers, namely the Gammasphere (Lee, 1990) and Eurogam/Euroball (Gerl and Lieder, 1992; Beausang *et al.*, 1992) arrays, contain up to 240 individual Ge elements and have sensitivities of better than 0.001% of the production cross section. Some of the detectors in these arrays are composites formed by closely packing several Ge detectors together as a unit. Two varieties of such units, called clover (Duchene *et al.*, 1999) or cluster (Eberth *et al.*, 1996) detectors, are nowadays the backbone of advanced arrays such as the YRAST Ball array at Yale University (Beausang *et al.*, 2000) or the planned Exogam and Miniball arrays in Europe (Simpson *et al.*, 2000). Even more powerful detectors, termed tracking detectors, are under development. These will be discussed below.

## 2. Counting Rates

It is informative to look at some of the numbers involved in a typical nuclear physics reaction carried out in the laboratory. Once again, we consider the example of the $^{48}$Ca + $^{108}$Pd reaction, which was used in the experiment in which the first superdeformed band in $^{152}$Dy was discovered (Twin *et al.*, 1986).

Typically, the beam intensity from an accelerator is about $10^{10}$–$10^{11}$ particles per second incident on a target. This corresponds to an electric current on the order of a few nano-Amperes (nA). About one beam particle in a million will actually strike a target nucleus and induce a nuclear reaction. Therefore, we expect about 100,000 reactions per second. About 20% of the reactions produce $^{152}$Dy and about 1% of these will populate the nucleus in the superdeformed state, corresponding to about 200 such events per second.

The array used in the original discovery of the superdeformed band in $^{152}$Dy, the TESSA3 array (Nolan, Gifford, and Twin, 1985), had a total photopeak efficiency of about 0.5%. Assuming that each superdeformed nucleus decays by emitting a cascade of ~25 gamma-rays, and that we require a coincidence between two detectors ($\gamma^2$) before accepting an event, we might expect to detect about 1 gamma-ray coincidence event per second originating from a superdeformed cascade. Since each cascade is ~25 transitions long, we expect about 1 count per gamma-ray transition every 20 sec or so. The background rate from other processes is many hundreds of times greater.

## 3. Doppler Effects and Segmentation

The lifetimes of the highest spin states populated via heavy-ion fusion-evaporation reactions are often comparable to, or shorter than, the stopping time of the recoiling nucleus (recoiling due to the momentum imparted by the incident beam nucleus that initiates the reaction) in the target material. Typical recoil velocities are on the order of a few percent the speed of light. Therefore, Doppler effects play a major role.

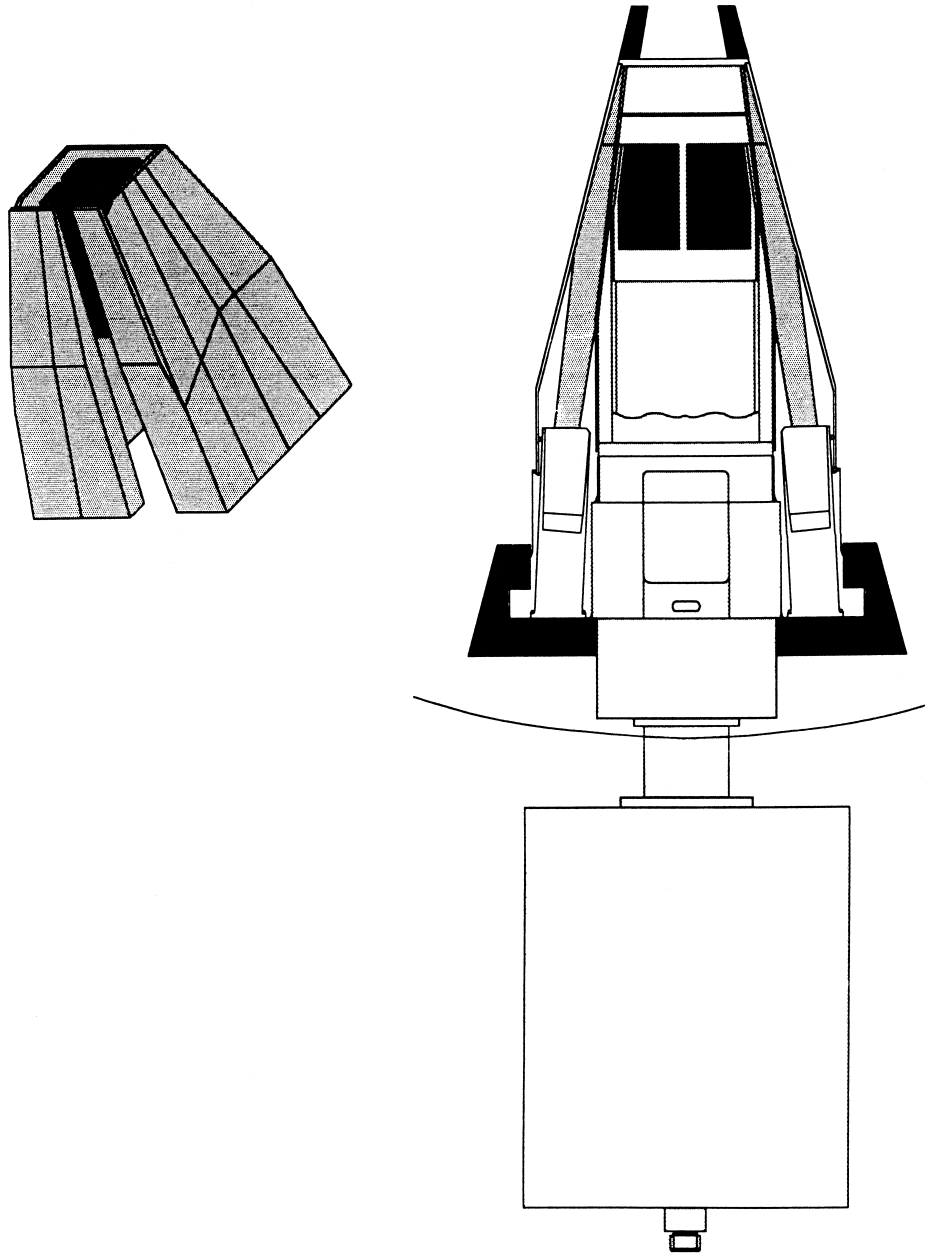


**FIGURE 5** Escape suppression spectrometer showing Ge detector and shield. This is the type of ESS used for clover Ge detectors in the Eurogam/Euroball array. The clover Ge detector position is indicated inside the suppression shield. The liquid nitrogen storage dewar is also shown.

The Doppler shifted energy of a gamma-ray emitted from a nucleus in flight is given by

$$E_\gamma = E_0\left[1 + \frac{v}{c}\cos\theta\right],$$

where $E_0$ is the unshifted energy and $\theta$ is the detector angle. If a detector records the same gamma-ray emitted from different nuclei having a wide range of velocities or traveling at different angles with respect to the beam direction, the resulting energy resolution can be very poor.

One solution is to use very thin targets in order to minimize slowing down effects and detectors that subtend only a small range of angles. Knowing the detector angles, one can correct for the Doppler shift and recover most of the resolution. The limit on detector resolution now becomes the finite opening angle of the Ge detector itself, in other words the uncertainty in knowing in which part of the Ge

detector the gamma-ray actually interacted. For a constant recoil velocity, the Doppler broadening is given by

$$dE = E_0 \frac{v}{c} \sin\theta \, d\theta,$$

where $d\theta$ is the opening angle of the Ge detector, typically about 5–10°. For experiments on very high spin nuclear states, the energy resolution is dominated by Doppler broadening effects and is often a factor of two or more worse than the intrinsic resolution of the Ge detector. For very high recoil velocities the problem is much worse. Because of the $\sin\theta$ dependence, the Doppler broadening is worst for detectors placed at $\theta = 90°$ to the beam direction (even though the Doppler shift is zero at 90°).

Various methods have been developed to minimize Doppler broadening effects. Most involve the concept of detector segmentation in which one determines in what part of a detector a given photon was detected. The development of the clover detector, for example, with four separate Ge detectors closely packaged in a single vacuum vessel, was driven by such concerns (Duchene *et al.*, 1999). A schematic diagram of a clover Ge detector is shown in Fig. 6. The idea is that by using four small detectors, one effectively has a much larger detector while preserving the smaller opening angle for each individual segment. Gamma-rays may interact in only a single element of the clover detector. In this case one takes the angle $\theta$ to be the center of this element and $d\theta$ is the opening angle of this segment. A gamma-ray may also scatter between two elements of the clover detector. In this case simulations and measurements have shown that the gamma-ray interaction usually takes place close to the boundary between the two crystals. Thus, one is justified in taking $\theta$ as the angle of the boundary. The other enormous advantage of the clover detector is that the energy is measured accurately, even for such scattering events. The energies measured in each separate crystal may be added together while preserving the good energy resolution of the individual crystals. Because of this add-back feature, the efficiency of a clover detector, consisting of four individual Ge crystals, is actually about six times the efficiency of the individual detector crystals.

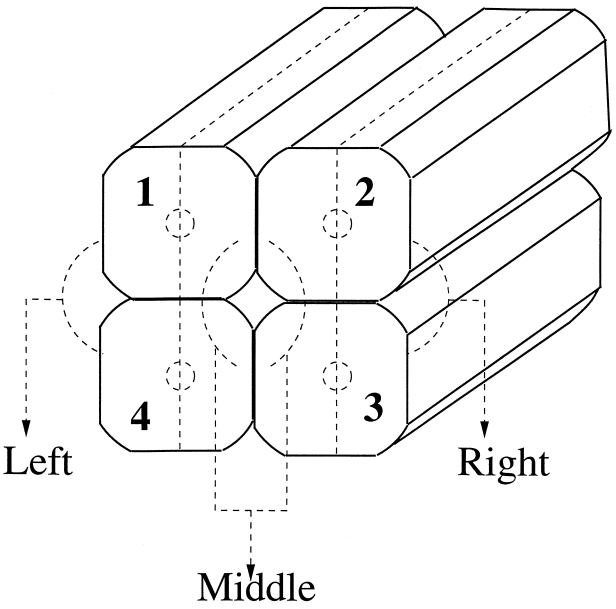



**FIGURE 6** Schematic diagram showing the four Ge crystals of a segmented clover Ge detector. In this type of clover detector, to further improve position information, signals are taken from the center contacts of each crystal (labeled 1–4) and also from the left, right, and middle parts of the outer electrical contacts.

## 4. Tracking Detectors

Recently, further advances in detector manufacturing technology allow the electronic segmentation of a single crystal into smaller elements, thus further localizing the interaction site within the volume of the detector. The ultimate goal of these developments is the development of a tracking detector array, which actually allows one to follow the trajectory of each individual gamma-ray as it traverses a detector, even if it undergoes multiple scattering events en route.

Ideally, such an array needs to cover all the available solid angle and localize each gamma-ray interaction to within 1–2 mm in three dimensions. A variety of tracking detectors are under development worldwide, including the Gamma-Ray Energy Tracking Array (GRETA for short) in the United States (Deleplanque *et al.*, 1999).

Such an array would be very efficient. Simulations for the proposed GRETA array indicate that it may be up to a thousand times more sensitive than the best of today's spectrometers. This sensitivity comes about because of the high-count rate capability (the relatively low-count rate in each segment is the limiting factor, rather than the high rate in the entire detector), excellent PT ratio, resolution, and efficiency.

A prototype detector for the GRETA array has already been extensively tested in Lawrence Berkeley National Laboratory, Berkeley, CA. One key test involved the determination of the gamma-ray interaction position by use of a closely collimated source. The interaction positions are determined by detailed measurements of pulse shapes on an event-by-event basis. A comparison of measured pulse shapes, with calculations show excellent agreement, which is a major first step in a proof of principle for the detector. The next step in this project is to purchase a mini-array of such detectors. These multiple detectors, assembled into a closely packed array, allow one both to

do physics and to prove the principle of practical gamma-ray tracking for the first time. The proposal to construct this array is currently awaiting funding.

### 5. Pair Spectrometers

High-energy gamma-rays (with energies, say, in the 1–10 MeV range) can interact with matter to produce positron-electron pairs. When the positron annihilates, two photons, each of 511 keV, are emitted at an angle of 180° to each other. When such a pair-production process occurs in a Ge detector, one or both of the 511-keV photons may escape from the detector without being detected. Hence, each gamma-ray transition leads to three peaks in the spectrum, the full energy peak plus the so-called single and double "escape" peaks. This proliferation of peaks can significantly complicate spectral analysis and adversely affect nuclear level scheme construction.

To improve such spectra one often uses a pair spectrometer, which, in essence, is the inverse of the anti-Compton shield spectrometer discussed earlier. Whereas in an anti-Compton shield any event detected in the shield is used to veto the coincident event in the Ge detector, in a pair spectrometer the simultaneous detection of 511-keV $\gamma$-rays on opposite sides of the Ge detector is used to positively trigger (i.e., to select) the double escape peak in the Ge detector spectrum.

## F. Measurements of Nuclear Level Lifetimes

Aside from the measurement of gamma-ray energies and intensities, and the determination of gamma-ray transition placements in nuclear level schemes by coincidence and Ritz Combination techniques, gamma-ray detectors can be used to measure another extremely important observable, namely, nuclear level lifetimes. These lifetimes are proportional to squares of quantum mechanical quantities called transition matrix elements and therefore can directly reveal insights into nuclear structure and the properties of nuclear excitations.

Two general classes of techniques are used: those involved in directly measuring the time difference between successive gamma de-excitations in a nucleus and those based on Doppler effects. The former has traditionally been limited by the rise time of voltage pulses from detectors to the nanosecond range, but advances using faster scintillation detectors have pushed the frontiers of electronic time measurements farther down to nearly the picosecond range. Doppler techniques are typically used to measure lifetimes from hundreds of picoseconds down to the few femtoseconds range. These techniques cover a range of lifetimes characteristic of a wide variety of nuclear decays.

### 1. Recoil Distance and Doppler Shift Attenuation Methods

Typical recoil velocities following heavy-ion fusion-evaporation reactions are a few percent the speed of light. The associated Doppler shifts of emitted gamma-rays can be used to obtain level lifetimes. The recoil velocity corresponds to an easily measured, maximum Doppler shift of about 20–30 keV in a 1000-keV gamma-ray. The fraction of the gamma-ray intensity which lies in the Doppler shifted peak can be proportional to the lifetime of the nuclear state. Two Doppler-based techniques are commonly used. The first, termed the Recoil Decay Method (RDM), utilizes two parallel foils separated by a distance $d$. The nuclei of interest are produced in the first, thin foil and recoil out of the foil with a well-defined recoil velocity. Having flown a distance $d$, they are rapidly stopped in the second, thicker stopper foil. If the nuclear state of interest decays while the nucleus is flying between the two foils, then the gamma-ray will be emitted with the appropriate Doppler shift. On the other hand, if the lifetime is long enough that the nucleus reaches the stopper foil and is stopped, the gamma-ray will be emitted from a nucleus at rest, without a Doppler shift. Changing the distance $d$ between the foils can access different lifetime ranges. Typically, the RDM technique is used to probe lifetimes in the nanosecond to picosecond range.

The Doppler Shift Attenuation Method (DSAM) is similar to the RDM in that two foils are used. However, in this case, the foils are placed in intimate contact with each other. Now the recoiling nuclei immediately enter the second foil and begin to slow down and stop. If the nuclear lifetime of interest is of the same order of magnitude as the slowing down time of the nuclei in the foil, around 1–2 ps, then the gamma-ray transitions will be emitted with a range of Doppler shifts, ranging from the maximum shift down to zero. Level lifetimes may be extracted by carefully analyzing the resulting, complicated peak shapes and comparing them to model calculations. Of course the calculations also have to include the slowing down process itself. The DSAM method is sensitive to level lifetimes on the order of picoseconds to femtoseconds, i.e., somewhat shorter than those accessible with the RDM method.

### 2. The GRID Technique

Another Doppler-based method is used at the ILL in Grenoble, referred to earlier (Koch *et al.*, 1980), using the ultra-high resolution crystal diffraction instruments GAMS4 and GAMS5. In this approach, a thermal neutron from a reactor is captured by a target nucleus which then emits a series of gamma-rays from the capture state (typically lying at an excitation energy of about 6 MeV) to

lower lying states. Each emitted gamma-ray carries a small linear momentum, $p = E/c$. Hence, the emitting nucleus recoils in a direction opposite to that of the gamma-ray. If the same nucleus then emits a subsequent gamma-ray prior to stopping in the target material, the second gamma-ray will be Doppler shifted.

The shifts are exceptionally small. Typical recoil energies are a few electron volts, and therefore, the technique relies of measuring Doppler broadening effects of this order using crystal diffraction techniques. Note that one measures a Doppler broadening rather than a shift because the gamma-ray emission of the ensemble of nuclei is effectively isotropic. The technique is known as the GRID (Gamma-Ray Induced Doppler) technique (Borner and Jolie, 1993) and, like the DSAM, is useful for nuclear level lifetimes shorter than or on the order of the stopping time, typically $\sim$1 ps.

### 3. Fast Timing Spectroscopy

The coincidence measurements with Ge detectors described above are typically used to establish nuclear level schemes. Such measurements utilize coincidence resolving times on the nanosecond time scale ($10^{-9}$ sec). However, the time response of $BaF_2$ scintillation detectors is much faster than that of Ge detectors and, with special care, can be reduced to the few picosecond range. Hence, coincidence timing can also be used to directly measure nuclear level lifetimes in the few tens of picoseconds range, which is typical of the lifetimes of many collective excitations in medium mass and heavy nuclei. In practice, the technique, called FEST (Fast Electron Scintillation Timing) [see Buescher *et al*. (1990) for a simplified discussion and for references to more technical literature], is most commonly used in $\beta$-decay experiments where the time is measured between the emission of a $\beta$-ray (detected in a thin, fast plastic scintillator) and the subsequent gamma-ray emission in the daughter nucleus.

The technique must be used with great care. One problem is that the $BaF_2$ detectors have very poor energy resolution ($\sim$10%). Additional gamma-ray selection, by coincidence with cascade gamma-rays using Ge detectors (with "normal" nanosecond timing), is normally needed to simplify the $BaF_2$ spectra to one of two gamma-rays at most. Therefore, most applications are in low multiplicity experiments. Another serious problem relates to the energy dependence of the timing. A gamma-ray moves at the speed of light and in 3 ps travels $\sim$1 mm. Since typical $BaF_2$ detectors have sizes on the order of centimeters, it is clear that the timing is sensitive to the exact position in the crystal where the gamma-ray absorption occurs. Hence, the time properties of such detectors are energy dependent and must be carefully calibrated. Nevertheless, the technique has proven to be quite useful in studies of nuclei off the line of nuclear stability in $\beta$-decay experiments.

## III. GAMMA-RAY SPECTROSCOPY AND NUCLEAR STRUCTURE

The atomic nucleus is a unique, many-body quantum mechanical system. When describing nuclei, numbers of the order of 100 seem to occur frequently. For example, the depth of the potential holding the protons and neutrons, collectively known as nucleons, together is about 50 MeV. The maximum angular momentum the nucleus can hold before centrifugal forces break it apart is about 100 $\hbar$, which occurs for nuclei around mass 100.

Typical nuclei have a few hundred constituent nucleons. This number implies that the nucleus occupies a unique position in the plethora of quantum systems found in nature. A few hundred particles grouped together is sufficient to allow one to contemplate macroscopic nuclear properties such as shape and surface area and thickness. One the other hand, it is few enough that the addition or subtraction of a single proton or neutron can radically change the behavior of the whole system. Indeed, one of the appealing features of the nucleus is that it is a many-body quantal system in which the number of interacting bodies can be precisely controlled, measured, and varied. We will see a stunning example of the microscopic nature of the nucleus below when we discuss the phenomenon of backbending. This mixture of macroscopic and microscopic behavior in a strongly interacting system (the nucleons are after all bound together in the nucleus by the effects of the strong force) is nearly unique in nature.

The behavior of the nucleons inside the nucleus can be likened to the behavior of a herd of wild animals. The herd clusters together for protection, defining a shape and form. (The Hungarian word for such a herd is gulyas, so the nucleus is a bit like a goulash soup of nucleons.) However, the behavior of a single animal can have dramatic effects on the collective motion of the whole system. In the following sections, we will describe some of the features of the excited atomic nucleus and attempt to describe a few of the many manifestations of its macroscopic and microscopic behavior.

Generally speaking, atomic nuclei can be excited from the "bottom up" using reactions such as Coulomb excitation, inelastic scattering, or direct reactions, or from the "top down" using $\beta$-decay, neutron capture, and heavy-ion fusion-evaporation reactions. The former approach most often excites states selectively, while the latter approach is much less selective, tending to populate most states along a myriad of possible de-excitation routes, subject only to constraints due to angular momentum

selection rules or phase space considerations. Gamma-ray spectroscopy is most often used in this second approach. When gamma-ray spectrometry is used in "bottom up" techniques, such as Coulomb excitation, it is exploited primarily as an indicator of the excitation probability of particular levels rather than as a study of de-excitation modes per se.

In this section, we will discuss a number of aspects of gamma-ray spectrometry. Although the distinction is a bit artificial, it is convenient, and historically pertinent, to break the discussion up into the study of low- and high-spin states.

## A. Low-Spin States

The study of the low-spin nuclear states dates back to the beginning of nuclear structure and is the basis for our understanding of the equilibrium structure of nuclei and its evolution with nucleon number. Low-spin states are typically populated following $\beta$-decay, neutron capture, Coulomb excitation, or photon scattering reactions.

### 1. Beta-Decay

Nuclei formed off the valley of stability decay back toward stable nuclei via $\beta$-decay (which includes the processes of $\beta^-$, $\beta^+$, and electron capture decay). Typically, $\beta$-decay populates several excited levels in the daughter nucleus. Half-lives near stability range from seconds to days. Production of $\beta$-decay parent nuclei can be achieved by simple reactions such as $(p, n)$ or by heavy-ion reactions. The simpler, lower energy reactions tend to form only one or a couple of parent nuclei, whereas heavy-ion reactions may form many times more, and, in that case, selection techniques are needed to select the decay products of interest.

A popular technique in $\beta$-decay is the use of moving tape collectors in which the activity is collected on a tape (e.g., movie reel tape or aluminized Mylar) for some period of time (typically $\sim$1.8 times the half-life for the desired $\beta$-decay). The tape is then moved to a low background area for detection of gamma-rays following decay. Collection of a new activity at another spot on the tape proceeds simultaneously.

Gamma-ray spectroscopy following $\beta$-decay was for many years in the 1950s–1970s a standard technique used to elucidate nuclear structure. Since $\beta$-decay itself carries off little or no angular momentum, the spin states accessible with this technique are generally those within $\pm 2$–$3\,\hbar$ of the parent (ground or isomeric) state.

In recent years the technique has enjoyed a renaissance with the use of arrays of much higher efficiency Ge detectors (e.g., clover or cluster detectors). Since the gamma-ray multiplicity following $\beta$-decay is low and there is no Doppler effect, the detectors can often be mounted in close geometry to maximize count rates and achieve considerable coincidence efficiencies.

One current setup for such studies is the Yale moving tape collector (Casten, 2000). Illustrated in Fig. 7,

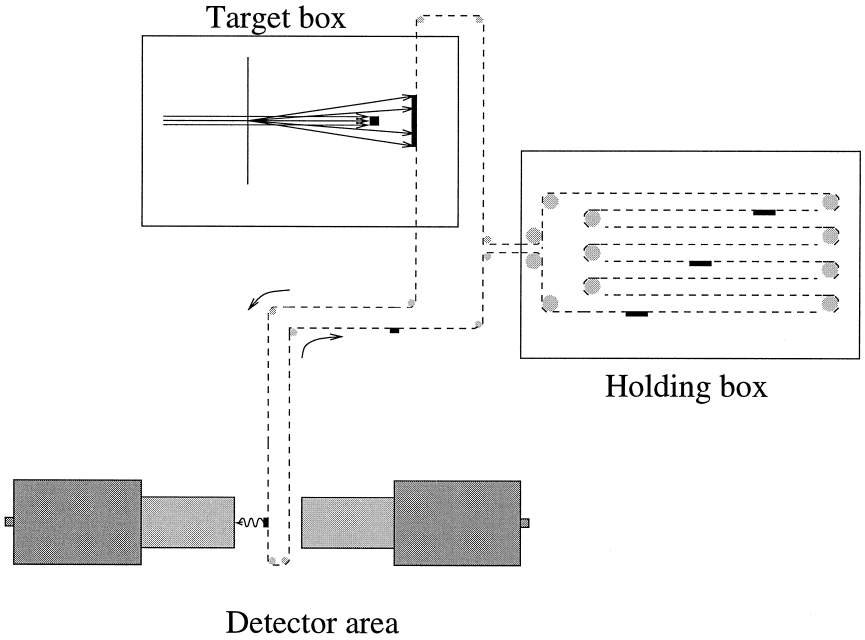



**FIGURE 7** Diagram of the Yale moving tape collector showing the target box, counting area, and tape holding box. Activity is deposited on the tape in the target box, with the beam entering from the left. It is then transported to the counting area. The holding box provides a delay to let unwanted extraneous activity decay away before the tape once more returns to the target box.

it uses up to four Compton suppressed clover detectors that can be positioned at any angle in a horizontal plane. Studies with this instrument have included searches for possible multi-phonon states in $^{162}$Dy and $^{164}$Er. Nuclei with ellipsoidal shapes can undergo vibrational oscillations (called phonons) of these shapes about their equilibrium position. In principle, it is possible to superpose two or more identical vibrations. However, the effects of the Pauli Principle acting on the particles in the nucleus may destroy such states. One test of their intact character is to study their gamma decay. If they have predominantly a two-phonon character, then they should decay to the one-phonon state. Experimental searches for weak gamma-ray decay branches to the single phonon excitation are being sought in these two nuclei.

Another application of $\beta$-decay exploits the fourfold segmentation of the clover detectors. In the Yale arrangement, four such detectors allow simultaneous coincidence measurements at a large number of different relative angles of emission between the two detected gamma-rays. These angular correlation measurements can be used to constrain spin arguments for levels in the gamma-ray cascade. With clover detectors situated at appropriate angles, it is also possible to exploit their segmentation to measure the linear polarization of the gamma-ray and thereby to deduce the parity relations of the nuclear levels involved (Duchene *et al.*, 1999).

Finally, $\beta$-decay measurements are also an important tool in mass measurements, since, often, the daughter or granddaughter mass is known but not that of the parent. Nuclear masses (that is, in effect, binding energies) are of importance in a number of contexts. The binding energy reflects the sum of all the nucleonic interactions. Differences of binding energies for neighboring nuclei give the separation energy of the last nucleon and are therefore sensitive to single particle energies of nucleons in a mean field nuclear potential, as well as to shape and structure changes from one nucleus to the next. Mass measurements are also important for understanding the astrophysical processes occurring in the interiors of stars that lead to nucleosynthesis. Recent studies of nuclei in the mass $A \sim 70$ region, for example, are helping to set constraints on the termination of the rapid proton capture process in certain classes of stars.

Nuclear mass measurements are carried out by measuring gamma-ray spectra in coincidence with $\beta$-particle detection in order to deduce the $\beta$-decay end point, that is, the maximum $\beta$-decay energy (where energy sharing with the simultaneously emitted anti-neutrino is insignificant). The end point energy directly gives the mass of the parent nucleus if the daughter mass is known. The gamma-ray coincidence is used to cleanly select the product nucleus of interest.

## 2. Coulomb Excitation

When a beam particle passes close to a target nucleus, one or both nuclei may be excited by the changing electromagnetic Coulomb field between them (without any nuclear reaction occurring). Usually, a series of low-spin levels of the target nuclei are excited. The excitation probabilities are deduced by observing the subsequent de-excitation gamma-rays. A typical Coulomb excitation experiment involves bombarding a target of the (stable) isotope to be studied with beams of particles (the beams used range from protons to very heavy ions) at beam energies of roughly 80% of the Coulomb barrier.

Coulomb excitation is a powerful technique to study nuclear structure. Since the excitation mechanism is purely electromagnetic, it is known and calculable. Therefore, one can extract nuclear information from the excitation probabilities. This is in contrast, for example, to inelastic scattering processes at beam energies above the Coulomb barrier where nuclear effects enter in both the excitation mechanism and the nuclear structure itself and must therefore be disentangled.

In typical Coulomb excitation experiments, to correctly account for Doppler effects, the gamma-rays are detected in coincidence with the scattered beam particle. As noted, the excitation probability is enhanced by smaller impact parameters, which often result in scattering at backward angles in the laboratory frame of reference. Hence, often, annular particle detectors are placed at back angles (say, $140^\circ \leq \theta \leq 170^\circ$). These detectors allow the beam to pass through and then selectively identify those scattering events most likely to have resulted in nuclear excitations.

## 3. $(n, \gamma)$ Reactions

Historically, an immense amount of critical data on medium and heavy mass nuclei came from the study of radiative neutron capture, or $(n, \gamma)$, reactions with reactor neutrons. Like other reactions, such as heavy-ion fusion-evaporation reactions or $\beta$-decay that populate nuclear levels from the top down, the process is non-selective and, therefore, gives access to a wide variety of nuclear states. Indeed, when used in the average resonance capture (ARC) mode, the technique can actually guarantee that all states in a given angular momentum and excitation energy range can be identified, thus providing very sensitive tests of models (Caston *et al.*, 1980). Such states can be directly observed from the so-called primary transitions that de-excite the capture state. The use of pair spectrometers is important here. When low-energy gamma-ray spectra are studied, one typically observes hundreds of transitions. Therefore, gamma-gamma coincidence techniques are crucial. Alternatively, many of the most important

$(n, \gamma)$ studies have used the ultra-high energy resolution GAMS crystal diffraction detectors (see Fig. 3b). Studies of nuclei such as $^{196}$Pt (Cizewski *et al.*, 1978) and $^{168}$Er (Davidson *et al.*, 1981) with $(n, \gamma)$ have provided some of the most comprehensive and complete level schemes ever produced and have provided key tests of nuclear models, such as the Interacting Boson Model. Today, most forefront $(n, \gamma)$ work is carried out using the GRID technique to measure lifetimes with GAMS detectors.

## B. High-Spin States

### 1. Backbending and the Pauli Principle

One of the fundamental questions to ask about a nucleus is: What is the shape? Is the nucleus spherical, like a soccer ball, or deformed, stretched out like an American football or perhaps flattened like a Frisbee? It turns out that some nuclei are spherical; some are deformed like footballs; and some are deformed like Frisbees. The excitation spectrum of deformed nuclei is particularly easy to understand. A deformed system has a defined orientation in space (it is not isotropic), and rotations of this shape can be observed. A quantum mechanical rotor has an excitation energy given by

$$E(I) = \frac{\hbar^2}{2J} I(I+1),$$

where $I$ is quantum number counting the angular momentum of the state ($I = 0, 2, 4, \ldots$ for the ground state rotational band, the odd spins are missing for symmetry reasons which are not relevant here) and $J$ is the moment of inertia of the nucleus. The gamma-ray energy (which is measured in the experiment) is just the energy difference between adjacent states.

$$\begin{aligned} E_\gamma(I \to I-2) &= \frac{\hbar^2}{2J}[I(I+1) - (I-2)(I-1)] \\ &= \frac{\hbar^2}{2J}[4I-2]. \end{aligned}$$

Thus, the gamma-ray energy increases linearly with angular momentum. For gamma-rays linking adjacent levels, the energy difference is given by

$$\begin{aligned} \Delta E_\gamma &= E_\gamma(I \to I-2) - E_\gamma(I-2 \to I-4) \\ &= \frac{\hbar^2}{2J}[(4I-2) - 4((I-2)-2)] \\ &= \frac{4\hbar^2}{J}. \end{aligned}$$

If the moment of inertia, $J$, does not change, then $\Delta E_\gamma$ is a constant, independent of spin. Usually, this is not the case in nuclei. A rare example of a nearly ideal rotational band, where the spacing between adjacent transitions is constant, is shown in Fig. 8 (the most intensely populated superdeformed band in $^{150}$Gd). However, usually, dramatic changes in structure occur (e.g., due to centrifugal forces or quenching of pairing) as a nucleus rotates faster and faster. These are manifest as deviations from the simple linear dependence outlined above. For example, a spectrum of the ground state rotational band of $^{158}$Er is illustrated in Fig. 9, where the lines indicate transitions linking states with increasing spin. Notice that at gamma-ray energies of about 400 keV the transitions double back on themselves. This phenomenon is called backbending and corresponds to a dramatic change in the internal structure of the nucleus.

The origin of this structural change lies in the effects of the familiar Coriolis force on the microscopic structure of the nucleus. As we have stressed, the nucleus is not a rigid body, but instead is made up of only a few hundred protons and neutrons that orbit the center of mass in orbits characterized by particular angular momenta. We know that many medium mass and heavy nuclei exhibit properties similar to those of a superconductor. In the ground state of an even-even nucleus, all of the protons are coupled pairwise, in identical but time-reversed orbits, so that the total angular momentum of each pair is zero. Similarly, the neutrons are also paired. Hence, the total angular momentum of the ground state of any even-even nucleus is zero. As an interesting aside, it follows that in an odd-proton or odd-neutron nucleus, the ground state spin and parity is usually determined by the quantum numbers of the final unpaired proton or neutron.

The question, therefore, becomes, what happens to these pairs of protons and neutrons as the nucleus as a whole begins to rotate? Just as a person walking on a merry-go-round experiences a force on the moving platform, the so-called Coriolis force, the nucleons in the nucleus also experience the effect of the rotating bulk. Just as with the merry-go-round, the Coriolis force increases the faster the nuclear rotation or the orbital velocity. Indeed, the size of the Coriolis force is such that at moderate nuclear rotational frequencies it perturbs the orbits of the particles sufficiently that the pairs of nucleons will begin to break apart. This has the effect of dramatically changing the excitation energies of the states and the gamma-ray energies for transitions between them. It is this breaking of the superconducting pairs that is responsible for the backbending observed in Fig. 9.

An illustration of the effects of the Pauli exclusion principle can be seen in the rotational spectra of odd-even nuclei. Figure 10 is a plot of the angular momentum as a function of rotational frequency for $^{133}$Pr, which has 59 protons and 74 neutrons. The two curves shown in Fig. 10 correspond to rotational bands in which the final

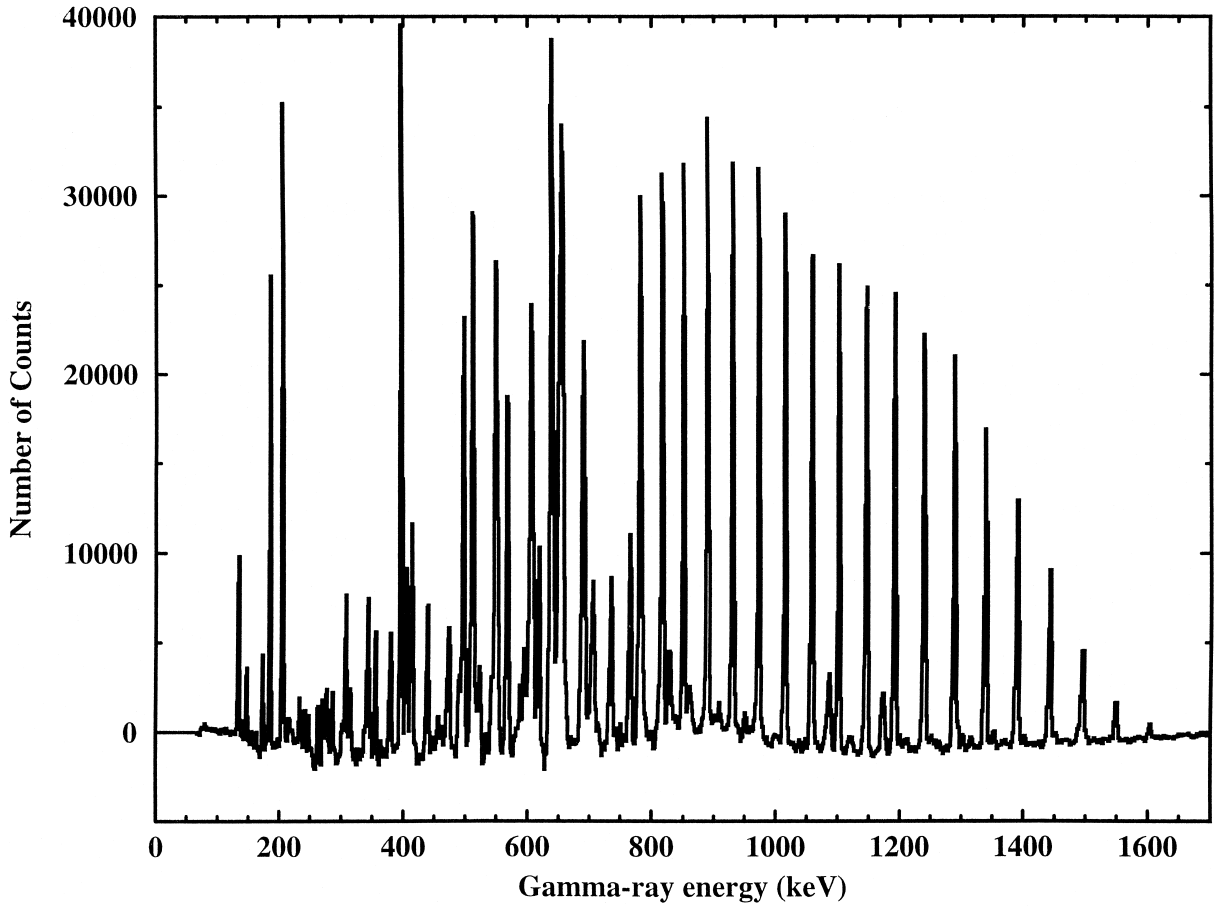



**FIGURE 8** Spectrum of the most intensely populated superdeformed band in $^{150}$Gd. In addition to the regular picket-fence pattern of gamma-rays associated with decays of superdeformed states, the spectrum also shows, at lower energies, the complex pattern of transitions depopulating the nearly spherical normal deformed states in $^{150}$Gd.

unpaired proton is in different orbits about the nucleus. In one of these cases the odd-proton acts like a spectator to the underlying even-even nucleus, and in this case the above backbending phenomenon occurs as before. In the other band, however, the odd-proton occupies one of the orbits of the pair of aligning nucleons. The pair breaking is therefore prohibited by the Pauli exclusion principle, and the backbending is delayed until higher rotational frequencies when it becomes possible to occupy higher lying orbits.

## 2. Superdeformation

One of the forefront areas of research in high-spin nuclear structure physics over the last decade has been the study of superdeformed (SD) nuclei. These states exist in a second minimum in the nuclear potential energy surface in which the nucleus takes on an ellipsoidally deformed shape which roughly corresponds to an integer ratio of major to minor axes, typically 2:1 or 3:2. The observation of the first high-spin SD bands in $^{152}$Dy and $^{132}$Ce, by the Liverpool University groups of Peter Twin and Paul Nolan, respectively (Twin *et al.*, 1986; Nolan *et al.*, 1985), sparked an enormous worldwide effort to discover additional examples of highly deformed nuclei and to characterize the properties of such highly stressed systems (stressed both by the application of very high angular momenta and by extreme values of deformation). Today, about 40 nuclei in four main mass regions, or islands, have been shown to exhibit SD behavior. Most of these nuclei have more than one known SD band.

Superdeformed rotational bands are generally characterized by extremely regular gamma-ray energy spacing. The energy spacing from one transition to the next in the rotational band is either constant or varying slowly and regularly from one transition to the next. For example, the strongest SD band in $^{150}$Gd is illustrated in Fig. 8. The regular picket-fencelike pattern of SD transitions is

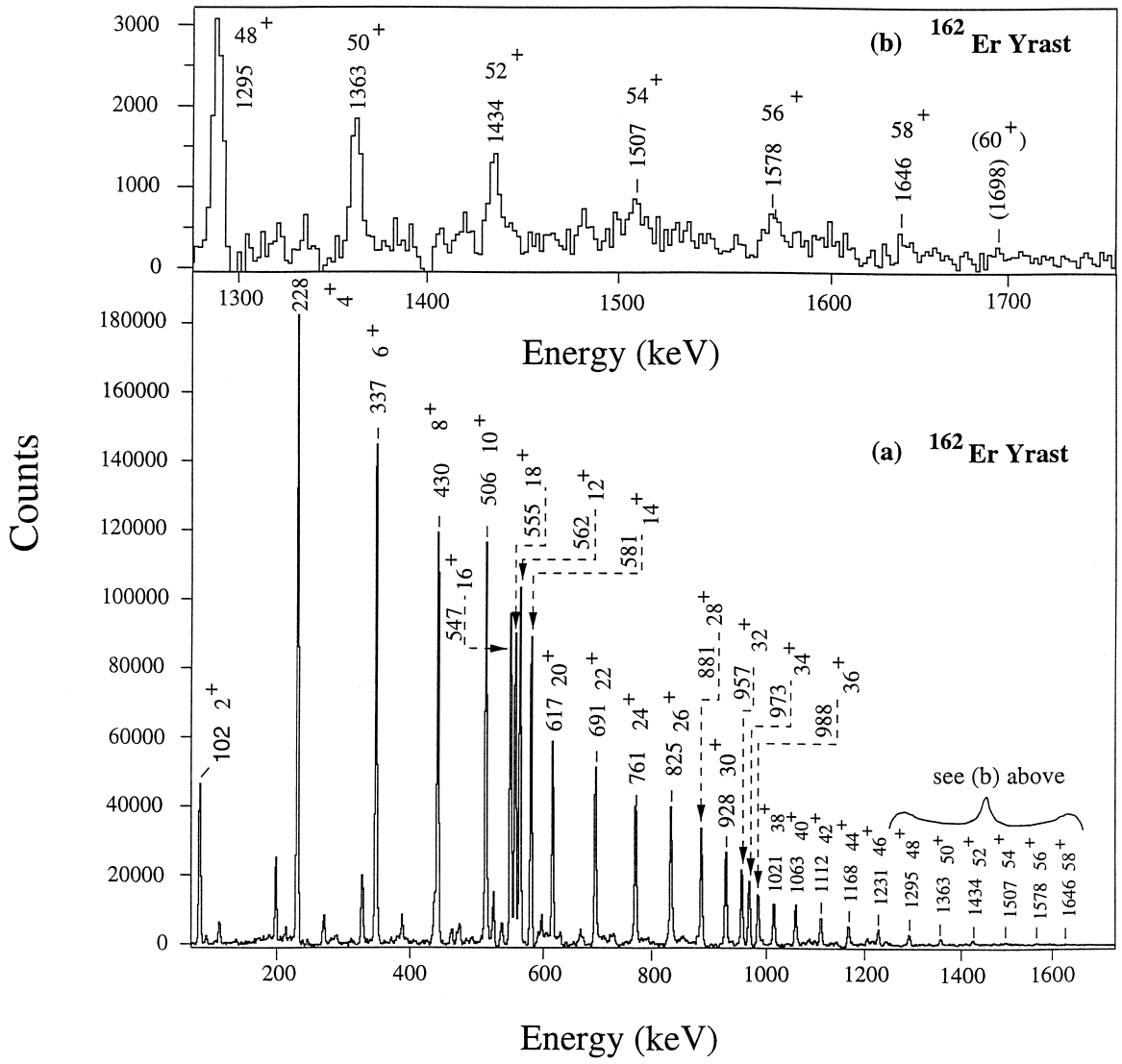



**FIGURE 9** Spectrum illustrating the ground state rotational band of $^{162}$Er illustrating the backbending phenomenon. (Figure courtesy of Mark Riley).

unmistakable in this spectrum as is the irregular pattern of transitions de-exciting lower lying normal deformed states ($^{150}$Gd is nearly spherical in its ground state).

Due to this regularity, which is the rule rather than the exception for SD bands, one can feel confident in predicting where transitions in a given band should occur. However, detailed measurements of the strongest SD band in $^{149}$Gd, using the Eurogam Ge detector array, revealed a very small deviation from this smooth behavior (Flibotte *et al.*, 1993), Indeed, it was found that every second energy spacing was larger/smaller than the average. The deviation, illustrated in Fig. 11, is very small, only about 0.25 keV, and is measurable only due to the very high quality spectra available from the Eurogam array. It is believed that the deviation is caused by alternate states in the rotational band being perturbed up and down in energy by very small amounts, on the order of 60 eV. This staggering essentially separates the rotational band into two $\Delta I = 4\ \hbar$ sequences. The origin of the perturbation, which affects states differing in spin by 4 $\hbar$, is still unclear. Several theoretical models have been proposed to explain this phenomena, none of which, however, can reliably predict which SD bands should exhibit staggering and which should not.

### 3. Magnetic Rotation and Chiral Symmetry

Interesting effects have also emerged from the study of near-spherical nuclei. One of the consequences of quantum mechanics is that the rotation of a spherical shape

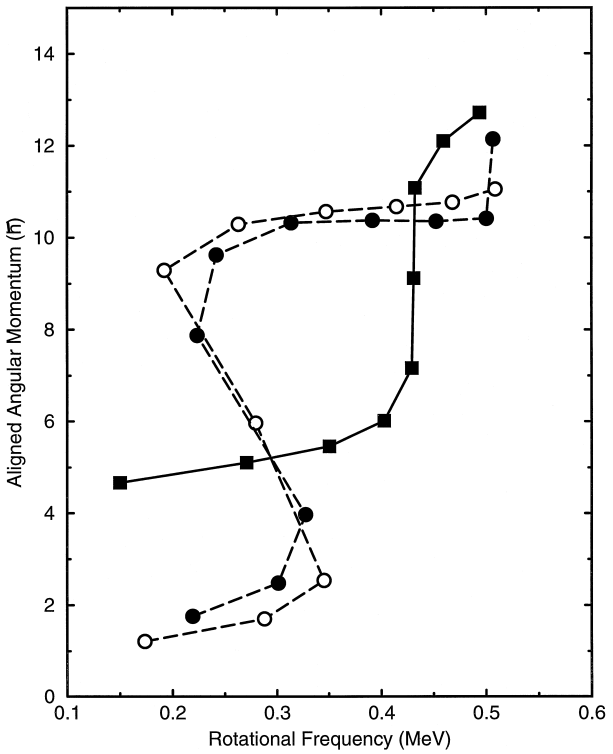



**FIGURE 10** Component of the angular momentum along the rotation axis vs rotational frequency for different rotational bands in $^{133}$Pr. Notice that the backbending observed at a frequency of ~0.25 MeV is completely missing in one of the bands. The absence of a crossing in this band is a dramatic illustration of the Pauli exclusion principle.

cannot be observed. How then does a spherical or near-spherical nucleus generate angular momentum? Rather than a collective rotation of the whole shape, it does so by rearranging the orbits of its constituent protons and neutrons, by single particle excitations to higher lying excited states with high values of angular momentum. Typically, such excitations have irregularly spaced energies resulting in a gamma-decay spectrum with many irregularly spaced peaks (see the lower energy portion of Fig. 8).

It was a surprise, therefore, when regularly spaced sequences of gamma-rays were observed in some almost spherical light Pb nuclei, near the doubly closed shell $^{208}$Pb. Furthermore, these apparently rotational-like cascades were found to consist of very strong magnetic dipole (M1) transitions which change the angular momentum by $\Delta I = 1\,\hbar$, with very weak, or unobserved, $\Delta I = 2\,\hbar$ electric quadrupole transitions (E2). In contrast, a rotational band in a well-deformed nucleus consists of a sequence of strong $\Delta I = 2\,\hbar$ E2 transitions. The absence of E2 transitions in these new bands is an indication of the near-spherical nuclear shape. However, the regularity of the new band structure implies a type of collective behavior.

The tilted axis-cranking model provides an explanation for these bands (Frauendorf, 1993). For certain near-spherical nuclei with proton and neutron particle numbers close to magic numbers, the angular momentum vectors of the unpaired proton particles and neutron holes prefer to align perpendicular to each other, with one vector pointing along the rotation axis and the other perpendicular to the rotation axis. The vector sum of these two angular momenta then lies at an angle to the nuclear symmetry axis. Furthermore, the vector accounts for almost all of the nuclear angular momentum, since the collective rotation of the near-spherical shape is small. Higher angular momentum states are generated by slowly closing these two angular momentum blades, or shears, pushing against the repulsive particle-hole nuclear interaction. The enhanced M1 transitions arrise because the magnetic dipole moment is proportional to the component of the individual proton and neutron angular momenta perpendicular to the total angular momentum.

An interesting extension of the idea of tilted axis cranking comes when we consider the possibilities in doubly odd deformed, triaxial nuclei. As for shears bands, for certain favorable particle numbers the angular momenta of the final unpaired proton and neutron align preferentially perpendicular to each other, along the nuclear rotation (short) and symmetry (long) axes. For a triaxial nuclear shape, considerations of irrotational flow indicate that the collective angular momentum should align preferentially with the intermediate length nuclear axis. Thus, the three angular momentum vectors can form either a

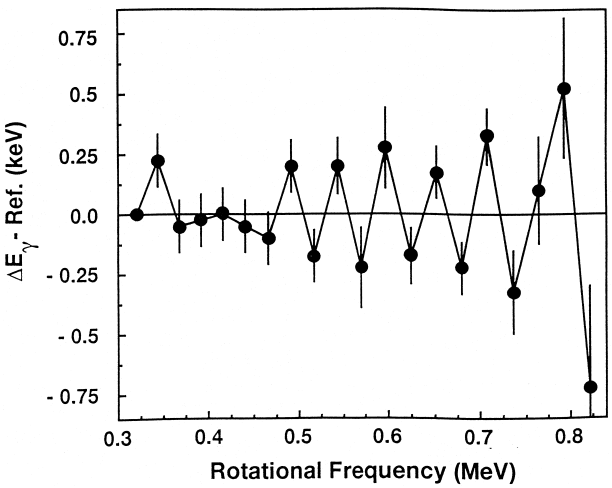



**FIGURE 11** Energy staggering in the strongest SD band in $^{149}$Gd (Flibotte *et al.*, 1993). The figure shows the deviation of the measured gamma-ray energy from a smooth reference as a function of rotational frequency. Notice that the deviation is extremely small, usually less than 0.25 keV. This deviation corresponds to a tiny perturbation in the nuclear energy levels of only about 60 eV.

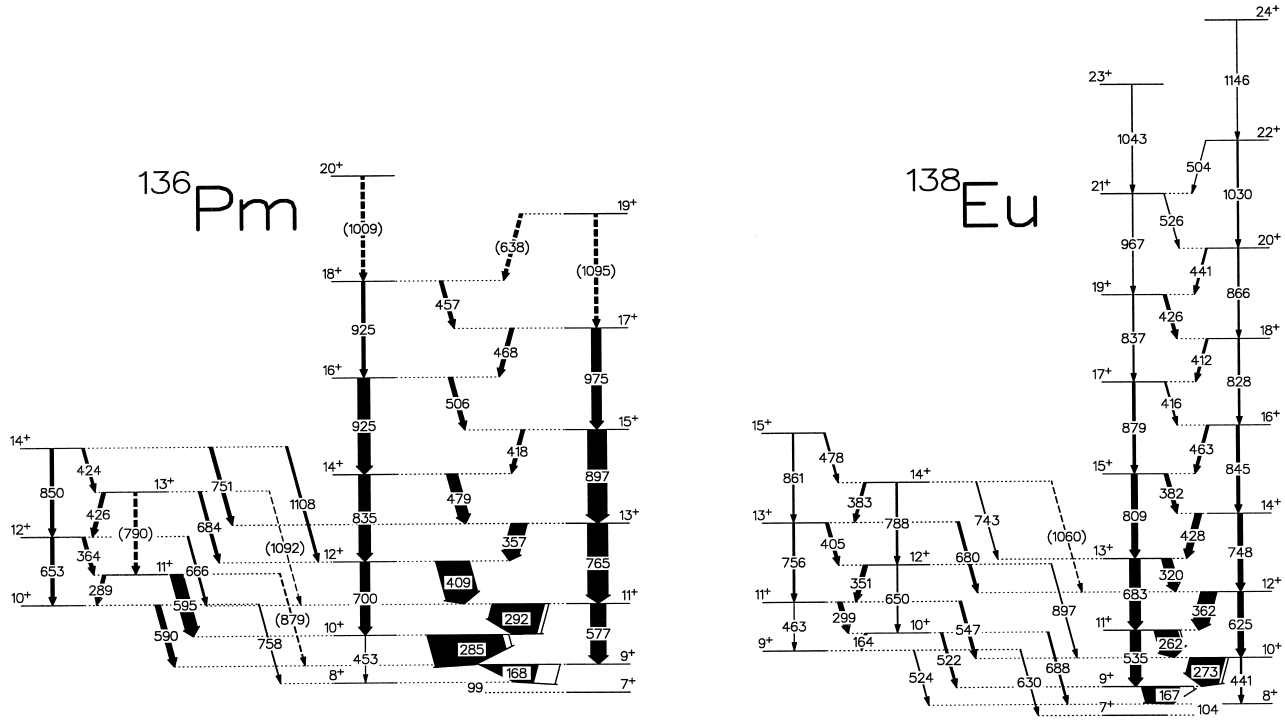



**FIGURE 12** A partial level scheme of the odd-odd nuclei $^{136}$Pm (61 protons and 75 neutrons) and $^{138}$Eu (63 protons and 75 neutrons). The proposed chiral twin bands are shown on the left of each level scheme.

right- or a left-handed coordinate system. The so-called 3d-tilted axis-cranking model, developed by S. Frauendorf and J. Meng (1997), addresses such a system and predicts a doubling of energy levels, one corresponding to each chirality or handedness. For complete symmetry, the levels of the same spin and parity would be degenerate. If the solutions for different chiralities mix, then the degeneracy will be broken, and one set of states, corresponding to a $\Delta I = 1\,\hbar$ rotational band, will be lifted with respect to the second band. Indeed, two $\Delta I = 1\,\hbar$ bands in the doubly-odd nucleus $^{134}$Pr have been proposed as a possible chiral candidate (Frauendorf and Meng, 1997). Following on this suggestion, several other candidate bands have been observed in nearby nuclei (Starosta *et al.*, 2001; Beausang *et al.*, 2001; Hecht *et al.*, 2001), while candidate bands have also recently been reported in doubly-odd $^{188}$Ir (Balabanski *et al.*, unpublished). The proposed chiral twin bands in $^{136}$Pm and $^{138}$Eu are shown in Fig. 12.

## C. Spectroscopy in Coincidence with Separators

A great deal of exciting new spectroscopy of nuclei far from stability or with very large $Z$ has been achieved over the last several years when large Ge detector arrays have been coupled to high-transmission magnetic separators. A magnetic separator is a device placed behind the target position which will selectively transport nuclei, produced in a reaction, to its focal plane where they can be detected and identified using a variety of different detectors. Residual nuclei that are not of interest, or scattered beam particles, will not be transmitted through the separator. Very small fractions of the total reaction cross section can be selected using this method. Nuclear structure information is obtained by detecting gamma-rays produced at the target position, in coincidence with recoils detected at the focal plane. One example of the use of this technique is illustrated here.

One of the goals of nuclear physics is to understand the limits of nuclear existence as functions, for example, of angular momentum, isospin, or indeed mass. For example, what are the heaviest nuclei that can exist? For many years now, various models have predicted that an island of superheavy nuclei should exist. However, most models disagree as to the exact proton and neutron numbers categorizing this island and indeed on the extent of the island. Recently, models have predicted that these superheavy nuclei might indeed be deformed. Therefore, it is very relevant to inquire as to what is the structure of the heaviest nuclei accessible to gamma-ray spectroscopy and to ask the simplest type of questions about them, for

example, are they spherical or deformed? Unfortunately, the production cross sections for superheavy nuclei are such that, even using very intense beams, only one or two nuclei are produced per week or two. These small numbers are clearly beyond what we can measure with existing gamma-ray facilities. Therefore, we cannot address the spectroscopy of the superheavy elements (yet). However, we can look at the structure of very heavy nuclei lying just below these unattainable regions.

Recently, groups at Argonne National Laboratory in the United States and at the University of Jyvaskyla in Finland carried out tour de force experiments to study the excitation spectrum of $^{254}$No (Leino *et al.*, 1999). With $Z = 102$, No is the heaviest nucleus for which gamma-ray spectroscopy has ever been carried out. The gamma-ray spectrum of transitions de-exciting states in $^{254}$No is shown in Fig. 13. A rotational band structure is clearly visible, indicating that $^{254}$No is in fact a deformed nucleus. A very surprising feature of the spectrum is that the rotational band is observed up to very high spins $\sim 18\,\hbar$, (an amazing number for such a heavy, fissile nucleus). The existence of a rotational cascade up to spin $\sim 18\,\hbar$, well beyond the classical fission barrier limit, indicates that $^{254}$No is held together primarily by microscopic shell effects, rather than macroscopic liquid drop binding, as in normal nuclei. Shell effects, for certain favorable proton and neutron numbers and for favorable deformation, can provide an additional 1–2 MeV of binding energy. It is this binding energy, which does not depend strongly on angular momentum, which holds $^{254}$No together to such high spin.

### D. Experiments with Radioactive Beams

Today, a new era in nuclear structure physics is opening up with access to a much wider selection of nuclei, extending far beyond the valley of stability and encompassing nuclei that are expected to be exotic in both proton/neutron composition and structure. The physics opportunities with such beams have been discussed elsewhere (RIA Physics White Paper, 2000) and need not be repeated here. What are relevant are the particular methods of carrying out gamma-ray spectroscopy on exotic nuclei. Basically, the techniques to be used will be familiar ones, such as $\beta$-decay, Coulomb excitation, and fusion-evaporation reactions. High-, medium-, and low-spin states will all present topics of interest.

Experiments with radioactive beams differ primarily in two critical respects from their stable beam siblings. First, beam intensities will often be much lower than with stable beams. Instead of beams of $10^{11}$ particles per second, many experiments will need to be carried out with intensities that are less than $10^6$ particles per second and, at the limits of accessibility, down to 1 particle per second or even less. Therefore, detectors will have to be correspondingly more efficient. Second, because the nucleus to be studied is sometimes the one produced as a beam by the radioactive beam facility, most experiments will be done in inverse kinematics in which the roles of beam and target are interchanged.

In inverse kinematics, $m_b > m_t$ where $m_b$ and $m_t$ are the masses of the beam and target nuclei. Therefore, the reaction products all go forward in the laboratory system. For $m_b \gg m_t$, this forward focusing results in a quite narrow cone of reaction products. For example, for elastic scattering of $^{62}$Ni on $^{12}$C, the maximum allowed scattering angle is $\sim 10^\circ$. This has two principal effects. First, measuring angular distributions of reaction products is much more difficult. Second, on the other hand, it is possible to capture much larger percentages of the reaction products in the acceptance angles of various types of charged particle spectrometers and mass separators, thereby enhancing counting rates. These considerations impose design constraints on gamma-ray detectors surrounding the target. First of all, ultra-high efficiency is needed. Second, generally, a forward angled cone needs to be left free of detectors.

The requirement of maximal gamma-ray counting efficiency generally means a close geometry and detectors that subtend large solid angles. However, Doppler effects can then be very large, especially when using inverse kinematics, and high detector granularity will generally be critical. This granularity can currently be achieved in two ways and considerable development in both directions is needed. One is the use of highly segmented tracking arrays such as GRETA discussed earlier. The other is the use of position-sensitive Ge detectors of the type developed by Glasmacher and colleagues for use in intermediate energy Coulomb excitation experiments at MSU (Muller *et al.*, in press). In these detectors a resistive readout at the two ends of a linear Ge crystal allows the localization of the $\gamma$-ray interaction to an accuracy of $\sim 2$ mm. This detector is capable of measurements even with beam intensities of $\sim 1$ particle per second or even less. An example of a gamma-ray spectrum (corrected for Doppler effects) from intermediate energy Coulomb excitation taken with an early generation detector system (using NaI(Tl) detectors) is shown in Fig. 14. This data was taken on $^{40}$S in order to test predictions of the underlying particle motion in exotic nuclei with a high excess of neutrons over protons (the heaviest stable isotope of sulfur has 20 neutrons, $^{40}$S has 24 neutrons). The Coulomb excitation was accomplished in this case using a $^{197}$Au target.

For specialized experiments, such as low-energy Coulomb excitation in inverse kinematics designed explicitly to excite only the lowest one or two states,

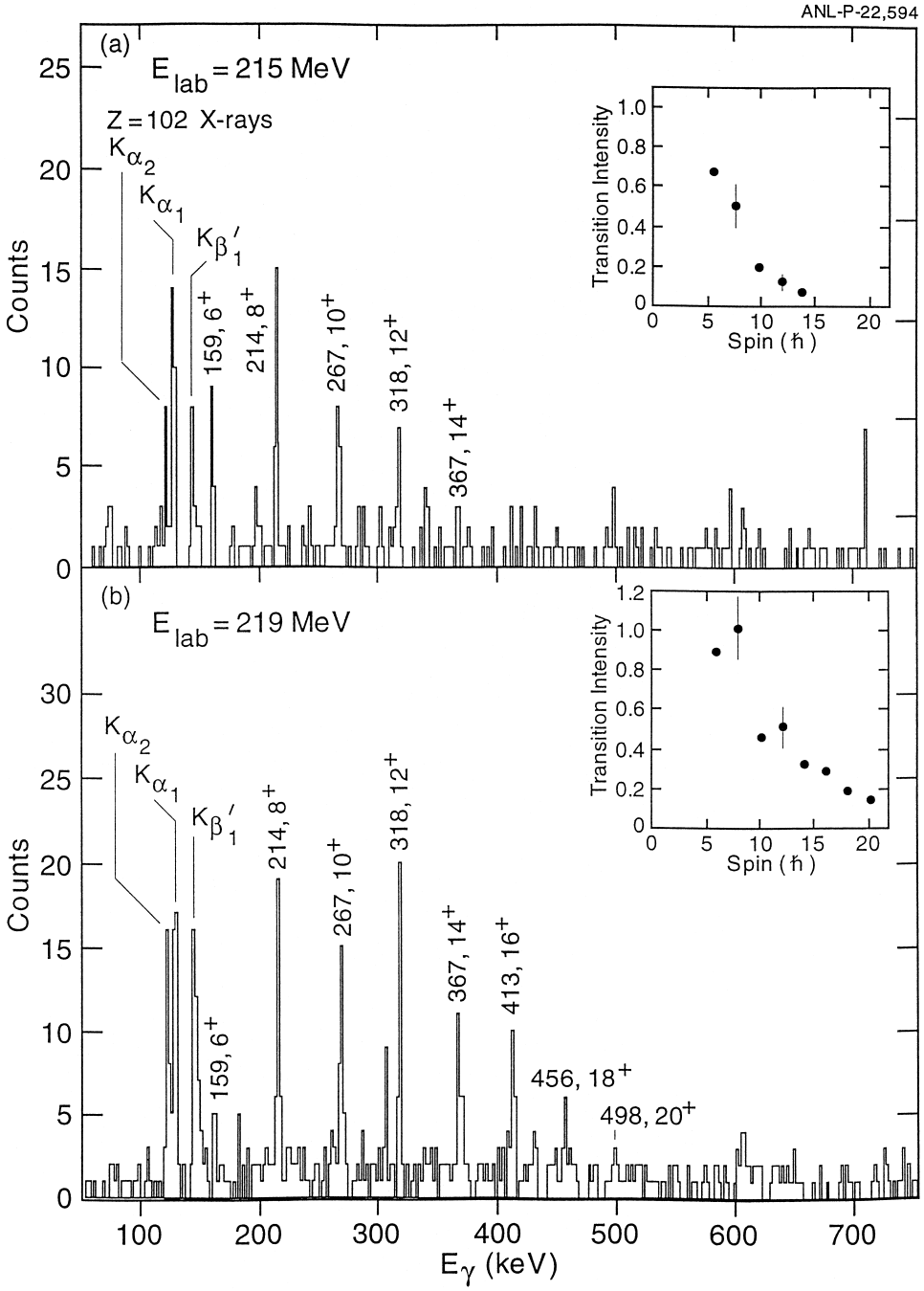



**FIGURE 13** Spectrum of gamma-rays depopulating excited states in $^{254}$No. With $Z = 102$, $^{254}$No is the heaviest nucleus for which gamma-ray spectroscopy has ever been accomplished. The gamma-rays are labeled by their transition energies in kilo-electron volts and also by the spin of the state they depopulate. The inserts show the population intensity as a function of spin for the two beam energies, 215 (top) and 219 MeV (bottom). More angular momentum is brought into the system at higher beam energies, and this is reflected in the stronger population of higher spin states in the lower spectrum (Leino *et al.*, 1999).

the gamma-ray spectra are particularly simple. Therefore, energy resolution is not a problem and high efficiency can be obtained with, for example, low-resolution NaI(Tl) detectors placed in close geometry. Here, Doppler effects are both unimportant and undetected. One typical design, the GRAFIK detectors (Sheit *et al.*, 1996) actually incorporates the target inside an annular hole in the detector, achieving ~80% of $4\pi$ solid angle coverage.

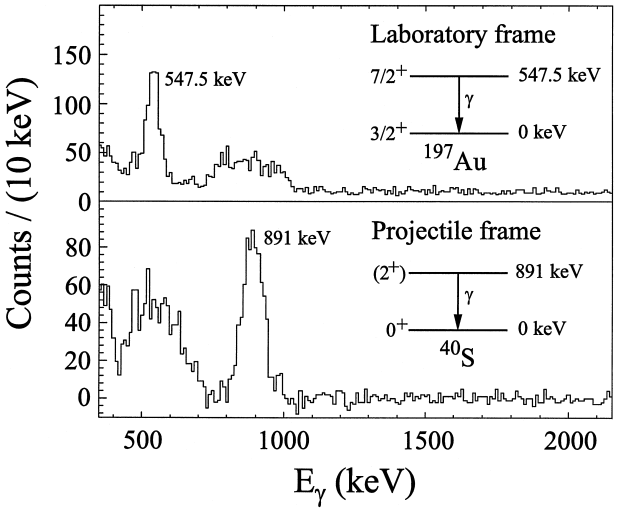



**FIGURE 14** Gamma-ray spectra following intermediate energy Coulomb excitation of a radioactive $^{40}$S beam on a $^{197}$Au target. The spectra are corrected for Doppler effects for gamma-rays emitted from nuclei at rest in the laboratory frame (top) or from nuclei moving at the beam velocity (bottom). The gamma-ray de-exciting the first excited state of $^{40}$S is clearly visible in the lower spectrum. [From Muller *et al.* (in press). *Nucl. Instrum. Methods.*]

## IV. CONCLUSIONS

Although the dominant interaction binding nucleons into nuclei is the strong force, the electromagnetic interaction, as manifested primarily in gamma-ray spectroscopy, provides an ideal probe of the structure and excitations of the nucleus. Indeed, gamma-ray spectroscopy in nuclear and astrophysics research is a broad and diverse field, utilizing a variety of detector systems that vary greatly (according to the needs of particular experiments) in resolution, efficiency, gamma-ray energy range, and other properties. These detectors are often used alone or in conjunction with auxiliary devices such as charged particle or neutron detectors. The areas of nuclear structure addressed with such instrumentation cover the whole gamut of nuclei spanning the entire nuclear chart and physics problems,, ranging from the motion of individual nucleons to collective flows (e.g., vibrations or rotations) of the nucleus as a whole.

## ACKNOWLEDGMENT

We are grateful to many colleagues for advice and discussions and, in particular, to Mark Caprio, Thomas Glasmacher, Hans Borner, and Mark Riley for providing the figures we have used. This work was supported in part by the U.S. DOE grant number DE-FG02-91ER-40609.

## SEE ALSO THE FOLLOWING ARTICLES

GAMMA-RAY ASTRONOMY • ION BEAMS FOR MATERIAL ANALYSIS • NUCLEAR PHYSICS • POTENTIAL ENERGY SURFACES

## BIBLIOGRAPHY

Barton, C. J., et al. (1997). *Nucl. Instrum. Methods* **A391,** 289.
Balabanski, D., et al., to be published.
Beausang, C. W., et al. (1992). *Nucl. Instrum. Methods* **A313,** 37; Beck, F. A., et al. (1992). *Prog. Part. Nucl. Phys.* **28,** 443; Nolan, P. J. (1990). *Nucl. Phys.* **A520,** 657c.
Beausang, C. W., et al. (2000). *Nucl. Instrum. Methods* **A452,** 431.
Beausang, C. W., et al. (2001). *Nucl. Phys*. **A682,** 394c.
Buescher, M., et al. (1990). *Phys. Rev.* **C41,** 1115; Mach, H., et al. (1990). *Phys. Rev.* **C41,** 1141.
Borner, H. G., and Jolie, J. (1993). *J. Phys.* **G19,** 217.
Casten, R. F., et al. (1980). *Phys. Rev. Lett.* **45,** 1077.
Casten, R. F. (2000). *Nucl. Phys. News Int.* **10,** 4.
Casten, R. F., and Nazarewicz, W. (2000). "White Paper for the RIA Workshop, Raleigh-Durham, North Carolina, July 24–26, 2000."
Cizewski, J. A., et al. (1978). *Phys. Rev. Lett.* **40,** 167.
Davidson, W. F., et al. (1981). *J. Phys.* **G7,** 443, 455.
Deleplanque, M. A., et al. (1999). *Nucl. Instrum. Methods* **A430,** 292.
Duchene, G., et al. (1999). *Nucl. Instrum. Methods* **A432,** 90.
Eberth, J., et al. (1996). *Nucl. Instrum. Methods* **A369,** 135.
Frauendorf, S. (1993). *Nucl. Phys.* **A557,** 259c.
Frauendorf, S., and Meng, J. (1997). *Nucl. Phys.* **A617,** 131.
Flibotte, S., et al. (1993). *Phys. Rev. Lett.* **71,** 4299.
Gerl, J., and Lieder, R. (1992). "Euroball III," GSI Darmstadt Report. Darmstadt, Germany.
Hecht, A., et al. (2001). *Phys. Rev.* **C63,** 051302(R).
Johnson, A., Ryde, H., and Sztarkier, J. (1971) *Phys. Lett.* **B34**, 605.
Koch, H. R., et al. (1980). *Nucl. Instrum. Methods* **175,** 401; Kessler, E. G., et al. (2001). *Nucl. Instrum. Methods* **A457,** 187.
Lee, I. Y. (1990). *Nucl. Phys.* **A520,** 641c; Deleplanque, M. A., and Diamond, R. M., eds. (March 1988). "The Gammasphere Proposal: A National Gamma-Ray Facility," LBL, Berkeley, CA.
Leino, M., et al. (1999). *Eur. Phys. J.* **A6**, 63; Reiter, P., et al. (1999). *Phys. Rev. Lett.* **82,** 509.
Morinaga, H., and Gugelot, P. C. (1963). *Nucl. Phys.* **46,** 210.
Muller, W. F., et al. (in press). *Nucl. Instrum. Methods*.
Nolan, P. J., Gifford, D. W., and Twin, P. J. (1985). *Nuicl. Instrum. Methods* **A236,** 95.
Nolan, P. J., et al. (1985). *J. Phys.* **G11,** L17.
Riedinger, L. L., et al. (1980). *Phys. Rev. Lett.* **44,** 568.
Sheit, H., et al. (1996). *Phys. Rev. Lett.* **77,** 3967.
Simpson, J., et al. (2000). *Heavy Ion Phys.* **11,** 159.
Starosta, K., et al. (2001). *Phys. Rev. Lett.* **86,** 971.
Twin, P. J., et al. (1986). *Phys. Rev. Lett.* **57,** 811.



# High-Pressure Research

**E. F. Skelton**
**A. W. Webb**
*Naval Research Laboratory*

I. Introduction
II. High-Pressure Research Environments
III. Measurement of Pressure
IV. Research at High Pressure

## GLOSSARY

**Bridgman anvil** Anvil design used originally with very hard metallic anvils and most recently with diamond anvils employing the principle of massive support.

**Diamond cell** Compact device used for generating pressures to 500 GPa in microscopic samples with the use of gem quality diamond anvils.

**Equation of state** Mathematical expression describing the relationship between the volume, mass, temperature, and pressure of a system under conditions of thermodynamic equilibrium.

**Hugoniot curve** The locus of points in either the shock velocity-particle velocity plane or the stress-volume plane derived from experimental measurements.

**Phase diagram** Graph combining two thermodynamical variables, such as pressure and temperature, of a material drawn so that a particular curve represents the boundary between two phases of the material.

**Shock compression** State in matter achieved by the passage of a very large amplitude mechanical pulse of disturbance through a material for necessarily brief durations.

**PRESSURE** is an important thermodynamical variable; it provides the most efficient means of altering interatomic distances while leaving the thermal energy of the system invarient. It therefore provides an important mechanism for testing theoretical models that are based upon interatomic separations and crystallographic configurations. Pressure can also be used along with temperature to assist chemical reactions or to bring about crystallographic phase transformations. New allotropes, formed under conditions of extreme pressure or temperature, may have physical properties that are significantly different from those of the material formed under normal conditions. A classic example is carbon: the hardness, electrical and thermal conductivities, and transparency of diamond, the phase of carbon formed at elevated pressures and temperatures, are significantly different from those of graphite, the phase of carbon that is stable under normal conditions. Another example of how pressure has been effective in producing a new and better polymorph is that of $Nb_3Si$. Empirical arguments have suggested that if this material could be formed in the cubic A15 structure, rather than in its normal tetragonal $Ti_3P$ phase, it would exhibit superior superconducting properties. High-pressure shock

treatments have been successful in producing this transformation. In this cubic phase, the superconducting transition temperature of $Nb_3Si$ is 18.5–19 K, as compared to 0.29 K in the tetragonal phase.

The role of pressure in understanding physical properties of materials can also be of importance. An example where it may be vital is in understanding the origin of superconductivity in certain organic charge transfer salts based on the cation molecule ditetramethyltetraselenofulvalenium (TMTSF), since all but one of these salts is only superconducting at elevated pressures.

# I. INTRODUCTION

## A. Definition of Pressure

Pressure is defined as the ratio of a force divided by the area over which that force acts; thus, the units of pressure, force per unit area, are newtons per square meter (=1 Pa), dynes per square centimeter ($=10^{-6}$ bars), or pounds per square inch (at sea level 1 atm pressure $= 1.0133 \times 10^5$ Pa $= 1.0133$ bars $= 14.696$ lb in.$^{-2}$). In this article we shall use the S.I. unit of pressure, the Pascal, abbreviated Pa. Since the pressure, or more correctly the stress state, will vary depending on the direction of the force relative to the area over which it is applied, in a strict sense, it is necessary to consider the six independent components of the stress tensor. In terms of a working definition, however, most researchers usually presume that they are dealing with hydrostatic pressure or something close to it, in which case the three diagonal elements of the stress tensor are the same value and all off diagonal terms, which represent the shear components, are zero.

In modern high-pressure research, the force is usually transmitted to the sample of interest via some medium. If that medium is a fluid, then a true hydrostatic pressure environment does exist, although today much high-pressure research is carried out in the range of tens of gauss Pascals pressure or above (1 GPa = 10 kbar), where few materials remain in the fluid state. Consequently, much work is done today under conditions of quasi-hydrostatic pressure (i.e., since the forces are transmitted through solidified fluids or relatively soft solids, shear stress components can be a problem).

Natural pressures found in our universe range from 100 kPa in the atmosphere at sea level, to about 100 MPa at the deepest part of the oceans (the bottom of the Marianas trench in the Pacific Ocean), to 0.36 TPa at the center of the earth, to tens of tesla Pascals at the center of our sun. Today, synthetic, or man-made, pressure environments can be produced in the laboratory to span this range. In terms of static pressures, the relatively large volume systems ($\sim 50$ cm$^3$) can achieve pressures up to 1 GPa, while the more recently developed diamond-anvil cells can be used to subject microscopic-size samples ($\sim 10^{-8}$ cm$^3$) to pressures in excess of 500 GPa. Each of these systems is discussed in more detail below.

Much higher pressures can be achieved for brief periods of time by using conventional or nuclear explosive-driven shock waves. Current research in this area also involves the study of shock waves produced by high-powered laser beams. These shock pressures are usually accompanied by significant elevations in the sample temperature. The pressures achieved by shock waves can be in range of a few tesla Pascals.

## B. Historical Review

The father of modern high-pressure research is Percy W. Bridgman, a man who dedicated his professional life to high-pressure experimentation. Working at Harvard University, he published over 200 research papers and, in addition to a wealth of basic scientific research, was also responsible for several important technological discoveries: the principle of unsupported area, known today as the Bridgman seal, and the principle of massive support, upon which the Bridgman anvil is based. In 1946 he was awarded the Nobel Prize for his pioneering work in this field.

The work following Bridgman can be divided into three parts based on the experimental technique employed: dynamic or shock pressures, large-volume studies, and diamond-anvil cell work. Each of these methods will be discussed in detail in the body of this article.

# II. HIGH-PRESSURE RESEARCH ENVIRONMENTS

## A. Static Pressures

### 1. Fluid Systems

The application of pressure by means of a fluid, whether liquid or gas, has generally proven to be the most satisfactory method. These hydrostatic environments will not subject a test specimen directly to shear stresses; however, oftimes the reaction in the sample may be anisotropic and internal shear components can exist. The maximum pressure obtainable in a fluid system is limited by the design and strength of the container, often referred to as a "pressure bomb." The ultimate strength of even compound cylinders poses a limit, but more often the sealing of such systems presents the defining pressure limit.

One of the best solutions to the sealing problem is the "unsupported seal" designed by Bridgman: in this

mechanism, the confining pressure is itself used to help seal the system. A mushroom-shaped piston or plug faces the system on its full diametrical face, but in turn transmits the total resultant force to soft annular packing that is placed between the plug and the driving rod. The smaller stem of the plug projects into a well in the rod for alignment purposes, but the well has sufficient depth that the packing must bear the full load. Therefore, the pressure on the packing is always larger than the fluid pressure by a constant factor. This factor is related to the relative areas of the piston face and the annulus. In practice, the packing material may vary from rubber, for lower-pressure systems, to stacks of softer and harder metals, with In, Pb, or even annealed Cu serving as the sealing agent.

Although the working pressures obtainable in fluid systems are generally below 1 GPa, even at these modest pressures, the viscosity of many fluids will have increased to the point where small-bore plumbing no longer allows equilibration of the pressure throughout the system. We shall return to this problem in the discussion of the diamond-anvil cells.

## 2. Means of Pressure Generation

Liquids can be pumped on by any of several methods. Hand-operated piston pumps have served for small-volume systems, although special modifications for the sliding seal and check valves may be needed as the ultimate pumping pressure increases. In some cases, double checks have been found to be effective, and all parts must be machined to extremely close tolerances and must receive fine finishes. For larger systems or very high pressures, these pumps are tedious to operate.

Another manually operated system employs a large hand-operated wheel driving a screw that forces a piston into the pump cylinder. This type of system is limited in maximum pressure by the relative volumes of the pump, the system, and the stroke of the piston.

Pumps adapted from manually operated units and driven electrically are common at hydraulic system pressures. Air-driven diaphragm pumps using normal air are convenient for the operation of relatively large systems to pressures approaching 200 MPa.

Liquid systems are also encountered in which the liquid is sealed in a small cylinder with a piston that is loaded with a hydraulic press and then clamped. Such systems can be made quite small and, for this reason, often find applications in research involving cryogenic temperatures or high magnetic fields.

Gas systems can be pumped directly to pressures approaching 100 MPa; pressures above this must be obtained indirectly. One of the oldest methods is to use a U-shaped, high-pressure chamber with Hg forced into one leg by gas pumped to the system capability. This flooded leg is then pumped with oil, thereby raising the pressure in the gas to the maximum attainable with the liquid system. This often takes several cycles of the system with appropriate valving, since the gas is significantly more compressible than the pump fluid.

A second method of increasing gas pressure is through the use of a stepped piston intensifier. In this device, a large oil-driven piston is mechanically linked to a smaller diameter piston, which uses the mechanical advantages of the ratios of the areas to pump the gas to pressures perhaps as much as 10 times that of the oil. As in the previous case, several cycles may be needed in order to attain the maximum effect.

## 3. Experimental Probes

One subject receiving extensive study is the fluids that serve as compression agents, as discussed above. Fusion curves, especially at low temperatures, are, of course, of considerable interest. The transformation from the Fluid to the solid state requires assumptions concerning the magnitude of the strains present and their effect on the pressure–temperature status.

Experiments designed to map out equations of state for selected fluids are more ambitious. These require simultaneous determination of the pressure, volume, and temperature ($P$, $V$, $T$) of a given system. The measurement of the volume is the most difficult, since even heavy walled containers will undergo some, albeit small, deformation under load. One technique is to seal a known amount of the fluid of interest in a system separated from the pressure fluid by a bellows. A sensing system, either internal or external, monitors the change in the extension of the bellows of the system and, thus, the volume as the pressure and temperature are varied.

If the pressure containers employed are fabricated of nonmagnetic materials, such as Be–Cu, then changes in the magnetic properties of the contained material can be assessed through the walls, thereby obviating the need for direct contact into the pressure chamber. Although this will greatly simplify the experiment by removing the need to pass electrical leads through the pressure walls, it can have the drawback of a large experimental error since the sample usually represents only a small portion of the external sensing coil volume.

The issue of making electrical contact to the contained sample is not a simple one. Insulated wires must be brought through the pressure wall into the fluid-filled volume without pinching off or extruding the wires themselves or destroying the insulation. Extrusion is usually controlled by placing a conical portion of the lead on the high-pressure side with an insulating sleeve that is only

driven tighter into its seat with the application of pressure. A preload is useful in setting the seal before the initial application of pressure. Liquids will not pass through the insulation as readily as gases, especially He, and therefore, this type of probe is often used in clamp-type cells.

Another sealing technique involves the use of leads swaged in high-temperature insulation with stainless steel jacketing. These leads are common with thermocouple assemblages; other special conductors can also be used. In these applications, the metal sheathing is silver soldered into a threaded plug that passes the leads into the pressure cavity. The interior end is sealed with epoxy.

A third technique that has been successfully used to pressures up to 1 GPa involves passing the leads from the high-pressure environment through a second length of high-pressure tubing, with the first being used to pressurize the system. A large U-bend is placed in this tubing that is filled with an oil and then submerged in a liquid nitrogen dewar, thus freezing the oil through which the leads pass. Care must be taken to insure that liquid oil is present above the frozen solid on the high-pressure side in gas-pressurized systems.

## 4. Piston-Cylinder Devices

*a. General.* Simpler, in principle, than fluid bombs are the piston-in-cylinder devices. In their simplest form, a hole is drilled in a block of solid material and, after plugging one end or using a blind hole, the test sample is inserted (see Fig. 1). A strong, close-fitting piston is then inserted and the pressure is applied. Solid materials are readily studied in such a device, unless they are very soft, such as Pb, In, or polymers, or have a low coefficient of friction, such as some of the transition metal dichalcogenides, for example, $MoS_2$ or $WSe_2$. In these latter cases, the aforementioned unsupported seal can be employed.

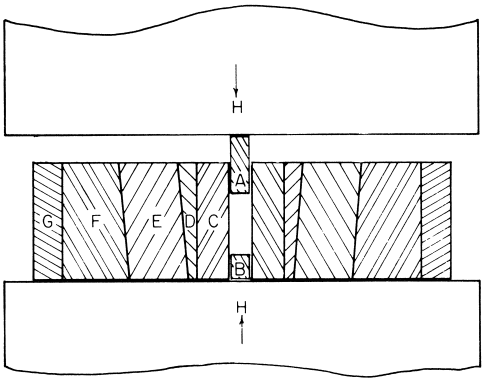



**FIGURE 1** Piston and cylinder device. A is the WC piston; B is the WC fixed piston or nib; C is the WC cylinder; D is the shim; E and F are the hardened steel compression binding rings; G is the soft steel safety ring; and H is the press plattens. D and E are machined with tapers giving an interference fit when pressed into the final assembly.

*b. Design principles.* Normally, a piston is ground with a very small clearance in the lapped cylinder, on the order of 0.0005 in. This is usually sufficient to seal most solids, except as noted above. For lower pressure applications, the piston and cylinder are made of hardened tool steel, with the piston made somewhat harder than the cylinder. With quality steels, pressures approaching 2.5 GPa can be attained in these systems. Higher pressures require harder materials, with WC being the most tractable. This is a brittle material, and while it does have a high compressive strength (~800,000 psi), it is weak in tension—its tensile strength is usually only about 2% of its compressive strength. Therefore, it needs support that is usually provided by two or three interference fit support rings or hardened steel. These will serve to prestress the WC and allow it to be worked to much higher pressures. These binding rings are designed to maintain a compressive load on the WC to its maximum working pressure.

*c. Failure modes.* Shortly after the internal pressure exceeds the sum of the interference compression and the modest WC tensile strength, radial tensile fracture occurs. If the problems with radial fracture can be forestalled, the next mode of failure is breakage along a plane perpendicular to the cylinder axis. The solution to this is to again prestress the WC, this time with end clamping of the cylinder. Some researchers have utilized sophisticated hydraulic clamping systems to adjust the end loading proportional to the sample pressure; however, the equipment needed for such work can be costly, and the ultimate pressures achieved can usually be attained more easily by other techniques. Heavy clamping bolts and support rings can also be used for such loading.

*d. Multi-staging.* Since the failure of the cylinder is due to the difference in pressure between the sample chamber and the outside environment, then the achievable internal pressure of the system could be elevated if a means could be found to pressurize the entire assemblage. Thus, if a device capable of generating, for example, 10 GPa were placed in a similar device, then theoretically a maximum pressure of 20 GPa could be attained. Of course, this requires a system large enough to contain the entire second system within its pressure cavity, and a third stage would then go within the second cylinder. Although simple in theory, application has proven difficult. Bridgman built an operating two-stage system and the High Pressure Institute, located near Moscow, Russia, has a very large press with three-stage operation.

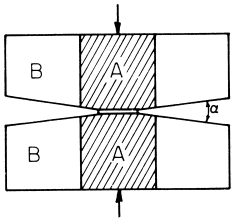



**FIGURE 2** Bridgman anvils: A is WC anvils with 5–20° angle on the outer portion, providing massive support on the load bearing center; B is the compression binding ring.

## 5. Other Uniaxial Systems

Second to the Bridgman anvils (Fig. 2), the "belt" is perhaps the most important apparatus used for large-volume, high-pressure work (see Fig. 3). This stems from its development, which provided the increased pressure and temperature conditions necessary for the original synthesis of diamonds. Although few research laboratories have need for this large sample volume, it is used by several manufacturers for the routine, commercial production of diamond grit.

As shown in Fig. 3, elements of both the piston-cylinder and the Bridgman anvil device are used in the belt apparatus. The cylindrical belt contains a large-volume sample, and the truncated concial pistons utilize the massive support and the compressive gasket concepts. With application of internal electrical resistance heaters, temperatures of 2000°C can be maintained at pressures up to 10 GPa. The belt anvils and cylinders are usually designed with a curvature, although some devices have utilized straight conical sections. These employ the same principles and are simply somewhat easier to machine.

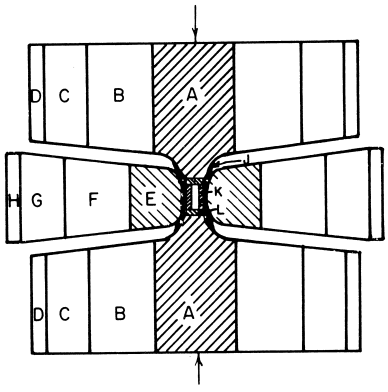



**FIGURE 3** Belt apparatus: A is WC tapered pistons; B, C, F, and G are hardened steel binding rings; D and H are soft steel safety rings; E is the WC belt cylinder; J is the compressible gasket; K is the cylindrical sample container; and L is the sample end caps. Binding rings are assembled with interference fits to give compressive support to the WC parts.

As noted above, the belt is principally a high-temperature, high-pressure device. It can also be used for electrical measurements at elevated temperatures by passing contact and thermocouple leads out through one or both of the gaskets. An equitorially split belt was once used for X-ray diffraction studies at temperatures up to 1000°C.

## 6. Multi-Anvil Devices

*a. Tetrahedral press.* In the hierarchy of pressure-producing apparatus, the tetrahedral press follows the uniaxial devices. It consists of four hydraulically driven rams that are designed to converge on the faces of a regular tetrahedron. The sample container is usually formed of the same material used to form the gaskets in the Bridgman anvils, and either the tetrahedron is formed with an edge face about 25% greater than the anvil face edges or gasket tabs are added. This system relies on compressible gaskets to contain the pressure and allow the ram some additional stroke for pressure generation after contact has been made. These anvils also utilize the principle of massive support.

The tetrahedral concept was employed with a uniaxial press by workers at the National Bureau of Standards by nesting three of the anvils in a cone, inserting a sample, and then driving the fourth anvil down. This had the effect of moving the lower anvils down and in at the appropriate rate.

These four-ram units are difficult to control, and in an effort to ensure equal advance of each ram, an anvil guide was developed. This consists of a linkage of heavy pins between holes in the anvil support plates and forms a nest that forces the system to open or close with complete synchronism.

The four rams of the tetrahedral press are generally mounted so as to be electrically isolated from each other. This allows access to the pressurized area by up to four leads, and as many as ten have been passed through the gaskets; however, as with any gasket passage, pinching off is not uncommon.

The pressure capabilities with tetrahedral presses run to about 10 GPa, although a much longer life of the WC anvils can be attained by limiting operations to below 7 GPa. As with the Bridgman anvils, the working piece is usually supported in an interference fit, high-strength steel retaining ring.

Temperatures in excess of 2000°C can be attained in these presses with internal electrical resistance graphite tube furnaces: By using smaller faced anvils and accepting a higher breakage rate of anvils, the mineral stishovite has been synthesized in this type of device. Stishovite is a metastable form of silica formed at pressures in excess of

9 GPa and temperatures above 2000°C. This is indicative of the upper limits of this system.

In addition to electrical resistance measurements, Mössbauer and X-ray diffraction studies have been carried out in tetrahedral presses. In either case, a portion of the tetrahedral sample container or the gasket is replaced by a light element (i.e., low X- or gamma-ray absorbing material) such as LiH, B, or B loaded epoxy. In some cases, the entire tetrahedron is made from these materials. Radiation reaches the sample either through a Be-plugged hole in the face of an anvil or through a gasket; the scattered radiation is scanned angularly through the other three or through one opposing gasket. In this mode, the device can be operated to pressure/temperature limits of about 8 GPa/600°C. Sample container sizes are typically about 2.5 cm on an edge.

*b. Cubic presses.* Following the tetrahedron, the next regular solid is the cube. Cubic symmetry is more easily implemented, and thus, there are a somewhat larger number of cubic presses in operation. As with the tetrahedral press, the six rams of the cubic press are linked together, either with massive tie-bolts or by hinges. Generally, the former offers easier access to the sample area, and, as with the tetrahedral press, the utilization of a guide mechanism speeds and simplifies operation. Other hexahedral presses have been built, generally with one sample axis slightly elongated. These are generally used in an attempt to extend the working volume at minimal cost.

The development of uniaxially powered cubic presses proceeded along two courses. In each case, the top and bottom anvils were driven by a uniaxial press, but they differed in how the four remaining anvils were to be powered. In one case, the top and bottom rams included large blocks with tapered internal faces that forced the side anvils in as the system closed. In the second system, often termed "DIA," large links were hinged to the rams and the four side anvils, thus generating the desired motion. This latter device has been favored in Japan in recent decades. Pressure and temperature capabilities are about the same as for the tetrahedral devices, while sample sizes range from a few millimeters on an edge to 5 cm. Higher pressures in most of these anvil devices can be attained with harder anvil tips; for this reason, some researchers are employing sintered diamond anvils.

*c. Multi-anvil sliding system (MASS).* Another large-volume concept that has been proposed but has found only limited application is that of the multi-anvil sliding system or MASS. The principle was independently proposed by R. Epain and M. Kumazawa. The principle of operation may be understood by examining the two-dimensional processes that are possible (Fig. 4).

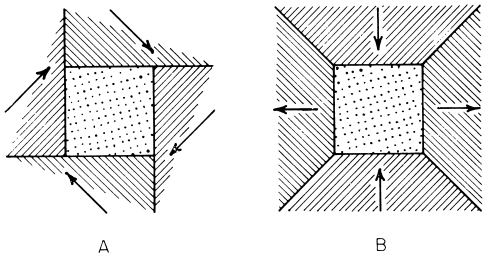



**FIGURE 4** Multi-anvil sliding system (MASS). The two basic forms are shown: A is rotational, where the anvils move tangentially to the central void enclosing a decreasing volume; B is irrotational, with some anvils retracting as the others advance to yield a decreasing volume.

A set of four anvils moves tangentially to enclose an ever-decreasing area. The same idea is carried over to three dimensions where a sample volume is contained: the anvil pieces must slide past one another easily and yet not allow extrusion of the sample. The tangential nature of the force application leads to the description of the first mechanism as rotational. The second mechanism controls the displacement of the two anvils during compression, and, since no rotation is involved, this concept is termed irrotational. Although there appears to be a definite mechanical advantage to MASS, implementation is not easy. Kumazawa has identified 48 possible MASS mechanisms, but most of these require complicated hydraulic rams and/or high load screw systems to operate. Further complicating the issue is that of extracting information from the pressure cavity during operation (i.e., the passage of electrical leads or radiation beam is not a simple matter).

*d. Split-sphere apparatus.* The last of the multiple-anvil systems to be described here also involves the concepts of massive support and compressive gaskets. The strength of the materials used in the anvils and support systems have been the limiting factor: ultimately, breakage occurs. In the case of the split-sphere apparatus, breakage is actually anticipated by cutting a sphere into six, eight, or more equivalent segments. Each segment is formed into an anvil, and the entire assemblage is reformed and held together with a flexible membrane (see Fig. 5).

Pressurization of the sample contained at the center of the split-sphere anvils is achieved by immersing the entire assemblage in a fluid that is, in turn, pressurized. The action of the pressurized bath is to force the advance of the anvils toward the sphere center. Very high pressures, on the order of 100 GPa, have been claimed with this apparatus. Although the pressures actually achieved are in dispute, the advantages of the technique are readily apparent. Power leads have been passed into the system to energize resistive heaters and diamond synthesis has been

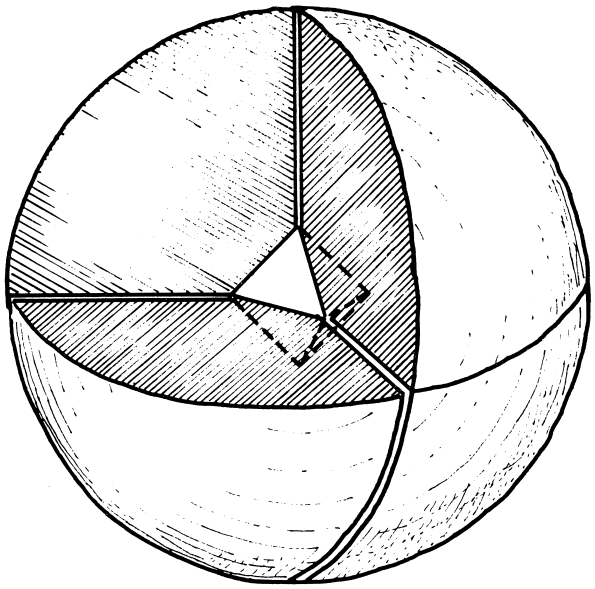


**FIGURE 5** Split-sphere apparatus showing a hardened metal sphere formed of eight equal volume segments, each having a triangular face impinging on the central, octahedral sample. One segment has been omitted for purposes of clarity in the figure.

achieved, thus indicating pressure/temperature conditions in excess of 6.7 GPa and 1500°C, respectively. Typically, the central cubic sample volume is ∼4 mm on an edge. Like the MASS, however, assemblage can often be a tedious task and the extraction of data from the pressure chamber can be difficult.

## 7. Diamond-Anvil Cells (DAC)

*a. Origin of the DAC.* The most powerful instrument for the performance of basic research with extreme static pressures is the diamond-anvil cell (DAC). It offers three distinct advantages over the older, larger systems.

1. It is compact—a typical DAC can be held in the hand and for this same reason it can be readily cooled or heated, as desired.
2. It is relatively inexpensive—a DAC can be constructed for a small fraction of the cost of the larger systems (for this reason, many modern high- pressure laboratories usually have several DACs in use).
3. Most importantly, the diamond anvils are themselves transparent to a broad spectrum of electromagnetic radiation; therefore, the pressure chamber can be readily probed in a variety of ways and samples can be readily studied *in situ*.

The concept of bringing flat surfaces of the hardest known material, diamond, into opposition for the purposes of creating a high-pressure chamber was employed independently and almost simultaneously by researchers at the National Bureau of Standards (NBS) and the University of Chicago in the late 1950s. During the following 15 years, extensive development took place at NBS as well as at other high-pressure laboratories, leading to a device that is capable of producing the highest static pressures, ∼500 GPa as of this writing.

Initially, the DAC was used to visually study phase changes in materials that were partially or totally transparent; even metals could be examined in reflection. This represented a natural extension of optical absorption investigations in the near infrared (IR) and the visible, often with the objective of quantifying earlier observations. The X-ray absorption cross section for carbon is 40,000 barns/atom for low-energy photons (energies $\leq$1 KeV), but it falls more than three orders of magnitude to $<$10 barns/atom for photons having energies $>$20 keV. The upshot of this is that high-energy X-ray photons can readily be used to study crystallographic structures at extreme pressures and, because the DAC can be so easily heated or cooled, over a wide range of temperatures as well.

Comparatively simple in design, the DAC consists of a pair of brilliant-cut, gem-grade diamonds with slightly enlarged culets, or tips (see Fig. 6). Differences in cell design have developed depending on the mechanism for anvil alignment, the method of force production and control, the temperature environment, and the experimental probe.

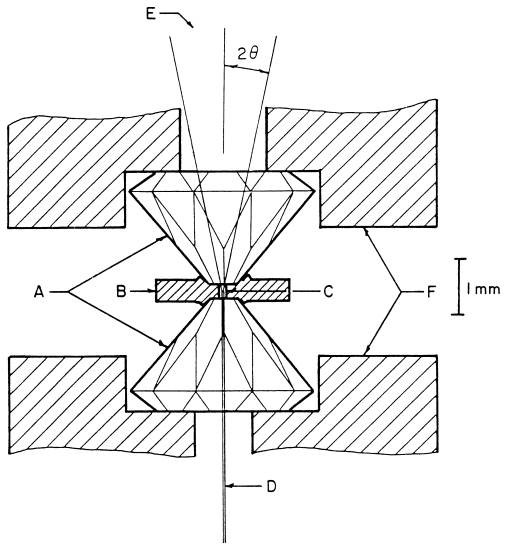



**FIGURE 6** Diamond-anvil pressure cell. A is the diamond anvils; B is the gasket; C is the pressure chamber; D and E are the incident and scattered radiation paths, respectively; and F is the hard WC support surfaces. The mechanism for advancing the diamonds and thus generating the pressure is not shown.

Some designs, such as a cell developed at the University of Rochester of single crystal X-ray studies, are quite simple, consisting of two triangular plates clamped together by three adjustable screws.

Critical in the design and operation of the DAC is the alignment of the anvils themselves. It is very important that the culet faces be parallel and normal to the uniaxial force direction. In earlier work, one anvil face was typically made much larger than the other, for example, 1.2 vs 0.6 mm diameters. This frequently led, at higher loads, to surface fracture of the larger diamond face due to penetration by the smaller. The modification to avoid this was to use a matched pair of diamond anvils with approximately equal surface areas and to align their axes as well as their planar surfaces.

The principle of massive support is being applied to the diamond anvils by the addition of culet faces that optimally make an angle of 5–10° with the culet flat; the load on these faces decreases as one travels radially out from the center. The highest pressure attained with these beveled anvils is on the order of 550 GPa, as of this writing.

In achieving these two alignment criteria, two designs have become somewhat generally adopted. One design has the anvils mounted in hemicylindrical rockers that can be translated along their cylindrical axes, which are set orthogonally in the cell body. Thus, axial alignment is achieved by appropriate translational adjustment, and facial alignment is achieved by appropriate rocking motions; both of these can be carefully controlled with small adjustment screws. In the other design, one diamond is mounted in a hemispherical support whose orientation can be carefully adjusted for the facal alignment; the other diamond is mounted in a flat plate whose position can be translated in a plane normal to the diamond axes, thus allowing the axial alignment.

It is important to maintain the alignment of the diamonds during operation of the DAC. This is usually accomplished by placing one diamond on a piston that slides in a closely fitting cylinder equipped with a guide pin to prevent rotation during loading. Some designs provide for alignment during operation, but this is generally unnecessary for sub-50 GPa work.

In operation, one of the anvils is usually kept stationary and the other is driven toward it. The magnitude of the force required for this is modest and can be developed in one of several ways. In the earliest design, a lever arrangement is used to provide a mechanical advantage of 3 or 4, and the force is manually generated through compression of a load screw, either directly of through a stack of Belleville spring washers. Similarly, in the three-bolt design referenced above, manual advancement of the screws will generate sufficient load to achieve the limits of the DAC. Other mechanical systems involving levers and pivots with screw adjustments, or a lever operated by a screw-driven wedge, have also proven effective. In order to achieve remote and/or programmable control of the load, the piston anvil can be driven by a hydraulically operated mechanism or by an electric stepping motor. One DAC designed for operation at cryogenic temperatures employs a metal bellows chamber pressurized with He gas cooled from room temperature; this system could achieve pressures of 10 GPa at temperatures as low as 30 mK.

As with Bridgman anvils, the pressure across the diamond faces is not uniform. If a sample is placed directly between the two anvils, as the load is applied, the sample will extrude laterally from the containment region until the frictional forces between the sample and the anvil faces become sufficient to contain the load. Under these conditions, the sample will be exposed to an extreme pressure gradient, ranging from a maximum near the center to almost atmospheric pressure at the periphery. This carries the added feature of subjecting the material under study to large shear strains. An advantage of this, however, is that the response of the material to a wide range of pressures can be examined at a single setting. In earlier work, phase transitions were detected optically with this technique. In more recent studies to the highest static pressures, researchers are using highly collimated beams of extremely intense radiation, for example, that produced with synchrotron storage rings, to study small portions of samples under pronounced pressure gradients.

It is more common in the operation of DACs, however, to employ a hardened metal gasket for containment of the sample. High-strength metals, such as stainless steel, Inconel, Waspaloy, or hardened Be–Cu, are often used. For purposes of alignment and extra hardening, the gasket is usually prestressed by compressing it between the anvils, frequently to about 50% of its original thickness. A hole with a diameter of $\frac{1}{2}$ to $\frac{2}{3}$ that of the culet face is then drilled in the center of the indentation. The indentation serves to allow reasonably rapid recentering of the gasket between the anvils. This cylindrical hole, whose linear dimensions are typically a few hundred micrometers, constitutes the sample chamber.

The sample of interest is then loaded in the gasket hole, frequently mixed with some standard material to serve as a pressure calibrant; NaCl or Au are frequently used calibrants for X-ray studies, while the wavelength shift of fluorescence from ruby is also a commonly used pressure gauge. Although the gasket will help in providing a more uniform distribution of pressure, if a truly hydrostatic pressure environment is required, the sample and calibrant must be immersed in a suitable fluid. A mixture of 16 parts methanol, 4 parts ethanol, and 1 part water will remain fluid to pressures just above 10 GPa at room temperature. For hydrostatic conditions at higher pressures, liquified gases are required. He, Ar, or $N_2$, condensed and

sealed at cryogenic temperatures, provide near hydrostatic conditions to pressures well above 10 GPa.

Another application of diamond anvils should also be noted here: researchers at Cornell University have been experimenting with a spherical diamond indentor with a tip radius of about 10 $\mu$m that is pressed into a flat diamond face. The sample is contained between the two diamond surfaces, and, as in the case of the ungasketed DAC, the sample will extrude to the point where frictional forces between it and the diamonds cause the sample to effectively form its own gasket. Since the contact area is controlled by the tip radius, contact force, and deformation of the diamonds, it is very small, and modest loads can produce extreme pressures: values in excess of 100 GPa have been claimed. One major drawback with this arrangement is that because the sample is small, it is difficult to detect. The reported pressures have been determined from the model used to represent the tip and flat, and therefore, reported pressures are vulnerable to the accuracy of the model.

*b. Temperature.* Elevated temperature studies with a DAC must be approached with the realization that the stable form of carbon at atmospheric pressure is graphite, not diamond. At normal temperatures though, the diamond-to-graphite transition is not observed because of the very high activation energy associated with this transformation. But, if sufficient thermal energy is provided, this activation barrier can be surmounted. Therefore, most high-temperature research with DACs is restricted to below 800°C.

High temperatures in a DAC have been attained by two methods: resistive heating and laser heating. In the former case either the gasket itself can be used as a heating element or, more commonly, resistive heaters are used to envelope the diamonds and gasket material. Temperatures are usually limited to below 200°C in the former and to below 800°C with the latter. Most recently, scientists from Los Alamos National Laboratory and the Naval Research Laboratory have collaborated to develop a DAC that is operated in a vacuum oven equipped with two concentric Ta heaters. Pressures in excess of 10 GPa have been attained at temperatures above 1200°C with this system.

The other heating technique employs a high-powered laser beam that is focused on the sample. Using a pulsed laser, a team at Cornell university has achieved temperatures in excess of 5400°C for brief periods of time, actually melting diamond, and using a pulsed YAG laser, temperatures of 2000°C have been sustained at pressures of 2.5 GPa.

Research has also been carried out with DACs at cryogenic temperatures. Researchers at the Naval Research Laboratory have attained temperatures down to 30 mK at pressures up to 10 GPa by coupling a DAC to a liquid-He dilution refrigerator. In this apparatus, the mixing chamber of the refrigerator is built directly into the DAC. A metal bellows chamber pressurized with He gas was used to generate the compressive force for the anvils.

*c. Experimental probes.* As noted above, the major advantage of the DAC is its virtual transparency to a broad spectrum of electromagnetic radiation. Some of the earlier studies in the DAC were made in the visible portion of the spectrum; refractive index changes, optical absorption, and birefringence are several types of measurements that were performed on samples as they underwent various pressure-induced phase changes.

Optical studies were quantified by incorporating the DAC into a spectrometer system, including appropriate focusing optics. Work in the infrared region is usually performed using type-II diamond anvils; these allow transmission studies to be carried out in the 1–4 and 5.5–15 $\mu$m regions, as well as the visible and near ultraviolet. Fluorescence and spectroscopic measurements, both absorption and Raman, have also been carried out in the DAC. An abundance of research has been performed in the higher photon energy regions as well: both angular and energy dispersive Bragg scattering measurements have been used to detect structural phase transitions, as well as measure thermal expansivities and compressibilities. Extended X-ray absorption (EXAFS) studies have been undertaken with limited success in the DAC, with the major difficulty being interference in the EXAFS patterns by Renninger scattering from the diamond anvils. Using appropriate gamma-rays, Mössbauer studies have also been performed in a DAC.

Magnetic susceptibility and microwave absorption measurements have been performed in a DAC; this work has been directed primarily toward superconducting materials. In this same context, some researchers have also equipped DACs with electrical leads extending into the pressure cavity, thereby permitting electrical resistance measurements under conditions of varied pressure and temperature. Additional details of the various measurements performed in a DAC are discussed in the later parts of this article.

## B. Dynamic Pressures

### 1. Fundamental Principles

The most obvious difference between static and dynamic pressures is that of duration: in most static systems, once the pressure is set, it is generally considered to be constant in time, whereas shock pressures result from large amplitude waves passing through matter and are necessarily of brief duration, typically on the order of microseconds. Another difference between the two techniques is that static pressures can be applied isothermally, whereas

shock waves are generally accompanied by large thermal excursions. Final, static pressures are three dimensional in character and often hydrostatic, whereas the shock pressure is often considered to be uniaxial, the result of a two-dimensional shock front passing through the sample. The development of pressure results from the inertial response of matter to a rapid acceleration. Although several different shock techniques will be discussed, all depend on the sudden application of force to a surface of the target for the initiation of a shock wave.

A shock wave represents a thin region, typically 1 nm wide, of a material over which there is a discontinuity in stress, density, material velocity, and internal energy. This region travels at supersonic velocity with respect to the material into which it progresses. Although impacts normally generate a region of rapidly increasing pressure, building to a maximum, a shock wave results because the pulse velocity increases with increasing pressure. This is a fundamental requirement for the establishment of a shock wave; it leads to a sharpening of the disturbing pulse to a steplike discontinuity. Since all forms of shock-generating force are short lived, or the target moves away from the disturbed region, a low-pressure region, or release wave, is launched into the target following the shock wave. The release wave travels in a denser material because of the preceding shock wave; therefore, it travels faster and eventually over takes and destroys the shock front.

Shock studies have depended heavily on the concurrent development of theoretical models or codes that allow the researcher to predict the effects of shock waves on specific materials. Materials are frequently treated as if they were liquids in these codes; however, the errors resulting from such approximations are generally small because of the extreme pressures involved (i.e., generally above 10 GPa).

The parameters that are typically measured in a shock experiment include the shock velocity, the particle velocity behind the shock front, and, more recently, the pressure and temperature. Within the experimental uncertainties, the shock velocity is usually found to be a linear function of the particle velocity.

Using the principles of conservation of mass, energy, and momentum across the shock front, a set of three equations can be derived that relate the final density, internal energy, and pressure (or stress) to the initial values for the material and the shock. The locus of values derived from a number of experiments forms a curve that defines the final states that can be reached for a given material in the shock velocity-particle velocity plane, or the stress-volume plane. This curve is generally referred to as the Hugoniot, since it defined the Rankine-Hugoniot equation of state for the material with respect to its initial state.

Loading methods will influence the precision of three parts of the shock experiment: (1) control of the pressure, (2) uniformity of the disturbance across the sample face, and (3) the decompression or release processes. The most common methods for the production of shock waves are detonation of explosives, either directly in contact or driving a flyer plate, and impact of a projectile from a gun. Exploding foils or wires and imploding magnetic fields have been used occasionally. Laser-generated shock waves are becoming more common, especially in the light of current national defense strategies. Electric rail guns promise interesting results at the upper pressure limits, but development has been sporadic. Finally, nuclear-driven shock studies have resulted in the highest dynamic pressures and have tested the extension of existing equation-of-state theories to pressures in the tens of tesla Pascal regime.

## 2. Experimental Techniques

*a. Explosively driven shock waves.* Interest in the effects of explosives on materials originated with military concerns for the consistent performance of these devices and later with the related issue of armor penetration. In the latter case, spalling became a major issue. This is the fracturing of part of the target on the opposite side from the impact due to a tensile failure in the interior where two release waves have met. Interest in this area was markedly increased by the need for predictable, well-controlled use of explosives in the original atomic bomb triggering sequence. It was this factor too that stimulated the accumulation of large amounts of data on the effects of shocks on various materials, and that forms the basis for many of the research programs currently underway.

There are two ways of explosively generating shock waves: one in which the explosive is in direct contact with the target and the other in which the explosive launches a flyer plate that then impacts the target. In the first case, it is important that the explosive impact reach all points on the front of the target simultaneously. Since detonation usually starts at a point, or along a line, the explosive burn will take place along a spherical or cylindrical front, respectively. Some commercial suppliers of explosives provide triangular sheet line generators that are perforated with an array of holes that serve to break up the curved shock front into a series of many smaller fronts that approximate a line.

One form of plane wave generator is called a "mousetrap" (see Fig. 7). It consists of a sheet of explosive material laid on a thin, inert (glass or metal) driver plate that is inclined above the main charge at an angle $\theta$, such that $\sin(\theta)$ equals $v/d$, where $d$ is the detonation velocity down the sheet over the driver plate and $v$ is the resultant velocity of the plate from the pressure generated by the detonated gases. Initiated at its upper edge, the driver plate strikes the main charge at all points simultaneously and

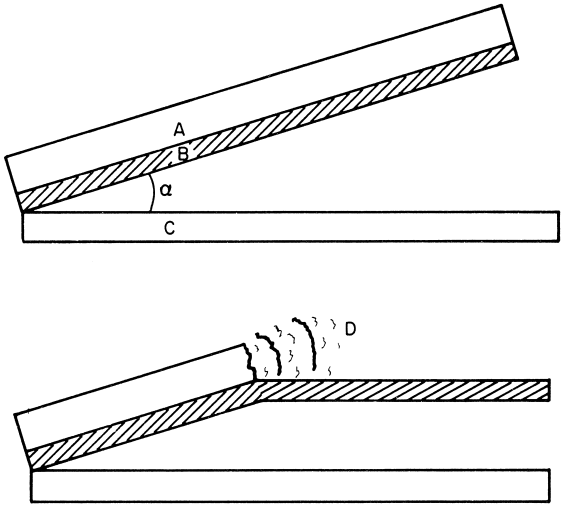



**FIGURE 7** Mousetrap plane shock wave generator. Explosive A is detonated along its upper edge and burns toward the hinge generating detonation products, D, and driving plate B toward the target, C, which may be a main explosive charge. The angle between the driver plate and the target is chosen such that its sine is equal to $v/d$, where $v$ is the velocity of the plate derived from the detonation of A at velocity $d$.

will initiate a plane detonation wave if the plate velocity is great enough. Edge effects and construction variations limit the planarity of this device.

Conical explosive lenses can be produced that will produce a simultaneity of detonation at the driving face to within 0.1 $\mu$sec, but the resultant impulse may be nonuniform. These lenses are formed either with a cone of explosive over an inert cone of larger angle or a similar inner cone of explosive with a slower detonation velocity (see Fig. 8). In the latter case, the base angle of the inner, slower explosive, $\alpha$, is determined by the ratio of the detonation velocities (i.e., $\sin\alpha = [d_{\text{out}}/d_{\text{in}}]$).

The target may be directly attached to the surface of the explosive, or it may be set a short distance away with a flyer plate attached to the explosive. The simplest form of the latter case is the mousetrap where the driver plate hits the target rather that a main charge.

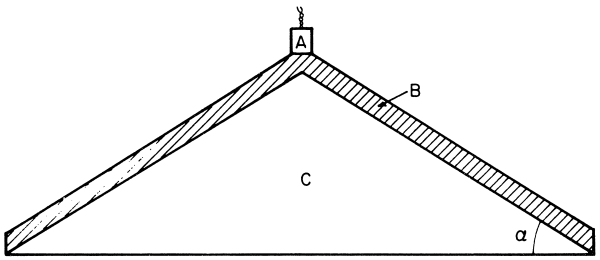



**FIGURE 8** Conical shock wave plate generator. The detonator, A, ignites the fast burning explosive, B. which in turn ignites the slow burning explosive, C. The conical angle is chosen such that its sine is equal to the ratio of the rate of C to the rate of B, resulting in the formation of a detonation front in C parallel to the base.

The pressures that can be obtained using these techniques range up to a few tens of gauss Pascals. The primary advantage is the relatively simple set up and correspondingly low cost.

*b. Guns.* Propellant guns, originally developed for military applications, provide a special example of the explosive-driven flyer plate. These also provide a somewhat more controlled mechanism than the explosive techniques described above. The resultant pressures derived from the impact of the flat-faced projectile on the target are somewhat greater than those attainable with the flyer plate.

Light gas guns offer greater control of the shock conditions. In these devices the projectile is propelled down the evacuated barrel toward the target by the expansion of a pressurized light gas, such as $H_2$ or He. The gas is suddenly released from its high-pressure reservoir by the rupture of a disk behind the projectile in the breech of the gun. These devices can also be operated in two stages, using a larger diameter projectile of compress gas for a second, smaller gun. The projectiles from these devices can achieve velocities as high as 7 km/sec, resulting in target pressures of over 150 GPa.

*c. Exploding wire or foil.* The force in this system is generated by exploding a thin metal foil by passage of an extremely high electrical current. The resistive heating of the metal will cause vaporization of the foil; this, in turn, accelerates an adjacent, thin dielectric plate to impact the target. There can be serious problems in the planarity alignment of the plate as it impacts the target with severe degradation of the resultant pressure. Associated electrical instrumentation is also heavily impacted by the electromagnetic noise generated by the system.

Exploding foils have been used to charge small guns with barrels only a few millimeters long. Using a flyer plate composed of plastic and metal, pressures in excess of 1 TPa have been achieved, with capabilities up to 5 TPa indicated. Such systems are simple in concept and relatively inexpensive to set up.

*d. Laser-driven shock waves.* High-energy laser beams impacting the surface of a target will cause very rapid heating, even to the point of forming a plasma. These very high temperatures are formed quite rapidly, and, because material or thermal flow is relatively slow, there is insufficient time to dissipate this energy with the result that shock waves are launched into the material. Although particle beams, for example, electrons, can also be used for this process, attention will be focused exclusively on laser-driven shocks.

One advantage of the laser initiation is that the energy deposited by the laser beam results in a shock wave that is initiated simultaneously over the irradiated area. Typically, the laser beam is focused to a spot size with a diameter of up to 1 mm. Irradiances greater than about $10^8$ W/cm$^2$ are required to ignite a plasma and produce a shock wave. These small impact areas lead to large energy losses due to two-dimensional expansion of the plasma. Edge effects will also be significant in evaluation of the shock. The facts that laser pulses normally last no more than a few hundreds of nanoseconds and the plasma dissipates very rapidly after the energy input ceases result in the launching of the aforementioned release wave, which rapidly overtakes and destroys the shock front. As a result of these constraints, effective sample thicknesses have been limited to a few tens of micrometers, although the pressures attainable with these systems can range up to 10 TPa. This pressure range is of considerable interest in the development of equations of state.

Not all of the laser energy is effective in the generation of the shock wave; some is lost in the formation of the plasma, some is carried off with the plasma, and some is reflected without effect. The absorption efficiency of most materials increases with the photon energy, ranging from about 30% in the near infrared to about 90% in the near ultraviolet. Short wavelengths offer another advantage, namely, the decrease in production of suprathermal electrons. This, in turn, lessens the electron preheat of the target prior to passage of the shock front. X-ray production, however, is enhanced at shorter wavelengths and for higher Z materials, which may cause some target preheating. Laser beams may have local fluctuations in energy density, or "hot spots," which can lead to nonplanarity of the shock front. These effects are minimized at longer wavelengths by thermal conduction, but are more pronounced at shorter wavelengths because the energy is deposited closer to the ablation surface.

The flyer plate concept has also been applied to laser-generated shocks. Carbon disks have been irradiated with a 3-nsec laser pulse, generating a shock pressure of 0.5 TPa and accelerating the disk to a velocity of 100 km/sec. Impact of this disk on a second disk delivers the energy in a much shorter time, resulting in the production of 2-TPa shock pressures.

A conventional shock diagnostic apparatus is electrical in nature. The intense electromagnetic storm generated by laser and particle beams is an extremely hostile environment of these sensitive detectors. Lasers, however, bring their own solution. Because the timing of the laser pulse can be precisely established, the incoming laser pulse can be used to trip optical diagnostics, or a portion of the laser beam itself can be used for diagnostic purposes at the target.

Plasmas also generate X-radiation with wavelengths varying according to the target composition. This radiation can be used either directly, by recording the shock on an X-ray streak camera, or indirectly, by exciting X-rays from a second target that are then used to monitor the shock passage. Shock velocities can be determined by monitoring the light generated when the shock reaches the back surface of the target using a streak camera.

*e. Nuclear-driven shock waves.* Nuclear explosions have been used to obtain equation-of-state data for many years. Pressures obtained with this method have approached 7 TPa, and although this seems to be less than those attainable with laser beams, the target area was nearly 30 cm in diameter and the nuclear device was detonated about 3.5 m away from the sample.

Optical techniques are preferred for the initial signal generation because the electromagnetic interference is extreme. Usually, the shock velocity is measured for a reference material and several samples mounted on the reference material target plate. Impedance matching is used to derive the particle velocities of the samples from the known equation of state of the reference material.

Access to nuclear tests is difficult, and although the technique appears to be comparatively simple, not only the sample, but most of the expensive signal processing equipment is lost in the process; thus, experiments are limited to materials of the greatest interest.

*f. Electric rail guns.* These devices represent a variation of the gas gun; however, here the force used to accelerate the projectile is electrical. The "gun" is made of two parallel, electrically conducting rails. The projectile, a conductor shorting the rails, rests between them. The system forms a linear dc motor, and the acceleration of the projectile is enhanced by the formation of a plasma behind it, thus aiding in its acceleration. Although several groups have experimented with this method, difficulties have prevented its fully successful implementation. Theoretical estimates indicate that ultimate impactor valocities as high as 40 km/sec should be achievable; these would lead to shock pressures up to 10 TPa.

## 3. Measurement Techniques

Evaluation of shock experiments requires the measurement of any two of the four shock variables: $U_p$, particle velocity; $U_s$, shock wave velocity; $P$, pressure; and $V$, specific volume. The three conservation laws (i.e., that of mass, linear momentum, and energy across the shock front) are then used to calculate the other two parameters:

$$\rho_0/\rho_1 = 1 - (U_p/U_s)$$
$$P_1 - P_0 = \rho_0 U_s U_p$$
$$E_1 - E_0 = [P_1 + P_0][V_0 - V_1]/2,$$

where $E$ refers to the energy, $\rho$ refers to the density, and the subscripts 0 and 1 refer to the variables ahead of and behind the shock front, respectively.

Shock velocities can be measured in several ways. The average valocity can be calculated from the sample thickness and by timing the impact and exiting of the shock front. Electrical pins that short upon the arrival of the shock front are used to obtain timing data. These pins can consist of an insulated conductor separated from a grounded metallic target or a metallic conductor in a plastic, insulated rod that is coated with metal forming a grounded sheath. The end of the rod is capped with a thin insulator, and the central conductor is charged with a dc voltage of up to several hundred volts. When the shock wave encounters the end cap, the center conductor is shorted to ground, generating a rapidly rising pulse on a line, which is then recorded. The separation of the bare contact from the grounded target or the thickness of the insulating cap is minimized, consistent with the need to prevent shorting for the level of readout voltage used. If closing time is to be no more than 100 nsec, then the gaps must be no more than 10 $\mu$m wide.

When electrical noise is a problem, optical pins, consisting of small microspheres containing a pressurized gas, such as Ar, are affixed to the end of optical fibers. As the shock front passes, the trapped gas emits light that is carried along the fibers to remote detection equipment. These pins can also be multiplexed onto a single streak camera record.

Pins of either type can be placed at various steps in the target or in holes of carefully determined depth. They should, of course, be separated by sufficient distance so that the release wave generated by one hole or step will not interfere with the readings of adjacent pins.

Electrically conducting targets can be used as one plate of a capacitor; in this case, the velocities are determined by measurement of the variations in capacitance as the target moves. This techniques provides information about the position, the free-surface velocity, and structure of the wave front.

Application of a strong, uniform magnetic field orthogonal to both the direction of shock propagation and to a conductor through which the shock will pass will generate an electrical voltage when the conductor is moved by the shock. The large magnetic field requirements limit more routine application of this technique.

The shock arrival at the rear surface can be detected optically for higher shocks because of the intense heating of the target; this produces a bright flash of visible radiation that can be detected with a streak camera. Since it may be difficult to establish the timing of the impact, a stepped target is often used that provides two signals of shock arrival with a well-defined separation, that, along with the timing information from the streak record, gives the shock velocity. This concept is also employed by using gaps that are prefilled with a gas, such as air, Ar, or Xe, or a liquid, such as $CCl_4$. These materials all emit light when impacted by a shock wave; the brightness of the light increases with the intensity of the shock, thus making this technique useful in the lower pressure regions. If lucite is used as the transparent rear cover for the flash gap, it becomes opaque when the shock reaches it, quenching the signal to the recording device.

Some polished surfaces change their reflectivity when shocked. This allows the use of mirrors at the impact and rear surfaces of the target. A streak camera is used to record the shock events. The shock-opacity principle can be applied here also: when the shock wave reaches the rear surface, the reflectivity of the polished surface is greatly altered and readily detected.

Optical measurements are frequently carried out in evacuated systems in order to minimize the effects of air shocks around the rear or free-surface of the target. Another type of optical technique involves using the mirrored rear surface of the target as an element in and optical lever system. With this method, a streak camera is used to record reflections of a series of point light sources recording the passage of a shock wave, or the motion of the image of a fine wire by the moving free surface. If a liquid cell is used, a thin foil may be suspended in it at some angle to the expected shock front. The shock velocity in the fluid can be determined, if the fluid remains transparent and the foil remains reflective when shocked.

Lasers, because of their brightness, monochromaticity, and coherence, have led to many new measurement techniques. They can be used as light sources in those techniques in which brilliance is important. Their monochromaticity allows them to be used as interferometers for velocity measurements. Interference fringe patterns are established when partially reflected light from the front surface of the target interferes with light reflected from the moving face at the rear. After the effects of the initial shock pass, the steady fringe reading corresponds to the particle velocity in the window material. The timing of the passage of the shock through the window material gives the shock velocity. These two pieces of data define a point on the Hugoniot curve.

Since X-ray interactions with matter generally involve low-level electronic processes, they are usually considered to be unaffected even by severe shock waves. Pulse X-ray sources are needed to study transient

phenomena. Laser-generated plasmas have been used as pulse sources.

Since pressure is also one of the primary variables, several methods have been used to obtain estimates of its value in the shock front. The variation of the electrical resistance of manganin wire has long been used as a means of measuring static pressures; the linearity of this variable with pressure has been shown to be valid to 30 GPa under dynamic conditions. The resistance of other materials will also vary with pressure, but their resistance–pressure curves are either not linear or they involve large temperature coefficients. Both tourmaline and quartz have been used as piezoelectric pressure transducers. They have been calibrated up to 2 GPa, probably the limit of their elastic behavior. Polyvinyl fluoride is a polymer film that can be poled to give a piezoelectric gauge, which is finding application. Ferroelectrics, such as lead zirconate titanate, have also been used.

Another electrical technique is based on the observation that many materials either lose polarization or become polarized upon the passage of a shock. In the first case, electrodes on the two faces of the material being shocked will generate a voltage through an external resistance; the value of the voltage will be proportional to the pressure and sample area and inversely proportional to the sample thickness.

Since each of these techniques requires electrical contact to sensors in the shock regime, they are each prone to many difficulties, including shorting by shock-induced conductivity in the gauge or support material, or loss of contact.

A number of other shock phenomena have been noted in shock studies; some of these are unique. Two examples are phase transformations, either reversible (and therefore present only in the compressed state) or metastable (such as the graphite-to-diamond transition), and changes in conductivity, particularly of semiconductors and insulators that become conducting in the vicinity of the shock. Shifts in the Curie point with both pressure and temperature lead to shock-induced demagnetization of ferromagnetic materials. Luminescence has been observed for some materials, while some transparent materials such as polymethyl methacrylate (PMMA) and NaCl become opaque during shock compression. In some cases, the mechanical strength of a substance increases markedly upon the passage of a shock. Post-shock studies span the full range of possibilities and are considered below.

## III. MEASUREMENT OF PRESSURE

### A. Absolute Pressure Scales

An absolute determination of pressure requires a knowledge of either the force and the area over which it is applied or the variation of the free energy with change in volume. The free piston gauge, or dead-weight piston gauge, makes use of the former. Presuming an exact knowledge of the piston area, the force is fixed by a set of calibrated weights balanced on the piston. Friction between the piston-cylinder interface is minimized by rotation of the piston.

Several techniques are used to compensate for distortions at elevated pressures: (1) extension of the cylinder under pressure is restricted by a separately, pressurized jacket surrounding the cylinder; (2) calculations of changes from engineering principles; (3) comparison of two identical systems using different materials. These systems are limited in pressure by the strength of the materials used in their fabrication; with tungsten carbide, measurements are routinely reported to 2.5 GPa, although some work has been carried to pressures as high as 6 GPa. In a recent program to compare pressure scales between 13 international metrology laboratories, variances in raw data amounted to no more than 78 ppm. Although this is within the combined uncertainities, the derived calibration constants (i.e., the slopes of calibration curves) disagreed in a more marked manner, suggesting fundamental differences, particularly between the controlled displacement and other systems.

A thermodynamic pressure scale, similar to the thermodynamically defined temperature scale, would be another approach to an absolute pressure scale. Although this has been proposed, it has not been implemented as of this writing. The volume can be defined as the variation of the free energy, $G$, with pressure at constant temperature. Using an electrochemical cell, this is also proportional to the variation of the cell potential $\phi$, with pressure; that is,

$$V = [(\partial G/\partial P)_T] = -n(\partial\phi/\partial P)_T,$$

where the parameter $n$ is dependent on the particular electrochemical cell employed. This method has been infrequently used. One of the more difficult problems is the variation in the ionic conductivity in the cell as the pressure is varied.

### B. Secondary, Practical Scales

Like its thermodynamical counterpart, temperature, pressure is most frequently determined from the variation of some previously calibrated state coordinate. Variations in crystallographic volume, electrical resistivity, or fluorescence wavelengths have all been used with success.

#### 1. Equations of State

The equation of state (EOS) of a material relates the three thermodynamical variables, pressure, volume, and temperature, over a range of each. Although work has been

underway to develop such equations from first principles calculations (i.e., from basic physical laws and interatomic potential functions), most materials today are treated with semi-empirical EOSs. These employ one or a few parameters derived from measurements, for example, the volume dependence of the Grüneisen parameter.

Shock wave data have been used for EOS development at much higher pressures and temperatures than those encountered in static experiments, for example, to thousands of Kelvin temperature and hundreds of gauss Pascals pressure. Researchers have carried out combined studies on the same materials at the upper end of available static pressures and the lower end of the dynamic pressure range to provide reliable calibration materials.

## 2. Calibrated Fixed Points

Structural phase transitions that occur in materials and can be readily detected from discontinuous changes in their physical properties are often used as specific pressure calibration points. For example, the freezing pressure of Hg at 0°C, which can be detected from a discontinuous change in the volume, has been accepted to be 0.75692 GPa for many years. Also, Bi exhibits discontinuous changes in its electrical resistivity at room temperature and 2.5499 and 7.7 GPa. One difficulty in employing these phase transitions is that they can sometimes be sluggish and can be affected by the hydrostatic nature of the pressure.

## 3. Other Pressure Scales

*a. Ruby fluorescence.* An excellent secondary pressure scale that has become very popular with the extensive use of the diamond-anvil cell is one based on the pressure-induced shift in the wavelength of the $R_2$ fluorescence peak from ruby. Since temperature will also cause a shift in this wavelength, a thermal correction factor of 0.068 Å/K must also be applied. Based on EOS studies with a number of materials (NaCl to 30 GPa and Ag, Cu, Mo, and Pd to 100 GPa), the ruby line shift is given by

$$P(\text{GPa}) = 380.8\left[(\lambda/\lambda_0)^5 - 1\right],$$

where $\lambda$ and $\lambda_0$ are the ruby fluorescence wavelength at elevated and atmospheric pressures, respectively. Based on Au and Cu EOS studies, the foregoing calibration has been found valid to 200 GPa to within 5%. Most recent experiments in a diamond-anvil cell have led to and extension of the ruby scale to ~500 GPa, although the ruby signal tends to be obscured by diamond fluorescence in the 150–300 GPa range.

*b. Electrical resistance.* Calibrated electrical resistance sensors have long been used as temperature gauges. Similar devices have also been introduced as pressure sensors. The ideal material for pressure calibration would have a small thermal coefficient of resistivity, a minimal hysteresis, and preferably a linear pressure response over as large a range as possible. Manganin, a Cu–Mn alloy, is a material that has long been used for this purpose. Unfortunately, the pressure coefficients tend to vary somewhat from sample to sample, and therefore, each lot must be separately calibrated. Manganin pressure gauges have been used in static systems to pressures of 6 GPa. They have also been used in shock work, but significant corrections for shock temperatures and errors derived from the creation of point defects tend to limit the accuracy of the results.

Other materials used for this purpose include Au–Cr with about 2% Cr and zeranin, formed of Cu, Mn, and Ge. These materials were originally used as resistance standards. The first has an appreciable temperature coefficient that limits its utility; the latter has not been in use for as long, but seems to have the same desirable qualities as manganin in addition to a more rapid recovery and a higher resistance to oxidation.

*c. Superconducting transition temperatures.* Early calibration techniques at cryogenic temperatures grew directly from interest in the effect of pressure on superconducting transition temperatures, $T_c$. These measurements can either be performed by monitoring the sample resistance, which drops to zero at $T_c$, or the magnetic susceptibility, which will change abruptly at $T_c$ due to the Meissner effect.

In, Pb, and Sn have all been calibrated and used in this regard. However, one difficulty in this work is the lack of other low-temperature calibrants against which to fix the $T_c$'s. Typically, a known fixed pressure is sealed in a bomb at room temperature, after which the system is cooled to determine the shift in $T_c$. The problem is that there is likely to be a change is sealed pressure due to differences in the thermal expansivity of the various components of the system. It is difficult to accurately correct for this change.

*d. Semiconductors.* Since the resistivity of semiconductors is also sensitive to pressure, they too can be used as pressure calibrants. These have the advantage that they will typically be significantly more responsive than metallic resistors; however, they also tend to have large temperature coefficients. In one case, Sb-doped Te has been shown to have an exponentially decreasing resistivity up to 0.65 GPa: specifically, the ratio of the resistance at pressure $P$ to that at atmospheric pressure has been reported to be $\exp[-1.122\,P(\text{GPa})]$. Doping GaAs has the effect of decreasing its temperature sensitivity while increasing its pressure sensitivity to a value approaching that of manganin.

*e. Capacitance.* Pressures can also be determined from variations in capacitance. In this case the dimensions of the solid dielectric used in the capacitor will change in both area and thickness as a function of increasing pressure. These gauges will typically have a resolution as small at 0.07/MPa. However, they too must be operated under conditions of extreme temperature control and necessitate at least three electrical contacts into the pressure chamber. These requirements, along with their relatively low range of applicability, have limited the use of such gauges.

## IV. RESEARCH AT HIGH PRESSURE

High-pressure research involves examining the response of materials to the environment of extreme pressures. This examination, usually in the form of a measurement of one or more physical parameters, can either be made *in situ* (i.e., while the pressure is changing) or upon completion of the pressure excursion. Although the latter experiments are easier to perform, since they do not require access to the high-pressure environment during pressurization, they have the undesirable drawback of requiring a metastable product.

The objective of most of this research is to assess the effect of altering interatomic distances on a variety of physical phenomena, for example, local, short range order; crystalline structure; electrical resistivity; and mechanical, magnetic, and optical properties. In short, most phenomena that can be measured at atmospheric pressure can also be studied at elevated pressures.

### A. Structural Measurements

One of the most fundamental properties of a condensed matter system pertains to the arrangement of atoms of which that system is formed. In the majority of materials, this is the crystallographic structure. A wealth of high-pressure research is directed toward the examination of changes in crystal structure under pressure. Such studies generally involve the application of basic X-ray diffraction techniques. A material is illuminated with either monochromatic or heterochromatic X-rays, and the scattered photons are analyzed to determine crystal structure. This can be accomplished through utilization of Bragg's equation:

$$Ed_{hkl} \sin\theta = hc/2,$$

where $E$ is the energy of the X-ray photon, $d_{hkl}$ is the interatomic planar spacing, $\theta$ is the Bragg diffraction angle, $h$ is Planck's constant, and $c$ is the speed of light. If monochromatic radiation is used, then $E$ is fixed and the scattered radiation must be analyzed over a spatial range to determine the values of $\theta$ for which the Bragg equation is satisfied. Alternately, if heterochromatic or white radiation is used, then the diffraction geometry (i.e., $\theta$) is fixed and the scattered photons are analyzed in terms of their respective energies.

An important experimental detail in this work is to provide a window into the pressurized region with a relatively low absorption coefficient for X-rays. As noted above, the diamond anvils of the DAC satisfy this requirements very well: the X-ray absorption cross section for carbon decreases by a factor of 4000 over the photon energy range from 1 to 20 keV. Other materials that also have been used as high-pressure X-ray windows are pryophyllite and $B_4C$. The latter material has the advantage of being amorphous and therefore will not contribute substantially to the diffraction pattern of the sample.

Nevertheless, it is often difficult to bring a sufficient number of X-ray photons in and out of the pressurized environment in a short time period. Consequently, X-ray diffraction experiments can require very long exposure periods. Typically, tens to hundreds of hours are required for a single measurement with a DAC when conventional X-ray tubes are employed.

One means of accelerating this process is to employ a much brighter source of radiation. In the past few years, most of the high-energy synchrotron storage rings in the world have been used for high-pressure structural research. Since the X-ray flux available with these machines is many orders of magnitude greater than that available with conventional radiation sources, the measurements can usually be completed in much shorter time intervals; exposure periods of minutes or seconds are typical. In addition to speeding up the entire process, these brighter X-ray sources also permit measurements that would not otherwise be feasible, for example, phase transition kinetics. Experiments performed recently at the Stanford Sychrotron Radiation Laboratory involved monitoring the structure of a number of alkali–halide salts in 1-min time intervals as they were driven through a first-order structural phase transition with increasing pressure. Parameters in the equations describing these phase transitions can be determined from these measurements.

In the quest for higher pressures, researchers using DACs are employing diamond anvils with beveled tips; the consequence of this, in addition to pressures in excess of 500 GPa, is a further diminution of the sample volume. Thus, even brighter X-ray sources will be needed for future work. Efforts have been made to employ devices inserted in the synchrotron rings to further increase the emitted photon flux, so-called wigglers and undulators, for materials studies on samples contained in a region only a few microns in diameter.

Another type of high-pressure X-ray measurement that has recently been advanced through the utilization of synchrotron radiation sources is EXAFS, or extended X-ray absorption fine structure. Analyses of X-ray absorption data is complementary to standard crystallography in that it allows determination of nearest and next-nearest distances of specific atoms, coordination numbers, and thermal vibrational properties. A difficulty encountered in EXAFS measurements performed with a DAC is the presence of Bragg or Renninger scattering in the absorption spectrum. To circumvent these problems, amorphous materials such as $B_4C$ are used to contain the pressure and provide and X-ray window.

These structural measurements are pursued for several scientific reasons: to determine compressibilities or equivalently bulk moduli, to detect structural phase transitions, and to identify new crystalline phases. Frequently, experimental work is closely coordinated with theoretical model calculations, the latter often predicting the existence of possible new and interesting phases.

## B. Electrical Measurements

### 1. Nonsuperconducting Materials

Electrical measurements are usually directed toward monitoring the electrical resistance of a sample under pressure, although other measurements may also require the incorporation of electrical connections to the pressure chamber. For example, the pressure dependences of elastic moduli have been measured ultrasonically with transducers bonded to the pressurized sample. The most difficult aspect of this work is providing a feed-through that will not significantly attenuate the electrical signal and will contain the pressure. In some cases the anvils themselves are used as probes, for example, as noted above, tetrahedral presses are constructed with each of the four rams electrically isolated, thus four-probe resistance measurements can readily be carried out.

The electronic properties of solids are related to the characteristics of the ground state and various excited state energy levels. Since these levels can be affected by pressure, much of the research in this area is concerned with how the states will move under pressure. Professor H. G. Drickamer of the University of Illinois has introduced the phrase "pressure tuning of electronic energy levels."

An area that is currently receiving considerable attention in this regard involves pressure-induced metallization. Although this phenomena has been studied in a number of materials, perhaps none holds more interest than the prospect of metallizing hydrogen. Originally predicted by Wigner and Huntington in the 1930s, recent model calculations indicate that at elevated pressures, hydrogen will undergo two transitions from its highly ordered diatomic insulator state, first to a nonmetallic monatomic structure, followed by a transition to a metallic state. Unfortunately, there are many possible high-density phases for hydrogen, all with very similar values for the free energy. Therefore, it is difficult to forecast, with certainty, the properties of metallic hydrogen or whether it may be metastable under normal pressure and temperature conditions once it is formed. The imagination and excitement of scientists throughout the world, however, have been captured by some serious theoretical predictions of very high temperature superconductivity in metallic hydrogen . . . if it can be created.

As a prelude to this, other systems that are more easily studied are being examined. Iodine, for example, at pressures of 13 to 17 GPa undergoes an insulator–metal transition, and at about 21 GPa it converts from a diatomic to a monatomic conductor. Similar studies on other halides would also seem appropriate (i.e., the metallization and subsequent dissocation in $Br_2$, $Cl_2$, and possible $F_2$). Materials that have been metallized under pressure include BrI, HI, CsI, BaTe, BaSe, BaS, and BaO.

Another interesting pressure-induced electronic phenomenon involves valence changes. The valence states in Yb, Eu, Pr, Sm, and Ce can all be shifted through application of pressure. In Ce, this valence change manifests itself in a very interesting, isomorphic phase transition. As the material changes valence state, there is a first-order phase transition with approximately a 15% volume change but no change in crystalline structure (i.e., each phase is face-centered cubic).

### 2. Superconducting Materials

The practical applications of superconducting materials are almost limitless. Power transmission lines; electrical motors and generators; magnetic levitation for, for example, high-speed transit systems, and computer electronics are but a few. A major problem limiting the utilization of superconducting wires is the simple fact that the best known superconductor, $Nb_3Ge$, has a transition temperature, $T_c$, no higher than 23.2 K. Therefore, any realistic application of superconductivity must also include an appropriate means for cryogenic refrigeration, and therefore, in most cases, the gains that would otherwise be realized are obviated by the temperature demands.

Since there is no theoretical reason why the cooperative electronphonon coupling mechanism that is required for normal (BCS) superconductivity cannot take place at higher temperatures, there are major research programs underway worldwide to discover new materials with higher $T_c$'s. Guidelines for these efforts can be

either empirical or theoretical. As an example of the former, in V-based $A_3B$ compounds that crystallize in the A15 structure, such as $V_3Sn$, $V_3Ge$, and $V_3Si$, it has been noted that a plot of $T_c$ vs $m_B^{-1/2}$, where $m_B$ is the mass of the B-atom, tends to be linear. On the basis of this it has been predicted that the $T_c$ of $Nb_3Si$ should surpass that of $Nb_3Ge$. Since the A15 phase of $Nb_3Si$ is expected to the denser than its normal $Ti_3P$ structure, it is expected that pressure should be favorable to conversion of $Nb_3Si$ to the A15 structure. Shock experiments have produced this transformation. Unfortunately, the measured $T_c$ is not as high as expected; the explanation is believed to lie in defects introduced during the shock conversion.

One of the most promising classes of materials in terms of potential high-$T_c$ superconductors is the recently discovered organic salts. These are often one dimensional or quasi-one dimensional and are promising because they do not involve the usual electron–phonon (BCS) mechanism for superconductivity. Rather, the electron–electron interactions are mediated through excitations of a Peierls ground state or excitons. The first organic superconductors are from the ditetramethyltetraselenofulvalenium $(TMTSF)_2X$ family, where $X$ represents a suitable anion.

At room temperature and ambient pressure, all of these salts, independent of the anion, exhibit the same crystallographic structure, electrical resistivity, and thermoelectric power. However, at low temperature and/or high pressure, a variety of ground states may exist. Many of these salts have a spin density wave at low temperature and ambient pressure, but at elevated pressures of 0.6–0.8 GPa, they become superconducting with a $T_c$ of 1.2 K. One of these salts, $(TMTSF)_2ClO_4$, becomes superconducting at ambient pressure with about the same $T_c$.

Other TMTSF-salts exhibit high-temperature metal–insulator transitions associated with anion ordering. However, it has been demonstrated that pressure can be used to effectively suppress these transitions, thereby leaving the material in the metallic state at low temperatures. In these cases the materials will undergo a superconducting transition. These salts also appear to pass through a glassy phase in which the resistivity coefficient is negative at intermediate temperatures; they then undergo superconducting transitions at lower temperatures. One of these, $(TMTSF)_2FSO_3$, exhibits the highest values of $T_c$ known to date for organic compounds, $\geq 3$ K.

The variables that determine under what conditions superconductivity occurs or the nature of the "glassy" phase are not at all understood. The facts that the proposed spin density wave ground state of these systems is very sensitive (1) to pressure and (2) to the type of anion used suggest that this is a structural issue. For this reason, it would be important to carryout extended high-pressure, single crystal, X-ray diffraction experiments on many of these organic salts at cryogenic temperatures. It is believed that such information would prove extremely useful in characterizing the nature of the origin of superconductivity in these materials.

### C. Melting/Freezing Phenomena

The transition between the condensed and liquid phases of matter is perhaps one of the most important and least understood in the field. It is a problem that touches a broad cross section of disciplines: condensed matter physics, rheology, metallurgy, and the geosciences. From a technological viewpoint, an understanding of these issues is very important for a variety of materials-related industries, for example, those dealing with semiconductor devices, ceramics, and optical components. For example, it is still not known why pressure will enhance the crystal growth rate in some systems.

The pressure dependence of the melting temperature, $T_m(P)$, has been measured for a wide variety of materials. It is generally found that $dT_m/dP$ is positive and the $T_m(P)$ curve is usually fit to some phenomenological or empirical relation, for example, the Lindemann equation. Although recent advances in first-principles calculations of Gibbs free energies from effective interatomic or intermolecular potential functions have been most encouraging, a microscopic theory explaining the solid–liquid transition is still lacking.

An area related to this concerns crystallization phenomena in amorphous solids, for example, metallic glasses. "Met-glasses" represent an important new class of materials with certain improved physical properties, for example, superior radiation and corrosion resistance. It has been demonstrated that hydrostatic pressure can raise the crystallization temperature of metallic glasses by 10–20°/GPa; this effect is still not understood. Moreover, detailed structural analyses of high-pressure crystallization phenomena have yet to be carried out. A number of outstanding questions in this area remain to be answered.

1. Why does pressure inhibit crystallization? Is this related to the instability of favorable crystalline phases thought to allow formation of metallic glasses from the melt?

2. Are the crystalline phases produced at high pressure different from those formed at ambient pressure, and, if so, what does this imply?

3. What are the changes that occur on cooling and release of pressure, and can they be understood in terms of more subtle structural changes that occur during heating?

4. What is the crystallization nucleation mechanism, and are there precrystallization phenomena that may shed light on the mechanisms involved?

### D. Materials Modification by Shock Treatments

It is recognized that the state of some materials can be significantly altered through application of shock pressures. Although no comprehensive theory has yet been developed to explain this phenomenon, it is presumed that the rapid, massive shear deformation induced by the shock waves and the concomitant defect state that follows is responsible for what has been perceived to be a unique state of matter in the post-shock material.

Anomalous behavior in post-shocked materials has been identified in a number of separate areas: enhanced reactivity in, for example, structural phase transitions, chemical reactions, and sintering processes; enhanced atomic migration as seen in radioactive tracers and thin films of materials deposited at interfaces; shock-induced polarization in polymers and ionic crystals; shock-induced opacity in optical materials; bleaching of color centers; formation of color centers; anomalous shifts in absorption bands; saturation of dislocation densities without deformations.

It is likely that many, if not all, of these anomalies are related to the massive defect state left in the wake of the shock wave. A wide variety of micro- and macrostructural effects have been observed. Dislocation multiplication, twinning, and void formation are examples of some of the small-scale effects; spallation and flow are prominent among large-scale effects of shock loading.

Shock effects influence the surface and near-surface characteristics of materials. The surface hardness of 2024 Al alloys was found to increase with increasing shock pressures produced by flyer plates. Explosive welding leads to hydrodynamic and thermodynamic interpenetration of colliding metallic surfaces. Compaction of powders, which have high strength and resist ordinary consolidation or are chemically unstable and thus cannot be sintered, can be accomplished with shocks. Interiors of the particles remain relatively cold, while the surfaces are heated to cause interdiffusion, welding, or melting. More ductile materials may be used to bond together stronger particles or fibers in order to form composite materials. A strong and oscillatory dependence on hardness of Au–Ge alloys has been observed for shock treatments with durations of 0.1–1 $\mu$sec.

Although the precise mechanisms are not yet understood, there is clear evidence to indicate that shock waves do, in fact, produce an altered state in many materials. The classic means of subjecting materials to extreme shock states is chemical explosive techniques: detonation of contacting explosives, impact by explosively driven flyer plates, or impact by gun-driven projectiles. However, shock conditions can also be achieved by subjecting materials to intense pulses of radiation, for example, from electron beam accelerators, X-ray or neutron sources, or high-intensity lasers. The pulsing methods hold an important advantage of relatively high repetion rate, in contrast to methods dealing with explosives.

Perhaps an even more important advantage of laser-driven shock studies is that the same laser pulse can be used to produce both a shock and a high-temperature plasma that will emit X-rays to probe the shock. Exquisite timing will be possible using time-of-flight delay methods. These advantages suggest that laser-driven shock states may be uniquely useful in some material processing or testing purposes. Hardening, welding, and compaction have been demonstrated using shocks generated by explosives; the possibilities of using lasers for these purposes has yet to be assessed.

## SEE ALSO THE FOLLOWING ARTICLES

DENSE MATTER PHYSICS • DIAMOND FILMS, ELECTRICAL PROPERTIES • HIGH-PRESSURE SYNTHESIS (CHEMISTRY) • PULSED POWER SYSTEMS • SUPERCONDUCTIVITY

## BIBLIOGRAPHY

Akimoto, S., and Manghnani, M. H. (1982). "High Pressure Research in Geophysics," Center for Academic Publications Japan, Tokyo.

Bridgman, P. W. (1958). "The Physics of High Pressure," G. Bell and Sons, London.

Drickamer, H. G., and Frank, C. W. (1973). "Electronic Transitions and the High Pressure Chemistry and Physics of Solids," Chapman & Hall, London.

Ferraro, J. R. (1984). "Vibrational Spectroscopy at High External Pressures," Academic Press, Orlando, FL.

Hazen, R. M., and Finger, L. W. (1982). "Comparative Crystal Chemistry: Temperature, Pressure, Composition and the Variation of Crystal Structure," Wiley, New York.

Jayaraman, A. (1983). "Diamond anvil cell and high pressure physical investigations," *Rev. Mod. Phys.* **55,** 65.

Minomura, S. (1985). "Solid State Physics under Pressure," KTK Scientific Publishers, Tokyo.

Schilling, J. S., and Shelton, R. N. (1981). "Physics of Solids under High Pressure," North-Holland, Amsterdam.

Skelton, E. F. (1978). "High Pressure Science and Technology in Japan," Office of Naval Research, Arlington, VA.

Spain, I. L., and Paauwe, J. (1977). "High Pressure Technology," Dekker, New York.



# Impedance Spectroscopy

**J. Ross Macdonald**
*University of North Carolina*

I. Short History of Impedance Spectroscopy
II. Categories of Impedance Spectroscopy: Definitions and Distinctions
III. Elements of Impedance Spectroscopy
IV. Applications

## GLOSSARY

**Admittance** A complex quantity usually symbolized by $Y = Y' + iY''$. It is the inverse of impedance and is sometimes called complex conductance. Here $i = +(-1)^{0.5}$, and the single and double primes denote in-phase and quadrature components, respectively.

**Complex dielectric constant** The ratio of the (complex) dielectric displacement to the small-signal AC electric field that induces the displacement. Conventionally written as $\varepsilon = \varepsilon' - \mathrm{i}\varepsilon''$. It is given by $Y/(i\omega C_C)$, where $C_C$ is the capacitance of the empty measuring cell.

**Complex forms** Impedance spectroscopy data may be expressed in two different forms. Rectangular: $I = I' + iI''$, where $I'$ and $I''$ are the real and imaginary parts of $I$, respectively; or Modulus: $I = |I|e^{i\phi}$, where $|I|$ is the modulus, or absolute value, of $I$ and $\phi$ is its phase angle, or argument. Note that the complex conjugate of $I$ is $I^* = I' - iI'' = |I|e^{-i\phi}$.

**Complex modulus** $M = M' + iM''$. It is the inverse of the complex dielectric constant and is also equal to $i\omega C_C Z$.

**Debye length** A characteristic length that determines the extent of a space charge region near a discontinuity. It depends on temperature, dielectric constant, and the valence numbers and bulk concentrations of the mobile charges present. The diffuse double-layer capacitance present near a non-ohmic electrode is inversely proportional to the Debye length.

**Immittance** A general term denoting any of the four basic impedance spectroscopy response quantities: $Y$, $Z$, $\varepsilon$, or $M$.

**Impedance** The ratio of a sinusoidal voltage, applied across two terminals of a measurement cell, to the sinusoidal component of the current flowing between the terminals that results from the applied potential difference. Unless the system is purely resistive, impedance is a complex quantity because the current will have a different phase from the applied voltage: $Z = Z' + iZ''$.

**IMPEDANCE SPECTROSCOPY (IS)** is a general term that subsumes the small-signal measurement of the linear electrical response of a material of interest (including electrode effects) and the subsequent analysis of the response to yield useful information about the physicochemical properties of the system. Analysis is generally carried out in the frequency domain, although measurements

are sometimes made in the time domain and then Fourier transformed to the frequency domain. Impedance spectroscopy is by no means limited to the measurement and analysis of data at the impedance level (e.g., impedance vs. frequency) but may involve any of the four basic immittance levels; thus, most generally, IS stands for immittance spectroscopy.

## I. SHORT HISTORY OF IMPEDANCE SPECTROSCOPY

Since IS deals directly with complex quantities, its history really begins with the introduction of impedance into electrical engineering by Oliver Heaviside in the 1880s. His work was soon extended by A. E. Kennelly and C. P. Steinmetz to include vector diagrams and complex representation. It was not long before workers in the field began to make use of the Argand diagram of mathematics by plotting immittance response in the complex plane, with frequency an implicit variable. Electrical engineering examples were the circle diagram introduced by C. W. Carter (1925) and the Smith-Chart impedance diagram of P. H. Smith (1939). These approaches were soon followed in the dielectric response field by the introduction in 1941 of the Cole–Cole plot: a plot of $\varepsilon''$ on the $y$ (or imaginary) axis versus $\varepsilon'$ on the $x$ (or real) axis. Such complex plane plots are now widely used for two-dimensional representation of the response of all four immittance types. Finally, three-dimensional perspective plots that involve a log-frequency axis were introduced to the IS area by the author and his colleagues in 1981; these plots allow complete response at a given immittance level to be shown in a single diagram.

Because IS analysis generally makes considerable use of equivalent circuits to represent experimental frequency response, the whole history of lumped-constant circuit analysis, which particularly flowered in the first third of the century, is immediately relevant to IS. Since then, much work has been devoted to the development of theoretical physicochemical response models and to the definition and analysis of various distributed circuit elements for use in IS-equivalent circuits along with ideal, lumped elements like resistance and capacitance. The preferred analysis method for fitting of IS data to either equivalent circuits or to a mathematical model is complex nonlinear least squares fitting (CNLS), introduced to the field in 1977 by Macdonald and Garber. In this procedure, all the parameters of a fitting model are simultaneously adjusted to yield an optimum fit to the data.

Early experimental work in the IS field is discussed in the book on IS listed in the bibliography (Macdonald, 1987). Here it will suffice to mention the work of Grahame on electrolyte double-layer response, the technique of AC polarography pioneered by D. E. Smith, and the electrolyte studies of Randles and Somerton, Sluyters and Oomen, R. P. Buck, and J. E. Bauerle. Since the late 1960s, IS has developed rapidly, in large part because of the availability of new, accurate, and rapid measuring equipment.

## II. CATEGORIES OF IMPEDANCE SPECTROSCOPY: DEFINITIONS AND DISTINCTIONS

There are two main categories of IS: electrochemical IS (EIS) and everything else. EIS involves measurements and analysis of materials in which ionic conduction strongly predominates. Examples of such materials are solid and liquid electrolytes, fused salts, ionically conducting glasses and polymers, and nonstoichiometric ionically bonded single crystals, where conduction can involve motion of ion vacancies and interstitials. EIS is also valuable in the study of fuel cells, rechargeable batteries, and corrosion.

The remaining category of IS applies to dielectric materials: solid or liquid nonconductors whose electrical characteristics involve dipolar rotation, and to materials with predominantly electronic conduction. Examples are single-crystal or amorphous semiconductors, glasses, and polymers. Of course, IS applies to more complicated situations as well, for example, to partly conducting dielectric materials with some simultaneous ionic and electronic conductivity. It is worth noting that although EIS is the most rapidly growing branch of IS, nonelectrochemical IS measurements came first and are still of great value and importance in both basic and applied areas.

In the EIS area in particular, an important distinction is made between supported and unsupported electrolytes. Supported electrolytes are ones containing a high concentration of indifferent electrolyte, one whose ions generally neither adsorb nor react at the electrodes of the measuring cell. Such an added salt can ensure that the material is very nearly electroneutral everywhere, thus allowing diffusion and reaction effects for a low-concentration ion of interest to dominate the AC response of the system. Support is generally only possible for liquid electrochemical materials; it is often, but not always, used in aqueous electrochemistry. Solid electrolytes are unsupported in most cases of interest, electroneutrality is not present, and Poisson's equation strongly couples charged species. Because of this difference, the formulas or models used to analyze supported and unsupported situations may be somewhat or completely different.

Another important distinction is concerned with static potentials and fields. In a material-electrode system

without an applied static external potential difference (p.d.), internal p.d.s and fields are, nevertheless, generally present, producing space-charge layers at interfaces. For solids such regions are known as Frenkel layers and arise from the difference in work function between the electrode and the material. Because the static fields and charge concentrations in the material are inhomogeneous, exact small-signal solutions for the impedance of the system are impossible and numerical methods must be used.

In an electrolytic cell such static space-charge regions are only absent when the external static p.d. is adjusted so that the charge on the working electrode is zero—the point of zero charge (PZC)—a flat-band condition. Such adjustment is impossible for systems with two symmetrical electrodes because an applied static p.d. increases the space-charge region at one electrode while reducing it at the other. But the use of a working electrode of small area and a large-area counter electrode ensures that the overall impedance of the system is little influenced by what happens at the counter electrode; in this situation the PZC can be achieved for the working electrode. In general, the current distribution near this electrode is frequency dependent and thus makes a frequency-dependent contribution to the overall impedance of the system, which is dependent on electrode geometry and character.

Figure 1 shows a flow diagram for a complete IS study whose goal is characterization of important properties of the material-electrode system from its electrical response, one of the major applications of IS. The experimental data are denoted by $Z_e(\omega)$, the impedance predicted by a theoretical fitting model by $Z_t(\omega)$, and that of a possible electrical equivalent circuit by $Z_{ec}(\omega)$, where $\omega = 2\pi f$ and $f$ is frequency. When an appropriate detailed model for the physicochemical processes present is available, it should certainly be used for fitting. Otherwise, one would employ an equivalent electrical circuit whose elements and connectivity were selected, as far as possible, to represent the various mass and charge transport physical processes thought to be of importance for the particular system.

Note that a complete IS analysis often involves more than a single set of measurements of immittance versus frequency. Frequently, full characterization requires that such sets of measurements be carried out over a range of temperatures and/or other externally controlled experimental variables. IS characterization may be used to yield basic scientific and/or engineering information on a wide variety of materials and devices, ranging from solid and liquid electrolytes to dielectrics and semiconductors, to electrical and structural ceramics, to magnetic ferrites, to polymers and protective paint films, and to secondary batteries and fuel cells. Other important applications of IS, not further discussed herein, have been made in the biological area, such as studies of polarization across cell membranes and of animal and plant tissues. Finally, the analysis techniques of IS are not limited to electrical immittance but apply as well to measurements of mechanical and acoustic immittance.

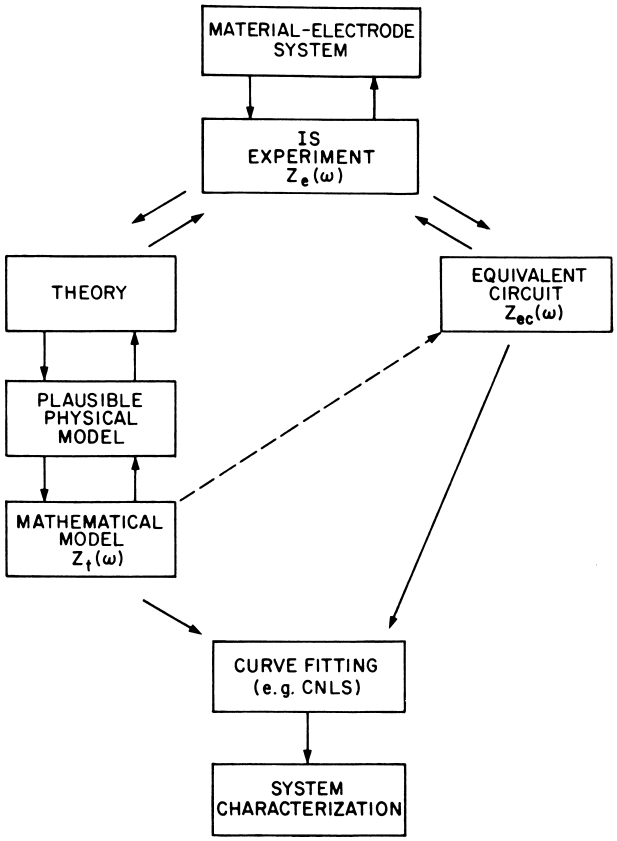



**FIGURE 1** Flow diagram for the measurement and characterization of a material-electrode system. (Reprinted by permission of John Wiley & Sons, Inc., from "Impedance Spectroscopy—Emphasizing Solid Materials and Systems," J. R. Macdonald, ed. Copyright © 1987, John Wiley & Sons, Inc.)

## III. ELEMENTS OF IMPEDANCE SPECTROSCOPY

### A. Measurement Methods

Although IS measurements are simple in principle, they are often complicated in practice. Part of the difficulty arises because the resistive and capacitive components of IS response have ranges, when one considers different materials, electrodes, and temperatures, that span 10 or more orders of magnitude. Measurements require comparison with standard values of these components and are thus only as accurate as the standards. Second, the IS frequency range may extend over 12 orders of magnitude or more: from as low as 10 $\mu$Hz for adequate resolution of

interfacial processes, up to 10 MHz or higher, sometimes needed to characterize bulk response of the material of interest.

Although IS measurements on solids or dielectric liquids usually involve cells with two identical-plane parallel electrodes, the situation is often much more complicated for measurements on liquid electrolytes. There, one usually employs one or more small working electrodes, a very small reference electrode, and a large counter electrode. Such an arrangement ensures that everything of interest (related to immittance) happens at or near the working electrode(s). Further, a rotating-disk working electrode is frequently used to control hydrodynamic conditions near the electrode.

Because the kinetics of electrode reactions often depend strongly on the static (dc) potential difference between the working electrode and the bulk, or, equivalently, the working electrode and the reference electrode, a potentiostat is needed to fix this p.d. to a known and controllable value. The simultaneous application of both ac and dc signals to a three- or four-electrode cell makes it particularly difficult to obtain accurate frequency-response results above 50 kHz or so.

Although a calibrated double-beam oscilloscope, or the use of Lissajous figures with a single-beam instrument, can be used to determine immittance magnitude and phase, such measurements are generally insufficiently accurate, are time consuming, and apply only over a limited frequency range. A superior alternative is the use of audio-frequency or high-frequency bridges. Several such bridges are discussed in the IS book. Of particular interest is the Berberian–Cole bridge, which can cover a wide frequency range and can allow potentiostatic dc bias control. Another important technique using a bridge and special error reduction procedures has recently been developed by Schöne and co-workers that allows potentiostatic control and yields very accurate impedance results up to 3 MHz. But manual balancing of a bridge is often disadvantageous because of its slowness, especially for corrosion studies where the properties of the system itself may be slowly changing.

Manual balancing is avoided in various automated network analyzers and impedance analyzers now commercially available. But the measuring instrument that has virtually revolutionized IS measurements and principally led to the burgeoning growth of the field in the past 20 years is the frequency-response analyzer (FRA). Typical examples are FRAs produced by Solartron and by Zahner. Although space does not allow a full description of their many features, such instruments allow potentiostatic control for three- or four-terminal measurements, they are highly digitized, they incorporate automatic frequency sweeps and automatic control of the magnitude of the applied ac signal, they can yield 0.1% accuracy, and they carry out measurements automatically.

Although FRAs such as the Solartron 1260 cover a frequency range from 10 $\mu$Hz to 32 MHz, impedance results using them are not sufficiently accurate above about 50 kHz when potentiostatic control is used. A typical FRA determines impedance by correlating, at each frequency, the cell response with two synchronous signals, one in phase with the applied signal and the other phase-shifted by 90°. This process yields the in-phase and out-of-phase components of the response and leads to the various immittance components. A useful feature is autointegration, a procedure that averages results over an exact number of cycles, with the amount of such averaging automatically selected to yield statistically consistent results. Recently, a *dielectric front end* has become available for FRAs. It has an extremely high input impedance and makes possible accurate measurements on dielectrics and on very-high-resistivity solids containing mobile charges.

## B. Analysis and Interpretation of Data

### 1. Graphics

Before carrying out a detailed analysis of IS immittance data, it is a good idea to examine the data graphically, both to search for any outliers and to examine the structure of the data, structure that will usually reflect, at least in part, the physical processes present that led to the data. From the experimental situation one will generally know whether one is dealing with an intrinsically insulating material, such as a nonconducting or a leaky dielectric, or whether the situation is of intrinsically conducting character: mobile charges dominate the response but may be completely or only partially blocked at the electrodes. For complete blocking, no DC can pass, a case that could be confused with dielectric response. In the intrinsically conducting situation, dielectric effects are generally minimal, and $Z$ and $M$ representations of the data are often most useful. In the nonconducting case, $Y$ and $\varepsilon$ are frequently most appropriate, but it is nevertheless a good idea initially to examine plots of the data for all four immittance levels, whatever the conducting/nonconducting situation.

When mobile charges are present, five principal physical processes may influence the data; these are bulk resistive-capacitive effects, electrode reactions, adsorption at the electrodes, bulk generation-recombination effects (e.g., ion-pairing), and diffusion. The double-layer capacitance is the reaction capacitance $C_R$, and the reaction resistance $R_R$ is inversely proportional to the reaction rate constant. It is important to distinguish $C_R$ from the usually much larger low-frequency psudocapacitance associated with the diffusion of mobile charge or

with adsorption at an electrode. Note that in general a process that dissipates energy is represented in an IS-equivalent circuit by a resistance, and energy storage is usually modeled by a capacitance. Detailed CNLS analysis of IS data can lead in favorable cases to estimates of such basic material-electrode quantities as electrode reaction and adsorption rates, bulk generation–recombination rates, charge valence numbers and mobilities, diffusion coefficients, and the (real) dielectric constant of the material.

There are many ways IS data may be plotted. In the IS field, where capacitive rather than inductive effects dominate, conventionally one plots $-\mathrm{Im}(Z) \equiv -Z''$ on the $y$-axis versus $\mathrm{Re}(Z) \equiv Z'$ on the $x$-axis to give a complex plane impedance plot. Such graphs have (erroneously) been termed Nyquist plots. They have the disadvantage of not indicating frequency response directly, but may, nevertheless, be very helpful in identifying conduction processes present. Another approach, the Bode diagram, is to plot $\log[|Z|]$ and $\phi$ versus $\log[f]$. Alternatively, one can plot $Z'$ (or any $I'$) or $-Z''$ (or $-I''$), or the logs of these quantities versus $\log[f]$.

An important IS building block is Debye response, response that involves a single time constant, $\tau$. A Cole–Cole plot of such response is shown in Fig. 2. The arrow shows the direction of increasing frequency. Debye response can be represented in complex form as $\varepsilon = \varepsilon_\infty + [\varepsilon_0 - \varepsilon_\infty]/[1 + (i\omega\tau)]$ and, in circuit form, involves a capacitance $\varepsilon_\infty C_C$ in parallel with the series combination of a resistor $R$, modeling dissipative effects, and a capacitor $C \equiv (\varepsilon_0 - \varepsilon_\infty)C_C$, representing stored charge. Finally, the time constant or relaxation time is given by $\tau \equiv RC$.

Three-dimensional perspective plots are particularly useful because they allow complete response to appear on a single graph. Figure 3 shows such plots at the impedance level for the analog of Debye response for a conducting system. By including projections of the 3-D curve of the response in all three perpendicular planes of the plot, one

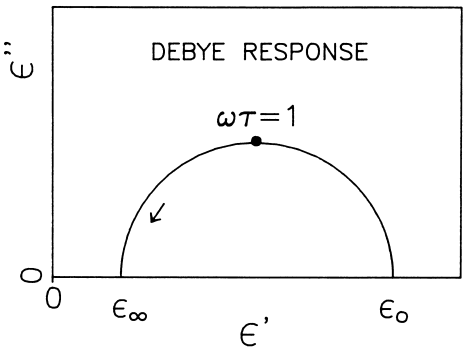



**FIGURE 2** Complex-plane plot of the complex dielectric constant for Debye frequency response.

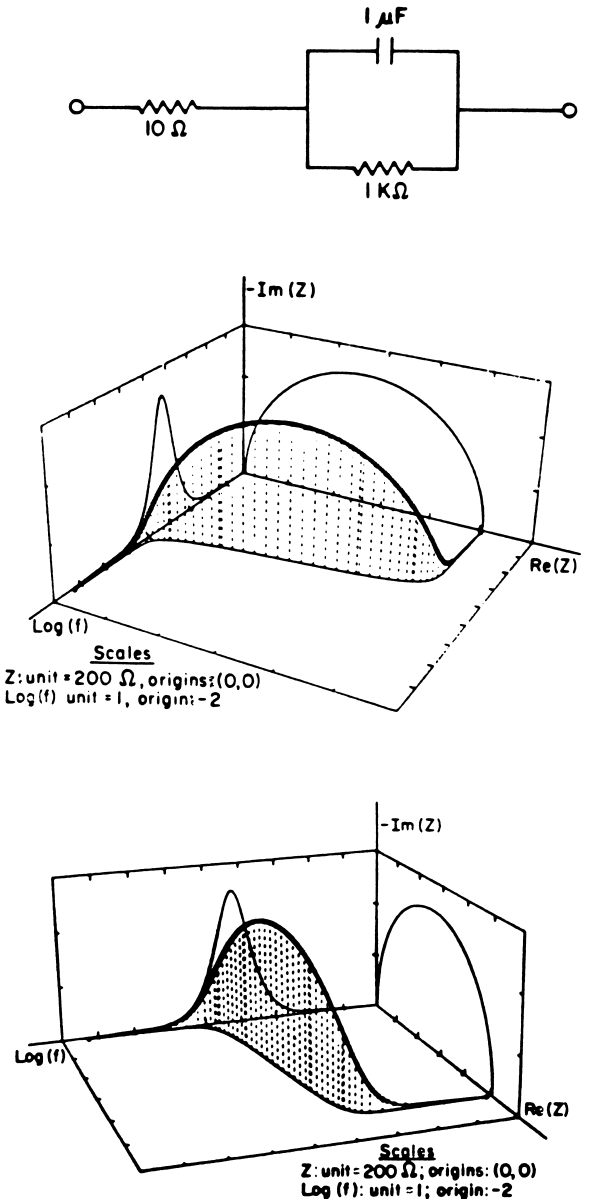



**FIGURE 3** A simple circuit and 3-D perspective plots of its impedance response. (Reprinted by permission of John Wiley & Sons, Inc., from "Impedance Spectroscopy—Emphasizing Solid Materials and Systems," J. R. Macdonald, ed. Copyright © 1987, John Wiley & Sons, Inc.)

incorporates all relevant 2-D plots in the same diagram. Note that the curve in the back plane, the complex-plane impedance plot, is just the usual Debye semicircle, one with its center on the real axis.

To demonstrate some of the power and weaknesses of 3-D plots, Fig. 4 includes three types of such plots, all for the same EIS data taken on single-crystal Na $\beta$-alumina. Graph A is an impedance plot and shows that only two out of the four curves indicate that the lowest frequency point is in error. In this plot, $\nu$ denotes frequency $f$. Clearly, one

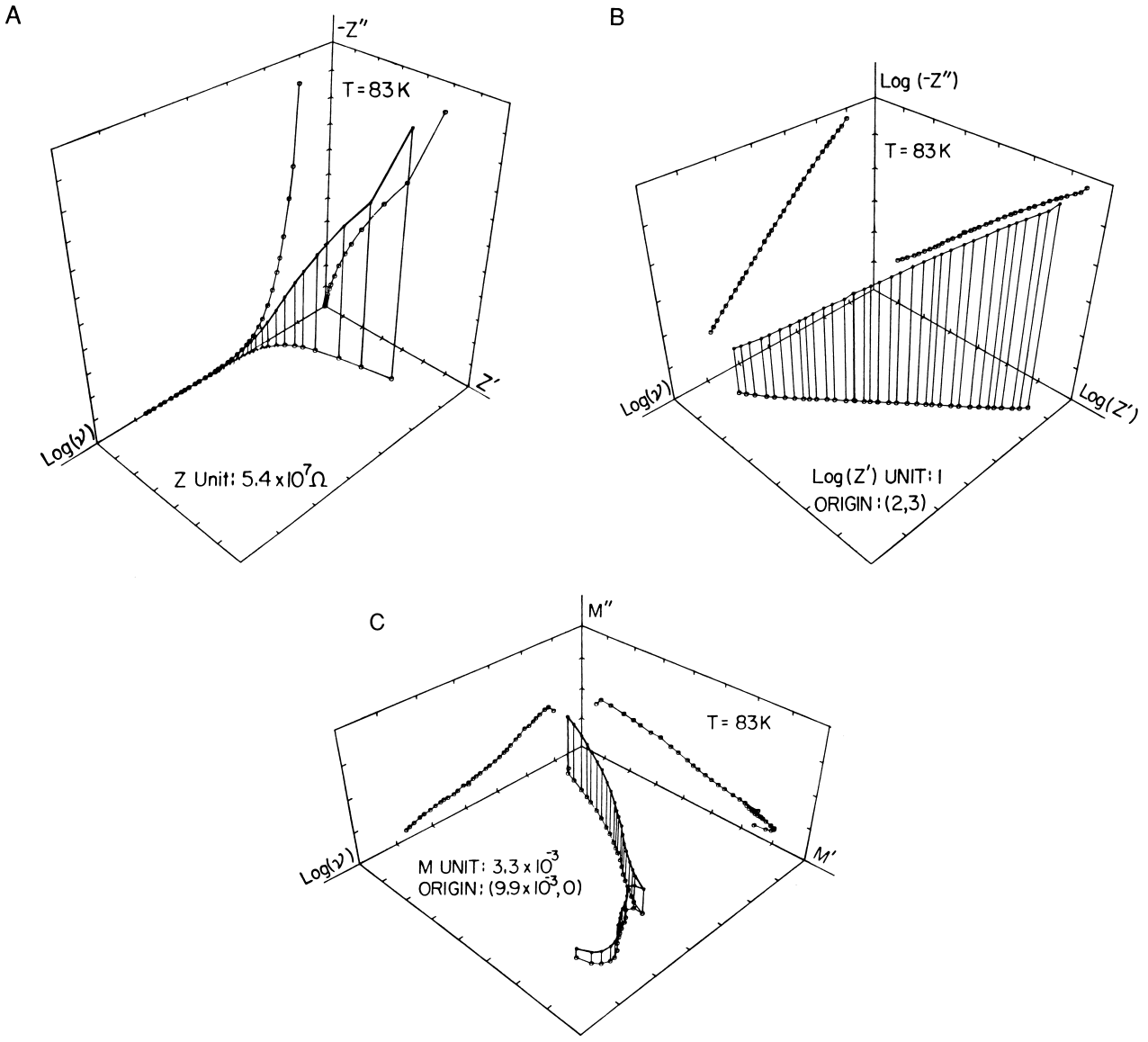



**FIGURE 4** Three-dimensional perspective plots of Na $\beta$-alumina data at (A) the impedance level, (B) log impedance level, and (C) complex modulus level. (Reprinted by permission of John Wiley & Sons, Inc., from "Impedance Spectroscopy—Emphasizing Solid Materials and Systems," J. R. Macdonald, ed. Copyright © 1987, John Wiley & Sons, Inc.)

should not rely on the conventional log[$f$] curves alone. Since the diagram shows that much high-frequency data are not resolved by this kind of plot, graph B involves the logarithms of the data. Although high-frequency response now appears, the error in the low-frequency point is nearly obscured by the reduced resolution inherent in a log plot.

Much improved results appear in graph C, a 3-D $M$ plot. Resolution over the full frequency range is greatly increased; the error in the lowest frequency point is clearly shown; a midfrequency glitch now appears that is not evident in the other plots and arises from a switch of measuring devices without adequate cross-calibration; and nonphysical behavior is now apparent at the highest frequencies. These results make it clear that even when 3-D plots are used, it is always desirable to explore the results of different transformations of the data and to pick the one with the best resolution.

## 2. Complex Nonlinear Least Squares Data Fitting

*a. Strengths and weaknesses.* Although graphic examination of IS data is an important analysis step, only

in the simplest cases can it be used to obtain even rough estimates of some system parameters. Since good parameter estimates are needed for adequate characterization of the material-electrode system, a fitting technique such as CNLS must be applied to obtain them. In doing so, the data, at any I level, are fitted to a mathematical model involving the parameters or to the response of an equivalent circuit. Such fitting models are discussed in Section IV.A. Not only does CNLS fitting yield estimates of the parameters of the model, but it also provides estimates of their standard deviations, measures of how well they have been determined by the data fit. These standard deviation values are valuable in deciding which parameters are crucial to the model and which are useless, or at least not well determinable from the data.

CNLS fits are produced by a program that minimizes the weighted sum of squares of the real and imaginary residuals. A residual is the difference between a data value at a given frequency and the corresponding value calculated from the model. The weights used are the inverses of the estimated error variance for a given real data value and that for the corresponding imaginary value. Weighting is the most subjective part of least squares fitting, yet it can often have crucial effects on the results of such fitting and is thus of prime importance.

Since individual error-variance estimates are usually unavailable, it has been customary to use simplified variance models to obtain values to use in the fitting. The simplest such model is to take all weights equal to one: unity weighting (UWT). Another popular and important choice is to set the error variance of each data value equal to the square of that value. Since the uncertainty of the value is then proportional to the value itself, this defines proportional weighting (PWT). It has recently been shown, however, that such weighting leads to biased parameter estimates; it should be replaced, when the fitting model is well matched to the data, by function-proportional weighting (FPWT), where the calculated rather than the direct data value is employed in the weighting.

PWT or FPWT is particularly important because the range of typical IS data can be as large as $10^3$ or even $10^6$. When UWT is used in fitting such data, only the largest parts of the data determine the parameter estimates, and the smaller values have little or no effect. Alternatively, with PWT or FPWT, which is equivalent to assuming a constant percentage error, small and large data values contribute equally to the final parameter estimates.

Figure 5 presents the results of PWT CNLS fitting of $\beta$-$PbF_2$ data using an equivalent circuit with a distributed element, the constant phase element (CPE). Both the original data and the fit results are shown in the 3-D plot. The figure indicates that seven free parameters have been quite well determined by the data, a remarkable result when one

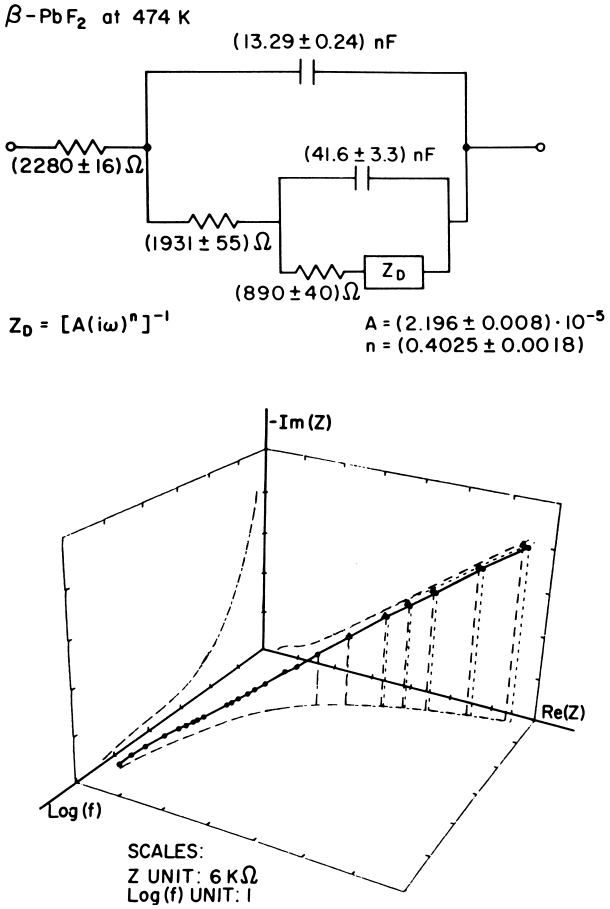



**FIGURE 5** Three-dimensional perspective impedance plot of $\beta$-$PbF_2$ data (—— ---) and fitted values and curves (— — —); the fitting circuit used and parameter estimates and estimates of their standard deviations. (Reprinted by permission of John Wiley & Sons, Inc., from "Impedance Spectroscopy—Emphasizing Solid Materials and Systems," J. R. Macdonald, ed. Copyright © 1987, John Wiley & Sons, Inc.)

considers the apparent lack of much structure in the data themselves.

A detailed physico-chemical model is always preferable to an equivalent circuit for fitting, especially since such modles often cannot be expressed in terms of an equivalent circuit involving standard elements. But most IS situations involve many-body problems currently insolvable at the microscopic level. Thus one must usually be satisfied with simpler continuum models, often expressed as equivalent circuits. One weakness of equivalent circuits involving only ideal elements is their ambiguity. The same elements may be interconnected in different ways and yet, with appropriate values, yield exactly the same frequency response at all frequencies. Thus, IS fitting cannot distinguish between the different possible structures, and only other measurements, such as IS fitting of data

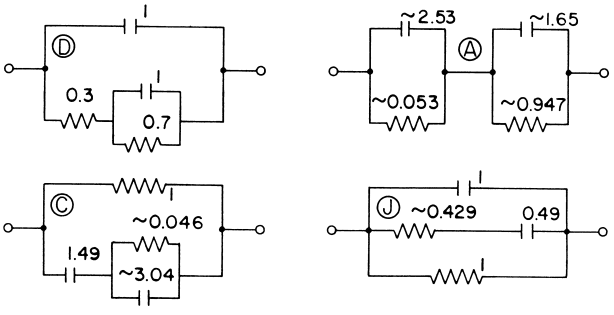



**FIGURE 6** Four two-time-constant circuits that exhibit the same impedance response over all frequencies. Units are MΩ for resistances and $\mu$F for capacitances.

over a range of temperatures and/or potentials, can help one establish which of the possible circuits is most physically reasonable.

Figure 6 shows all possible potentially equivalent conducting circuits involving two resistances and two capacitances. Specific parameter value choices that make them all have exactly the same response are also indicated. Here the values for circuit D were taken exact, and approximate values for the other elements are denoted with a $\sim$ sign. Let the units of these elements be MΩ for resistances and $\mu$F for capacitances. Note that the two RC time constants of circuit A, a series connection, differ by less than 17% and are thus very close together. Can IS procedures resolve such a situation? Figure 7 shows the exact complex-plane response of these circuits at both the $Z$ and the $M$ levels, compared to single time-constant Debye response. The $M$ curve shows much better separation of the two response regions than does the $Z$ curve. Thus, adequate graphical resolution is indeed possible. Further, it turns out that CNLS fitting of synthetic data calculated from any of these circuits with appreciable proportional random errors added still yields excellent parameter estimates. In fact, with reasonably good data, CNLS can resolve response

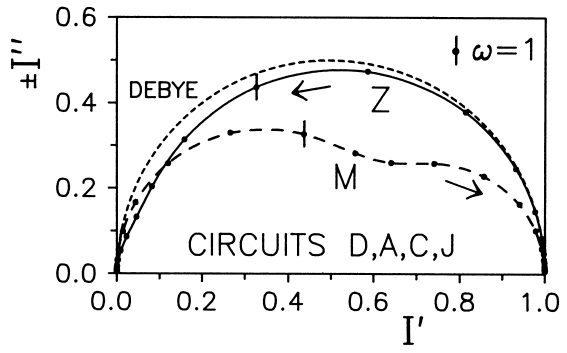



**FIGURE 7** Complex-plane immittance responses, at the $Z$ and $M$ levels, of the circuits of Fig. 6.

involving considerably closer time constants than those involved here.

Although several CNLS fitting programs now exist for use on personal computers, two commercially available ones have been especially tailored for the IS field. The first, EQUIVCRT, can be obtained from Dr. B. A. Boukamp, Twente University, P.O. Box 217, 7500 AE Enschede, The Netherlands: the second, LEVM, is now available at no cost, and both its extensive manual and program, including source code, may be downloaded from this home page: http://www.physics.unc.edu/~macd/. The programs to some degree complement each other, but LEVM is more general and flexible in many ways and incorporates much more sophisticated weighting possibilities.

*b. Recent developments.* Currently, the capability of using various types of weighting involving model predictions instead of data values exists only in LEVM, first released in the summer of 1989. Although weighting such as FPWT is somewhat more complicated than PWT because it varies with each nonlinear least squares iteration as the parameter estimates change during the fitting procedure, its bias reduction potential makes such complexity often worthwhile. Although LEVM allows the fitting of real or imaginary parts of the data separately, fitting both together, as in CNLS, ensures that the best use is made of all the data in determining the parameter estimates and is thus preferred when both parts are available.

Real IS data often have independent random errors that have both an additive term and one that depends on the true model predictions. A rather general error-variance model incorporating these possibilities is included in LEVM. For a specific angular frequency $\omega_j$, the real and imaginary parts of $\nu_j$, the error variance used in determining the weighting, may be written as

$$\nu'_j = U^2 + |F'(\omega_j)|^{2\xi}$$

and

$$\nu''_j = U^2 + |F''(\omega_j)|^{2\xi},$$

where $U$ is associated with the additive random errors and $\xi$ is an arbitrary positive fractional exponent.

When $U = 0, \xi = 1$, and $F$ is a data value, one has PWT; whereas when $F$ is a model prediction the result is FPWT. Another widely used weighting, modulus weighting, follows when the same values of $U$ and $\xi$ are used but both $F'$ and $F''$ in the above equations are replaced by $|F|$. It is usually inconsistent, however, with the types of errors likely to be present and generally leads to appreciably more bias in parameter estimates even than PWT. CNLS fitting yields a standard deviation $S_f$ of residuals, which is a measure of the overall goodness of fit. For proportional random errors having a proportionality constant of $\sigma_r$, $S_f$

is an unbiased estimate of $\sigma_r$ for FPWT, is nearly so for PWT, and is appreciably biased on the low side for FMWT and MWT, types of modulus weighting.

When the data involve one or more inductive-type loops, such as may arise from adsorption of a mobile charge at an electrode, it is desirable to use a nonzero $U$ along with PWT or FPWT or to use modulus weighting. Otherwise, because values of the imaginary part of the data may become very small and even pass through zero in the loop region, PWT or FPWT alone can strongly overemphasize the effect of these values near zero and thus lead to poor fitting.

Although $U$ and $\xi$ may be given fixed values during CNLS fitting, a way has been found to incorporate them as free parameters in LEVM least squares fitting. When this is done, the data themselves determine the most appropriate weighting for their fit, thus removing an appreciable part of the subjective element present in prior weighting approaches. Further, Monte Carlo fitting studies have shown that the statistical uncertainties of $U$ and $\xi$ in CNLS fits with them both taken free, or with only $\xi$ free, are usually quite small compared to their estimates, and their biases are even smaller. Thus, their fit estimates may be used with confidence.

Although for much IS data one would expect that $\xi$ would be close to unity, this need not always be the case. Consider, for example, a set of real data arising from the sum of the radioactive decay of two different species. Now the statistics of such decay follows a Poisson process, one for which $\xi = 0.5$. The radioactive background count will also involve such a process. Thus the appropriate variance model would involve $U = 0$ and three terms, each with $2\xi = 1$. The first two would be the two exponential decays and the last, the background. In such a fitting situation, where $\xi$ is known absolutely, it should be held fixed at its proper value.

## IV. APPLICATIONS

### A. Basic Analysis of Material Properties and Electrode Effects

#### 1. Bulk and Reaction Response

Although IS is of great value for the characterization of the electrical properties of material-electrode systems, its use for this purpose requires that connections be known between model and/or equivalent circuit parameters and the basic characterization parameters. One must be able to pass from estimates of macroscopic quantities, such as resistances and capacitances, to estimates of average microscopic quantities. Here only a brief overview will be given of some of the large amount of theoretical IS-related work of the past 40 years. More details appear in the IS book.

Because of the charge decoupling present in a supported situation, it is often an excellent approximation to treat the effects of the various physical processes present independently. On the other hand, for unsupported conditions where strong coupling is present, a unified treatment of all the processes together is necessary. The most complete such theory, which incorporates all five of the processes mentioned in section III.B.1, was published by Franceschetti and Macdonald in 1978. It is a continuum (i.e., averaged, not microscopic) theory, includes intrinsic and extrinsic charge effects, and applies to either ionic or electronic conduction conditions. Even though it strictly applies only to flat-band conditions, its results are still sufficiently complicated that only in simplified cases does it lead to responses that may be modeled by an equivalent circuit.

It is useful to separate the electric processes present into bulk- and electrode-related groups whenever possible, something which is usually indeed possible using CNLS fitting. The first group includes bulk resistance and dielectric effects, the homogeneous reactions associated with dissociation and recombination of the charges present, and even possible dispersive response. It is generally associated with response effects at the high end of the frequency range, while significant electrode effects often occur near the low end, possibly at very low frequencies. Bulk resistance and capacitance are extensive quantities, dependent on the effective separation between electrodes.

The second group involves what happens in the neighborhood of the electrodes (within a few Debye lengths of them) and is thus intensive. No net charge is transferred to an electrode if it is completely blocking for all mobile charges. The next simplest EIS situation is that where a mobile metallic ion is of the same species as the atoms of a metallic electrode: a parent-ion electrode. Then, in a symmetrical-electrode situation there is a sink/source of ions at each electrode, since electron transfer at an electrode can transform ions into atoms and vice versa, depending on the polarity of the electric field at the electrode. Such a reaction can be written

$$\mathrm{Me} \rightleftharpoons \mathrm{Me}_{z+} + z\mathrm{e}^-,$$

where Me denotes a metal atom and $z$ the number of electrons transferred. An example of a symmetrical cell of this type is Ag|AgCl|Ag.

Particularly important for the aqueous electrolyte area is the redox electrode, where charge crosses the interface at the electrode only in the form of electrons. The species Red and Ox are usually soluble in the electrolyte, satisfy

$$\mathrm{Red}^{(z-n)+} \rightleftharpoons Ox^{z+} + n\mathrm{e}^-,$$

and involve the forward and reverse reaction rate constants $k_f$ and $k_r$, respectively. If $z = n$, the Red species is uncharged and may diffuse in the electrode, or may evolve if it is a gas.

### 2. Distributed Circuit Element Response

*a. Diffusion.* Since diffusion is not localized at a point in space but is distributed over a finite region, it leads to electrical response characteristic of a distributed circuit element (DCE). Such elements cannot be described by means of a finite number of ideal elements, such as resistances and capacitances. The response of several DCEs important to IS will be discussed.

In addition to possible diffusion of uncharged species within an electrode, diffusion of mobile species in the electrolyte may contribute significantly to the impedance of a IS system. Generally, diffusion response is neither intensive nor extensive. At sufficiently high frequencies, diffusion effects are confined to the immediate neighborhood of the electrode (or within a hydrodynamic boundary layer at a rotating electrode) and so are intensive; whereas at low enough frequencies, diffusion occurs throughout the material between electrodes, and the response becomes extensive as the frequency decreases and the effective diffusion length $l_d$, proportional to $(\omega)^{-0.5}$, becomes comparable to the size of the cell.

The diffusion impedance, appropriate when there is a fast electrode reaction, is of the form

$$Z_W(\omega) = Z_W(0)\left[\tanh\left\{i(l/l_d)^2\right\}^{0.5}\right]/\left\{i(l/l_d)^2\right\}^{0.5},$$

where $l$ is the separation between symmetrical electrodes and $Z_W(0)$ is a resistance proportional to $l$ and thus is extensive. Such response is known as finite-length Warburg behavior. At high enough frequencies that the tanh term goes to unity, $Z_W(\omega)$ becomes proportional to the intensive quantity $l_d$ and is termed (ordinary) Warburg response.

As the electrode reaction rate decreases toward zero, a more complicated expression for $Z_W(\omega)$ must be used, but it reduces to the form

$$Z_W(\omega) = Z_C\left[\mathrm{ctnh}\left\{i(l/l_d)^2\right\}^{0.5}\right]/\left\{i(l/l_d)^2\right\}^{0.5}$$

when the electrode is completely blocking (open-circuit diffusion). Here $Z_C$ is given by $(l/l_d)^2/\omega C_{DOC}$, and $C_{DOC}/C_C$ is the effective low-frequency limiting dielectric constant associated with the process.

For general unsupported situations, those with positive and negative charged species mobile and having diffusion coefficients of $D_n$ and $D_p$ and valence numbers of $z_n$ and $z_p$,

$$(l_d)^2 = (4D_nD_p/\omega)[(\eta_n + \eta_p)/(\eta_nD_n + \eta_pD_p)].$$

No diffusion effects appear when only charge of a single sign is mobile; this often is an excellent approximation for solid electrolytes.

For supported conditions, matters are different. Consider a single species with diffusion coefficient $D$ and valence number $z$ (possibly zero). Then $(l_d)^2 = 4D/\omega$, a result that follows from the above expression when one sets $D_n = D_p = D$ and $z_n = z_p = z$. Further, when both positive and negative charges are mobile, diffusion under unsupported conditions leads to a single expression involving tanh, as above, but for supported conditions, as in the redox case, two such terms appear, one for each species—in keeping with the lack of coupling between the species.

*b. Other DCEs.* A characteristic signature of diffusion is $(i\omega)^{\pm 0.5}$ response, but IS data more often exhibits CPE response $(i\omega\tau)^{\pm\psi}$, where $0 \le \psi \le 1$. But such response is not physically realizable over all frequencies, and so other DCEs have been introduced that approximate such behavior over a limited frequency range. They may be written as impedances or complex dielectric constants, depending upon which $I$ level is appropriate. Here they will be given at the $Z$ level.

An empirical DCE of the above type is Havriliak–Negami (HN) response, written as

$$Z_{HN}(\omega) = R_{HN}/[1 + (i\omega\tau)^{\alpha}]^{\beta},$$

where $0 \le \alpha \le 1$ and $0 \le \beta \le 1$. It reduces to Cole–Davidson response when $\alpha = 1$ and to Cole–Cole response (termed *ZC* response at the $Z$ level) when $\beta = 1$. The first of these yields an asymmetric arc in a complex-plane plot and the second one a symmetric arc. Both shapes appear often in practice, and *ZC* fitting is frequently used to represent data that yield an arc of a semicircle with its center below the real axis. Such behavior is usually ascribed to the presence of a distribution of some physical quantity in space, time, or energy. Rough electrodes are one example. Although fitted values of $\alpha$ and/or $\beta$ often show appreciable temperature dependence, there exists no theory yielding such dependence for HN response.

Another important DCE is that of Kohlrausch–Williams–Watts (KWW) response. It yields a stretched exponential in the time domain, response proportional to $\exp[-t/\tau]^{\psi}$, with $0 \le \psi \le 1$. Here there are, if anything, too many different theories yielding such response, but again they do not usually predict the temperature dependence of $\psi$. The corresponding frequency response is very difficult to calculate accurately, but an excellent approximation for it is available in LEVM. Complex plane plots of KWW response yield an asymmetric arc for any $\psi < 1$ until $\psi = 1$, when a Debye semicircle appears.

Another DCE category is associated with the presence in a material of a distribution of activation energies (DAE).

Such distributions are likely in IS materials and may be expected even in single crystals when there are many competing possibilities for the individual motion of mobile charges. Both Gaussian and exponential distributions have been considered in detail and can lead to either symmetric or asymmetric complex-plane arcs. But only an exponential DAE yields CPE-like fractional-exponent frequency response over a finite frequency region. This exponent, $\phi$, is not limited to the range from 0 to 1 but satisfies $-\infty < \phi < \infty$. Further, unlike the other DCEs considered, an exponential DAE predicts temperature dependence of $\phi$ in good agreement with many experimental results.

## 3. Equivalent Circuits

Many different equivalent circuits have been proposed over the years for IS fitting, and no one circuit structure is appropriate for all situations. Figure 8 shows a circuit, however, that has been found useful for a variety of materials and experimental conditions. Bulk properties are represented by $C_g$, the geometrical compacitance, and $R_\infty$, the high-frequency limiting resistance. $C_R$, associated with an electrode reaction, is the double-layer capacitance (possibly including both a compact inner-layer capacitance and a diffuse double-layer capacitance), and $R_R$ is the reaction resistance. Finally, $C_A$ and $R_A$ are associated with adsorption at an electrode. The $Z_D$ elements, when present, are DCEs. Also, not all the other elements need be present; for example, in the absence of adsorption $C_A$ and $R_A$ would not appear.

For an unsupported, fully dissociated material with charges of only a single sign mobile, the Fig. 8 circuit with all $Z_D$s absent has been found to yield an accurate representation of the impedance resulting from a flat-band theoretical analysis of the situation. Since only $R$s and $C$s are involved, ambiguity is present, and many other circuit structures with the same elements and the same frequency response are possible. Nevertheless, the present hierarchical ladder-net-work connection is more physically reasonable than the others for homogeneous material. It ensures that bulk charging and conduction effects take place before reaction/adsorption ones. For polycrystalline materials, however, circuits involving series rather than hierarchical connection of parallel $RC$ subcircuits are often found appropriate.

For the conditions above, no diffusion DCE element is present. The $Z_W(\omega)$ one discussed earlier appears, however, in the $Z_{D3}$ position when charges of both signs are mobile and at least one of them reacts at an electrode. When static fields are present in the material, either intrinsic and/or externally produced, numerical analysis of the nonlinear transport equations governing the IS response shows that the Fig. 8 circuit still applies to good approximation but elements such as $C_R$ and $R_R$ then depend appreciably on the static p.d. present. Finally, the circuit of Fig. 8 has often been found appropriate for the fitting of data for supported conditions as well as unsupported ones.

The Fig. 8 circuit is particularly appropriate for analyzing electrode–interface effects in low-resistivity EIS situations. For high-resistivity solids, however, such as ion-conducting glasses, one is usually more interested in isolating and inerpreting bulk dispersion behavior. Figure 9 shows a circuit appropriate for such materials, where te DCE block is a conductive-system dispersive DE, and the DED block is a dielectric-system dispersive one. The circuit thus allows one to account for electrode effects when important, and either conductive dispersion, or dielectric dispersion, or both.

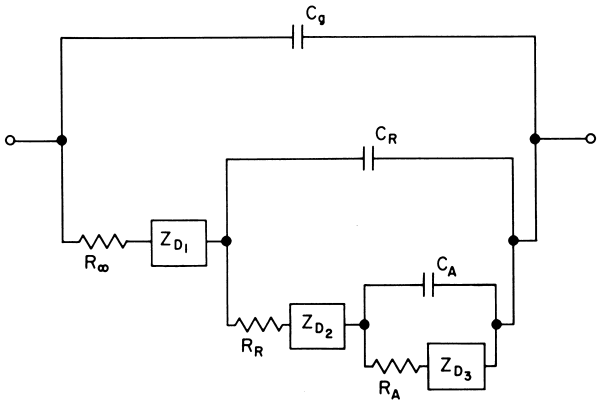



**FIGURE 8** An equivalent circuit of hierarchical structure useful in fitting much IS data. (Reprinted from "Interface Effects in the Electrical Response of Non-Metallic Conducting Solids and Liquids," J. R. Macdonald, *IEEE Trans. on Electrical Insulation,* Vol. EI-15, pp. 65–82, Fig. 3. Copyright © IEEE 1981.)

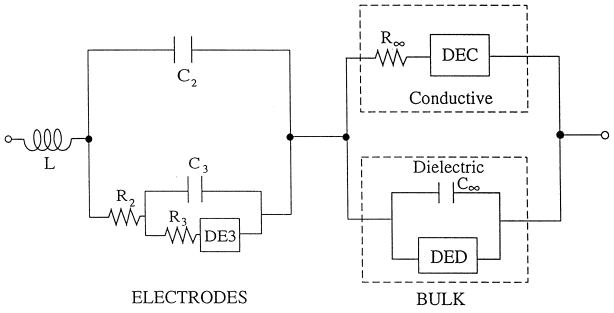



**FIGURE 9** An equivalent circuit available in LEVM that is appropriate for analyzing the frequency response of high-resistivity materials. As usual, "DE" indentifies possible distributed circuit elements. (Reprinted by permission of Elsevier Science B. V., from "Power-law exponents and hidden bulk relaxation (the word "relaxation" was erroneously printed as "relation") in the impedance spectroscopy of solids," J. R. Macdonald, J. Electroanalytical Chem., Vol. 378, pp. 17–29, 1994.)

## B. Uses of IS for Evaluation and Control of Electrochemical Processes of Industrial Interest and Importance

### 1. Corrosion and Surface Protection

Corrosion of metallic structural materials leads to immense damage each year (an estimate for the United States for 1988 is $200 billion); thus its control and amelioration are of tremendous economic importance. EIS has played and is playing a valuable role in quantifying and mitigating corrosion effects. For example, it has been successfully applied to complicated corrosion systems to determine corrosion rates as well as the mechanisms and efficiency of corrosion inhibitors. The use of EIS has broadened the range of corrosion phenomena that can be studied using electrochemical techniques and has been particularly valuable in evaluating the corrosion behaviors of polymer-coated metals and anodized aluminum alloys. In addition, it has been incorporated into a quality control test for anodized aluminum surfaces and for chromate-conversion-coated aluminum alloys.

The application of EIS techniques has resulted in a great deal of information on methods of corrosion protection that are difficult or impossible to study with traditional dc techniques, such as conversion and polymer coatings, anodic films, and inhibitors. Not only can EIS measurements provide greater sensitivity and more information about the processes investigated than can conventional dc methods, they are particularly appropriate when impedances are high and/or when low-conductivity media are used.

EIS measurement and analysis has been used to provide fast and sensitive information on the protection properties of chromated galvanized steel. Such measurements may be used as a quality control procedure, since the charge transfer resistance has been found to be well related to the corrosion rate. EIS has been used to detect corroding areas of large structures accurately and has been applied for corrosion monitoring of steel reinforcing bars in concrete to yield a nondestructive estimate of the amount of corrosion damage.

Since the roughness of an electrode surface is reflected in the results of EIS measurements involving the electrode, EIS may be used to identify surface inhomogeneities produced by corrosion. It provides (averaged) information on surface morphology on a much smaller scale than does even electron micrography. EIS has been employed as a means of nearly continuous evaluation of localized corrosion processes such as pitting, crevice corrosion, stress corrosion, cracking and fatigue corrosion, abrasion, and corrosion under a porous surface layer.

EIS measurements over a relatively wide frequency range have been found to yield valuable detailed information about the properties of aluminum oxide layers formed under different anodizing and sealing conditions. Discrimination was possible between properties of the dense barrier layer and the porous outer layer, and changes arising from aging and from the effects of natural environmental conditions were reflected in the results.

An EIS monitor has been used for the detection of paint degradation under atmospheric exposure. A model is being developed to help predict the lifetime of protective organic coatings on steel based on short laboratory tests. The model includes the steps of defect formation, transport of corrodents, loss of adhesion, and corrosion. EIS helps elucidate how these four processes interact and depend on coating processes and environmental effects.

Although IS analysis should properly be carried out only on time-invariant data—data obtained from a system whose properties are independent of time—some of these properties are often not time-invariant during measurement of a corroding system. If the change is slow compared to the required measurement time and/or if it is approximately linear in time, imporved results may be obtained by making a set of measurements from low to high frequencies immediately followed by one from high to low frequencies. Averaging of the results will then eliminate much of the variation with time.

A test of time invariance can be made by analyzing the data with the Kramers–Kroning (K–K) relations, integral transforms connecting real to imaginary parts of the data and vice versa. They are only applicable for time-invariant systems. All useful fitting models and equivalent circuits are minimum phase and so automatically satisfy the K–K relations. Thus, a good fit is evidence of time invariance. Strong failure of the K–K relations for a given set of data is immediate evidence of unwanted time variation, and, unlike CNLS fitting, no model or circuit is required to carry out such a test. Although ordinary K–K analysis requires much computation and uncertain extrapolations as well, an alternate program available in LEVM avoids such difficulties, uses only measured data, and caan readily test for time-invariance as well as estimating the imaginary-part response associated with real-part data, or vice versa.

### 2. Batteries and Fuel Cells

EIS studies have been made of the kinetics of the insertion reaction in solid-state batteries based on such reactions. A single EIS experiment allows information to be obtained about the electrode-interface reaction and diffusion in the electrolyte or electrode. Measurements at different battery voltages to determine the dependence of the results on the charge of the battery have led to increased understanding of the discharge process and thence to improved battery design.

An interesting EIS study has been carried out on electrochemically impregnated Ni electrodes from four different manufacturers of Ni/H cells. The EIS measurements were made in KOH electrolyte, and large differences were found in the impedance behavior of the electrodes from the different manufacturers. The results indicated a probable correlation between impedance parameters and cell life and performance.

EIS studies of molten carbonate fuel cells have increased the understanding of processes going on under operating conditions of the cell. In particular, they have helped identify and elucidate the reactions that occur both at the anode and at the cathode.

### 3. Other Devices and Techniques

Electrolyte-insulator-semiconductor sensors meld integrated circuit technology with traditional chemical technology. They can be used to monitor pH changes, for example, and can be constructed with ion-selective membranes to make them sensitive to a specific ion. IS measurements and analysis can yield, in favorable cases, information on the electrical characteristics of the electrolyte, the insulator, the semi-conductor, and the various interfaces and on interface states. The IS approach allows very low surface-state densities at the insulator-semiconductor interface to be determined. Measurements have shown, however, that it is the electrolyte–insulator interface that responds to pH changes.

Solid-electrolyte chemical sensors are electro-chemical cells designed to measure the concentration or pressure of chemical species in gases or fluids; for example, zirconia-based solid electrolytes have been used to measure oxygen concentration. Such sensors are employed to measure the oxygen concentration in steel melts and the air–fuel ratio in automobile engines. EIS has been found very useful to study (and to help optimize) electrode materials and appropriate pretreatment preparation for such sensors.

In recent years a number of variants on and extensions of IS have been developed. An important one is electrohydrodynamic impedance. Here the speed of a rotating-disk electrode is modulated sinusoidally, resulting in modulation of the mass transport in a liquid electrolyte. Such modulation allows the minimization of the coupling with interfacial kinetics. Modulation of numerous other quantities in an IS experiment is also possible, such as light, temperature, or magnetic field. Thus analysis of other transfer functions, cause and effect relations that go beyond potential and current, can add valuable additional information to IS studies. It is likely that considerable future development will be concerned with such possibilites.

## SEE ALSO THE FOLLOWING ARTICLES

BATTERIES • CORROSION • ELECTROCHEMICAL ENGINEERING • ELECTROCHEMISTRY • ELECTROLYTE SOLUTIONS, THERMODYNAMICS • ELECTROLYTE SOLUTIONS, TRANSPORT PROPERTIES • FUEL CELLS, APPLICATIONS IN STATIONARY POWER SYSTEMS

## BIBLIOGRAPHY

Archer, W. I., and Armstrong, R. D. (1980). The application of A. C. impedance methods in solid electrolytes. *Electrochemistry* **7,** 157–202.

Armstrong, R. D., Bell, M. F., and Metcalfe, A. A. (1978). The AC impedance of complex electrochemical reactions. *Electrochemistry* **6,** 98–127.

Franceschetti, D. R., and Macdonald, J. R. (1978). Theory of small-signal AC response of solids and liquids with recombining mobile charge. *J. Chem. Phys.* **68,** 1614–1637.

Gabrielli, C., ed. (1990). Proceedings of the First International Symposium on Electrochemical Impedance Spectroscopy. *Electrochim. Acta* **35,** 1483–1670.

Macdonald, D. D. (1991). Mechanistic Analysis Using Electrochemical Impedance Spectroscopy, Proceedings of the Symposium on High Temperature Electrode Materials and Characterization, 91-6, pp. 1–43. The Electrochemical Society, Inc., Pennington, NJ.

Macdonald, D. D., ed. (1993). Proceedings of the Second International Symposium on Electrochemical Impedance Spectroscopy, *Electrochim. Acta* **38,** 1797–2143.

Macdonald, J. R., ed. (1987). "Impedance Spectroscopy—Emphasizing Solid Materials and Systems," Wiley–Interscience, New York.

Macdonald, J. R. (1987). Impedance spectroscopy and its use in analyzing the steady-state AC response of solid and liquid electrolytes. *J. Electroanal. Chem.* **223,** 25–50.

Macdonald, J. R., and Potter, L. D., Jr. (1987). A flexible procedure for analyzing impedance spectroscopy results: Description and illustrations. *Solid State Ionics* **23,** 61–79.

Macdonald, J. R. (1999). A full list of categorized JRM publications, with titles, is available in *http://www.physics.unc.edu/~macd/*.

Mansfeld, F. (1988). Don't be afraid of electrochemical techniques—but use them with care. *Corrosion* **44,** 856–868.

Schöne, G., Wiesbeck, W., and Lorenz, W. J. (1987). High-frequency impedance spectroscopy of fast electrode reactions. *J. Electroanal. Chem.* **229,** 407–421.

Sluyters-Rehbach, M., and Sluyters, J. H. (1984). AC Techniques. *In* "Comprehensive Treatise of Electrochemistry" (E. Yeager, J. O'M. Bockris, B. E. Conway, and S. Sarangapani, eds.), Vol. 9, pp. 177–292. Plenum Press, New York.



# Incommensurate Crystals and Quasicrystals

**Uwe Grimm**
*The Open University*

**Max Scheffer**
*Chemnitz University of Technology*

I. Aperiodic Crystals
II. The Structure of Quasicrystals
III. Physical Properties of Quasicrystals
IV. Concluding Remarks

## GLOSSARY

**Aperiodic crystal** Crystalline structure without three-dimensional lattice periodicity.
**Approximant phase** Periodic crystalline phase that approximates a quasicrystalline phase.
**Commensurate** Having a rational ratio.
**Crystal** Ordered structure with essentially pure Bragg diffraction pattern; the current definition includes both periodic and aperiodic crystals.
**Decagonal phase** Quasicrystal with one periodic direction and a 10-fold rotational symmetry.
**Dodecagonal phase** Quasicrystal with one periodic direction and a 12-fold rotational symmetry.
**Icosahedral phase** Quasicrystal with icosahedral symmetry and no periodic direction.
**Icosahedral symmetry** The symmetry of the regular icosahedron. It is the largest symmetry group of a three-dimensional regular polyhedron and comprises two-, three-, and fivefold rotational symmetry axes.
**Incommensurate** Having an irrational ratio.
**Incommensurate phase** Aperiodic crystal whose structure is based on periodic lattices. The aperiodicity may be due to an incommensurate modulation of the periodic lattice or an incommensurate combination of several periodic lattices.
**Lock-in transformation** An incommensurate-to-commensurate phase transition.
**Modulated structure** Result of a (small) periodic distortion of a periodic pattern. It is periodic when the two periods are commensurate; an incommensurate modulation results in an aperiodic structure.
**Octagonal phase** Quasicrystal with one periodic direction and an eightfold rotational symmetry.
**Phason** Degree of freedom associated with relative phases of incommensurate waves.
**Quasicrystal** Aperiodic crystal without an underlying periodic lattice structure, usually showing noncrystallographic symmetries that cannot occur in periodic crystals.
**Quasiperiodic** A particular type of aperiodicity where the Fourier transform is pure point, i.e., consists only of

Bragg peaks, and is supported on the (usually dense) set of integral linear combinations of a finite set of vectors in Fourier space.

**Quasiperiodic tiling** Quasiperiodic space-filling tiling that serves as the analogue of the periodic lattice in structure models of perfect quasicrystals.

**Random tiling** Space-filling tiling in which tiles are arranged randomly. Random tilings are believed to be a more realistic description of the structure of real quasicrystals than perfect quasiperiodic tilings.

**SOLIDS** have traditionally been classified as either crystalline or amorphous. The basic property that distinguishes a crystal from an amorphous or glassy material is the long-range positional order of its microscopic constituents. Classical crystallography deals with lattice-periodic structures that can be described by a space-filling periodic repetition of a single microscopic building block, the so-called unit cell. However, order does not imply periodicity, and over the last decades it has become evident that aperiodically ordered materials not only are theoretically possible, but commonly realized in nature. Aperiodic crystals can be classified into incommensurate crystals, known since the 1950s, on the one hand, and quasicrystals, discovered in the early 1980s, on the other. In the former, aperiodicity is due to the combination of several periodic structures with incommensurate periods. Quasicrystals are, in a sense, a more radical manifestation of aperiodic order, as the atomic positions cannot be interpreted in terms of underlying periodic lattices in three dimensions. They are usually identified by symmetries that are incompatible with lattice periodicity and hence forbidden in classical crystallography, such as icosahedral symmetry. The beautiful symmetry, the peculiar aperiodic order, the rather intricate and subtle atomic structure, the unique (and only partially understood) physical properties, and, last but not least, the quest for technical applications have made quasicrystals an important topic of crystallography, mathematics, physics, chemistry, and materials science.

## I. APERIODIC CRYSTALS

Crystals, in our common perception, are characterized by their morphology, their faceted shape, and have traditionally been be classified according to their symmetry. Obviously, the regularity of crystals reflects the underlying order of their microscopic structure. It can be visualized by the beautifully ordered patterns of sharp diffiraction spots, so-called Bragg peaks, as observed, for instance, in X-ray diffiraction experiments. For a long period of time, it was taken for granted that the microscopic structure of crystals, apart from defects that exist in any real solid, is periodic in space. In other words, associated to a crystal there is a three-dimensional periodic lattice, and this lattice also determines the possible symmetries that may be apparent in the crystal shape.

### A. Periodic Crystals

A "conventional," periodic crystal is thus characterized by a periodic lattice. Once the distribution of atoms in a single fundamental domain of the lattice, a unit cell, is known, the entire structure is determined by periodicity. In addition to this translational symmetry, crystalline structure may have other symmetries such as a rotational symmetry with respect to a certain axis. However, these symmetries have to be compatible with each other, and this restricts the possible symmetries of a periodic three-dimensional crystal to one of 230 crystallographic space groups classified by Schoenflies and von Fedorow in the late 19th century. In particular, the crystallographic restriction only concedes two-, three-, four-, and sixfold rotational symmetry. Other symmetries, such as fivefold rotational symmetry or icosahedral symmetry, cannot be reconciled with lattice periodicity in three-dimensional space, and thus cannot be accommodated in periodic crystals.

In general, the atomic structure of a material need not manifest itself in the shape of its surface; for instance, the structure of gold cannot be guessed by the morphology of a gold nugget. Only if the surfaces correspond to special planes, for instance, planes parallel to the faces of a cube in a cubic crystal, does the shape of the crystal reflect its atomic structure. This happens if special surfaces of a crystal are energetically favored and if the growth velocity of these surfaces is high. For example, minerals frequently develop beautiful facets.

### B. Diffraction

Direct access to the underlying atomic structure and its long-range order is provided by diffraction experiments. An X-ray, electron, or neutron beam is scattered by the sample, and interference gives rise to a diffraction pattern that can be recorded, providing information about the structure of the material. Let $\rho(\mathbf{r})$ denote the density of scatterers in space, and $\mathbf{q} = \mathbf{k}_{\text{out}} - \mathbf{k}_{\text{in}}$ the scattering vector, i.e., the momentum diffierence between the incoming and the scattered radiation. Provided that the scattering is elastic and that multiple scattering can be neglected, the measured intensity $I(\mathbf{q})$ is proportional to the Fourier transform $\hat{g}(\mathbf{q})$ of the pair correlation or Patterson function

$$g(\mathbf{r}) = \int d^3\mathbf{r}' \rho(\mathbf{r} + \mathbf{r}')\rho(\mathbf{r}'). \qquad (1)$$

In many situations, $I(\mathbf{q})$ can also be expressed directly in terms of the absolute square of the Fourier transform $\hat{\rho}(\mathbf{q})$ of the scattering density,

$$I(\mathbf{q}) \sim |\hat{\rho}(\mathbf{q})|^2. \tag{2}$$

If the scattering density $\rho$ is periodic with respect to the three-dimensional lattice spanned by the base vectors $\mathbf{a}_1$, $\mathbf{a}_2$, and $\mathbf{a}_3$, the diffraction intensity $I(\mathbf{q})$ is concentrated on a lattice, the reciprocal lattice in momentum space. This lattice is spanned by the three dual base vectors $\mathbf{a}_1^*$, $\mathbf{a}_2^*$, and $\mathbf{a}_3^*$. The diffraction pattern thus consists of sharp peaks, so-called Bragg peaks, at positions

$$\mathbf{q} = \sum_{j=1}^{3} h_j \mathbf{a}_j^* \tag{3}$$

in momentum space, where $h_j$, $j = 1, 2, 3$, are integer numbers indexing the diffraction spots.

Incommensurate crystals and quasicrystals, unlike amorphous solids, also display sharp Bragg diffraction patterns, but in contrast to periodic crystals, the positions of the diffraction spots do not lie on a periodic lattice in three-dimensional momentum space. Indexing the Bragg peaks as linear combinations of three base vectors as in Eq. (3) would require irrational coeficients. However, upon enlarging the number of vectors $\mathbf{a}_j^*$, $j = 1, 2, \ldots, D > 3$, one recovers an indexing scheme

$$\mathbf{q} = \sum_{j=1}^{D} h_j \mathbf{a}_j^* \tag{4}$$

with integer coefficients $h_j$, but now $D > 3$ integers are required to index a Bragg spot. Assuming that the base vectors $\mathbf{a}_j^*$ are linearly independent with respect to integral linear combinations (otherwise one could do with fewer vectors), one finds that the set of integral linear combinations of the $\mathbf{a}_j^*$, in general, densely fills space. The generalization from Eq. (3) to Eq. (4) may appear innocent, but the question remains of how the aperiodic order in real space that looks like that produces such diffraction patterns.

## C. What is Aperiodic Order?

At first sight, the term "aperiodic order" may appear paradoxical. However, there exists a wealth of possible structures that, in a sense, are intermediate between the periodic order of a perfect crystal and the disorder that one may find realized in amorphous intermetallic alloys. Aperiodic order is realized in incommensurate crystals and quasicrystals, which are discussed in detail below; while the discovery of these structures came quite as a surprise at the time, it is now apparent that these are not merely rare caprices of nature, but, on the contrary, occur quite commonly.

The kind of aperiodicity encountered here is known as quasiperiodicity, and may be most easily understood by a one-dimensional example. The paradigm of a periodic function is the trigonometric function $\sin(x)$, which is periodic with period $2\pi$, i.e., $\sin(x + 2\pi) = \sin(x)$. Now, consider the sum of two sine functions

$$f(x) = \sin(x) + \sin(cx) = 2 \sin\left(\frac{1+c}{2}x\right) \cos\left(\frac{1-c}{2}x\right), \tag{5}$$

where $c$ is some fixed number. Is the function $f(x)$ periodic? Well, this depends on the values of $c$. If $c$ is a rational number, $c = m/n$ with coprime integers $m$ and $n$, then the periods $2\pi$ and $2\pi c = 2\pi m/n$ are commensurate, and the function is periodic with period $2\pi n$ because $\sin[c(x + 2\pi n)] = \sin(cx + 2\pi m) = \sin(cx)$. However, if $c$ is irrational, say $c = \sqrt{2}$, the two frequencies are incommensurate, and $f(x)$ is aperiodic. This can also be seen from the product form of the function $f(x)$ that is also given in Eq. (5). Looking, for instance, at the set of solutions of $f(x) = 0$, we see that the sine and the cosine functions in Eq. (5) each contribute zeros at equally spaced positions $x_k = 2\pi k/(1 + c)$ and $x_\ell = 2\pi(\ell + 1/2)/(1 - c)$, respectively, but the two spacings are incommensurate if $c$ is irrational. Still, the function $f(x)$ retains a lot of its regularity—after all, it is just the sum of two sine functions. In fact, it is almost periodic in a sense that may be inferred from Fig. 1, which shows two different sections

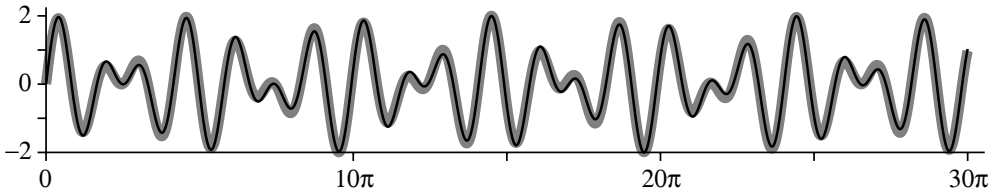



**FIGURE 1** Graphs of $f(x)$ (gray) and $f(x + x_0)$ (black) on the interval $0 \le x \le 30\pi$, with $x_0 = 58\pi/(1 + \sqrt{2})$. The grey curve has been widened to a "tube" of four times the width of the black line, indicating the size of the deviation between the two curves on the interval shown.

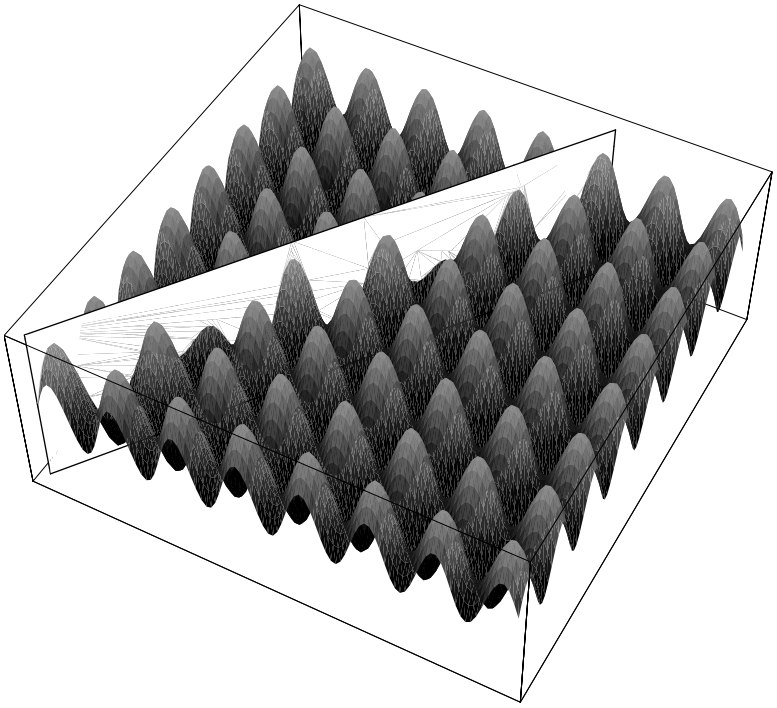


**FIGURE 2** The quasiperiodic function $f(x)$ of Eq. (5) as a cut of the two-dimensional periodic "egg carton" function $F(x, y)$, Eq. (6), for $c = \sqrt{2}$.

of the graph of the function $f(x)$ that almost agree on a large interval.

In Fourier space, the function $f(x)$ is represented by two frequencies, either commensurate or incommensurate. In general, a function is periodic if all its frequencies in Fourier space are situated on a periodic lattice, thus are integer linear combinations of $d$ base vectors in $d$ space dimensions. A quasiperiodic function is a generalization of this concept where, again, all frequencies are integer linear combinations of $D$ vectors, but this number may be larger than the spatial dimension, $D \geq d$, where equality implies periodicity. Thus, the Fourier transform of a quasiperiodic structure that is not periodic will be supported on the set of all integer linear combinations of $D > d$ vectors in $d$-dimensional Fourier space. Thus, the diffraction pattern of a quasiperiodic structure consists of Bragg peaks located on a particular dense set of points in Fourier space, and the Bragg peaks can be indexed by $D$ integer numbers as in Eq. (4). In mathematical terminology, the Fourier transform is finitely generated over the integers; its support is a module of rank $D$. One can also construct ordered structures whose Fourier transforms consist of Bragg peaks that cannot be indexed by a finite number of integers; however, such structures have not yet been observed in nature.

It is instructive to think of the function $f(x)$ as a cut of the two-dimensional periodic function

$$F(x, y) = \sin(x) + \sin(y) \tag{6}$$

along the line $y = cx$, i.e., $f(x) = F(x, cx)$. This is shown in Fig. 2. The quasiperiodic function emerges as a section through a higher dimensional periodic function along a direction that induces the incommensurability. It is precisely the same idea that underlies the higher dimensional description of incommensurate crystals and perfect quasicrystals.

## D. Incommensurate Crystals

Incommensurate structures in crystals have been known since the 1950s. These are magnetic crystals that exhibit a helical ordering of the spins which is incommensurate with the underlying periodic lattice structure. A sketch of such a situation is shown in Fig. 3a. The existence of these systems may not seem too surprising, as the incommensurability occurs due to an additional degree of freedom, the spin, in an otherwise perfectly periodic crystal. However, it did not take long until evidence for incommensurability of the structure itself was found in form of so-called satellite reflections in diffraction patterns. The satellite peaks show up next to the main reflctions, hence the name, but their coordinates with respect to the lattice of main peaks are not simple fractions, and may even depend continuously on temperature. In particular this continuity invalidates the

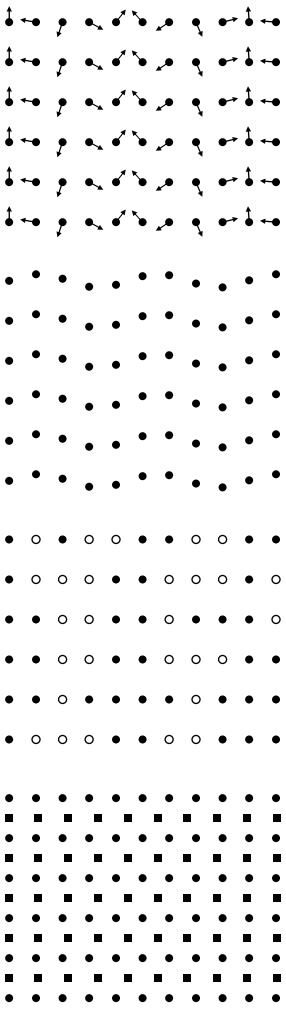


**FIGURE 3** Examples of incommensurate structures. (a) Incommensurately ordered degree of freedom. (b) Displacive modulation. (c) Occupational modulation. (d) Composite structure.

interpretation of the satellite peaks in terms of structurally stable periodic superstructures, which refers to another periodic structure with a large unit cell that is superimposed on the original lattice.

Sketches of several possible scenarios of incommensurability in a crystalline structure are compiled in Fig. 3. Most important are the modulated structures depicted in Figs. 3b and 3c. These crystals are characterized by a periodic deviation, the modulation, from their underlying periodic lattice structure. The structure is incommensurate if the period of the modulation does not match the lattice periodicity. In Fig. 3b, the modulation is displacive, i.e., the positions of the atoms are shifted. Figure 3c shows another scenario where the deviation is occupational. In this case, the periodic modulation determines the occupation probability of the perfect lattice positions. Finally, Fig. 3d shows an example of an incommensurate composite structure which consists of two subsystems, indicated by two different symbols, each of which by itself is perfectly periodic, but which are incommensurate with each other.

For clarity, the modulations in Fig. 3 are chosen to be one dimensional, i.e., the deviation from the perfect periodic structure occurs in one direction only. For such systems, one would need $D = d + 1$ base vectors in Eq. (4) to describe their Fourier transform. The additional vector accounts for the periodic modulation in one direction of space, which need not coincide with any lattice direction of the basic structure. There exist many examples for one-dimensional modulations; higher dimensional modulations also occur in nature, where the dimension of the modulation may be defined as the number of additional vectors $D - d$ in Eq. (4) that are needed in order to describe the diffraction pattern of the structure.

## E. Quasicrystals

Quasicrystals entered the scene only in the early 1980s when icosahedral symmetry was found in a selected-area electron diffraction analysis of a rapidly cooled Al–Mn alloy (see Fig. 4). Figure 5 shows the two Platonic solids with icosahedral symmetry, the icosahedron and the dodecahedron. The 20 faces of the icosahedron are equilateral triangles; they meet in 30 edges and 12 vertices. The dodecahedron consists of 12 faces that are regular pentagons, and comprises 30 edges and 20 vertices. Both polyhedra show the same symmetry. There are six fivefold axes, connecting opposite vertices of the icosahedron or the centers of opposite pentagons of the dodecahedron, respectively. The 10 threefold axes connect the centers of opposite faces of the icosahedron or opposite vertices of the dodecahedron, respectively; in both cases, the 15 twofold rotational axes connect midpoints of opposite edges. The three different types of symmetry axes and their relative orientations in space are perfectly recovered in the diffraction pattern of Fig. 4.

Apparently, the diffraction pattern consists of sharp spots, and thus the structure must be ordered. However, as icosahedral symmetry is incompatible with lattice periodicity, it cannot be a conventional periodic crystal. Even though incommensurate crystals had been known for about 30 years and an explanation of the observed diffraction pattern in terms of a quasiperiodically ordered structure was readily available, this discovery came as a surprise and gave rise to prolonged controversy. However, it soon became clear that alternative interpretations of the results in terms of periodic crystals were either inconsistent or required periodic structures with huge unit cells comprising thousands of atoms. Strictly speaking, this question can never be resolved completely because the peaks observed in experiments cannot be infinitely sharp

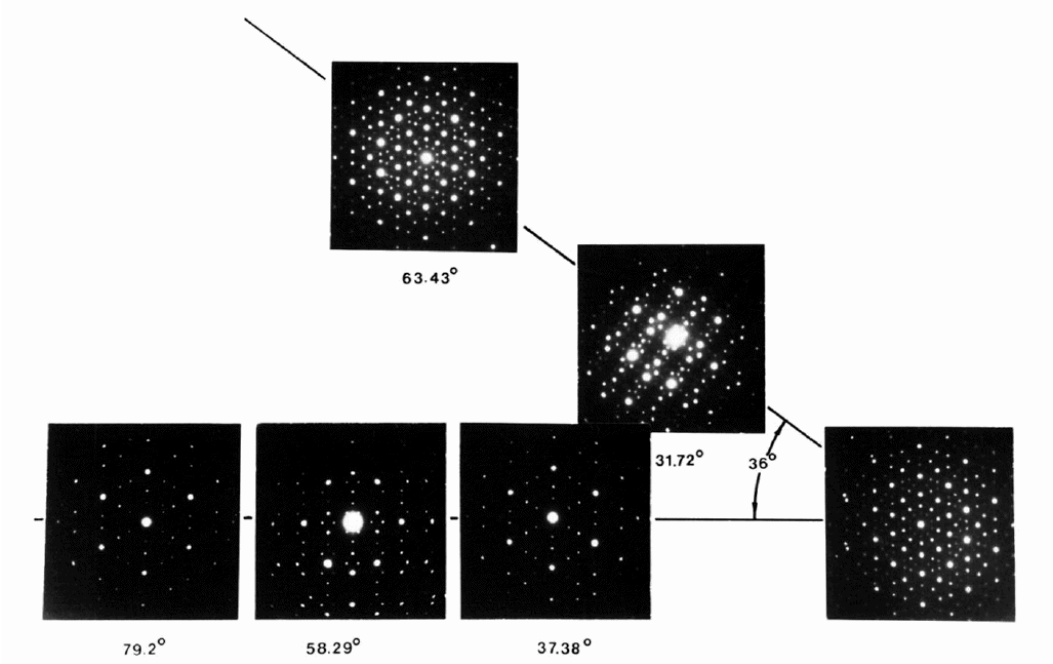



**FIGURE 4** The first reported experimental evidence of a quasicrystal: Selected-area electron diffraction patterns of a rapidly cooled Al–Mn alloy showing icosahedral symmetry. Different symmetries were observed by tilting the sample by the angles given in the figure, exactly corresponding to the angles between different symmetry axes of isosahedral symmetry. [From Shechtman, D., Blech, I., Gratias, D., and Cahn, J. W. (1984). "Metallic phase with long-range orientational order and no translational symmetry." *Phys. Rev. Lett.* **53,** 1951–1953. Copyright 1984 by the American Physical Society.]

due to the finite size and disorder of the crystal and due to the limited resolution. Thus, one may always describe the experimental data in terms of a hypothetical periodic structure; when the number of atoms in a unit cell becomes too large, however, the description as an aperiodic crystal is not only more elegant and appealing, but also much simpler. In particular, it can easily account for the symmetry of quasicrystals (Fig. 6).

Soon after the discovery of icosahedral quasicrystals, intermetallic alloys with further crystallographically forbidden symmetries were found, showing either a 12-, 10-, or 8-fold symmetry axis. The corresponding diffraction patterns reveal that these quasicrystals are periodic in one direction of space, which coincides with the symmetry axis, i.e., they consist of a periodic stacking of planes with 12-, 10-, or 8-fold rotational symmetry. Accordingly, they are referred to as dodecagonal, decagonal, and octagonal quasicrystals.

## II. THE STRUCTURE OF QUASICRYSTALS

One would like to reduce the atomic structure of a quasicrystal to a small number of basic building blocks analogous to the unit cell for a periodic crystal. Thus, structure models of quasicrystals consist of two parts: The geometric arrangement of the building blocks, which takes care of the quasiperiodic long-range order, and the location of the atoms within each building block, their "decoration." The geometric arrangement of building blocks is conventionally encoded in a space-filling tiling with a finite number of prototiles. To some extent, these tilings can be visualized directly by high-resolution electron microscopy of quasicrystals. Whereas much is known about quasiperiodic tilings of space and their symmetry properties, the actual distribution of atoms in quasicrystalline solids remains largely unknown. Diffraction data alone do not suffice to derive the atomic density unequivocally, and electron microscopy methods are just on the verge of reaching the required atomic resolution. For several systems, sophisticated models have been proposed, although many details, for instance, the kind and the amount of inherent disorder, need to be unraveled. For this reason, the following discussion focuses on the geometric part.

### A. A One-Dimensional Quasicrystal

It is worth starting with a one-dimensional example of a quasiperiodic structure, not merely because it is easy to understand, but because it is, in fact, reflected in higher

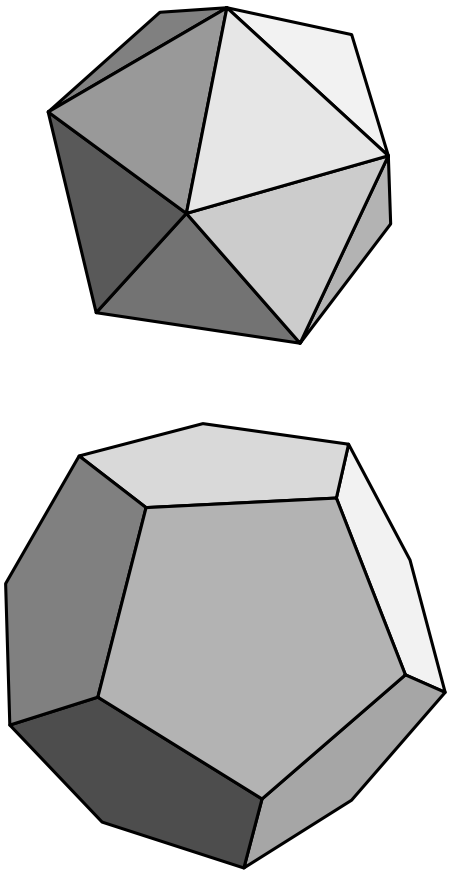


**FIGURE 5** Regular polyhedra with icosahedral symmetry: The icosahedron (left) and the dodecahedron (right).

dimensional quasicrystalline structures. The paradigm, dwelled on in almost any introductory text on quasicrystals, is the ubiquitous Fibonacci sequence, related to the celebrated Fibonacci numbers $f_n$ defined by the simple recursion

$$f_{n+1} = f_n + f_{n-1}, \qquad f_0 = 0, \qquad f_1 = 1. \quad (7)$$

The Fibonacci sequence consists of two symbols, say L and S, and can be constructed by successive application of the substitution rule S $\rightarrow$ L, L $\rightarrow$ LS, starting, for instance, with the single letter S. The resulting sequence

$$\text{LSLLSLSLLSLLSLSLLSLSL}\ldots \quad (8)$$

is aperiodic; the ratio of L's and S's in the sequence tends to the golden number $\tau$,

$$\tau = \lim_{n\to\infty} \frac{f_{n+1}}{f_n} = \frac{1+\sqrt{5}}{2}, \quad (9)$$

which is irrational, $\tau = 1.61803\ldots$. The sequence can be made into the one-dimensional "tiling" of Fig. 7a by associating two intervals of different length to the two letters.

Much as the quasiperiodic function $f(x)$ of Eq. (5) was recovered as a cut through the two-dimensional periodic function $F(x, y)$, Eq. (6), the Fibonacci quasicrystal can be constructed from the two-dimensional square lattice as shown in Fig. 8. Here, the shaded strip corresponds to the region swept out by a unit square of the lattice when moved along a line of irrational slope $1/\tau$. The lattice points within the strip are projected onto that direction, yielding a binary one-dimensional tiling of long and short intervals of length ratio $\tau$. This tiling in "physical space" coincides, apart from a shift that depends on the location of the strip, with the Fibonacci tiling obtained from the two-letter substitution rule.

By construction, it is clear that the Fourier transform of this one-dimensional structure will have pure Bragg peaks, located on the projected points of the dual lattice, which again is a square lattice. However, when projecting all lattice points, and not just those within a certain strip, the projected points are dense on the line, and one arrives at a dense set of Bragg peaks which can be indexed by two integers. Nevertheless, a measurement would yield a diffraction pattern of peaks that appear to be well separated, as shown in Fig. 7b. This apparent contradiction is resolved by realizing that only those Bragg peaks that carry more than a certain minimum intensity will be visible, and however small the minimum intensity is chosen, the set of peaks with larger intensity is discrete.

This one-dimensional Fibonacci structure can also be found in experiment; a beautiful example is given in Fig. 9a. As shown in Fig. 9b, the ordering is not always perfect, but the sample exhibits a disorder mechanism commonly known as "phason defects." These are deviations from the ideal quasiperiodic structure that can be interpreted in terms of local flips in the sequence of Fig. 7 or in terms of a slight deformation of the strip in the projection of Fig. 8. In analogy to phonons that describe the motion of atoms from their ideal positions, the motion of the projection strip perpendicular to the "physical space" can be described in terms of quasi-particles called phasons whose experimental verification and characterization is a topic of current research.

## B. Quasiperiodic Tilings

The projection approach of Fig. 8 can be generalized in a straightforward way to construct quasiperiodic tilings of space in any dimension. In particular, three-dimensional tilings with icosahedral symmetry and planar tilings with 8-, 10-, or 12-fold symmetry can be obtained. Some of these, such as the celebrated Penrose tiling shown in Fig. 10, had already been known to mathematicians and theoretical physicists before quasicrystals were discovered. As an example, the diffraction pattern of the

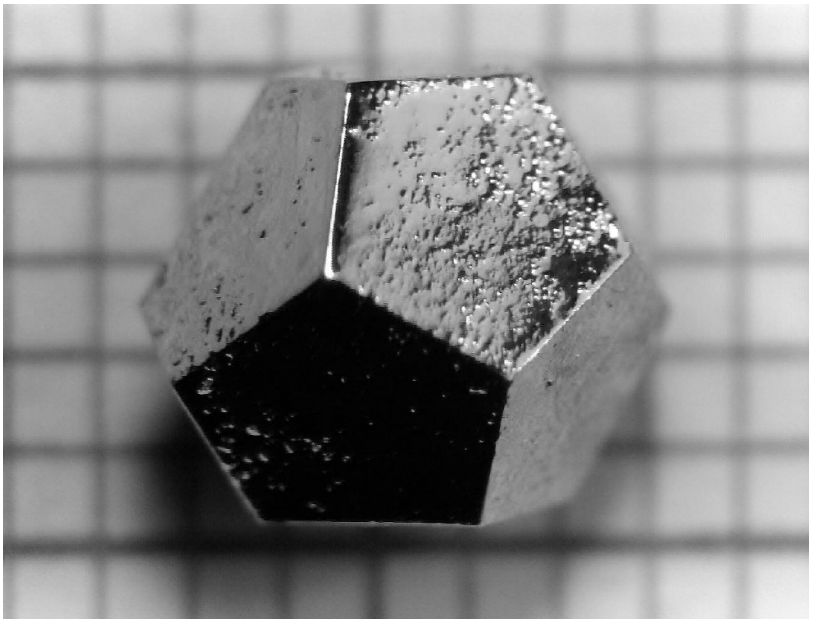


**FIGURE 6** A single-grain icosahedral HoMgZn quasicrystal, grown from the ternary melt, on a millimeter-scale background. [Figure courtesy of P. C. Canfield and I. R. Fisher; reprinted with permission from Fisher, I. R., Cheon, K. O., Panchula, A. F., Canfield, P. C., Chernikov, M., Ott, H. R., and Dennis, K. (1999). "Magnetic and transport properties of single-grain *R*–Mg–Zn icosahedral quasicrystals [R = Y, $(Y_{1-x}\ Gd_x)$, $(Y_{1-x}\ Tb_x)$, Tb, Dy, Ho, and Er]." *Phys. Rev. B* **59,** 308–321. Copyright 1999 by the American Physical Society.]

pentagon tiling is shown in Fig. 11, exhibiting perfect 10-fold symmetry.

It is worth mentioning that the golden mean $\tau$, and in fact the Fibonacci sequence itself, reappears in 10-fold and icosahedral tilings as well as in their diffraction patterns. This is related to the fivefold rotational symmetry common to these structures because $\tau = 2\cos(\pi/5)$. In the diffraction pattern of Fig. 11, it can be recognized as the length ratio of distances between peaks of similar intensity. This is related to a rescaling symmetry of the quasiperiodic tiling, the so-called inflation/deflation symmetry. In essence, it is the higher dimensional version of the substitution rule that was used to construct the Fibonacci sequence: in an inflation step, each tile is dissected into several parts of tiles such that a tiling emerges whose tiles are just scaled copies of the original prototiles and that, upon rescaling, is equivalent to the original tiling. Deflation is the reverse process, in which a number of tiles is replaced by a larger tile. For the Penrose tiling, the linear rescaling factor associated to this symmetry is just $\tau$ again. This property means that the quasiperiodic ordering is of the same kind no matter at what length scale it is probed.

Besides this nice property, the Penrose tiling has another "magic" property that it shares with a number of other tilings used in description of quasicrystalline structures. It is the existence of perfect matching rules, which means that there exist a marking of the basic prototiles and a set of local rules that determine the possible local neighborhoods of a marked tile such that, if the marked tiles are assembled as in a jigsaw puzzle, the resulting tiling is aperiodic and, in fact, is indistinguishable from a Penrose tiling. However, these matching rules do not provide a constructive instruction to produce a perfect quasiperiodic tiling: in general, after assembling a number of tiles, one meets the situation that it is impossible to add a tile without violating the rules, and there is no information on the location where the arrangement of tiles needs to be altered to rectify the problem.

Quasiperiodic tilings are a natural generalization of periodic lattices that can account for the noncrystallographic symmetries found in diffraction experiments. The

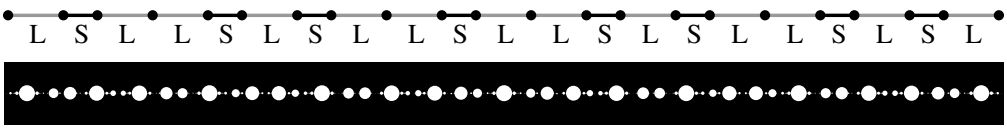



**FIGURE 7** (a) The Fibonacci quasicrystal and (b) its diffraction pattern.

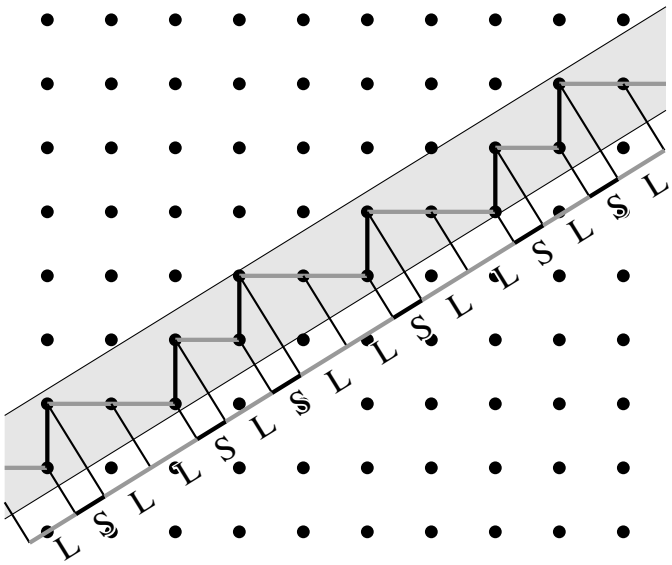



**FIGURE 8** Projection of the Fibonacci quasicrystal from the two-dimensional square lattice.

beautiful example of several difierent modifications of the decagonal phase in AlCoNi is shown in Fig. 12. Unfortunately, the information contained in the diffraction pattern is not sufficient to reconstruct the structure, as the phase information is lost. This is why holographic methods that can access the phase information have attracted increasing interest. For quasicrystals with one periodic direction, however, high-resolution transmission electron microscopy along the periodic direction yields direct information on the spatial distribution of structural units. An example for a decagonal AlCoNi quasicrystal is shown in Fig. 13a. Connecting similar contrasts in the image, one arrives at the tilings of Fig. 13b, at difierent length scales connected by multiplication with $\tau$. This observation is related to the inflation/deflation symmetry of the Penrose tiling mentioned above. For this particular alloy, the tilings obtained in this way are very close to ideal quasiperiodic tilings. This is corroborated by Fig. 13c, which shows the projection into "internal space" corresponding to the cross section of the strip in the projection of Fig. 8. If the tiling were perfect, all projected points would come to line inside the decagons. Indeed, for the three experimentally derived tilings of Fig. 13b, only few points fall outside the decagons. For this reason, this particular high-temperature AlCoNi phase is referred to as a highly perfect quasicrystal.

## C. Clusters and Coverings

Right after the discovery of quasicrystals, the idea emerged that highly symmetric atomic clusters are the basic constituents of their structure. These may be conceived as particularly stable local configurations of atoms, and a simplistic picture of a quasicrystal would be a conglomerate of such clusters held together by "glue atoms." While this approach can easily be combined with an underlying tiling picture, an alternative concept has recently attracted growing attention, according to which a covering of space rather than a tiling is employed.

The difference between a covering and a tiling is the possibility of overlaps. An example of a covering by a regular decagon is shown in Fig. 14. Here, the possible overlaps are restricted by the markings of the decagon and the resulting structure is in fact completely equivalent to the Penrose tiling of Fig. 10. This shows one of the advantages of the covering picture: only a single "quasi unit cell" is needed, whereas at least two different tiles are required in a Penrose tiling. Furthermore, the shape of the quasi unit cell resembles the typical motives observed in electron microscopy, whereas the tiles usually have to be imposed artificially. Last but not least, the picture is very much reminiscent of interpenetrating atomic clusters, which is rather appealing from the physical point of view. In fact, the structures that are equivalent to Penrose tilings are characterized by a maximal density of clusters.

An example of a structure model of a decagonal AlCoNi quasicrystal based on the decagon covering is shown in Fig. 15. The model was chosen such that it fits the features seen in atomic resolution *Z*-contrast scanning transmission electron microscopy. Here, the size of the basic decagonal cluster is 2 nm. Note the asymmetric decoration, matching the asymmetric contrast in the *Z*-contrast image.

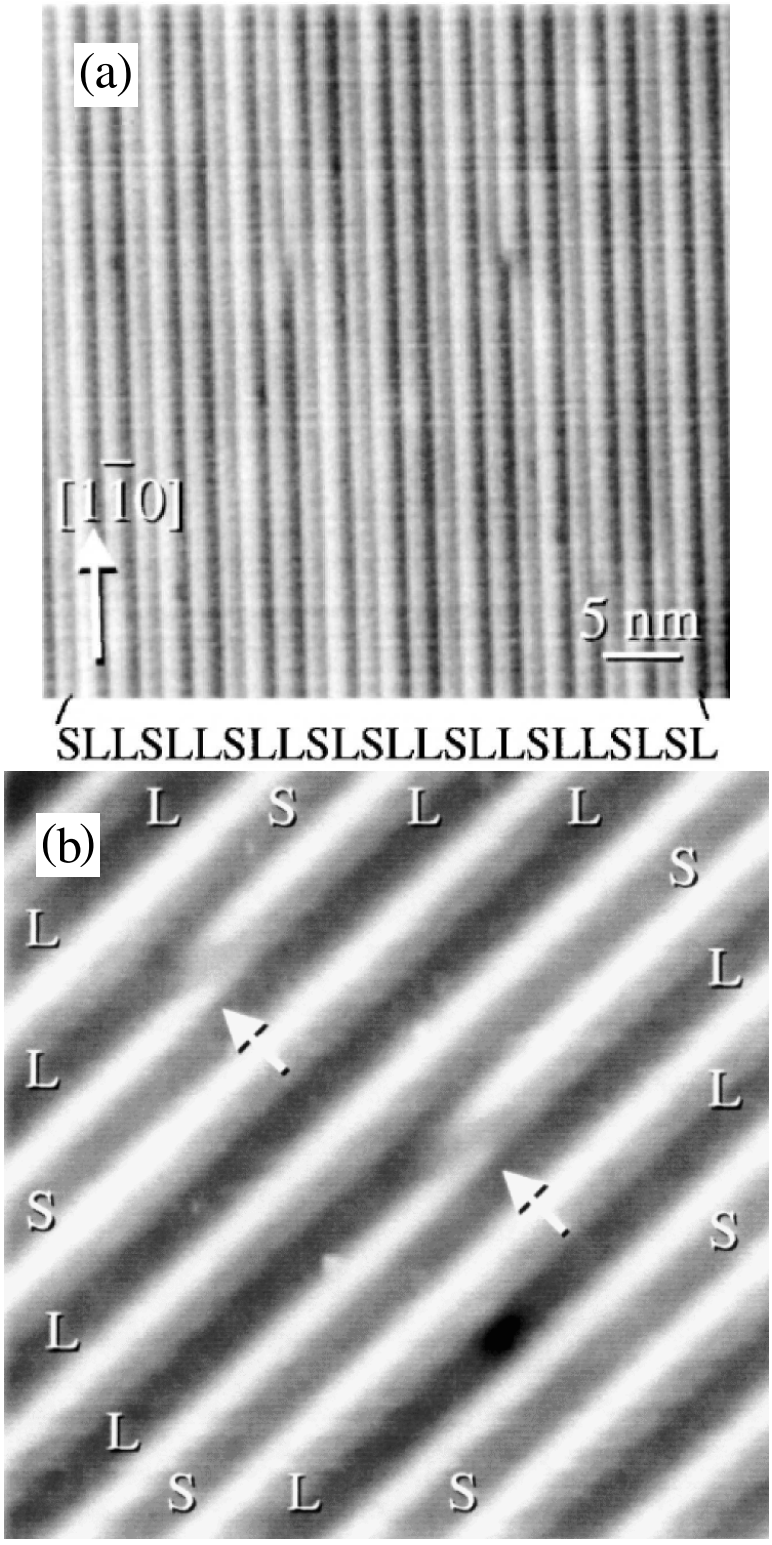



**FIGURE 9** (a) Scanning tunneling microscopy image of a 1.5-nm silver film on a GaAs(110) surface and (b) a detail showing two phason defects marked by arrows. [Reprinted with permission from Ebert, Ph., Chao, K.-J., Niu, Q., and Shih, C. K. (1999). "Dislocations, phason defects, and domains walls in a one-dimensional quasiperiodic superstructure of a metallic thin film." *Phys. Rev. Lett.* **83,** 3222–3225. Copyright 1999 by the American Physical Society.]

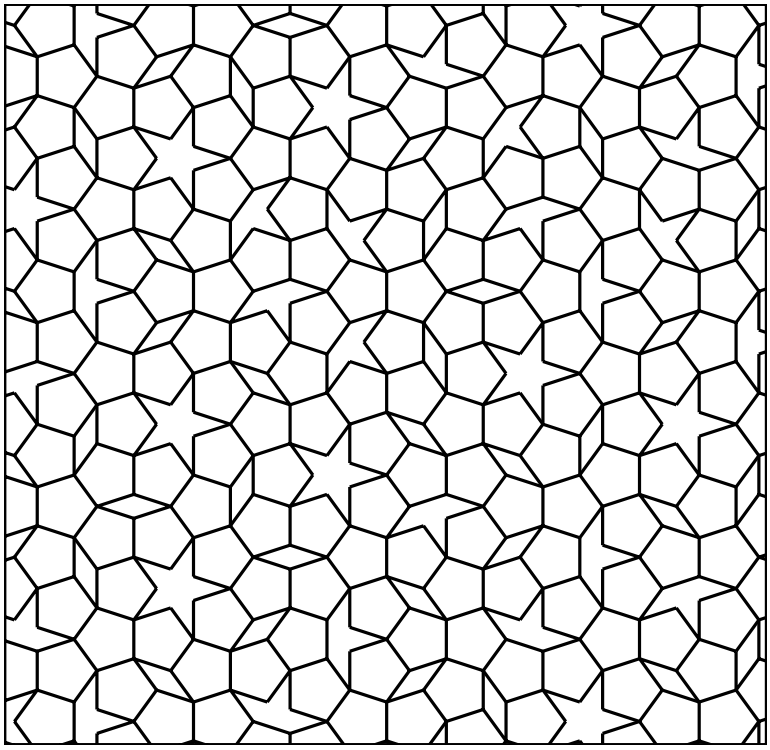


**FIGURE 10** The Penrose pentagon tiling.

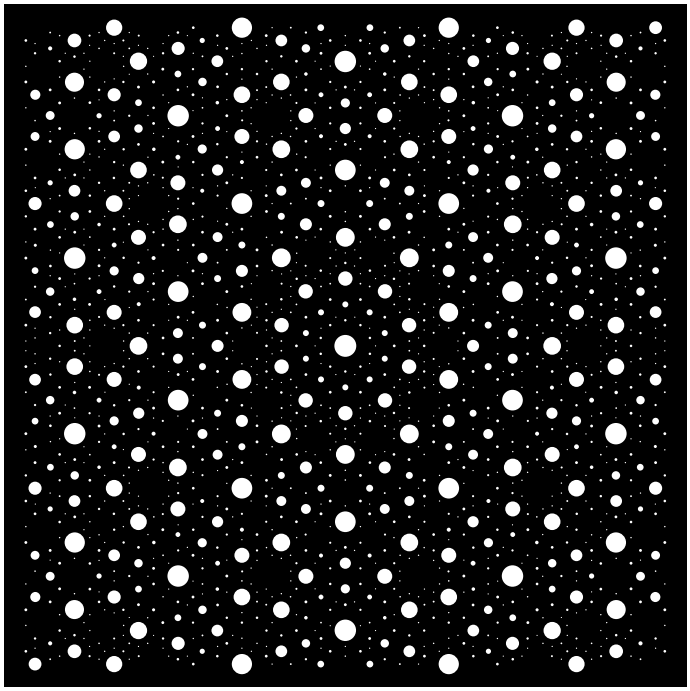


**FIGURE 11** Fourier transform of the Penrose pentagon tiling. The diffraction peaks are represented by disks whose areas are proportional to the intensity.

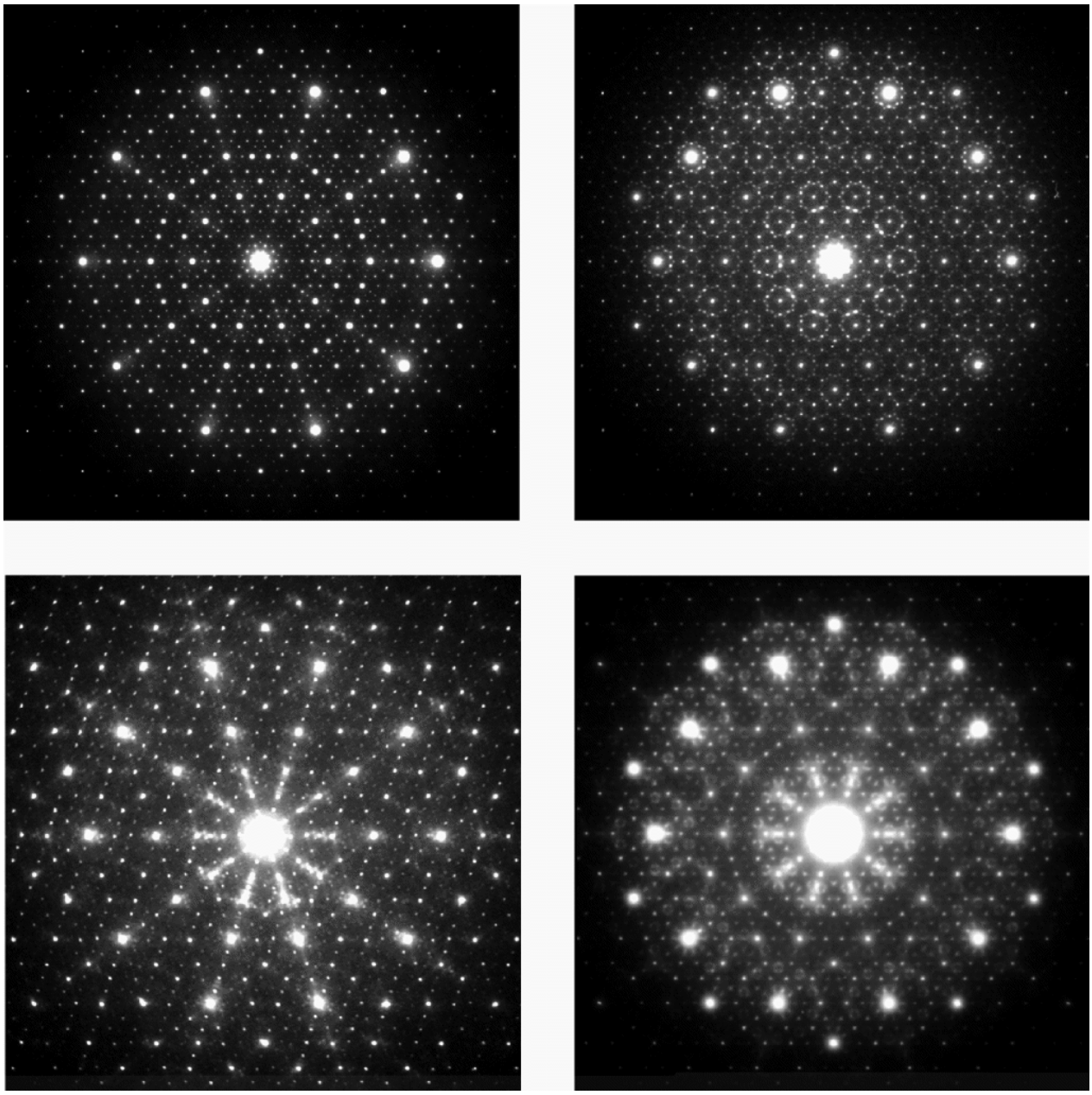
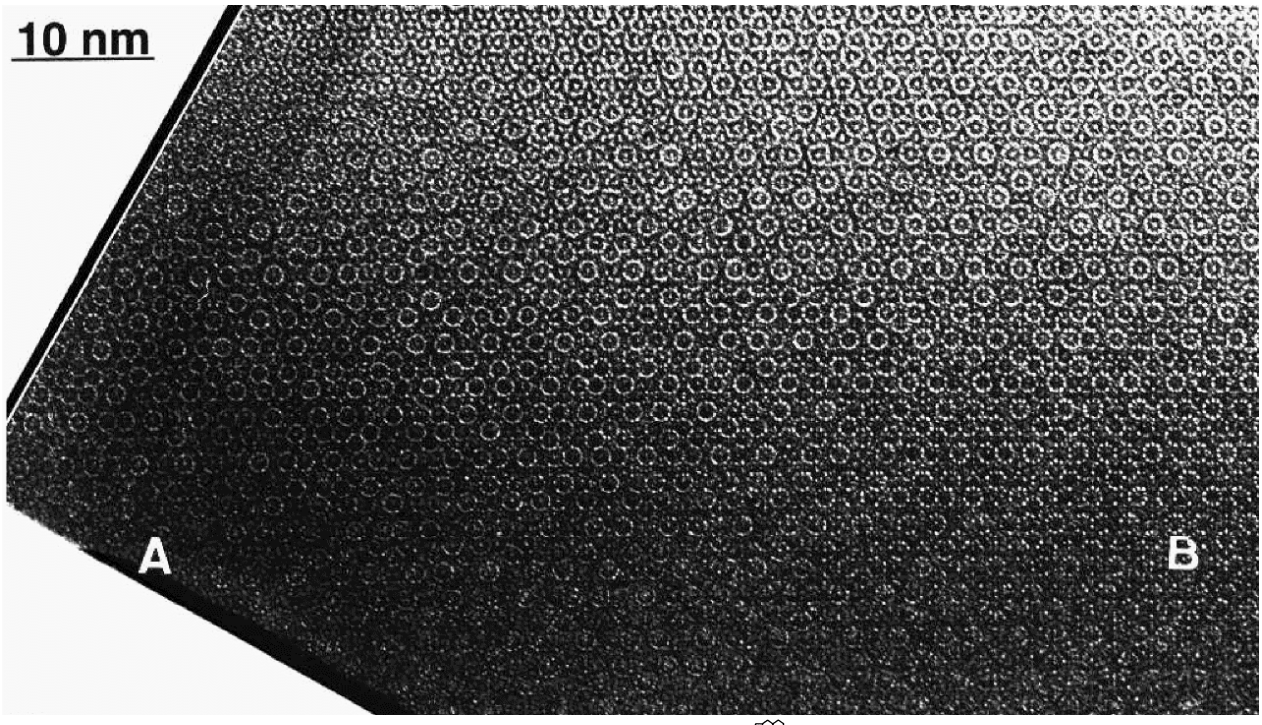


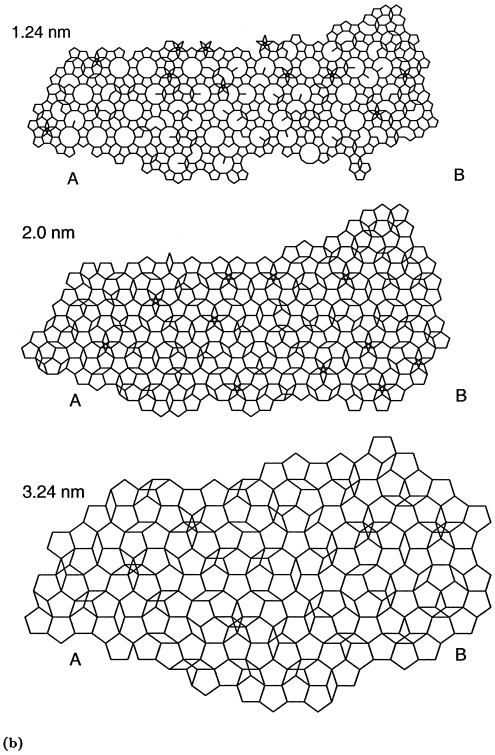
**FIGURE 12** Electron diffraction images of four of the eight known structural modifications of the decagonal phase in AlCoNi. [From Ritsch, S., Beeli, C., Nissen, H.-U., Gödecke, T., Scheffer, M., and Lück, R. (1998). *Phil. Mag. Lett.* **78,** 67–75.]

## D. Disorder and Randomness

The decagonal AlCoNi quasicrystal is a rather special example in the sense that, at high temperature, its structure is well represented by a perfect quasiperiodic tiling or covering of the plane, whereas most other structures show a large amount of configurational disorder. The fact that the most perfect structures are found at higher temperatures, and the large compositional ranges where metastable quasicrystals can be obtained by rapid solidification of liquid alloys, are indications that entropy plays an important role in the stability of quasicrystals.

One way to incorporate configurational entropy is by considering random tilings rather than perfectly ordered quasiperiodic tilings as the basis of structure models. In a random tiling, all possible space-filling arrangements of a certain set of prototiles are taken into account. A perfect tiling such as in Fig. 10 can be randomized by local




(b)

**FIGURE 13** (a) High-resolution electron microscopy image of a decagonal AlCoNi quasicrystal, and tiling analysis (b) in physical and (c) in internal space. [From Ritsch, S., Beeli, C., Nissen, H.-U., Gödecke, T., Scheffer, M., and Lück, R. (1996). "Highly perfect decagonal Al–Co–Ni quasicrystals." *Phil. Mag. Lett.* **74,** 99–106.]

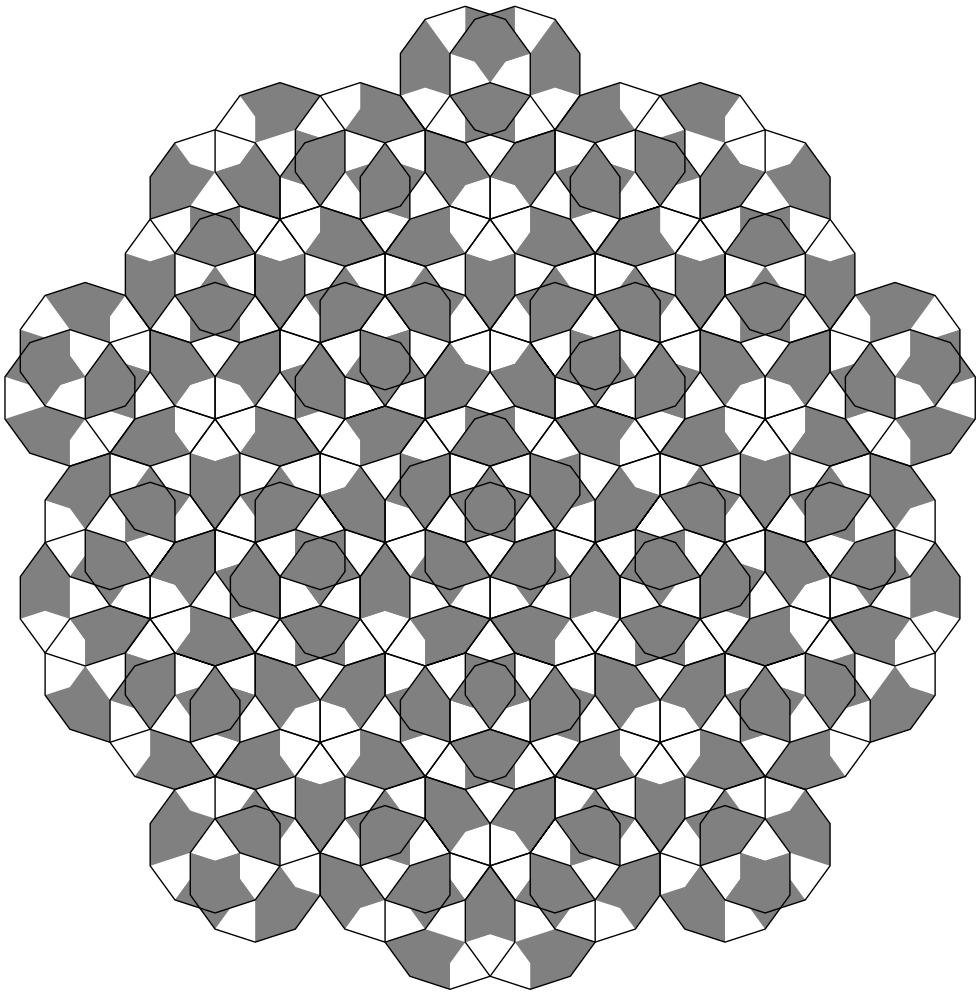


**FIGURE 14** This covering of the plane by marked regular decagons was shown to be equivalent to the Penrose tiling of Fig. 10 by P. Gummelt.

flips to obtain a random tiling such as shown in Fig. 16. The disorder may not be apparent on first view, but, for instance, the star-shaped tiles are distributed in a less regular way and may appear at short distances in the randomized tiling, while they are separated by at least two pentagons in the perfect tiling. In other words, the random tiling comprises local configurations that are absent in the perfect case. In the projection setup, this can be interpreted as an arbitrary deformation of the projection strip of Fig. 8. It turns out that arrangements that lead to the highest statistical symmetry are entropically favored, which may be interpreted as an entropic mechanism that stabilizes quasicrystals. Although the tilings are stochastic, their diffraction images are still believed to be point-like. Moreover, most electron microscopy investigations of the local order in quasicrystals support this stochastic picture.

As a further benefit, the random tiling picture can also resolve the somewhat mysterious growth mechanism of quasicrystals. The problem with quasicrystal growth is that a perfect quasiperiodic structure such as the Penrose tiling in Fig. 10 cannot be grown by local growth rules. Thus, even though the perfect matching rules would allow for an assignment of energies to local configurations that result in a perfect quasiperiodic ground state, the random tiling scenario seems to be much more realistic, with the system choosing among many possible local configurations that just differ slightly in energy.

## III. PHYSICAL PROPERTIES OF QUASICRYSTALS

In contrast to the incommensurate crystals, quasicrystals constitute a fairly coherent class of materials which share

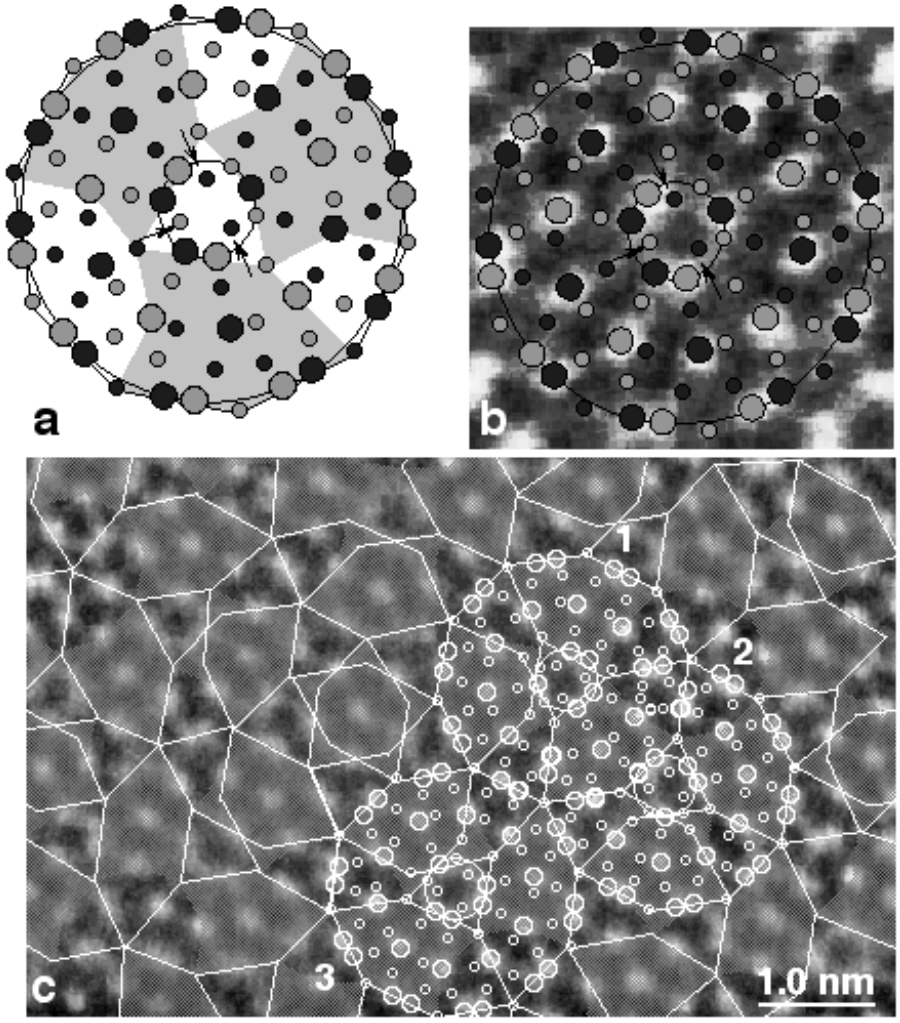



**FIGURE 15** Structure model of decagonal AlCoNi and *Z*-contrast images. (a) Decoration of the Gummelt decagon by transition metal atoms (large circles) and aluminum atoms (small circles). Dark spots refer to positions $c = 1/2$ along the periodic axis, light symbols to $c = 0$. The arrows denote positions that moved significantly during a first-principles relaxation of the structure. (b) The structure superimposed on a *Z*-contrast image. (c) Lower resolution *Z*-contrast image, several clusters, and their overlaps. [From Yan, Y., and Pennycook, S. J. (2001). "Chemical ordering in $Al_{72}Ni_{20}Co_8$ decagonal quasicrystals." *Phys. Rev. Lett.* **86,** to appear. Copyright 2001 by the American Physical Society.]

similar physical properties. These are briefly summarized below.

### A. Appearance of Quasicrystals in Nature

One may wonder why it took about 30 years after the first investigations of incommensurably modulated crystals until quasicrystals were finally discovered. Arguably, one reason for this is the natural appearance of incommensurate crystals. Whereas incommensurate phases are found in minerals like plagioclase feldspars, the predominant number of the known quasicrystals occur in intermetallic systems, which have to be prepared synthetically. A compilation of the concentrations of the components and the temperature ranges where such structures exist is provided by phase diagrams, which are particularly important for sample preparation. Most of the systems forming quasiperiodic and incommensurate crystals show, in addition, a variety of phases with different structures. Since the complexity of phase diagrams increases with the number of phases, phase diagrams of such systems are generally intricate. An example is given in Fig. 17, which shows a cut through the aluminum-rich part of the three-dimensional phase diagram of the

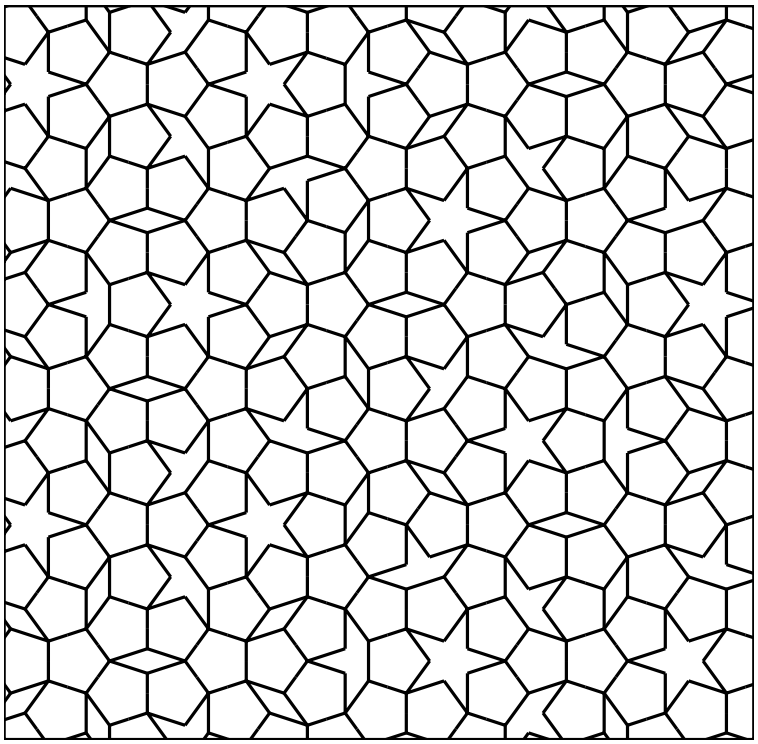


**FIGURE 16** A randomized pentagon tiling.

ternary Al–Pd–Mn system at a constant temperature of 873 K.

### B. Morphology

In certain cases, it is possible to grow single-grain quasicrystals from the melt. An example is shown in Fig. 6. The pentagonal surfaces of the dodecahedral crystal are perpendicular to the fivefold axis of the icosahedral structure. The existence of other planes with different symmetries was corroborated by the investigation of voids in quasicrystalline alloys. The magnificent electron microscopic picture in Fig. 18 depicts the surface of a hole inside an icosahedral quasicrystal. A variety of different polygons emerge, indicating twofold, threefold, and even more complex surfaces. For decagonal quasicrystals, a prismatic morphology prevails. These needle-shaped crystals, which often show a decagonal cross section, form as a result of an anisotropic growth of the quasicrystalline grains, which usually grow considerably faster in the periodic direction than in the quasiperiodic planes.

### C. Mechanical Properties

The mechanical properties of metallic alloys are strongly influenced by the type and the concentration of structural defects. For instance, the plastic deformation of the material happens by the migration of defects. Quasicrystalline structures possess special kinds of defects not existing in crystalline structures. Besides dislocations, one finds structural rearrangements, so-called phasons, which do not generate structural misfits but destroy the perfect quasiperiodic order. Quasicrystals are mostly very hard and brittle, a very common property of intermetallic alloys. For example, the Vickers hardness of Al-based quasicrystals is comparable to the hardness of steel and slightly lower than the hardness of silicon. The brittleness of the quasicrystalline alloys is expressed in their low toughness, which is around 40 times lower than that of other Al-based alloys. This circumstance changes at higher temperatures, above about 900 K, where a brittle-to-ductile transition was experimentally observed in which a softening of the material occurs. For the explanation of this behavior, the knowledge of the structure and the kinetics of the defects is required. Due to the high symmetry of icosahedral quasicrystals, a higher isotropy of the mechanical properties compared to crystals was expected and was verified by experiments.

### D. Electronic Properties

As it is the electronic interaction of the charged constituents that holds a solid together, the electronic structure

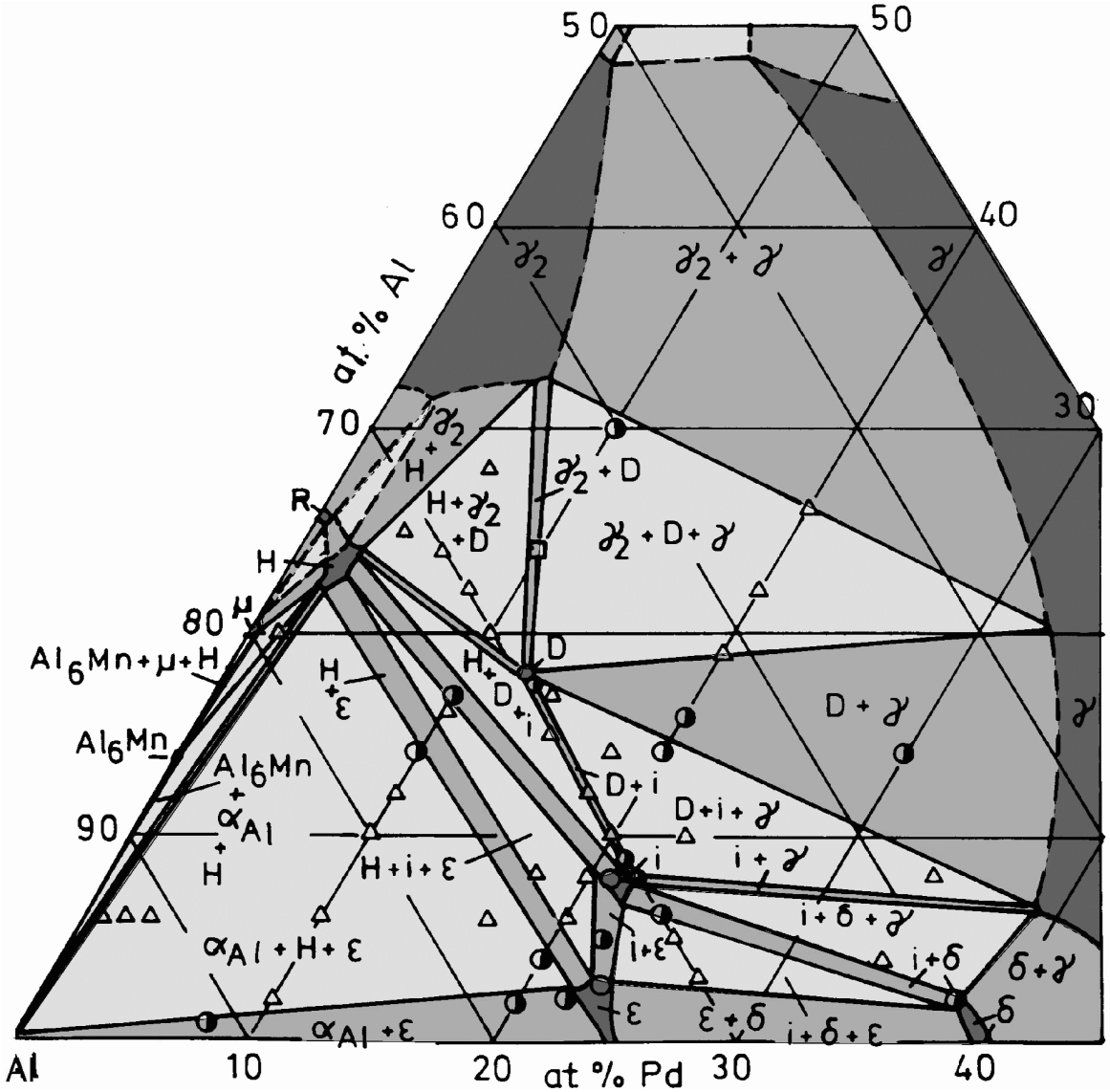



**FIGURE 17** Cut through the Al-rich part of the three-dimensional phase diagram of the ternary system Al–Pd–Mn at a constant temperature of 873 K. The concentration of the three components is represented by a plane. Lines in the diagram separate regions of thermodynamic equilibria of one or more phases, depending on the concentration. For most compositions, no structure can exist (light and middle gray). Those alloys will decompose into several phases with different compositions. Regions where only a single phase exists (dark gray) are always separated by regions in which several phases are coexistent. Remarkable is the coexistence of a decagonal (D) and an icosahedral (i) phase in the two-phase region (D + i) in one ternary system. [Figure courtesy of T. Gödecke and R. Lück; From T. Gödecke and R. Lück (1995). "The aluminum–palladium–manganese system in the range from 60 to 100 at. % Al." *Z. Metallkd.* **86,** 109–121.]

of a solid is important for stability. If, in turn, the spatial arrangement of the atoms influences the electronic structure, a complex interplay between the electrons and the structure will result. This is often observed in intermetallic phases, where, under special conditions, the electronic system favors special atomic structures. For many quasicrystals, a so-called Hume–Rothery stabilization is assumed. Its main fingerprint is the development of a pseudogap in the electronic density of states at the Fermi level, which may also explain the transport anomalies observed in quasicrystals. For instance, the electric conductivity is very low, so quasicrystals are poor conductors. The conductivity of quasicrystals, contrary to the conductivity of metals, decreases enormously as the

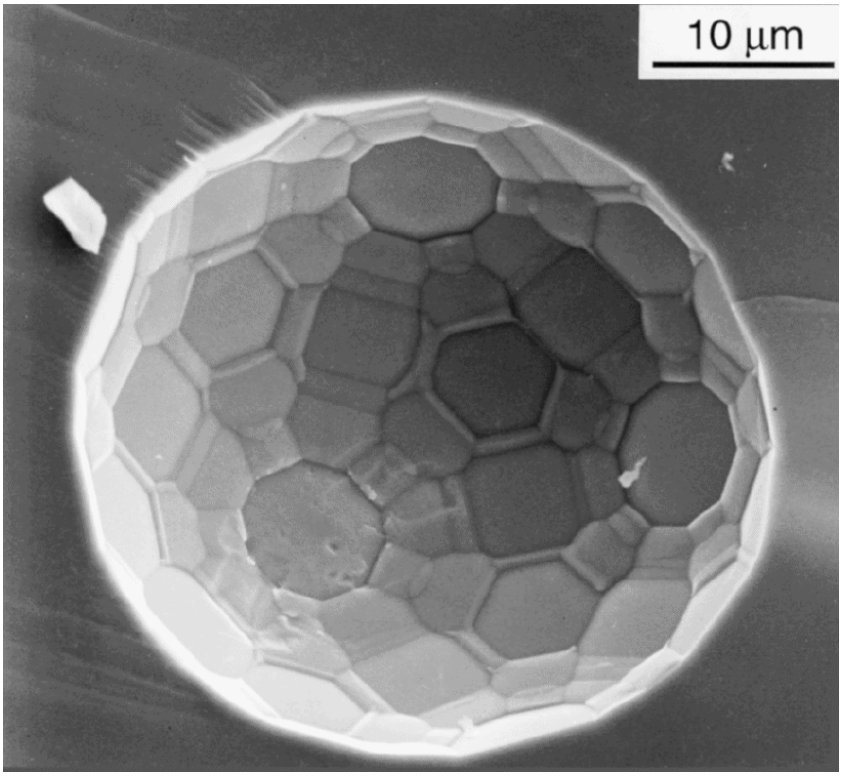



**FIGURE 18** A beautifully faceted hole in an icosahedral quasicrystal. [From Beeli, C., Gödecke, T., and Lück, R. (1998). "Highly faceted growth shape of microvoids in icosahedral Al–Mn–Pd." *Phil. Mag. Lett.* **78,** 339–348.]

temperature is lowered, and it also appears to decrease with increasing structural perfection of the sample. Similar anomalies are also observed in other transport properties, such as thermal conductivity, Hall coefficients, and thermopower.

## E. Magnetic Properties

As mentioned above, magnetic moments can form incommensurable phases, even if the moments are situated on a periodic lattice. Ordering phenomena of magnetic moments in quasicrystals could be rather interesting due to the geometric frustration that may be caused by the aperiodic structure. Experimentally, magnetic properties were investigated mainly for Al-based quasicrystals and for quasicrystals with the composition ZnMgRE, where RE denotes a rare earth metal. Besides the approximately 70 at % Al, many Al-based quasicrystals contain transition metals such as Mn, Fe, Ni, and Co. In the pure metals, these atoms show magnetic moments, which originate from partially occupied 3*d* states. As a consequence of the changed electronic structure in the quasicrystal, however, these magnetic moments vanish together with the partial occupation of the 3*d* states. Thus, high-quality Al-based quasicrystals often show diamagnetic behavior even though they contain a fair proportion of transition metal atoms. Concerning the ZnMgRE quasicrystals, the situation is different. In contrast to the 3*d* states of the transition metals, the 4*f* states of the rare earth metals cannot be filled by the electrons of the other constituents, and the magnetic moments survive. However, neutron diffraction experiments show that these moments are only short-range ordered; no long-range magnetic order in a quasicrystalline alloy has been found. At very low temperatures, around approximately 4 K, these phases behave like spin glasses, which means that the short-range order of the spins becomes frozen.

## F. Applications

An exceptional property of some quasicrystalline phases is their very low surface energy, which results in a wetting of the surface which lies in between PTFE (Teflon) and normal metals. Their high resistivity against scratches makes quasicrystalline materials well suited for coatings. The hardness as well as the low weight of quasicrystalline materials can be exploited in composite materials, in which advantageous properties of the components can be

combined. Icosahedral quasicrystals based on titanium can store up to two hydrogen atoms per metal atom, which makes them good candidates for use in hydrogen technology.

## IV. CONCLUDING REMARKS

Aperiodic crystals not only form a fascinating chapter of modern crystallography, but are also of importance for a variety of scientific disciplines. On the mathematical side, one is interested in the aperiodic ordering and its mathematical description as well as in a characterization of the plethora of possible structures that still may be found to exist. As far as the physics of quasicrystals is concerned, the understanding of the physical properties on the basis of their structure is at the center of the interest. However, this may first require a more detailed knowledge about the structure than is available to date, and, in particular, a thorough account of the type and amount of the inherent disorder in quasicrystals. Current technological applications of quasicrystals, partly still in a preliminary stage, look promising, and further research should be rewarding.

Aperiodic crystals, like other surprising discoveries, have again taught us that even long-held beliefs in science may eventually prove wrong. Who knows—even though it appears improbable today, maybe some day someone will come up with a sevenfold quasicrystal.

## SEE ALSO THE FOLLOWING ARTICLES

CRYSTAL GROWTH • CRYSTALLIZATION PROCESSES • CRYSTALLOGRAPHY

## BIBLIOGRAPHY

Axel, F., and Gratias, D. (eds.). (1995). "Beyond Quasicrystals," Editions de Physique, Les Ulis, France, and Springer, Berlin.

Baake, M., and Moody, R. V. (eds.). (2000). "Directions in Mathematical Quasicrystals," AMS, Providence, RI.

Blinc, R., and Levanyuk, A. P. (eds.). (1986). "Incommensurate Phases in Dielectrics," North-Holland, Amsterdam.

Grünbaum, B., and Shephard G. C. (1987). "Tilings and Patterns," Freeman, New York.

Hippert, F., and Gratias, D. (eds.). (1994). "Lectures on Quasicrystals," Editions de Physique, Les Ulis, France.

Janot, C. (1994). "Quasicrystals: A Primer," 2nd ed., Clarendon Press, Oxford.

Janssen, T., and Janner, A. (1987). "Incommensurabilty in Crystals." *Adv. Phys.* **36,** 519–624.

Moody, R. V. (ed.). (1997). "The Mathematics of Long-Range Aperiodic Order," Kluwer, Dordrecht.

Patera, J. (ed.). (1998). "Quasicrystals and Discrete Geometry," AMS, Providence, RI.

Senechal, M. (1995). "Quasicrystals and Geometry," Cambridge University Press, Cambridge.

Stadnik, Z. M. (ed.). (1999). "Physical Properties of Quasicrystals," Springer, Berlin.

Steinhardt, P. J., and Ostlund, S. (eds.). (1987). "The Physics of Quasicrystals," World Scientific, Singapore.

Yamamoto, A. (1996). "Crystallography of Quasiperiodic Crystals." *Acta Cryst. A* **52,** 509–560.



# Interacting Boson Model

**Bruce R. Barrett**
**Philip Halse**
*University of Arizona*

I. Description of the Model
II. Interacting Boson Model-1 (IBM-1)
III. Neutron–Proton Interacting Boson Model (IBM-2)
IV. IBM-3 and IBM-4
V. Interacting Boson–Fermion Model
VI. Boson–Fermion Symmetries
VII. Other Extensions of the IBM
VIII. Microscopic Interpretations of the IBM

## GLOSSARY

**Atomic weight or nuclear mass number ($A$)** Integer equal to the sum of the number of protons $Z$ and neutrons $N$.

**Boson** Particle possessing integer angular momentum (or spin) and satisfying Bose–Einstein statistics (that is, symmetric under particle interchange).

**Fermion** Particle possessing half-odd-integer angular momentum (or spin) and satisfying Fermi–Dirac statistics (that is, antisymmetric under particle interchange) and thereby the Pauli exclusion principle.

**Isospin** Vector operator relating to the charge of particles. For the nucleon, the total isospin is $T = \frac{1}{2}$, and the third component is $T_3 = +\frac{1}{2}$ for the proton and $T_3 = -\frac{1}{2}$ for the neutron.

**Parity** Symmetry of a wave function under inversion of the coordinate system: $r \rightarrow -\mathbf{r}$. The wave function either remains unchanged (even or + partiy) or changes sign (odd or − parity).

**Seniority ($v$)** Integer equal to the number of nucleons in a nucleus not coupled pairwise to zero.

**FOR OVER** 30 years, nuclear structure physics has been dominated by two models, the single-particle shell model, developed by Maria Goeppert-Mayer and J. H. D. Jensen, and the collective model, developed by Aage Bohr and Ben Mottelson. The shell model is successful in explaining the so-called magic numbers (or closed shell values) for protons and neutrons that lead to highly stable nuclei. It is also able to describe the properties of light nuclei and of nuclei near closed shells. However, because of the large number of possible states, shell-model calculations

for medium-mass and heavy mass nuclei away from closed shells are prohibitively difficult. On the other hand, the collective model is phenomenologically successful in treating the nucleus as a liquid drop, whose excitations are taken to arise from rotations and small oscillations about an equilibrium shape, with the modes corresponding to quadrupole (angular momentum two) deformations dominating. On quantization, this model can be expressed in terms of angular-momentum-two phonons (that is, bosons).

Although considerable effort has been made to unite these two models since their development, these investigations have met with only partial success. In 1974, Akito Arima and Francesco Iachello introduced a new model, the interacting boson model (IBM), which is an algebraic model and offers the real possibility of providing the missing link between the single-particle shell model and the collective model, in that it contains features of both. Although the IBM was first developed for medium-mass to heavy mass nuclei with an even number of protons and an even number of neutrons (so-called even–even nuclei), it has now been extended to describe odd-mass nuclei (even–odd and odd–even nuclei) and odd–odd nuclei, the latter being the most difficult to understand. For historical as well as practical reasons, the IBM for even–even nuclei will be described first.

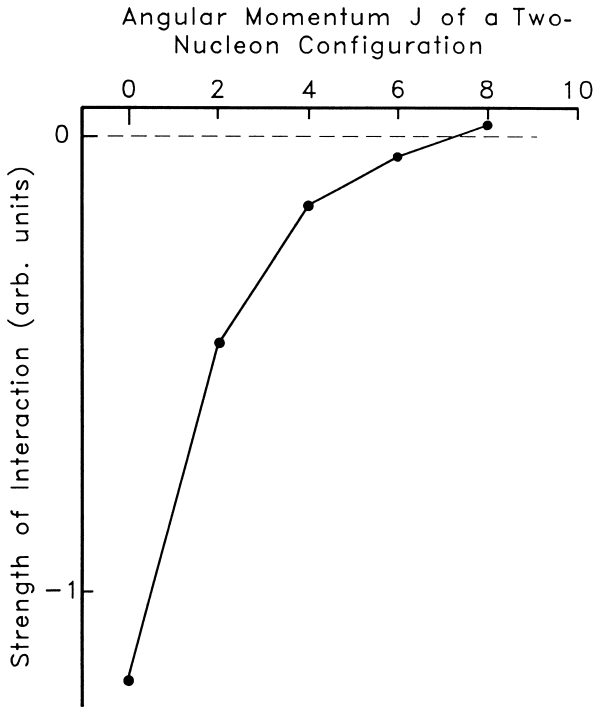



**FIGURE 1** Strength of the two-alike nucleon interaction versus the total angular momentum $J$ (taken from an analysis of pairs in a $j = \frac{9}{2}$ level).

## I. DESCRIPTION OF THE MODEL

The shell model treats the nucleus as a system of neutrons and protons interacting through the strong interaction. Neutrons and protons are collectively referred to as nucleons and are fermions, because they have an intrinsic spin angular momentum of one-half. As fermions, they satisfy Fermi–Dirac statistics and obey the Pauli exclusion principle, which states that no two fermions can occupy the same state in the same system, that is, they cannot have the same set of classifying quantum numbers. The Pauli exclusion principle leads to the filling of shells (or levels) produced by the mean field of the nucleons. As in atoms, the filling of a shell leads to a highly stable structure, with all the angular momenta of the nucleons in the shell summing up to zero. In the shell model, such structures are assumed to be inert, and nuclear properties are described in terms of the remaining nucleons (that is, the valence nucleons) moving outside the closed shells.

When two alike nucleons occur outside a closed shell, it is observed that their angular momenta couple to zero in the nuclear state of lowest energy, that is, the ground state. In fact, it is found empirically that the ground-state angular momenta ($J$) of all even–even nuclei are zero. The physical explanation of this result is the short range of the attractive strong interaction, that is, oppositely aligned angular momenta of the alike nucleons produce maximum overlap of the nucleons' wave functions and so the largest interaction. By similar reasoning, the next lowest energy states of the two alike nucleons are $J = 2$, then $J = 4$, etc. Figure 1 shows the relative strength of the alike-nucleon interaction versus the total angular momentum $J$ of the nucleons. For two alike nucleons in the same level, only even total angular momenta can occur because of the Pauli exclusion principle, which requires their total wave function to be antisymmetric under particle interchange. The above empirical observation suggests that building blocks of nucleon pairs of angular momentum zero and two may play an important role in determining low-lying nuclear properties. A system of fermion pairs is symmetric under the interchange of any two pairs. Consequently, such pairs are boson-like objects. These observations, together with the known phenomenological usefulness of angular-momentum-two bosons in the geometrical model, provide the motivation for the IBM.

The basic assumption of the IBM is that an even–even nucleus of $N_p$ valence protons and $N_n$ valence neutrons in the shell model can be treated as a system of $n_p = N_p/2$ valence proton bosons and $n_n = N_n/2$ valence

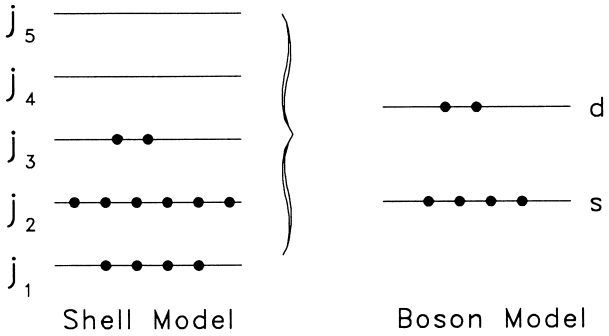



**FIGURE 2** Example of the truncation in the number of levels and the number of particles involved in replacing the shell-model problem by the IBM problem. The 12 nucleons in 5 shell-model levels (left-hand side) become 6 bosons in 2 levels, *s* or *d* (right-hand side) In reality, the boson configuration shown would correspond to a superposition of many shell-model configurations.

neutron bosons, each having angular momentum zero or two. For reasons related to the naming of angular momenta in atomic physics, the angular-momentum-zero bosons are called *s* bosons, and those with angular momentum two are called *d* bosons. Since the number of bosons is directly related to the number of valence nucleons, the number of IBM bosons is strictly conserved. The neutron bosons can be in *s* states (their number given by $n_{sn}$) or *d* states (their number given by $n_{dn}$), such that $n_n = n_{sn} + n_{dn}$, with a similar relation for the protons. This relationship is indicated in Fig. 2. It is assumed that bosons of higher angular momenta, for example, *g* bosons of angular momentum four, are less probable, because the corresponding fermion pairs are less tightly bound, for the reason given earlier (see Fig. 1).

The IBM is a model, instead of a theory, because it is known that the nucleus is made up of fermions and not bosons. However, the IBM can be a successful phenomenological model without defining or understanding the fermionic structure of the bosons. At the present time, the exact nature of this underlying structure is not known. Attempts to associate a microscopic structure with the IBM bosons will be discussed in Section VIII.

By building the valence structure of the nucleus from *s* and *d* bosons, one gains a twofold truncation of the shell-model problem. First, the bosons exist in only two states, *s* and *d*, while the fermions may occupy several single-particle levels with various large angular momentum values, and second, the number of interacting particles is cut in half, as shown in Fig. 2. This double truncation can reduce a shell-model problem involving $10^{12}$ or $10^{14}$ states to a boson problem in $10^2$ or $10^3$, which can be easily handled on a computer. Thus, it is noted that the interacting boson model is actually a shell model for bosons, but it is much simpler to apply to heavier mass nuclei.

## II. INTERACTING BOSON MODEL-1 (IBM-1)

The original version of the IBM does not distinguish between proton and neutron bosons; there are simply $(n_p + n_n)$ *s* and *d* valence bosons. This form of the model is referred to as the IBM-1. If one assumes that only one-body and two-body terms are important in describing the interactions among the bosons, one can easily write down a boson Hamiltonian involving all possible interactions to this order. This empirical Hamiltonian contains nine independent terms, only six of which are needed to define a spectrum for each value of $N$. The strength parameters of these terms can be easily determined by fitting experimental data for a given nucleus, a procedure also often used in shell-model calculations.

In their early papers, Arima and Iachello noted that the IBM-1 Hamiltonian possesses three symmetry limits, which could be related to geometrical descriptions in the collective model. Physicists feel that symmetries in nature are very fundamental, since they are often related to conserved quantities and basic principles. In the case of the IBM, the largest symmetry is the unitary group in six dimensions, U(6). The six dimensions come from the one *s* boson and the five possible states of the *d* boson [that is, the five possible orientations of its angular momentum ($J = 2$) along a given axis].

This overall U(6) symmetry for the *s* and *d* bosons can be broken in three distinct ways that contain the conserved rotation group SO(3), giving rise to three dynamical symmetry chains, as indicated in Eq. (1):

$$\begin{aligned} &\mathrm{U}(6) \supset \mathrm{U}(5) \quad \supset \mathrm{SO}(5) \supset \mathrm{SO}(3) \\ &\mathrm{U}(6) \supset \mathrm{SU}(3) \supset \mathrm{SO}(3) \\ &\mathrm{U}(6) \supset \mathrm{SO}(6) \supset \mathrm{SO}(5) \supset \mathrm{SO}(3) \end{aligned} \tag{1}$$

A dynamical symmetry comes from breaking the larger symmetry by the Casimir operators of groups making up one of the subgroup chains in Eq. (1). The nuclear-physics phenomenology corresponding to the U(5) and SU(3) chains was already known. The U(5) chain is related to a spherical vibrator, while the SU(3) chain displays aspects of rotational motion. Figure 3 shows the spectrum of a nucleus exhibiting SU(3)-like structure. The SO(6) chain was a new prediction, which was later verified by experiment and shown to represent what are known as $\gamma$-soft or $\gamma$-unstable nuclei. In these symmetry limits, exact analytical solutions can be obtained to the IBM-1 Hamiltonian. Moreover, the general IBM-1 formalism provides numerical solutions for cases between the symmetry limits, known as transitional nuclei.

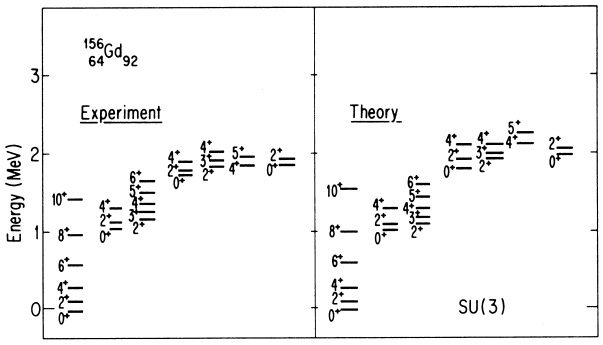



**FIGURE 3** Comparison of the experimental spectrum of $^{156}_{64}$Gd (experiment) with that corresponding to the SU(3) limit of the IBM-1 (theory). Each energy level is labeled by the value of the total angular momentum $J$.

## III. NEUTRON–PROTON INTERACTING BOSON MODEL (IBM-2)

The key feature missing from the IBM-1 approach is the relationship of the bosons to the underlying fermionic structure of the shell model. The work of Igal Talmi has shown the importance of the interaction between valence protons and valence neutrons in producing nuclear deformations. For this reason, the IBM was later expanded to treat separately the proton boson and neutron boson degrees of freedom. This proton–neutron interacting boson model is known as the IBM-2. A completely general one- and two-body IBM-2 Hamiltonian would contain 30 independent terms, so it is usually simplified to basically two components, a term that splits the energies of the $s$ and $d$ bosons, related to the pairing interaction (see Section I), and a quadrupole–quadrupole interaction between the proton bosons and the neutron bosons. The latter term is the lowest order interaction that can mix states containing different numbers of $s$ bosons and $d$ bosons, thought to be appropriate for nuclear deformation. The parameters of the IBM-2 Hamiltonian have been determined for a wide range of medium-mass to heavy mass nuclei, mainly in the rare-earth region, and the model has enjoyed considerable success in describing the low-lying properties of these nuclei.

Because the IBM-2 contains separately proton and neutron boson degrees of freedom, it yields not only states that are totally symmetric in both the charge and $sd$ spaces (corresponding to the IBM-1 solutions), but also states of mixed symmetry in both spaces. These mixed symmetry states in the IBM-2 lead to predictions regarding new forms of collective motion. The degree of symmetry of the IBM-2 states can be classified according to a quantity known as $F$-spin, which treats the proton and neutron bosons as two charge states of a single particle in the same way that the isospin $T$ treats protons and neutrons as two charge states of one particle, the nucleon. The states of maximum $F$-spin are totally symmetric and correspond to the IBM-1 states. States with $F$-spin less than the maximal value are of mixed symmetry and are believed to lie higher in energy as their $F$-spin value decreases. For highly deformed nuclei (i.e., in the SU(3) limit), the lowest energy mixed symmetry (or F maximal minus one) state should have a signature of angular momentum 1 and parity plus with a strong magnetic-dipole gamma transition, for which the orbital component is more important than the intrinsic spin, to the ground state. Numerous states with these properties have now been observed in rare-earth nuclei, supporting this prediction of the IBM-2. The theory also predicts mixed symmetry states in other mass regions, including those where the U(5) and O(6) limits are appropriate. States with characteristic signatures for these limits have now been seen. The detailed study of these states provides us with new information about the structure of nuclear collective states.

## IV. IBM-3 AND IBM-4

In most medium-mass to heavy mass nuclei, the protons and neutrons fill different major shells. In this mass region, the nuclear interaction strongly favors the proton–proton and neutron–neutron like pairs instead of pairs of neutron–proton structure. In light nuclei (mass number $A$ less than 100), the protons and neutrons often fill the same shells. In such cases, it is equally likely to form pairs constructed from a proton and a neutron. A neutron–proton pair can be either symmetric ($T = 1$) or antisymmetric ($T = 0$) in its charge state $T$. The symmetric neutron–proton state has the same space–spin structure as the proton–proton and neutron–neutron pairs, so that together they form a triplet of equivalent states (i.e., the three $T = 1$ states). The interacting boson model constructed from these three bosons is called the IBM-3. If the antisymmetric neutron–proton charge state is included as a fourth boson, one obtains the IBM-4. The IBM-3 and IBM-4 were developed by J. P. Elliott *et al*. and have been successfully applied to light nuclei, mainly for $18 \leq A \leq 46$.

Recent experiments with radioactive ion beams have produced many previously unobserved proton-rich nuclei and have now extended the region of observed $N = Z$ nuclei as far as $^{94}$Ag ($N = Z = 47$). The IBM-4 may be important in understanding the structure of these newly observed $N = Z$ nuclei, because the antisymmetric $T = 0$ proton–neutron pair appears to play a significant role. For nuclei where they are appropriate, the inclusion of proton–neutron bosons allows the IBM-3 and IBM-4 to describe

odd–odd nuclei and $\beta$-decay, which cannot be done with the IBM-2.

## V. INTERACTING BOSON–FERMION MODEL

The empirical success of the IBM for even–even nuclei encouraged its developers to expand it to odd–even and even–odd nuclei by the addition of one fermion. This odd-$A$ version is known as the interacting boson–fermion model or IBFM. The IBFM Hamiltonian contains a boson term, a fermion term, and a third term representing the interaction (or coupling) between the bosons and fermion. As in the IBM, the IBFM can be discussed in symmetry limits, as in Eq. (1), in which the odd fermion is coupled to the valence bosons in either the U(5), SU(3), or SO(6) limits. The different limits can be related to particular cases in the collective model, such as the strong-coupling or weak-coupling limits.

The addition of an odd fermion greatly increases the number of possible parameters in this model; the number of possible states also greatly increases. For these reasons, the IBFM has been applied mainly to special cases, such as a nucleon in a single $j$ level or in several $j$ levels with the imposition of some boson–fermion symmetry (see Section VI). This model has also been used for studies of $\beta$-decay between odd–even and even–odd nuclei.

## VI. BOSON–FERMION SYMMETRIES

In 1980, Iachello observed that in certain cases new symmetries, corresponding to simultaneous transformations of the boson and fermion systems, can be introduced. This is possible if some groups in the fermion classification (that is, group chain) coincide with some groups in the boson classification [see Eq. (1)]. Combined bose–fermi groups can then be introduced corresponding to particular couplings of the bosons and fermions.

It was, of course, known that the conservation of the total angular momentum requires the combined system to be invariant under the total angular momentum operator (the sum of parts acting on the bosons and on the fermions), generating $SO^{BF}(3)$. But it was found that spectra are often closer to those associated with combining the larger groups in Eq. (1), such as SO(6), with their fermion counterparts, implying the conservation of less well-understood quantities.

Figure 4 illustrates the spectrum of an odd-$A$ nucleus related by such a bose–fermi symmetry to the spectrum of its even–even neighbor. It was claimed that cases such as those shown in Fig. 4 are examples of supersymmetric structure in nuclei. However, supersymmetry conventionally refers to a description involving a superalgebra, which is an algebra containing operators that transform bosons into fermions and vice versa. In fact, the cases and examples given then and in later work are for bose–fermi symmetries rather than true supersymmetries. Nevertheless, the fact that the properties of certain neighboring even–even and odd-$A$ nuclei can be related by the same group-theoretical chains and the same Hamiltonian is of significant interest and provides new insight into the structure of complex nuclei. Present investigations regarding high-spin superdeformed bands indicate that superdeformed bands in certain neighboring even–even and odd-$A$ nuclei may prove to be the best examples of bose–fermi symmetries in nuclei. Recent work with superalgebras indicates that examples of real supersymmetries may exist in nuclei.

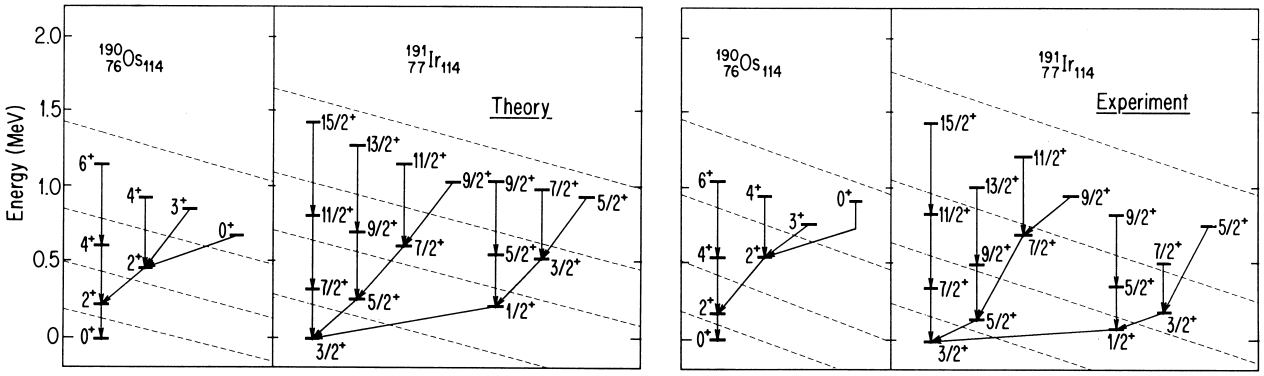



**FIGURE 4** Comparison of the theoretically predicted (theory) and observed (experiment) energy spectra for $^{190}_{76}$Os and $^{191}_{77}$Ir, as an example of a bose–fermi symmetry in nuclei. Each energy level is labeled by the value of the total angular momentum $J$. The dashed line enclose levels of the same symmetry. The solid lines indicate levels between which strong electromagnetic radiation occurs.

## VII. OTHER EXTENSIONS OF THE IBM

By its basic assumptions, the IBM is a model for low-excitation nuclear structure. For this reason, a number of expanded versions have been developed, so as to describe other nuclear properties. Angular momentum one ($p$) and three ($f$) bosons have been introduced to explain negative parity states in nuclei, and aligned pairs coupled to large values of the angular momentum have been used to describe high-spin states. Procedures have also been developed for treating configuration mixing in nuclei, such as the mixing between vibrational-like and rotational-like states.

Particle-like bosons can be combined with hole-like bosons in a similar manner as proton bosons were combined with neutron bosons in the IBM-2, thereby leading to states of different F-spin. In the case of particle and hole bosons, the states of different particle-hole symmetry are classified by I-spin, which can connect states in different nuclei.

## VIII. MICROSCOPIC INTERPRETATIONS OF THE IBM

The success of this formalism involving rather abstract bosons suggests that they might represent real objects within the nucleus, in particular that they may be interpreted in terms of the valence protons and neutrons of the shell model (Section I). Investigations of this possibility make up the largest single area of research arising from the IBM.

Since a pair of fermions is bosonlike (a similarity which improves as the number of fermion states increases), a natural proposal is that the $s$ and $d$ bosons are modeling pairs of nucleons coupled to angular momentum 0 and 2 denoted as $S$ and $D$, respectively; indeed, this idea was used to motivate our discussion of the IBM and is commonly seen as part of the IBM per se. However, concluding that the validity of this interpretation follows from the equality of statistics and angular momentum alone would be a non sequitur. Moreover, many other situations are possible, such as the bosons representing quartets of nucleons (IBM results are generally not very sensitive to the number of bosons), or even having no interpretation of the bosons singly, necessitating a more complicated many-boson–many-nucleon correspondence.

In fact, there can be no automatic answer to the question of what the bosons represent, since their interpretation must depend on the phenomena they are used to describe. For instance, $s$ and $d$ bosons could in principle be used to describe the giant quadrupole resonance; any shell model interpretation of such bosons would have to be very different from one designed to reflect the description of low-energy phenomena with which the model is associated in practice.

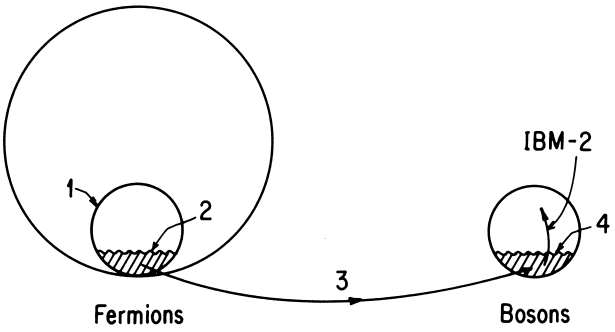



**FIGURE 5** Pictorial description of one microscopic IBM procedure. The full fermion space (large circle) is truncated to an $S$ and $D$ collective-pair space, 1. An appropriate subset, 2, is associated through the mapping (represented by label 3) to the corresponding states, 4, in the boson space. The IBM-2 interaction can now be computed microscopically, completing the boson picture.

Nevertheless, as described above, an interpretation of the bosons as fermion pairs is almost always the basic postulate of such investigations. It is apparent that the structure of the pairs associated with the bosons must be collective in nature, because of the transition from the nucleon shell model space to the collective-pair space is already a significant truncation.

Most attempts to develop a microscopic interpretation have centered around a search for the appropriate collective pairs in the fermion (that is, nucleon) space to be mapped onto (that is, related with) the bosons. This is indicated schematically in Fig. 5. Here, the large circle represents the full fermion shell model space, which is then truncated to the subspace constructed using the $S$ ($J = 0$) and $D$ ($J = 2$) pairs. Then, some subset, 2, restricted for computational reasons to consist of only those states with a small number of $D$ pairs, is associated through mapping (represented by label 3) to the corresponding set of states, 4, in the boson space and used to determine the boson operators, corresponding to those of interest in the shell model.

One prescription for the collective fermion pairs is to solve the shell-model problem for two alike nucleons for $J = 0$ and $J = 2$ and to equate the lowest $J = 0$ eigenstate with the collective $S$ pair and the lowest $J = 2$ eigens state with the collective $D$ pair. This procedure follows ideas suggested by I. Talmi regarding his work on generalized seniority. Other procedures have been proposed for constructing the collective fermion states to be associated with the IBM states, but there is no general agreement regarding an ideal choice.

However, the impossibility of performing shell model calculations for heavy nuclei (Section I), itself a rationale

for use of the IBM, means that the validity of the pair interpretation cannot be definitively tested. It simply is not known whether the many-pair condensate analogues of the many-boson IBM states would indeed allow a reasonable approximation to the shell model eigenstates, nor whether the shell model operators mentioned above would in fact reproduce the data. As discussed above, the shell model interpretation of the bosons must be appropriate for the levels the IBM is used to model: for instance, the interpretation for a description of all the low-energy rotational bands of $^{156}$Gd (Fig. 3.) would be different from that appropriate for a description of the first, fourth, and fifth bands only. (For example, if the strength of the interaction in the theoretical boson calculations were increased by around 50%, then the second and third model bands obtained would correspond to the fourth and fifth bands in the experiment, while the observed second and third bands would have no IBM counterparts). A possible failure of the $SD$ pair interpretation is then apparent in exact calculations for lighter nuclei, where it is found that the many-pair states describe only some of the levels that the IBM would be used to model. A similar conclusion has been obtained in an approximate calculation for $^{156}$Gd itself. If this situation does indeed persist in heavy nuclei, it would have to be concluded that the simple interpretation of the bosons as pairs is inconsistent with the use of the IBM to model all the collective low-energy levels, as is invariably the case (Fig. 3).

There is much controversy in this area, which only further research can resolve. A truly valid shell model interpretation of the elegantly simple IBM would reveal a correspondingly simple latent structure amid the complexity of realistic shell model calculations.

This said, it is worth noting that after 25 years the Interacting Boson Model approach to describing the properties of medium-to-heavy-mass nuclei has held up extremely well and has proven itself to be quite versatile and robust.

## SEE ALSO THE FOLLOWING ARTICLES

GROUP THEORY • NUCLEAR PHYSICS • PARTICLE PHYSICS, ELEMENTARY

## BIBLIOGRAPHY


Arima, A., and Iachello, F. (1984). "Advances in Nuclear Physics" (J. W. Negele and E. Vogt, eds.), Vol. 13. Plenum, New York.

Barrett, B. R. (1984). "Nucleon–Nucleon Interaction and Nuclear Many-Body Problems" (S. S. Wu and T. T. S. Kuo, eds.). World Scientific, Singapore.

Bonatsos, D. (1988). "Interacting Boson Models of Nuclear Structure," Clarendon Press, Oxford.

Casten, R. F. (ed.) (1993). "Algebraic Approaches to Nuclear Structure: Interacting Boson and Fermion Models." *Contemporary Conc Physics*, **6**. Harwood Academic Publishers.

Casten, R. F. and Feng, D. H. (1984). Nuclear dynamical supersymmetry. *In* "Physics Today," Vol. 37. American Institute of Physics. New York.

Casten, R. F., and Warner, D. D. (1988). The interacting boson approximation. *In* "Reviews of Modern Physics," Vol. 60. The American Physical Society, New York.

Dieperink, A. E. L., and Wenes, G. (1985). "Annual Review of Nuclear and Particle Science," Vol. 35. Annual Review Inc., Palo Alto, California.

Iachello, F., and Arima, A. (1987). "The Interacting Boson Model." Cambridge Univ. Press, London and New York.

Iachello, F., and Talmi, I. (1987). Shell-model foundation of the interacting boson model. *In* "Reviews of Modern Physics," Vol. 59. The American Physical Society, New York.

Iachello, F., and Van Isacker, P. (1990). "The Interacting Boson-Fermion Model." Cambridge University Press, London and New York.

Mizusaki, T., and Otsuka, T., (ed.) (1996). Microscopic Study of the Interacting Boson Model. *In* "Progress of Theoretical Physics: Supplement," Number 125. Yukawa Institute and the Physical Society of Japan, Kyoto.

Scholten, O. (1985). "Progress in Particle and Nuclear Physics," Vol. 14. Pergamon, Oxford.

Talmi, I (1993). "Simple Models of Complex Nuclei: The Shell Model and Interacting Boson Model." Contemporary Concepts in Physics, Vol. 7. Harwood Academic Publishers.



# Liquid Crystals (Physics)

**Paul Ukleja**
*Southeastern Massachusetts University*

I. Brief History
II. Description of Liquid Crystal Phases
III. Properties of Liquid Crystals
IV. Applications

## GLOSSARY

**Amphiphile** Material in which each molecule has one part that is attracted to water and another part that rejects water.

**Cholesteric** Liquid-crystalline state of matter in which the molecules align in a helical structure that has a regular pitch. Over distances much smaller than the pitch, a cholesteric has a nematic structure.

**Director** Direction (often denoted by a unit vector, **n**) about which the long axes of molecules or aggregates of molecules fluctuate in liquid-crystalline phases.

**Hexagonal phase** Lyotropic phase in which the amphiphilic molecules aggregate into parallel cylinders that pack into a hexagonal array.

**Homeotropic** Alignment of liquid crystals in which the director is uniformly aligned perpendicular to the opposite, parallel surfaces of a thin, flat container.

**Lamellar phase** Smectic lyotropic phase commonly consisting of alternate flat layers of water and amphiphile.

**Lyotropics** Liquid crystals that form in solutions and change phases primarily with concentration.

**Nematic** State of matter in which molecules or aggregates of molecules align along a common direction, the director, but are otherwise fluid.

**Order parameter** Parameter that indicates the degree to which the molecules of a liquid crystal align with the director.

**Polymer dispersed liquid crystals (PDLC)** Composite material made of tiny droplets of liquid crystal in a polymer matrix. The droplets normally scatter light, but an applied electric field aligns them, switching the material from opaque to transparent.

**Smectics** Phases in which the molecules have orientational order and partial positional order, generally in layers.

**Supertwist LCD** Similar to the twisted nematic display but with a twist angle of 270°, giving it a faster response time. Cholesteric material is used to create the large twist angle in the relaxed state.

**Surfactant** Amphiphilic material; the molecules tend to arrange themselves on surfaces with water on one side.

**Thermotropics** Liquid crystals that change phase with changes in temperature.

**LIQUID CRYSTALS** are materials that have properties and characteristics of both liquids and crystalline solids. For many liquid-crystalline materials the liquid-crystalline phases, also called mesophases, occur in a

range of temperatures between those at which the materials are normal liquids and solid crystals. For other liquid crystals, it is mainly the concentrations of different components of a solution that determine the phases. Many solutions or pure compounds form several distinct phases at different temperatures or concentrations. Several thousand compounds have liquid-crystalline phases in their pure states. Many biological materials, such as cell membranes, also display liquid-crystalline behavior. Since liquid crystals have both the ability to flow and the anisotropy of crystals, they display many properties of great interest and, often, of practical importance.

## I. BRIEF HISTORY

The study of liquid crystals began with observations by an Austrian botanist, Friedrich Reinitzer, in 1888. Reinitzer found that cholesteryl benzoate melted from a solid at 145°C into a liquid having a murky or turbid appearance. At 179°C the liquid cleared. The color of the turbid liquid changed from red to blue as the temperature increased. On cooling, the reverse occurred. Reinitzer sent some of this material to O. Lehmann, who was able to make further studies with polarized light on a microscope equipped with a heating stage, which allowed him to vary the temperature of samples being observed. Lehmann discovered that the turbid liquid actually displayed optical anisotropy or birefringence, as do solid crystals. The combination of the ability of the material to flow like a liquid and yet retain the anisotropic optical properties of a crystal led Lehmann to coin the name "liquid crystal" to describe this state of matter.

The turbid appearance, resembling that of a colloidal solution, gave rise to early ideas that the liquid crystal was no more than such a solution, but it was later found that the liquid-crystalline state is a distinct phase of matter with fixed transition temperatures into the solid and normal or isotropic liquid states and that the molecules in a liquid crystal have orientational order.

In 1922, G. Friedel proposed a classification system that is still used extensively, dividing liquid crystals into three classes: smectic, nematic, and cholesteric. In a smectic phase, the molecules are arranged in sets of parallel planes. The smectic phases generally have high viscosity and a soapy appearance; hence the name was derived from the Greek word for "soap." The word "nematic" is derived from the Greek word for "thread." A nematic liquid crystal often shows a thread-like pattern when placed between crossed polarizers and viewed through a microscope. The third class was named "cholesteric" since the molecules forming these phases commonly contained cholesterol. A cholesteric has a characteristic iridescent color, which can change dramatically with changes in temperature or other aspects of the environment.

Early attempts to explain the turbid appearance of the nematic liquid crystals included the idea that the molecules in these materials grouped into swarms. The boundaries between swarms would represent variations in the optical properties of the medium that could scatter light. It was later found that nematic liquid crystals, which are most often formed from rod-like molecules, are generally homogeneous throughout, with the long axes of the molecules lining up parallel to one another. Long-wavelength thermal fluctuations in the direction of the alignment scatter visible light.

Classification and identification of liquid crystal phases were first done by using a polarizing microscope with a heating stage. Observations of the textures and how they changed from phase to phase were useful in determining some of the properties of the crystals. By mixing two compounds in various proportions and determining the resulting phase transitions, it was possible to compare the phases they formed in their pure states. The present names for the different smectic states (A, B, C, etc.) were merely assigned as they were observed and do not necessarily bear a logical relation to their structures. For many of the states formed in lyotropic materials (those whose concentrations determine the phases), there are several names arising from different lines of research.

The study of liquid crystals continued fairly strongly into the 1930s, including measurements of viscosity coefficients and development of theoretical models for the elastic and flow properties of nematics. The renewed interest in liquid crystals in this country owes much to Glenn H. Brown of Kent State University, who organized a series of international conferences starting in the 1960s and founded the Liquid Crystal Institute at Kent State University. Low-power electronic liquid crystal displays (LCDs) are now found in a very wide range of devices. Further developments in displays and other uses of liquid crystals have stimulated a wide range of investigations in this country and abroad into the nature of these phases.

## II. DESCRIPTION OF LIQUID CRYSTAL PHASES

The most obvious difference between liquids and solids is the ability of a liquid to flow or to adapt its shape to its container under the influence of small external forces. On the molecular level, a crystalline solid has long-range order, that is, a strong correlation between the positions and orientations of molecules that are far apart, whereas such

coordination extends to only a few neighboring molecules in a liquid. One result of this is that the physical properties of normal liquids show no distinctions among different directions; that is, liquids are isotropic. Crystals, on the other hand, often have properties that vary with direction; they are anisotropic. Light or sound, for instance, may travel faster in some directions than in others.

Liquid crystals are characterized by partial ordering; that is, one or more degrees of freedom, but not all, will have long-range order. In particular, the molecules will retain some ability to move throughout the medium, although the motion may be restricted in extent or direction. The molecules may be as little restricted as in the nematic phase where there is a preferred direction for the long axis of a molecule but the other axes and overall location are free to vary. The translational motion may also be restricted as in the smectic phases in which the molecules, besides aligning their long axes, arrange themselves in parallel planes. In plastic (as opposed to liquid) crystals the molecules are well ordered translationally but not orientationally.

Pure compounds which display liquid crystal phases as the temperature is changed are called thermotropic liquid crystals. Homogeneous mixtures of these compounds are generally also thermotropic liquid crystals. When certain substances, such as soaps, are dissolved in a suitable solvent, such as water, liquid crystalline phases are observed. These phases are called lyotropic phases from the Greek root "lyein," to dissolve. For lyotropic phases, concentration is the main physical variable, although temperature changes can also effect phase changes. Solutions of polymers also display liquid crystalline order. Friedel's terminology of nematic, smectic, and cholesteric is still used to describe the main classes of liquid crystals.

## A. Thermotropic Liquid Crystals

The thermotropics (see Table I) are the most studied and perhaps best understood of the liquid crystals. It is thermotropic liquid crystals that are used in liquid crystal displays in wristwatches, computers, and televisions. Cholesterics, sensitive to temperature, are used to make very thin thermometers and films that change color with temperature. Most thermotropics are formed from organic molecules with a rod-like or lath-like shape. The majority of the molecules having liquid crystalline mesophases have planar and rigid nuclei, typically including two or more benzene rings. Table I shows examples of typical thermotropic liquid crystalline compounds and the temperature ranges over which they are liquid crystalline. Table II lists the most common thermotropic phases along with some of their properties.

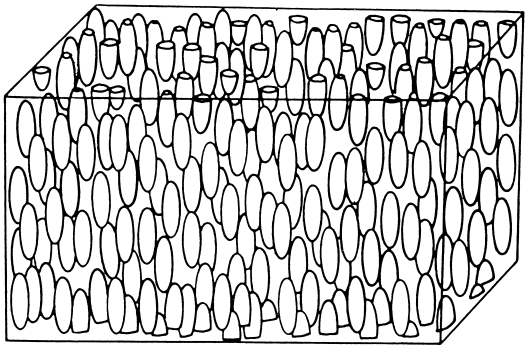


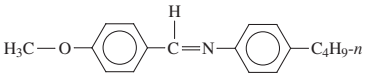
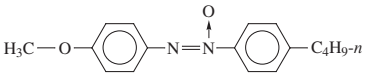
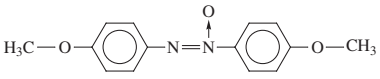
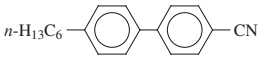
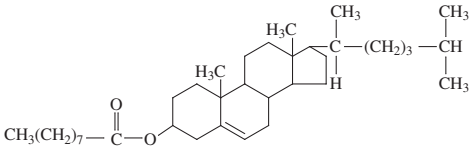
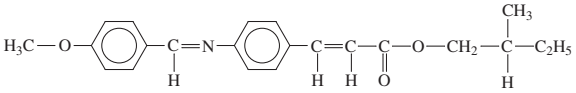
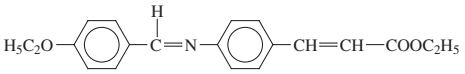
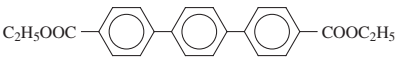
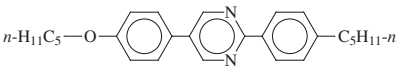
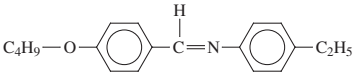
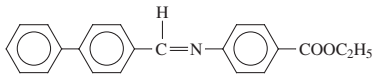
**FIGURE 1** A nematic liquid crystal. The oval shapes in this and the next two figures represent the positions and average orientations of molecules. The instantaneous directions of the long molecular axes fluctuate about the average direction (the director) by angles as large as 40°. [Courtesy of Nuno Vaz.]

### 1. Nematic and Cholesteric Phases

*a. Ordinary nematic.* This is the simplest of the liquid crystal phases. In the nematic phase the long axes of the molecules have a preferred orientation, a director, about which they fluctuate rapidly. In Fig. 1, the average orientations of the molecules are represented by cigar-shaped forms. An instantaneous picture of the molecules would show the long axes tilted at angles up to 40° away from the director. The molecules behave like a fluid in that they can move freely from point to point in the medium. The director responds to very weak external forces and usually varies from point to point in the medium. As the nematic is optically birefringent, this variation in the director over distances on the order of magnitude of the wavelength of visible light is what gives the nematic its turbid or cloudy appearance. It is possible to make a uniformly aligned sample through the use of external magnetic or electric fields or interactions with treated surfaces. This makes it possible to use nematics for displays in which electric fields are used to realign the director and thus change the optical behavior of the sample. Some nematics may be composed of groups of hundreds of molecules, called cybotactic groups, in which the molecular centers are arranged in layers.

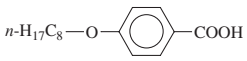
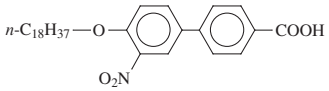
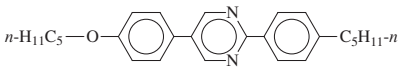
*b. Cholesteric.* A cholesteric phase is similar to the ordinary nematic but with a natural twist in the director (Fig. 2). Thus the long axes of molecules that are side by side prefer to align at a small angle. On a local scale (distances on the order of tens or hundreds of molecules), the ordering is essentially nematic. Thermodynamically, the cholesteric behaves like a nematic, as the energy of the twist is small compared to the energy associated with parallel alignment of the molecules. Mixtures of cholesterics having opposite twists form cholesterics of

**TABLE I Some Thermotropic Liquid Crystalline Compounds[a]**

| Formula | Name | Liquid crystalline range (°C) |
|---|---|---|
| 1. Nematic liquid crystals | | |
| A. Some ordinary classic nematic liquid crystals | | |
|  | *p*-Methoxybenzylidene-*p*′-*n*-butylaniline(MBBA) | 21–47 |
|  | *p*-Methoxy-*p*′-*n*-butylazoxybenzene | 19–76 (mixture of isomers) |
|  | *p*-Azoxyanisole (PAA) | 117–137 |
|  | *p*-*n*-Hexyl-*p*′-cyanobiphenyl | 14–28 |
| B. Cholesteric–nematic liquid crystals | | |
| 1. Cholesteric esters | | |
|  | Cholesteryl nonanoate | 145–179 |
| 2. Noncholesteryl, chiral-type compound | | |
|  | (−)-2-Methylbutyl-*p*-(*p*-methoxybenzylideneamino) cinnamate | 76–125 |
| II. Smectic liquid crystals | | |
| A. Structured smectic liquid crystals | | |
| Smectic B | | |
|  | Ethyl *p*-ethoxybenzal-*p*′-aminocinnamate | 77–116 |
| Smectic E | | |
|  | Diethyl *p*-terphenyl-*p*-*p*”-carboxylate | 173–189 |
| Smectic G | | |
|  | 2-(*p*-Pentylphenyl)-5-(*p*-pentyloxyphenyl)pyrimidine | $S_G$79–103 |
| Biaxial $S_B$ (or smectic H) | | |
|  | 4-Butyloxybenzal-4-ethylaniline | 40.5–51 |
| B. Unstructured smectic liquid crystals | | |
| Smectic A | | |
|  | Ethyl *p*(*p*′-phenylbenzalamino)benzoate | 121–131 |

*continues*

**TABLE I** (*Continued*)

| Formula | Name | Liquid crystalline range (°C) |
|---|---|---|
| Smectic C<br> | *p*-*n*-Octyloxybenzoic acid | 108–147 |
| Smectic D<br> | *p*′-*n*-Octadecyloxy-3′-nitrodiphenyl-*p*-carboxylic acid | 159–195 |
| Smectic F<br> | 2-(*p*-Pentylphenyl)-5-(*p*-pentyloxyphenyl)pyrimidine | $S_F$ 103–114 |

[a] From Brown, G. H. (1977). *J. Colloid Interface Sci.* **58,** 534.

infinite pitch, which correspond to nematics. A cholesteric is formed by adding a small amount of a cholesteric, or even a substance that is not in itself liquid crystalline but is optically active, to a nematic substance. The pitch of a cholesteric can vary sensitively with temperature, vapor pressure of certain substances, and other influences. When the pitch corresponds to the wavelength of visible light, the scattered light is highly colored. A cholesteric also rotates the direction of linearly polarized light, that is, is optically active. This activity is roughly 1000 times stronger than the activity of an ordinary optically active substance such as quartz.

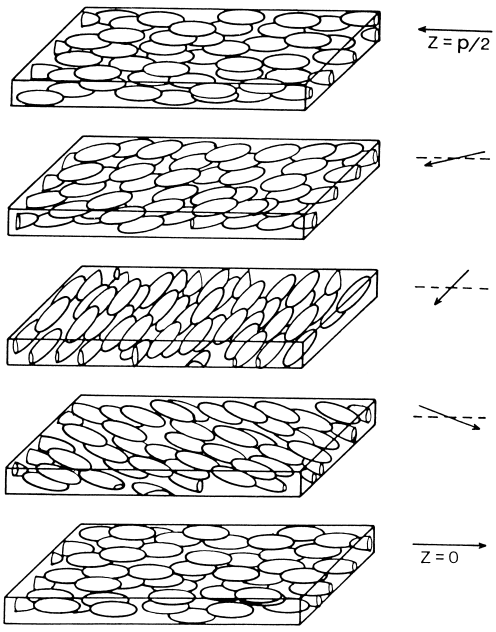



**FIGURE 2** A cholesteric or twisted nematic. On a small scale, the molecules behave as in the nematic phase. Over longer distances the director rotates along a helix whose pitch is sensitive to changes in temperature, pressure, etc. [Courtesy of Nuno Vaz.]

*c. Blue phases.* Some cholesteric compounds exhibit a phase between the cholesteric and isotropic phases, usually in a narrow temperature range. The local molecular orientation shows a three-dimensional periodicity and is perhaps a stable lattice of defects in the uniform cholesteric structure. Mechanically, the behavior is similar to that of a cubic crystal with a large resistance to shear (a shear modulus of several hundred to several thousand ergs per cubic centimeter). Several blue phases exist.

*d. Discotic.* Disklike molecules may have a discotic phase in which the molecules are stacked aperiodically, forming liquidlike columns having nematiclike ordering of the symmetry axes of the molecules. Some of these materials have smectic phases similar to the smectic A in which the columns form a hexagonal array. These phases thus exhibit translational order in two dimensions, but not in the third. The appearance under a microscope is similar to that of an ordinary nematic.

## 2. Smectic Phases

The various smectic phases have, in addition to the orientational order found in nematics, different degrees of positional and, in some cases, bond-orientational ordering (Fig. 3). On the basis of appearance under a polarizing microscope, miscibility with known phases, and X-ray scattering, at least nine thermotropic smectic phases have been identified, although not all are truly liquid crystals. Of these nine phases, eight have a characteristic packing of the molecules in layers. The ninth phase, known as smectic D (the letters used to denote the different phases were assigned in chronological order as the phases were first observed), has a cubic packing. In the smectic A, B, and E phases, the molecules align with their long axes

**TABLE II Characteristics of Thermotropic Liquid Crystals[a]**

| Class | Optical properties | Textures | Structure | Examples |
|---|---|---|---|---|
| I. Nematics | | | | |
| Ordinary nematic | Uniaxially positive | Schlieren; threaded marbled; pseudoisotropic; homogeneous | Parallelism of long molecular axes | *p*-Azoxyanisole; *p*-methoxybenzylidene *p*-*n*-butylaniline |
| Cholesteric nematic | Uniaxially negative; or isotropic optically active | Focal conic with Grandjean steps; homogeneous; isotropic | Nematic packing in planes; superimposed twist in direction perpendicular to long axes of molecules | Cholesteryl nonanoate |
| II. Structured smectics | | | | |
| Smectic B | Uniaxially or biaxially positive | Mosaic; stepped drops; pseudoisotropic; homogeneous; schlieren | Layer structure; molecular axes orthogonal or tilted to layers; hexagonal arrangement within layers | Ethylethoxybenzylideneaminocinnamate; terephthal-bis-butylaniline |
| Smectic E | Uniaxially positive | Mosaic; pseudoisotropic | Layer structure; molecular axes orthogonal to layers; ordered arrangement within layers | di-*n*-Propylterphenyldicarboxylate |
| Smectic G | Uniaxially positive | Mosaic | Layer structure with ordered arrangement within layers | 2-(4-*n*-Pentylphenyl)-5-(4-*n*-pentyloxy-phenyl)pyrimidine |
| III. Unstructured smectics | | | | |
| Smectic A | Uniaxially positive | Focal conic (fanshaped or polygon); stepped drops; homogeneous; pseudoisotropic | Layer structure; molecular axes orthogonal to layers; random arrangement within layers | Diethylazoxybenzoate |
| Smectic C | Biaxially positive | Broken focal conic; schlieren; homogeneous | Layer structure; molecular axes tilted to layers; random arrangement within layers | Dodecyloxyazoxybenzene |
| Smectic D | Isotropic | Isotropic; mosaic | Cubic structure | 4′-Octadecyloxy-3′-nitrodiphenyl-4-carboxylic acid |
| Smectic F | Uniaxially positive | Schlieren; broken focal conic with concentric axes | Layer structure | 2-(4-*n*-Pentylphenyl)-5-(4-*n*-pentyloxy-phenyl)pyrimidine |

[a] From Brown, G. H., and Wolken, J. J. (1979). "Liquid Crystals and Biological Structure," pp. 30–31, Academic Press, New York.

perpendicular to the layers. In the tilted smectics, C, F, G, H, and I, the long axes are at an angle with the layer normals. The smectic A, C, D, and F phases are unstructured smectics: They do not show an ordered arrangement of molecules within layers. The smectic A, B, and C phases are the best known of the smectic phases.

One variation on the ordinary smectic structure can occur in optically active compounds having a tilted smectic phase or in a tilted smectic phase to which a small amount of a chiral compound has been added. A macroscopically chiral structure can form in which the directors of adjacent layers of molecules form a small angle, giving an uniform twist and resulting in a strongly optically active substance. Some of these chiral smectics are ferroelectric and have the potential for applications in displays with fast response times (microseconds).

Extremely thin films, down to one molecular layer thick, have been made with smectics. These films are being used to probe surface effects as well as new thermodynamic phase behavior.

In the discussion that follows, the smectic phases are treated in alphabetic order, which does not always correspond to the sequence of phases observed on heating or cooling.

*a. Smectic A.* Smectic A liquid crystals (Fig. 3) are the least ordered of the untilted or orthogonal smectic phases. The molecules are arranged in layers with the director perpendicular to the layers. Except over short distances, the molecules show no correlations in position within the layers. Studies of X-ray scattering by smectic

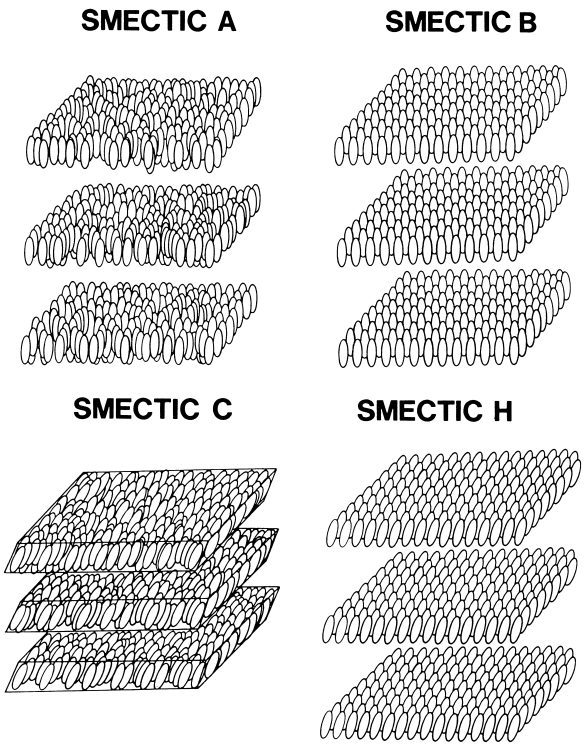



**FIGURE 3** Molecular arrangements in four smectic phases in which molecules form layers. In the A and C phases molecules are aligned but their positions are not ordered. In the C and H phases molecules are aligned in a direction tilted with respect to the layers. [Courtesy of Nuno Vaz.]

A materials show that the spatial density is best described not in terms of sharply defined layers of molecules but as a one-dimensional sine wave in a three-dimensional fluid with the density wave along the director. The higher spatial harmonics of the density wave are surprisingly weak. Furthermore, the correlation in the positioning of the layers dies away algebraically (as an inverse power of the distance) rather than being constant as in a true crystal. In liquid crystals having both nematic and smectic A phases, the nematic phase is the higher temperature phase. It is possible to make well-aligned single crystals of smectic A by aligning the director in the nematic phase, for instance, with a magnetic field and then cooling into the smectic phase. Once "frozen" into this alignment, the director can no longer reorient to align with the field as it is fixed perpendicular to the layers. Such a sample does not have the turbid appearance of the nematic, as the director does not have the variations in direction characteristic of the nematic.

*b. Smectic B.* The structure of this phase (Fig. 3) consists of layers of molecules having hexagonal packing within the layers. The director is perpendicular to the layers. The smectic B phases in some materials may not, in fact, be liquid crystals at all, but have long-range order in all three dimensions. There may be a hexatic smectic B phase having short-range positional order combined with long-range bond-orientational order in the plane of the layers.

*c. Smectic C.* The smectic C phase (Fig. 3) is similar to the smectic A phase except that the director makes an angle, called the tilt angle, with the normal to the layers. The layer thickness deduced from X-ray scattering data is less than the molecular length and the phase is optically biaxial, unlike the nematic and smectic A. Tilt angles up to 45° have been observed and can vary with temperature. Because the orientation of the director can change while the tilt angle is kept constant, the director can change from point to point as in the nematic, and as a consequence there is strong light scattering. Cooling to smectic C from nematic or smectic A phases does not create single crystals. As is the case for the ordinary nematic, the addition of optically active molecules can give a twist to the smectic C phase. Pure compounds with this structure have also been observed. In a compound having the A, B, and C smectic phases, the sequence on cooling is A, C, and then B (Table III).

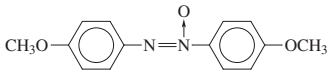
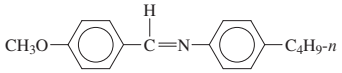
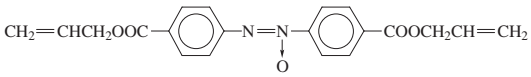
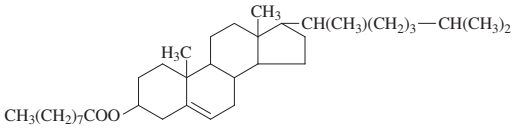
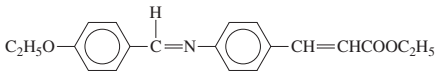
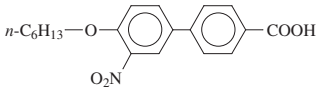
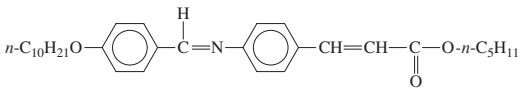
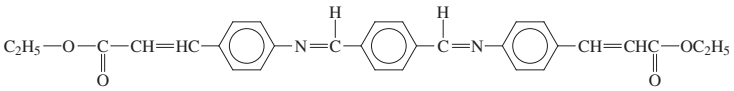
*d. Smectic D.* The smectic D phase does not have the characteristic layers of the other smectics and is optically isotropic. The overall structure has cubic symmetry. One model of the structure has molecules packed hexagonally into roughly spherical shapes, which are then packed into a cubic framework. This kind of structure is also seen in cubic lyotropic phases.

*e. Smectic E.* The director is perpendicular to the layers in the smectic E as in the A and B phases. Within the layers, the molecules show correlations in both position and the orientation of the benzene ring. The molecules pack into a herringbone type of pattern.

*f. Smectic F.* This phase is similar to the smectic C phase but is more ordered, having a short-range hexagonal order within the layers.

*g. Other smectics.* The smectic G and H phases (Fig. 3) correspond to smectics B and E, respectively, differing in that the directors are tilted at an angle with respect to the layer normals. The smectic I is yet another smectic phase with hexagonal correlations in the layers but with a tilt that is uniform with respect to neighboring molecules. New smectic phases have been found in materials in which the molecules have electric dipole moments. These include antiferroelectric phases.

**TABLE III Typical Examples of Polymorphic Forms of Thermotropic Liquid Crystals[a]**

| Polymorphic form[b] | Example |
|---|---|
| N |  *p*-Azoxyanisole |
| N |  4-Methoxybenzylidene-4′-*n*-butylaniline |
| A |  Diallylazoxybenzene-4, 4′-dicarboxylate |
| Ch,A |  Cholesteryl nonanoate |
| A,B |  Ethyl-4-ethoxybenzylidine-4′-aminocinnamate |
| N,A,C |  4′-*n*-Hexyloxy-3′-nitrobiphenyl-4-carboxylic acid |
| A,C,B |  *n*-Amyl-4-*n*-decyloxybenzylidene-4′-aminocinnamate |
| N,A,C,B |  Diethyl terephthalylidene-bis-(4-′aminocinnamate) |

[a] From Brown, G. H., and Wolken, J. J. (1979). "Liquid Crystals and Biological Structure," pp. 32–33, Academic Press, New York.

[b] Key: N, nematic; Ch, cholesteric; A, smectic A; B, smetic B; C, smectic C.

## 3. Polymorphism

Many thermotropic liquid crystals display more than one mesomorphic phase on heating from the solid to the isotropic liquid phase. These substances are said to be polymorphous (Table III). The usual sequence of phases on heating is as follows: solid, smectic B, smectic C, smectic A, nematic, isotropic. A more complete listing of the sequence of smectic phases on heating is E, H, G, F, I, B, C, D, and A. When one or more of the given phases are not present, the remaining phases appear in the established order. When a cholesteric phase is present, it takes the place of the nematic phase in the above sequence. Except for the action of external forces, twisted and ordinary nematic phases do not occur for the same (pure) compound.

For discotics, a tentative sequence on heating is as follows: crystal, discotic nematic, columnar (biaxial), columnar (uniaxial), and isotropic.

Most of the transitions between liquid crystalline phases are reversible or enantiotropic. That is, the transitions reverse direction on switching from heating to cooling. Supercooling is not uncommon for many of the phase transitions. Some phase transitions are apparently monotropic;

that is, one of the phases forms only on cooling. One example is the smectic phase of cholesteryl nonanoate.

A variety of reentrant phase transitions have been discovered in which the samples "leave" a phase and then "reenter" again as the temperature is continuously raised. Some examples of sequences found with increasing temperatures are as follows: smectic A, nematic, smectic C, smectic A, nematic, isotropic; smectic C, nematic, smectic C, smectic A, nematic, isotropic; smectic C, smectic A, nematic, smectic A, nematic, smectic A, nematic, isotropic; and cholesteric, smectic A, cholesteric, isotropic.

## B. Lyotropic Phases

### 1. General Characteristics

Mixtures of two or more components that change phase with changes of concentration are called lyotropic. Although compounds that form thermotropic phases are by no means uncommon, lyotropic mixtures are very familiar; soap-and-water being one of the prime examples of a two-component system with lyotropic phases. Mixtures of three or more components are also common, one example being the use of an emulsifier to mix the oil and water or vinegar of a salad dressing. The soap-and-water mixture, of course, is most useful when the soap molecules allow grease to "dissolve" in the water, forming a three-component system. Although water is a very common component of lyotropics, other solvents can be used as well. Most lyotropic phases involve the solution of rod-like molecules or aggregates of molecules in a normally isotropic solvent such as water. It is principally the rod-like entities that become ordered.

Soaps are a simple example of a whole class of molecules, called surfactants or amphiphiles, which form lyotropic phases in water (Fig. 4). These molecules are called amphiphiles (from the Greek *amphi*, meaning "of both kinds," and *philo*, meaning "loving"). One part of an amphiphilic molecule, the polar "head," has an affinity for polar solvents such as water (hydrophilic), while the other end, the organic "tail," is relatively insoluble in water (hydrophobic). The result of these opposite tendencies is that the molecules prefer to organize themselves into surfaces (hence the name "surfactants") with the polar heads pointing toward the water. Such systems may form a number of possible structures, depending on the concentrations of the components and the shapes of the molecules involved (Fig. 5). The amphiphilic molecule is usually represented in a figure with a circle as the polar head and a wiggly line as the organic tail. Simple soaps often have a string of hydrocarbons forming a hydrophobic tail attached to a polar head. A second type of molecule has two such tails attached to a polar head.

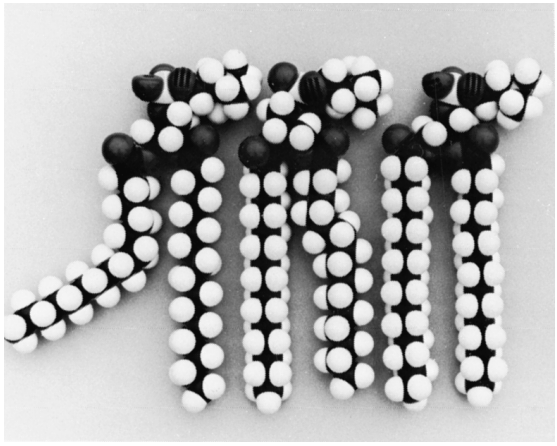


**FIGURE 4** Three molecular models of the same lipid in different configurations. The lipid, dipalmitoyl phosphatidylcholine, consists of two hydrocarbon tails, $(CH_2)_{14}CH_3$, linked to a head group, $(CH_3)_3N^+(CH_2)_2PO_4^-$, by ester linkages and a glycerol backbone, $(OCO)_2(CH_2)_2CH$. The molecule on the left has a single jog (gauche rotation) in one tail, while the other tail is all trans. The middle molecule has a kink (gauche–trans–gauche sequence) in one tail. [Courtesy of H. L. Scott, J. F. Nagle, and the American Institute of Physics; from "Biomembrane phase transitions," *Phys. Today* **31**(2), 38–47 (1978).]

When a crystalline amphiphile is added to water, several mesophases can be observed, ranging from a true solution to the crystal state. Some of the phases may show smectic or lamellar packing (in layers) or even cubic or hexagonal molecular packing. The amphiphile can also aggregate into structures such as spherical or cylindrical micelles. Micelles have the ability to solubilize an otherwise insoluble chemical by encapsulating it. This is what soap does to dissolve grease. Similarly, water can be dissolved in oil by inverse micelles formed by surfactant molecules with the tails pointing outward instead of inward. One common sequence of mesophases obtained on adding water is as follows: solid, lamellar liquid crystal, cubic liquid crystal, hexagonal liquid crystal, micellar, homogeneous solution. Table IV gives some of the properties of some of the common lyotropic phases.

### 2. Lamellar

These phases, also called "neat" or G phases, correspond to the thermotropic smectic phases; that is, they are characterized by layers having a well-defined thickness but no structure within the layers (Fig. 6a). Different lamellar phases have been found in the same system. In the phase designated $L_\alpha$, the $\alpha$ is used to indicate that the hydrocarbon chains in the tail of the amphiphile are fluid or flexible. It is possible to observe a transition from

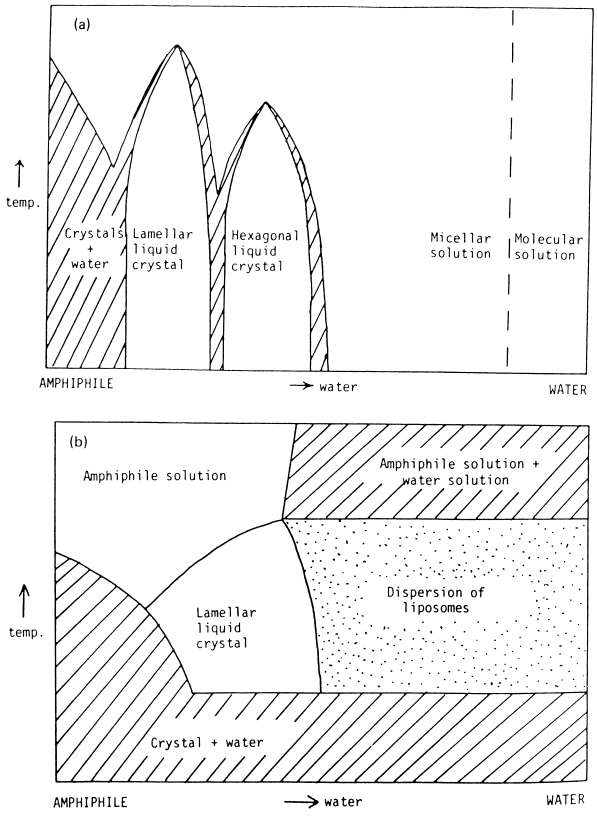



**FIGURE 5** Characteristic phase diagrams of amphiphile–water systems. Two-phase regions are shown shaded. (a) Strongly polar amphiphiles (e.g., soaps, alkyl sulfates, quaternary ammonium salts, and lysolecithins). (b) Amphiphiles with relatively large hydrophobic regions (e.g., monoglycerides and lecithins). [From Friberg, S., and Larsson, K. (1976). "Liquid crystals and emulsions." *In* "Advances in Liquid Crystals," Vol. 2, Academic Press, New York, by permission.]

one lamellar phase, such as $L_\alpha$, to another in which the tails "freeze," that is, lose much of their flexibility. The water and surfactant molecules in the lamellar phase are in alternate layers with the surfactant molecules forming double-thickness layers called bilayers in which the hydrophobic tails are separated from the water layers by planes of polar heads. Single bilayers of lipid molecules form the underlying structure of biological membranes (Fig. 7).

### 3. Hexagonal

As water is added to a lamellar phase, the layer structure can be replaced by one in which the surfactant molecules apparently form cylindrical structures with the polar heads forming the outer shell. The cylinders line up in hexagonal arrays with the water between (Fig. 6b). This phase is also called the middle, or $M_1$, phase. At concentrations of surfactant greater than that of the lamellar phase, some systems also form an inverted hexagonal phase, also called the inverse middle, or $M_2$, phase, in which the tails point away from the centers of the cylinders. The water occupies the centers of the cylinders.

### 4. Cubic

A cubic structure, also referred to as the viscous isotropic or $V_1$ phase, sometimes forms at amphiphile concentrations between those producing lamellar and hexagonal phases. Ordinary optical observations show only an isotropic structure. X-ray diffraction studies show that the surfactant molecules pack into spheres, which then pack into a face- or body-centered-cubic lattice. The inverted structure ($V_2$) can also form between the lamellar and inverse hexagonal phases. Another viscous isotropic phase ($S_{1c}$) has been observed at concentrations of amphiphile lower than that of the hexagonal phase.

### 5. Nematic

Lyotropic nematics have been observed for which the optical axes are easily oriented, as is the case for thermotropic nematics. The basic units that align are not single molecules, but aggregates of molecules whose sizes are comparable to those of micelles—20 to 100 Å.

## C. Polymeric Liquid Crystals

Examples of liquid-crystalline order have been found in fluid polymer melts and solutions. Such polymers may play an important role in the spatial organization of biological macromolecules, for instance, in the packaging of DNA in chromosomes and in the aggregation of microtubules, which form the structural framework of cells. A fiber formed from the liquid crystalline phase of a polymer, Kevlar, is an ultra-high-strength polymer that has a stiffness comparable to that of steel with a much lower mass density.

Examples of both lyotropic and thermotropic polymeric phases have been studied. Solutions of synthetic polypeptides can form a helicoidal cholesteric structure in which the polymers form twisted rods with pitches between $10^{-7}$ and $10^{-3}$ m. The rods may be separated by several millimeters. Unlike the monomeric thermotropics, however, the rods can be untwisted into a nematic phase by changing temperature. Polymers having thermotropic liquid crystalline mesophases have been made by adding single liquid crystalline molecules to polymer chains to form either comb-like or linear polymers. In a comb-like polymer, the liquid crystal monomers are attached by flexible links to

**TABLE IV Some Properties of Lyotropic Systems Composed of an Amphiphile and Water[a]**

| Suggested structural arrangement |  |  |  |  |  |  |
|---|---|---|---|---|---|---|
| Percent water[b] (approximate range) | 0 | 5–22–50 | 23–40 | 34–80 | 30–99.9 | >99.9 |
| Physical state | Crystalline | Liquid crystalline, lamellar | Liquid crystalline, face-centered cubic | Liquid crystalline, hexagonal compact | Micellar solution | Solution |
| Gross character | Opaque solid | Clear, fluid, moderately viscous | Clear, brittle, very viscous | Clear, viscous | Clear, fluid | Clear, fluid |
| Freedom of movement | None | Two directions | Possibly none | One direction | No restrictions | No restrictions |
| Microscopic properties (crossed nicols) | Birefringent | Neat soap texture | Isotropic with angular bubbles | Middle soap texture | Isotropic with round bubbles | Isotropic |
| X-ray data | Ring pattern 3–6 Å | Diffuse halo at about 4.5 Å | Diffuse halo at about 4.5 Å | Diffuse halo at about 4.5 Å | | |
| Structural order | Three dimensions | One dimension | Three dimensions | Two dimensions | None | None |

[a] From Brown, G. H., and Wolken, J. J. (1979). "Liquid Crystals and Biological Structure," pp. 30–31, Academic Press, New York.

[b] The different percentages of water show that different amphiphiles require different amounts of water. For soaps, the lamellar structure generally occurs between 5 and 22% water; with some lipophiles the water may be as high as 50%. The cubic structure generally occurs between 23 and 40%.

the main chain, like the teeth of a comb. All the common thermotropic phases have been obtained in this way, with the possibility of locking the structure by quenching (rapid cooling) in the presence of applied magnetic or electric fields. Linear polymers formed by linking liquid crystals end to end have been formed with stable liquid crystalline phases in the range 100–400°C. The phases have properties in common with the monomeric liquid crystals, but the response to external stimuli can be much slower.

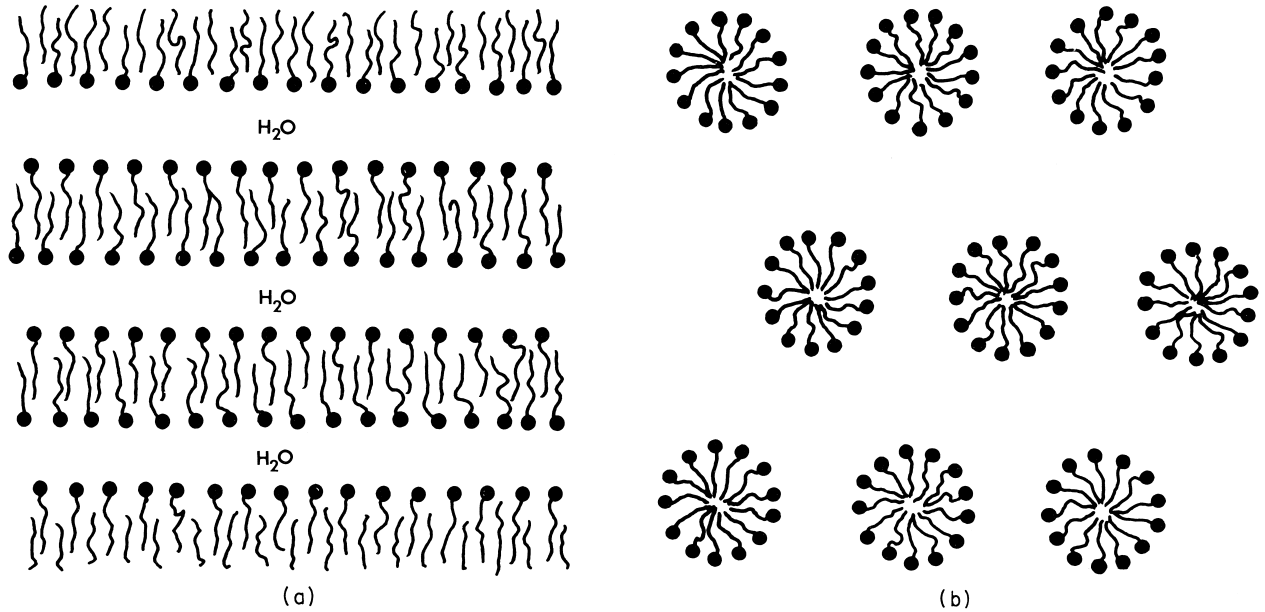



**FIGURE 6** Molecular arrangements in two lyotropic phases, shown in cross section. (a) A lamellar phase, in which the amphiphile molecules form bilayers with their hydrophobic tails toward the layer centers, away from the water. (b) The hexagonal phase, showing cross sections of rod-like structures having their axes perpendicular to the plane of the drawing. Water is in the region between the cylinders. [Courtesy of Ging-Sheng Yu.]

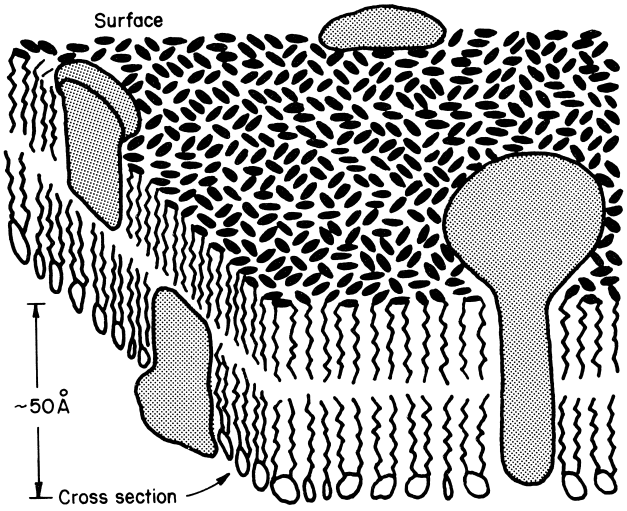



**FIGURE 7** Main features of a biomembrane: the bilayer of lipid (fat) molecules and the proteins (shaded). At low temperatures the hydrocarbon tails of the lipid molecules appear as regular zigzag lines, as the cross section indicates. The tails are joined in pairs by a backbone, shown on the top surface. Omitted from the top but shown on the bottom layer are the head groups that are attached to the backbones of the lipid molecules. [From Nagle, J. F., and Scott, H. (1978). "Biomembrane phase transitions," *Phys. Today* **31**(2), 38–47 by permission.]

## III. PROPERTIES OF LIQUID CRYSTALS

### A. The Director and the Order Parameter

For many experiments and in many mesophases, a useful model of the orientational motions of a molecule separates the motions into the following classes:

1. Rapid rotations (librations) about the long axis.
2. Rapid fluctuations of the long axis about a local director, designated by a unit vector, $\mathbf{n}(\mathbf{r}, t)$.
3. Fluctuations of the local director, which represent collective motions of many molecules and are correspondingly slower than the individual molecular motions.

Due to external influences, the average director may then vary over macroscopic distances (Fig. 8). An instantaneous snapshot would show molecules with their long axes at angles with the local director that can average as much as 40° in a nematic. A nematic would also show symmetry with respect to alignment of molecules parallel or antiparallel to the director. One measure of the degree to which molecules align with the director is given by the order parameter $s$, which is defined by the following equation:

$$s = \tfrac{1}{2}\langle 3\cos^2(A) - 1\rangle = \langle P_2(\cos(A))\rangle, \qquad (1)$$

where $A$ is the angle between the local director and the instantaneous molecular long axis, $P_2(x)$ the second Legendre polynomial, and $\langle\ \rangle$ indicates the average value. A value for $s$ of 1.0 corresponds to perfect order and a value of 0.0 would indicate complete disorder, as in an isotropic liquid. For a nematic, the values commonly range from 0.4 near a nematic-isotropic transition to as high as 0.8.

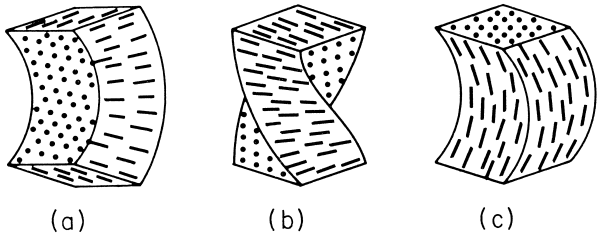



**FIGURE 8** Static deformations of the director in a nematic liquid crystal showing pure (a) splay, (b) twist, and (c) bend modes. [Courtesy of Nuno Vaz.]

### B. Elastic Properties

Although a liquid crystal may appear at first glance to be simply a turbid liquid like milk or a very viscous gel like petroleum jelly, the flow properties can be very complicated. For instance, an aligned smectic A in a test tube looks liquid when tilted one way but nearly solid when tilted in another direction. Properties such as elasticity and viscosity are not simple scalar parameters but depend on the relation between the direction of the director, layer normals in smectics, and the distortion or motion.

In the nematic, unlike the solid, there are no permanent forces opposing the change of distance between two molecules or small volume elements. There are, however, torques that oppose the curvature of the director. The assumption that the restoring torques are linearly proportional to the curvature strains then gives rise to a free-energy density that is a quadratic function of the curvature strains. The linear components of curvature are grouped into three modes of deformation: splay, twist, and bend (see Fig. 8). Splay can be demonstrated by the fingers of one's hand when spread out, that is, "splayed." The fingers then diverge from a point; in fact, the divergence of the director is the mathematical expression for splay. In the twist mode, the director changes as one moves along a line perpendicular to the original director. The change in director is perpendicular to both directions, unlike the splay deformation. Bend can be displayed by curling the fingers of one's hand. In this mode, the director changes in direction as one proceeds along it. Taking the $z$ axis along the local director, the first-order derivatives of the director are classed as follows:

| | | | |
|---|---|---|---|
| splay: | $\delta n_x/\delta x$ | and | $\delta n_y/\delta y$ |
| twist: | $\delta n_y/\delta x$ | and | $\delta n_x/\delta y$ |
| bend: | $\delta n_x/\delta z$ | and | $\delta n_y/\delta z$. |

Making use of the symmetries of the nematic phase, that is, invariance under rotations about the director and the equivalence of $\mathbf{n}$ and $-\mathbf{n}$, the free energy density of a nematic in bulk can be written to second order as follows:

$$F = \tfrac{1}{2}\{K_1(\boldsymbol{\nabla}\cdot\mathbf{n})^2 + K_2(\mathbf{n}\cdot\boldsymbol{\nabla}\times\mathbf{n})^2 + K_3(\mathbf{n}\times\boldsymbol{\nabla}\times\mathbf{n})^2\}. \quad (2)$$

The constants $K_1$, $K_2$, and $K_3$ are referred to as the Oseen–Frank or splay, twist, and bend elastic constants.

The elastic constants can be determined by several methods. One method involves setting up a uniform alignment of directors by strong anchoring at the parallel boundary surfaces. A magnetic field (or electric field) is then applied at right angles to the director. The alignment remains undeformed until a critical field $B_c$ is reached, at which point there is a transition, called the Frederiks transition, to a state in which the director varies throughout the thickness of the sample. The transition can be detected with a polarizing microscope. The critical field in certain geometries of this experiment is simply related to the elastic constants, sample thickness, and anisotropy of the magnetic susceptibility. For instance, for a homeotropic alignment, the critical magnetic field is proportional to the square root of $K_3$, the coefficient for bending. In practice, the strength of the anchoring must be taken into account. Measured values are of the order of $10^{-11}$ N. The twist elastic constant $K_2$ is generally the smallest of the three, typically $3.0{-}4.0\times10^{-12}$ N.

Thermal fluctuations in the local director can be described in terms of the continuum theory and hence depend on the elastic constants. As a rigid rotation of the director requires no energy, the energy required to create a long-wavelength fluctuation in the director is small, and the relaxation time for such a fluctuation is long compared to the period of visible light. The result is significant fluctuations in the local optical properties. These fluctuations give rise to light scattering, which can then be studied to give information on the elastic constants. For instance, for scattering at an angle $A$ from an incoming beam polarized at right angles to the sample director and then analyzed with a polarizer at right angles to the original polarization, the differential cross section per solid angle is roughly given by the following:

$$d\sigma/d\Omega \sim \cot^2(A/2) + K_1/K_2, \quad (3)$$

allowing the ratio $K_1/K_2$ to be determined.

Some of the concepts and results of the continuum theory can be applied to other phases. In the smectic A phase, for instance, the smectic layers are easily bent, corresponding to a splay deformation, so $K_1$ has values similar to those found in nematic phases. Twist and bend deformations of the director, on the other hand, are nearly ruled out, as they require changes in the layer thickness comparable to the compression of a normal liquid. One would expect $K_2$ and $K_3$ to increase anomalously in a nematic phase as a smectic A phase is approached, especially when the transition is nearly second order, that is, when the transition has a small latent heat.

### C. Flow

The coupling between directors and flow complicates the theoretical and experimental studies of flow properties, even in the "relatively simple" nematic phase. The angles between the local director, flow velocity, and velocity gradient (shear) all affect the flow, and the orientational and translational motions of the molecules are linked. The formulation of the dynamical properties by Leslie and Ericksen is most commonly used in studies of the flow properties of the nematic state. In this formulation the viscous stress tensor is decomposed into the sum of six tensors with coefficients having the dimension of a viscosity (Leslie coefficients). The effective viscosities measured in different experiments are then analyzed in terms of these coefficients, five of which are independent. Experimentally, the direction of alignment must be controlled (by electric fields, typically) and measured. Because of the turbidity, optical measurements are restricted to thin samples. The hydrodynamics and electrohydrodynamics of nematics often lead to many interesting and potentially useful instabilities.

In a technique used as early as the 1930s to study the viscous properties of nematics, a strong magnetic field was used to align the director of a sample in which a shear flow was set up. That is, the velocity of the liquid was directed along the $x$ axis and the shear directed along the $y$ axis. With the director along one of the three orthogonal axes, the measurement of the ratio of shear stress to shear gave the apparent viscosity for that geometry. As might be expected from the elongated shape of typical liquid crystalline molecules, the measured viscosity was least when the director was parallel to the direction of flow (Fig. 9a) and greatest when the director was parallel to the shear (Fig. 9b). Typical values for viscosity coefficients in nematic phases at around 130°C are 1–2 cP for director parallel to the flow direction, 8–9 cP for director parallel to the velocity gradient, and 2–4 cP for director perpendicular to both flow and gradient (1 cP = 1/100 poise = 1/100 kg × m$^2$/sec).

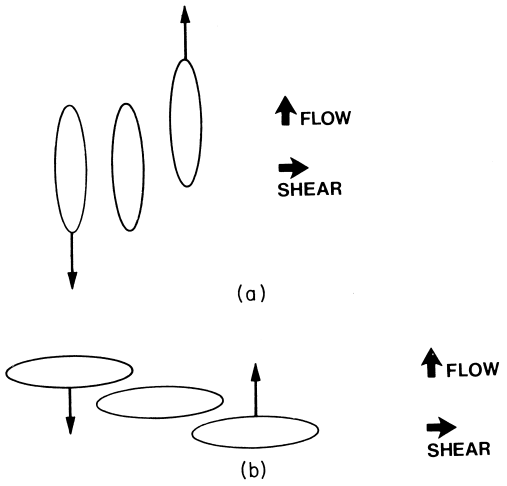



**FIGURE 9** Two examples of shear flow in which the flow of molecules and the velocity gradient (shear) are at right angles. (a) Director parallel to the flow; (b) director parallel to the shear. [Courtesy of Jie Zhu.]

A rotational viscosity can be determined by measuring the torque required to rotate the director of a cylindrical sample of nematic liquid crystal placed in a magnetic field at right angles to the axis of rotation, the latter also being the cylinder axis. Disregarding anchoring effects and the contribution from the bottom of the sample, at low rotational velocities the magnetic and viscous torques balance when the director makes a uniform angle with the magnetic field. In a variation of this method, a cylindrical nematic sample in a magnetic field is twisted suddenly, rotating the director away from the direction of the magnetic field. The director then relaxes to the original orientation, with a characteristic time constant depending on the anisotropy of the magnetic susceptibility, the strength of the magnetic field, and the rotational viscosity of the liquid crystal. A nuclear magnetic resonance (NMR) signal is used to determine the orientation of the director.

Several other experiments give information about the flow of liquid crystals. The time dependence of the fluctuation of the nematic director can be studied to yield information about the viscosity coefficients. Two such experiments are measurements of the frequency modulation of light scattering and the dependence of the NMR longitudinal relaxation time ($T_1$) on (1) the frequency and (2) the angle between the director and the magnetic field. The reflection of ultrasonic shear waves and the attenuation of such waves as a function of the angle between the wave vector and the director has also been used to determine viscosity coefficients. The sudden application of an electric or magnetic field to cause a Frederiks transition (see above) is sometimes accompanied by flow, often in a complicated way. As this effect (especially in a homeotropic sample) limits the behavior of certain types of liquid crystal displays, it is not infrequently of practical importance.

Although a cholesteric liquid crystal behaves locally like a nematic, its flow properties are vastly different. Its apparent viscosity increases by as much as a million times as the shear rate drops to very low values. On the other hand, for some geometries in which the flow is perpendicular to the helical axis, the apparent viscosity is approximately of the same order of magnitude as in a nematic. Apparently, at low shear rates, flow in the direction of the pitch axis takes place along a fixed helical structure, with the molecules constrained to twist as they move along the axis.

The apparent viscosity of a smectic is very high and, like that of a cholesteric, depends drastically on the shear rate, with typical values rising from 10 to $10^4$ poise as the shear rate (velocity gradient) is reduced from 100 to 0.01 $\text{sec}^{-1}$. Furthermore, for acoustic waves a smectic liquid crystal will generally have more than one mode of oscillation at a given frequency, observable from acoustic studies or from Brillouin scattering. One branch is associated with density oscillations and has a velocity essentially independent of direction. The second branch, similar to second sound in superfluids, has a velocity that depends on the sine of twice the angle between the director and the direction of propagation. As is the case with elastic constants, some nematic viscosity coefficients diverge as a nematic–smectic transition is approached.

### D. Director Alignment

The typical effect of an electric or magnetic field on an isotropic liquid is weak, as the external forces work on the molecules individually and the thermal motions dominate; that is, the energy difference between the alignments of a molecule parallel and perpendicular to an electric or magnetic field is much smaller than $kT$. On the other hand, in solid crystals the molecules are fixed in position and orientation and the effect of an aligning field is a torque on the whole crystal. Liquid crystals, having both the fluidity characteristic of liquids and the collective behaviors of crystals, respond in unique and sometimes useful ways to external stimuli. The ability to align the director of a liquid crystal and, in turn, affect its optical or mechanical behavior by using external fields, surface interactions, and flow gives rise to many of the interesting applications of liquid crystals such as displays and high-strength materials. The interactions of the director with external stimuli are most important for the nematics, especially the ordinary nematic, which is fluid in all three dimensions. Tilted smectics such as the smectic C can also respond to external fields, subject to the constraint of having a fixed tilt

angle. Most of the remarks below refer to nematic liquid crystals.

### 1. Surface Interactions

Surface forces are often strong enough to impose a well-defined direction to the director at boundaries of the liquid crystal with other materials. By treating a cleaned glass surface with certain detergents, it is possible to align the nematic director perpendicular to the surface. It is thus easy to prepare a sample of nematic liquid crystal between two parallel plates of glass that has a "homeotropic texture," that is, a single-domain crystal with its optical axis perpendicular to the walls. Other surface treatments, such as rubbing the glass with a tissue or evaporating films at oblique angles of incidence, can give rise to other anchorings of the director at the surface. in some cases, as with the free surface of MBBA (the first example of Table I) in the nematic phase, there is a continuous set of directions that the director can take on, such as the cone of directions that make a constant angle with the surface normal. In such a case, transitions in the anchoring have been observed on changing the temperature of a sample with a thickness on the order of 50 $\mu$m.

Surface alignment of the nematic director is used to form the twisted-nematic liquid crystal used in many LCDs. A nematic liquid crystal having a macroscopic twist is created by placing an ordinary nematic between two surfaces, each of which has been treated to align the director parallel to a particular direction in the surface (see Fig. 17). The two surfaces are parallel to one another, but the alignment directions are at an angle of 90°, causing the director to twist slowly through a right angle.

### 2. Magnetic Fields

While the interaction between an isolated molecule of liquid crystalline material and a magnetic field of, say, 1 T is several orders of magnitude smaller than thermal energies, even a field 10 times smaller will align a sample of nematic liquid crystal. The molecules in the nematic phase line each other up so that the field acts collectively on a large number of molecules. This provides a method for aligning liquid crystals that have a nematic phase. The time scale for the orientation of the director in a 1-T field depends strongly on the viscosity of the sample and can range from milliseconds to hours. A nematic sample in a cylindrical container, when rotated on its axis with the axis perpendicular to the magnetic field, can align with its local directors sampling all the directions perpendicular to the rotation axis. The competition between the aligning effects of a surface and a field can give rise to a number of geometries. For instance, a simple twist can be created close to a surface at which the director is anchored in one direction contained in the plane, for example, along the $z$ axis of an $x$–$z$ plane. If the magnetic field is applied along the $x$ axis, the director aligns along a right- or left-handed twist, eventually lining up with the magnetic field far from the surface. In a typical case, this distance can be about 3 $\mu$m with a field of 1 T. It is interesting that a weak external perturbation can be used to create a distortion on a scale approaching an optical wavelength. A magnetic field applied at right angles to the helical axis of a cholesteric can distort the structure and eventually "untwist" the cholesteric into a nematic structure.

### 3. Electric Fields

The director of an insulating nematic liquid crystal tends to align either parallel or perpendicular to an electric field, depending on the structure of the molecule. For a typical nematic that aligns parallel to the electric field, an electric field of about 1 V/cm is equivalent in effect to a magnetic field of 1 G (roughly the strength of the earth's magnetic field). In some LCDs an electric field is used to switch from the twisted nematic configuration set up by the surface alignment to a nearly homeotropic alignment when the field is turned on. A potential difference of only a few volts is sufficient to cause this realignment. The devices use very little energy as only a rotation of molecules is involved. Two display devices that make use of the alignments due to surfaces and electric fields are discussed in Section IV (see Figs. 17 and 18).

### 4. Flow Alignment

Flow of a liquid crystal affects the alignment and, conversely, a realignment caused, for instance, by the application of an electric or magnetic field can set up a flow of material. The effects of flow are quite different in different phases. Because of their characteristic elongated shapes, the molecules in a nematic flow more readily along the director than perpendicular to it; thus, if a nematic sample is sheared between two glass plates, the tendency is for the director to align in the direction of the shear. If the same thing is done to a cholesteric having no overall alignment of the helical axes, a planar or "Grandjean" texture arises in which the axes line up perpendicular to the direction of shear. A smectic flows best in directions perpendicular to the planar normals so that a shear tends to align the smectic layers with their normals perpendicular to the direction of the shear.

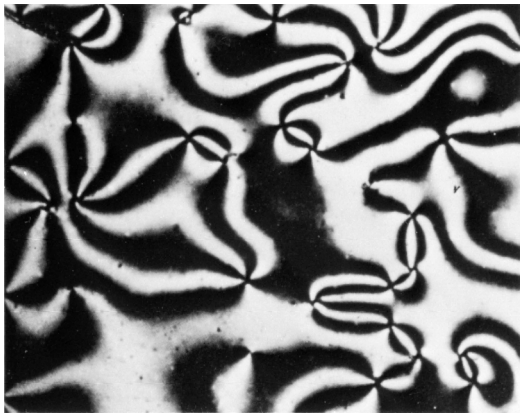


**FIGURE 10** A schlieren texture obtained in a thin sample of nematic liquid crystal. The points from which alternating light and dark areas (brushes) radiate are the end points of line singularities of the director. The picture here is produced in a microscope, as in Figs. 11–16, by viewing the light transmitted through a sample placed between crossed polarizers. [Courtesy of J. W. Doane, Liquid Crystal Institute.]

## E. Optical Properties

The optical behaviors of liquid crystals give rise to useful and often spectacular effects such as the optical switching characteristics used for liquid crystal displays and the vivid temperature-dependent colors of the cholesterics. The appearance of a bulk sample of liquid crystal in ordinary light ranges from transparent through translucent and from turbid to brightly and iridescently colored. By using polarized light, a very wide range of appearances or textures can be created by means of surface, flow, and field alignment of the various phases (see Figs. 10–16). Transitions from one phase to another can often have a remarkable appearance.

The microscopic basis for the many optical effects is the elongated shape and electronic structure of the typical

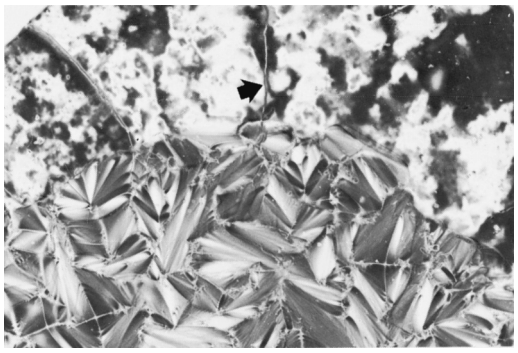


**FIGURE 11** Sample undergoing a transition from a nematic (top) to smectic A (bottom). The temperature is lower at the bottom of the picture. One of the "threads" from which the nematic takes its name is indicated by an arrow in top half of the picture. [Courtesy of Dr. Mary E. Neubert.]

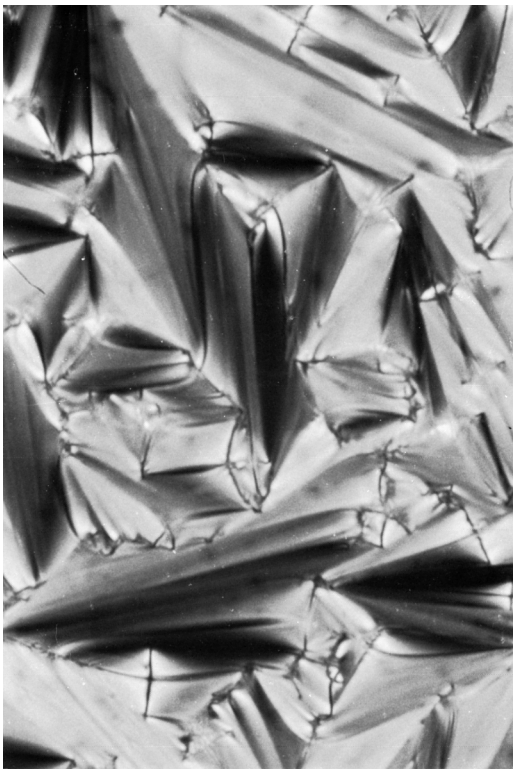


**FIGURE 12** Focal conic fan-shaped texture in a smectic A liquid crystal. The fans in different regions may appear in different colors. [Courtesy of Dr. Mary E. Neubert.]

liquid crystalline molecule together with the tendency for the long axes of the molecules to align with each other along the director. This makes the electric polarizability and, in turn, the index of refraction of the medium anisotropic. Phases such as the ordinary nematic and the untilted smectic A are optically uniaxial with the optic axis along the director. In fact, the cylindrical symmetry of the phases implies that any macroscopic physical properties, optical ones included, have identical values when measured in any orientation perpendicular to the director. For light propagating along the optic axis, all directions of polarization are equivalent, so there is no birefringence. For light traveling along other paths, birefringence is observed. Most commonly, the uniaxial phases have positive birefringence; that is, the refractive index is at its maximum for light polarized along the director (also the optic axis).

The colors that are produced with liquid crystals can arise in several ways.

1. When the pitch of a cholesteric is in the range of wavelengths of visible light, Bragg reflections occur for visible light, similar to the Bragg scattering of X-rays from solid crystals.



**FIGURE 13** Fan-shaped texture in a smectic E. The striations often bound areas having different colors. [Courtesy of J. W. Doane, Liquid Crystal Institute.]

2. Dichroic dye molecules such as methyl red, when dissolved in liquid crystals, tend to orient along the director. These molecules will absorb light polarized in one direction. By aligning the director of such a solution in different directions with respect to the polarization of a light beam, different colors may be produced.

3. For optical studies of the nature of phases and the temperatures of transitions, liquid crystals are placed between crossed polarizers. The beautiful and revealing textures that are observed are due to the interference characteristic of birefringent materials.

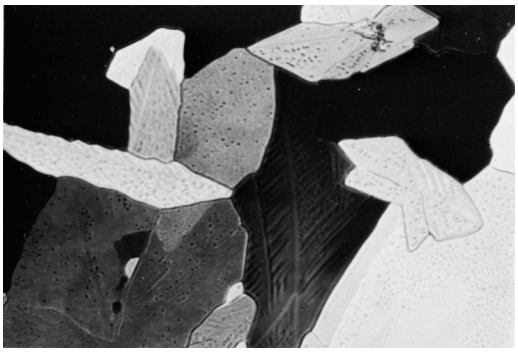


**FIGURE 14** Mosaic textures in a smectic H. The different areas are usually of different colors. [Courtesy of Dr. Mary E. Neubert.]



**FIGURE 15** A cholesteric with large, visible pitch, created by adding an optically active material to a normally nematic material. In the dark areas the molecules are perpendicular to the plane of the picture. [Courtesy of J. W. Doane, Liquid Crystal Institute.]

A cholesteric is uniaxial on a local scale of tens or hundreds of molecules in length with the optic axis rotating to describe a helix identical to that described by the director. Light incident along the pitch axis can be thought of as a sum of two waves, one with electric field rotating in the opposite sense to the helix and one with the electric field rotating in the same sense. The first wave behaves as it would in a normal medium, having an effective index of refraction equal to the average of the refractive indices for light polarized along and perpendicular to the optic axis. The second wave shows anomalous behavior, with nearly perfect reflections in a narrow band of wavelengths close to the pitch of the cholesteric. This band is typically only 25 nm ($25 \times 10^{-9}$ m) wide, producing a very pure colored appearance similar to the colors sometimes seen on beetle wings. In the blue phase of cholesterics an unusual platelet structure can be observed when different crystal domains have different faces alighned with the sample surface. Distinct colors are reflected from each domain, corresponding to the wavelengths satisfying the conditions for Bragg scattering.

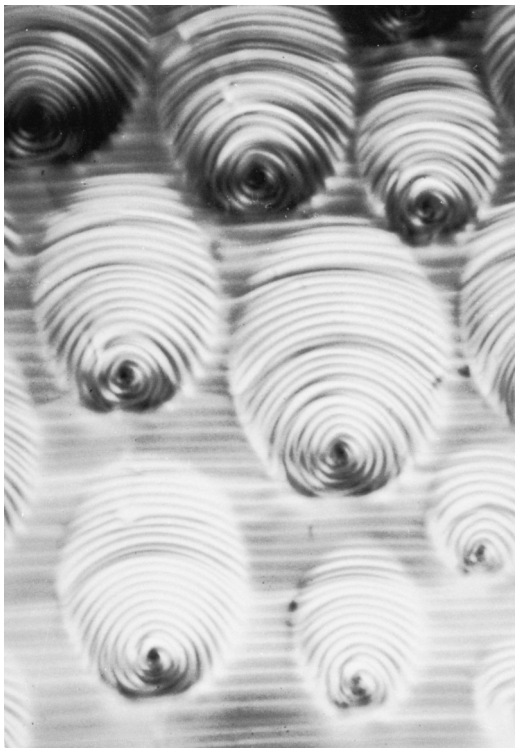


**FIGURE 16** Focal conic defects in a chiral smectic C liquid crystal. The striped bands are due to the *c* director helix and run parallel to the layers. The typical focal conic ellipse-shaped line defects with the circular layers nested about the control line defect are evident. [Courtesy of N. A. Clark, University of Colorado, Boulder.]

The colorful interference patterns displayed by many birefringent liquid crystals observed between crossed polarizers are useful in determining the symmetry of unknown phases and the kinds of allowable defects. When a beam of linearly polarized monochromatic light travels through a birefringent material, the portion of the beam polarized along the optic axis travels at a speed different from that of the portion polarized perpendicular to that direction. Depending on the retardation in phase of one component with respect to the other, the two components combine to give elliptical, circular, or linearly polarized light at various points along the beam. For instance, the beam returns to its original state of polarization whenever the retardation is a multiple of 360°. It will be absorbed in the second polarizer. When the retardation in phase is 180°, the beam is again linearly polarized, but not in the original direction, and is not entirely absorbed in the second polarizer. In a typical liquid crystal sample, the relative retardation depends on the color of the light and the sample thickness and varies with the alignment of the director with respect to the light beam. Thus, with incident white light, different regions of the sample appear to have different colors, which can be changed by rotating the sample with respect to the polarizers, rotating one or both polarizers, or changing the alignment of the sample by external fields or flow. A uniaxial sample having a homeotropic alignment with the director (and thus the optic axis) parallel to the beam does not exhibit this behavior and appears uniformly dark.

A nematic or smectic A sample of constant thickness, prepared with its director everywhere at a constant angle to the beam of light, will have a uniform colored appearance except at defects, which can then easily be seen. An unaligned sample will show gradual changes in color or brightness except at boundaries between domains or director singularities, the shape and motion of which can be used to probe the nature of an unknown phase. A thin sample of nematic having nearly homogeneous alignment displays a characteristic schlieren texture (see Fig. 10) with point-like singularities that are the end points of thread-like singularities (known as disclinations) at which the directors are undefined. The nature of a singularity, one example of which is a radial pattern of directors leading away from the point, can be deduced from its shape, its movement when the sample is rotated, and the result if it combines with another singularity. Although planar in nature, the smectic C can also form a schlieren texture since the component of the director in the plane of the layers can vary smoothly in direction. Smectics and cholesterics that are not uniformly aligned usually exhibit a form of focal conic texture displaying numerous elliptical or fan-like structures that scatter light strongly in all directions and have a strong depolarizing effect (see Figs. 11–13 and 16).

There are many other possible textures, from mosaic to fingerprint in appearance, some having very regular grid-like appearances. Disruptions in uniform structures can be produced by flow, giving rise to patterns such as the sets of parallel lines or even of feathery "chevrons" seen in Williams domains formed by patterns of flow in nematics caused by electric fields of certain strengths and frequencies. At higher voltages these patterns give rise to turbulence, which is accompanied by intense light scattering. This "dynamic scattering" is used in some display devices.

In bulk, the nematic liquid crystal scatters light strongly due to small-amplitude collective modes of orientational fluctuations with significant components near visible wavelengths. Mention has already been made of the study of light scattering to determine elastic and flow parameters. Light scattering is also measured in the vicinity of phase transitions to study fluctuations and critical phenomena. Interesting studies have been made of nonlinear light scattering in which the intensity of light scattered from an aligned sample was not proportional to the intensity of the original beam of

light. Two laser beams of the same frequency can be directed at a liquid crystalline sample to produce a "phase grating," which causes a spatial variation in the refractive indices and, in turn, diffracts some of the original light. This phenomenon may be usable for holographic imaging.

## F. Homologous Series

Since it is ultimately intermolecular forces that give rise to the formation of mesophases and determine which of them form and at which temperatures, it is important to study the relationships between physical properties and chemical structure. Although the shape of a molecule has a large bearing on the phases it forms and the temperatures at which it transforms, other considerations, such as the rigidity of the bonds and the imbalances in attractions between different parts of neighboring molecules, play strong roles. These roles can be investigated by studying series of compounds differing in, for instance, the nature of the terminal groups or the polarizability of different substituents. One of the chemical variables that lends itself to such investigation is the length of the alkyl chains that commonly terminate liquid-crystalline molecules. The compound MBBA is one of a series of molecules differing only in the number of carbon atoms appearing in the end chains. Another molecule in the series is the 10th example of Table I, 4-butyloxybenzal-4-ethylaniline. Typically, the earlier members of such a series of molecules, called a homologous series, will have nematic phases, the later members will have smectic phases, and intermediate members will display both smectic and nematic phases. Plots of quantities such as transition temperatures versus carbon number will often show an even–odd effect. For instance, a plot of the nematic-to-isotropic transition temperature versus carbon number will show an alternation between two curves, one for odd numbers of carbons and the other for even numbers. The separation of the two curves tends to diminish at higher carbon numbers. This suggests that the attraction between ends of molecules plays a role in the stability of nematic versus isotropic phases, since besides affecting the length of the molecule, the addition of one carbon to an alkyl chain changes the orientation of the last carbon–carbon bond with respect to the molecule's long axis.

## G. Other Studies

Many of the tools available to the physicist and chemist have been used to study the properties of liquid crystals and determine the nature of the phases. Many have been mentioned above. Others include measurements of refractive indices, dielectric constants and relaxation, magnetic susceptibilities, neutron scattering, and fluorescence recovery after photobleaching (FRAP, used to study diffusion). Besides the study of optical textures, X-ray scattering, NMR, and thermal measurements are of particular importance in determining the structure and nature of phases and phase transitions and are discussed briefly.

### 1. X-Ray Scattering

As in the study of solids, X-ray scattering is used to determine the symmetries of the phases as well as to measure the separation of the planes of smectics, intermolecular distances, packing of molecules, and degree of long- and short-range molecular order. For many thermotropic liquid crystals it is possible to study oriented samples, but for the lyotropics the studies are more often restricted to powder methods. The diffraction patterns seen are neither the sharp Bragg peaks of monodomain solid crystals nor the diffuse reflections characteristic of isotropic liquids. In oriented samples one finds combinations of sharp and diffuse rings, arcs of rings, and spots.

The recent availability of synchrotron sources of high intensity and low line width (typically under $10^{-3}$ $Å^{-1}$ as opposed to 0.04 $Å^{-1}$ for a rotating-anode generator) has enabled studies of free-standing films of liquid crystals with thicknesses down to several molecules. This has yielded information relevant to the study of two-dimensional phases and phase transitions and, in fact, evidence that the thin films have the expected properties of a two-dimensional crystal.

### 2. Nuclear Magnetic Resonance

The phenomenon of NMR offers a number of tools useful in elucidating the nature of phases and the motions of the molecules in those phases. The spectra obtained from the hydrogen nuclei in most liquid crystal phases are not made up of sets of well-defined absorption lines, as in isotropic liquids, but are typically as broad as 40 kHz, with widths proportional to the order parameter, $s$. This is due to dipole–dipole interactions between neighboring hydrogen nuclei. In isotropic liquids, such interactions are rapidly modulated by the tumbling of the molecules, rapid enough on the NMR time scale ($10^{-7}$ sec, say, for proton NMR at 100 MHz) to average to zero. In most liquid crystals the tumbling is not isotropic, and the interactions are not averaged out, although they are somewhat reduced compared to the crystalline solid. In the smectic D and viscous isotropic phases, chemical shift spectra are obtained as in isotopic liquids, apparently because the molecules diffuse rapidly between areas in which they have different orientations.

The NMR spectra and relaxation times observed in liquid crystalline phases depend on the alignment of the sample in the magnetic field. Especially in a smectic A phase, it is possible to create a single-domain liquid crystal for which the director can be rotated with respect to the magnetic field; this drastically changes the spectrum and even reduces it to a single line when the director makes a "magic angle" of 54.74° with the magnetic field. In a tilted smectic, such as smectic C, the limited freedom of the director to reorient can be seen in the variation of the NMR signal (the free induction decay in a pulsed NMR experiment) when the sample is rotated in the magnetic field. The directors can align in any of a cone of directions centered on the layer normal. Thus, when rotating a sample originally having a uniform alignment of directors but not necessarily layers (this is accomplished by cooling from a nematic phase while in a strong magnetic field), the directors "follow" the field, although not to the extent seen in the nematic phase. Unless the sample consists of a single domain (all the layers are parallel), the final spectrum will be made up of a superposition of spectra corresponding to the parts of the sample having various angles between the director and the magnetic field.

For a nematic sample that has been suddenly rotated, the realignment of the director can sometimes be observed in the changing NMR signal, especially near room temperature. An effective rotational viscosity can be measured with this effect.

The NMR signals from other nuclei are also studied. The signals from deuterium nuclei that have replaced hydrogen are especially useful, as the difference in chemical behavior is usually slight and the spectra from aligned samples show many individual, in some cases nonoverlapping, lines that can be assigned to the nuclei at particular locations in the molecules. This allows detailed investigation into the alignment of the molecules in various phases, including the more solidlike smectic phases, in which even the rotations of the molecules about their long axes may be restricted.

Measurements of spin–lattice relaxation times ($T_1$'s) have been a rich source of information pertaining to the motions of the molecules in liquid crystal phases. For nematics at temperatures above roughly 50°C, for instance, the relaxation of hydrogen nuclei is dominated by motions due to cooperative fluctuations in the local director, giving a characteristic frequency and angular dependence to the measured times. Similar effects can be seen in the rotating-frame relaxation rates in smectic liquid crystals. With deuterated samples (samples in which deuterium atoms have replaced some hydrogen atoms), it is possible to study the motions of different parts of the molecules.

It is possible to study translational motions of molecules by using pulsed magnetic field gradients to "label" the nuclei of molecules according to their positions. Such experiments have allowed direct measurements of self-diffusion coefficients in a few nematics and smectics as well as the diffusion coefficients of probe molecules such as benzene or tetramethylsilane (TMS) dissolved in liquid crystals. Since the method is sensitive to the direction of molecular displacements, anisotropies can and have been measured. It is also possible to vary the diffusion times, allowing investigations into nonlinear diffusion, for instance, that occurring between restrictions such as cell walls.

### 3. Thermal Measurements

Most of the measured values of properties of liquid crystals depend on temperature. Studies of these temperature dependencies are of the utmost importance in determining the correctness of models used to predict them. Determinations of phase diagrams and measurements of heat capacities and latent heats are also necessary to discover the nature of phases and the differences between them. Because of the large number of mesophases and possible parameters that can be manipulated to effect transitions, studies of liquid crystals allow investigations into many predictions of statistical physics, even extending to studies of the phases of two-dimensional systems. For example, by mixing a compound that has nematic, smectic A, and smectic C (or only nematic and A) phases with various proportions of another that has only nematic and smectic C, one obtains a phase diagram (with temperature and concentration as variables) having a "multicritical point," where the nematic–A, nematic–C, and A–C phase boundaries meet. As the point is approached by varying concentrations, the nematic–C transition entropy decreases to zero, as does the "bump" in the specific heat curve near the nematic A transition, while the corresponding bump near the A–C transition increases. Although the phase diagrams for different mixtures differ, high-resolution investigations close to the multicritical point have shown universal features in the shapes of the diagrams. It has been difficult to make theoretical predictions for these shapes because, among other reasons, the smectic A and C phases lack long-range translational order. As confirmed with X-ray scattering, fluctuations in layer positions diverge logarithmically with the sample size in those phases.

## IV. APPLICATIONS

### A. Liquid Crystal Displays

Liquid crystal displays are used in a wide number of applications, from clocks to oscilloscopes. Their popularity is due to the conveniently thin, flat shape and the very low

power required. In these devices, a thin layer of liquid crystal (usually nematic) is sandwiched between parallel cell walls, which have been treated to control the alignment of the liquid crystal director. When a potential difference of several volts is applied to transparent electrodes on either side of the liquid crystal, the resulting electric field causes a realignment of the molecules and a change in the optical behavior of the layer. Many different approaches have been tried. In an early type of display, the dynamic scattering display, turbulence was set up in the liquid crystal, causing it to scatter light. Observed with backlighting, the turbulent area would appear darker than the surrounding areas, while it would appear lighter than the surroundings when observed by reflected light. Polarizers are not required for this kind of display.

In a second type of display, a dichroic dye is dissolved into the liquid crystal. The dye molecules, which act like polarizers, are lined up by the liquid crystal molecules so that the application of an electric field changes the direction of polarization and the amount of light absorbed. A polarizer is required in this device.

The twisted nematic display, shown in Fig. 17, is the one commonly used for digital watches and other small displays. The surfaces of the cell are treated so that, in the absence of an electric field, the local directors are all coplanar, but twist through 90° as shown in the top part of the figure. Light entering the cell (the wide arrow going down at the top of the figure) is polarized parallel to the director at the top surface. The polarization follows the twist in the director and the light passes through the polarizer at the bottom. It is reflected by a mirror and reverses its path to emerge at the top surface. This area appears bright. In an area in which the electric field is turned on, the directors align with the field throughout most of the sample. Now the beam's polarization is not rotated by the liquid crystal and the light is absorbed by the second polarizer. This area appears dark. The flow that accompanied the change of state in early designs typically made the response time of these devices too long to allow the twisted nematic cell to be used in televisions or oscilloscopes.

The essential features of one of the first practical color displays to use liquid crystals are shown in Fig. 18. Red and green light is emitted by the phosphors of a cathode ray tube (CRT). A pair of color polarizers is used so that the light incident on the liquid crystal consists of red light polarized vertically and green light polarized horizontally. As with the twisted nematic cell, if the light passes through the liquid crystal cell when the electric field is on, it arrives unchanged at the final polarizer. In this case the red light is absorbed and the green is transmitted. On the other hand, with the electric field off, the molecules relax toward the configuration favored by the surface interaction, in which the directors lie along a curve similar to a parenthesis, (.

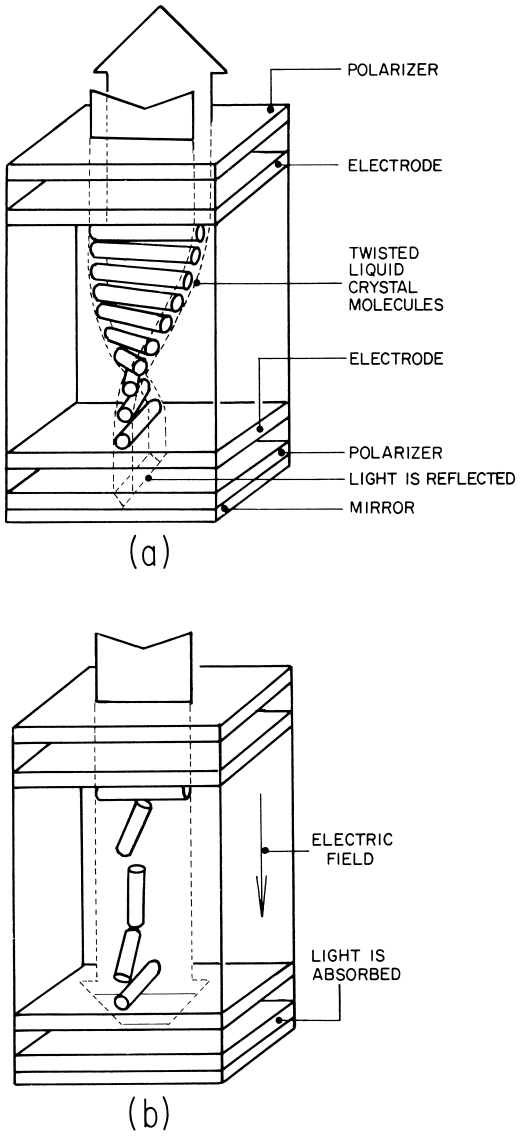



**FIGURE 17** Operation of a twisted nematic LCD. (a) In the off state, the molecules align perpendicular to the incoming light (indicated by wide arrows) and with a twist from top to bottom that rotates the direction of polarization of the light so that it passes through the second (crossed) polarizer and is reflected, giving a bright appearance. (b) With the electric field on, the molecules line up with the field, except for layers very close to the treated surfaces. The incoming light's polarization is not rotated, and the light is absorbed in the second polarizer, making the area appear dark. [Courtesy of Jie Zhu.]

The thickness and birefringence of the liquid crystal cell are such that the direction of polarization of the light is rotated by 90°, with the result that the red light is transmitted and the green absorbed. The device switches states in milliseconds and produces a multicolored display with excellent contrast, performing well in high ambient light.

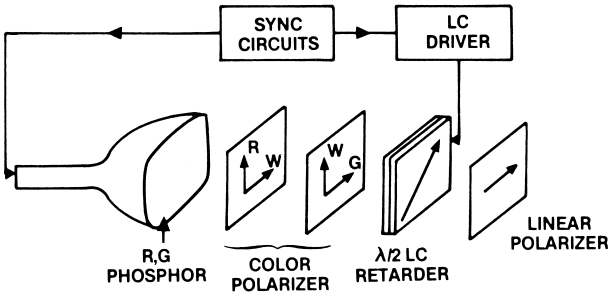



**FIGURE 18** Liquid crystal CRT sequential color display. Red and green information is sequentially written on a multicomponent (R, G) phosphor screen. Color polarizers orthogonally polarize the red and green light emitted. The LC switch sequentially rotates colored information into the transmission axis of a linear polarizer. [Courtesy of P. J. Bos, Tektronix.]

Other display devices using liquid crystals include a screen for a miniature color television, storage displays using thin-film transistors, and displays that store information by altering the state of the liquid crystal locally (e.g., destroying the alignment in a small area by heating it with a laser beam). Bistable electro-optic switches with switching times under 1 $\mu$sec have been constructed with ferroelectric smectic C liquid crystals. Cholesterics have been used in several bistable displays. In one kind, the cholesteric is initially in a planar configuration. A low-frequency voltage pulse disrupts the alignment into a light-scattering focal-conic structure, which persists after the end of the pulse. A short, higher frequency pulse restores the initial alignment. Alternatively, the planar alignment can be disrupted with heat from an infrared beam. Liquid crystals encapsulated in epoxy have been used to create rugged, fast-switching displays with high contrast.

## B. Other Commercial Applications

The pitch of a cholesteric liquid crystal, and thus its colored appearance, is sensitive to such things as temperature, pressure, electric and magnetic fields, and impurities. Cholesterics are used to create continuous maps of temperatures on various surfaces, for instance, to locate circuit board or welding faults and to detect radiation and carcinoma of the breast. In such applications, a coating of liquid crystal can be painted on the area. The range of temperatures to which the coating responds can be widely varied by choice of the liquid crystals used.

Ultrasonic waves have been detected with cholesteric liquid crystals in which the pitch is altered by local heating or by the direct effect of high-intensity waves. In another application the ultrasonic waves directly cause a change from one stable director configuration into another. Such detectors may be usable in sonar devices.

The development of Kevlar, a high-strength polymer competitive with steel on a weigh-per-strength basis, has stimulated the study of liquid crystal polymer phases, from which the fibers are spun. Graphitic fibers formed from discotic phases form another class of strong, light materials.

Lyotropic phases are not without applications either. Everyone is familiar with the usefulness of detergents in everyday life. The correct use of systems—formed from water, surfactants, and oil—may help to recover more of the oil left in the ground after the primary methods of oil recovery have been exhausted.

## C. Biological and Medical Uses

Structures having liquid-crystalline order occur in many biological systems. One of the prime examples is the cell membrane, a representation of which is shown in Fig. 8. The lipid bilayer in the membranes has the same basic structure as the lamellar structure found in lyotropic liquid crystals. More knowledge of the nature of such materials should help in understanding the operation of the cell membrane and is actively being sought.

Liquid crystals are a factor in several diseases. The cells in sickle-cell anemia have a liquid crystal structure. Hardening of the arteries is due to the deposition of liquid crystals made from molecules containing cholesterol. It may be possible to convert the material forming gallstones into liquid crystals, which can then be passed out of the body.

The uses of cholesteric liquid crystals to determine temperature distributions have been mentioned above. The ability of these liquid crystals to convert temperatures into a visual pattern provides a unique diagnostic tool that has been useful in studying abnormalities in venous patterns, detecting primary or metastatic carcinoma in the skin, and locating the placenta of a fetus.

## D. Statistical Mechanics

The diversity of liquid crystalline phases, the range of materials that form those phases, and their accessibility to a wide range of experimental studies make the study of liquid crystals a rich source of information about statistical mechanics. Many of the behaviors have analogies in, for instance, studies of ferromagnetism, superconductivity, or superfluidity. The blue phase that appears in some cholesterics may be one rare example of a thermodynamically stable array of defects. Even the relatively simple smectic A phase is not completely understood. The study of transitions from smectic A to smectic C or nematic phases has yielded much information relevant to modern ideas of statistical mechanics, such as the concept of spontaneously broken symmetries and the

corresponding appearance of new hydrodynamic modes. Studies of the importance of fluctuations for different symmetries, ranges of interactions, and the spatial dimensions of ordering are useful in understanding critical phenomena.

Near a phase transition, many properties of materials should change in ways characteristic of the symmetries of the phases involved as opposed to the specific materials being used. This universality can be tested in many ways with liquid crystals. An example is the shape of phase diagrams for mixtures of a compound that has nematic and smectic A phases in its pure form with another compound that has nematic and smectic A and smectic C phases. On a plot of temperature versus composition, the curves forming the boundaries between nematic, smectic A, and smectic C regions meet at a multicritical point. A second example of this behavior has been found in systems that have a reentrant nematic phase; that is, they display the sequence of phases nematic–smectic A–nematic on heating at constant pressure. Although a phenomenological model for this behavior is successful in predicting certain properties of the phase transitions, the microscopic model to account for such behavior is less certain.

Free-standing films of liquid crystals have been used to investigate theories of two-dimensional phase transitions and have applications to studies of membrane biology and chemical catalysis. Even narrow strands formed by columnar liquid crystals have been studied. These experiments may be of relevance to possible one-dimensional nematic phases.

## SEE ALSO THE FOLLOWING ARTICLES

FERROMAGNETISM • LIQUID CRYSTAL DEVICES • LIQUIDS, STRUCTURE AND DYNAMICS • MACROMOLECULES, STRUCTURE • NUCLEAR MAGNETIC RESONANCE • SURFACTANTS, INDUSTRIAL APPLICATIONS • ULTRASONICS AND ACOUSTICS

## BIBLIOGRAPHY

Chandrasekhar, S. (1993). "Liquid Crystals," 2nd ed., Cambridge Univ. Press, Cambridge, UK.

Chigrinov, V. G., and Blinov, L. M. (1996). "Electrooptic Effects in Liquid Crystal Materials," Springer-Verlag, Berlin.

Collings, P., and Hird, M. (1997). "Introduction to Liquid Crystals: Chemistry and Physics," Taylor & Francis.

Demus, D., ed. (1998)."Handbook of Liquid Crystals," Wiley, New York.

Dong, R. Y. (1997). "Nuclear Magnetic Resonance of Liquid Crystals," 2nd ed., Springer-Verlag, Berlin.

Kumar, S. (2000). "Liquid Crystals: Experimental Study of Physical Properties and Phase Transitions," Cambridge Univ. Press, Cambridge, UK.

Lebedev, V. V., and Kats, E. I. (1994). "Fluctuational Effects in the Dynamics of Liquid Crystals," Springer-Verlag, Berlin.

Mark, H. F., ed. (1987). "Encyclopedia of Polymer Science and Engineering: Liquid Crystalline Polymers to Mining Applications," 2nd ed., Vol. 9, Wiley, New York.

Vij, J. K., ed. (2000). "Advances in Chemical Physics: Advances in Liquid Crystals," Vol. 113, Wiley, New York.

Vill, V. (1995). "LiqCryst: Liquid Crystal Database," Springer-Verlag, Berlin.

Virga, E. G. (1995). "Variational Theories for Liquid Crystals," CRC Press, Boca Raton, FL.



# Luminescence

**J. N. Demas**
**S. E. Demas**
*University of Virginia*

I. Introduction
II. Origins of Luminescences
III. Excited State Types
IV. Methods of Studying and Characterizing Excited States
V. Processes Affecting Luminescence
VI. Types of Luminescence

## GLOSSARY

**Color centers** Absorbing sites in solids caused by lattice defects, trapped electrons or holes, or the formation of new chemical species.
**Excimer** Excited complex that does not exist in the ground state and is formed between one excited and one ground-state molecule of the same type.
**Exciplex** Excimer formed between two molecules of different types.
**Fluorescence** Luminescence characterized by very short lifetimes; typically a spin-allowed process.
**Hole** In solids, an electron-deficient center that frequently can move through the lattice.
**Internal conversion** Relaxation of a system from an upper state to a lower one of the same spin multiplicity.
**Intersystem crossing** Conversion of a system from a state of one spin multiplicity to another.
**Laser** Acronym for light amplification by stimulated emission of radiation—a stimulated emission device that produces intense, highly directional, coherent, monochromatic optical radiation.
**Luminescence** Emission of ultraviolet (UV), visible, or infrared (IR) radiation of excited materials.
**Phosphorescence** Luminescence characterized by a long lifetime; frequently a spin-forbidden process.
**Quenching** Deactivation of an excited state by a nonemissive pathway.
**Stimulated emission** Photon emission from an excited species promoted by the presence of other photons.
**Trap** Lattice defect or chemical center in solids that can trap an electron or a hole.

**LUMINESCENCE** is the emission of ultraviolet (UV), visible, or infrared (IR) radiation from materials and arises from a radiative transition between an excited state and a lower state. The classification of the luminescence depends on how the excited state was derived. Photoluminescence arises following excitation by the absorption of a photon of light. Electroluminescence and cathodoluminescence arise from electric current flow in solids or solutions or in gases during an electrical discharge. Chemiluminescence arises during chemical

reactions, and bioluminescence is chemiluminescence in biological systems. Radioluminescence arises from the passage of ionizing radiation or particles through matter. Thermoluminescence occurs during gentle sample heating.

## I. INTRODUCTION

Since the beginning of recorded history, and undoubtedly much earlier, individuals have been fascinated by luminescence. The cold bioluminescences of glowworms, rotting wood, and sea creatures and the spectacular light shows of the aurora borealis have been particularly intriguing, and a great deal of effort has been made to understand their origins. Until the advent of quantum mechanics, however, the fundamental origins of these emissions could not be satisfactorily explained.

There were numerous ingenious attempts to quantify luminescence phenomena using photographic and manual recording of emission behavior; however, especially for broad molecular emissions, the major breakthroughs in luminescence studies tended to parallel instrumental developments. In particular, the high-sensitivity commercial photomultiplier tube marketed in the 1940s and low-cost spectrofluorimeters of the 1950s can be credited with much of the modern information, theories, and applications of luminescence. More recently, lasers and nanosecond and subnanosecond decay time instruments have revolutionized the types of information that can be extracted.

Any study of luminescence should address the following key questions:

1. What is the molecular and atomic nature of the origin of the luminescence?
2. What are the detailed paths of molecular excitation and deactivation?
3. What are the structures of excited states?
4. Can one rationally design systems with specific and useful properties or exploit existing properties?

This article is concerned primarily with the phenomenological aspects of each type of luminescence rather than the theoretical underpinnings of the subject. The origins and factors affecting luminescence are described. Some experimental methodologies for studying luminescences are examined. Finally, applications of the various luminescences are described. Emissions of very high-energy photons from nuclear or inner-electron-shell transitions or from the nonspecific incandescence of hot solids or plasmas are excluded.

## II. ORIGINS OF LUMINESCENCES

Specifically considered are emissions that arise by radiative transitions between two states of atomic, molecular, or extended molecular systems. A radiative transition is one in which the energy is released as a photon. The nature of the emission depends on the nature of the initial and final states and the route to the excited state.

First, types of excited states are categorized, the factors that influence excited state emission are described, and then the methods of excited state population that define the nature of the emission are discussed.

Figures 1 and 2 show some of the wealth and complexities of atomic and molecular emissions. Figure 1 shows the absorption and emission (relative intensity as a function of wavelength) spectra of anthracene. There are two distinct emissions. The high-energy band at 400 nm is characterized by a short luminescence lifetime of a few nanoseconds, while the lower-energy emission at 700 nm can be characterized by millisecond lifetimes. The overlap between the lowest energy absorption and the high-energy emission is characteristic of this type of system. The regular progression of peaks on both emission systems is also common to many molecular systems.

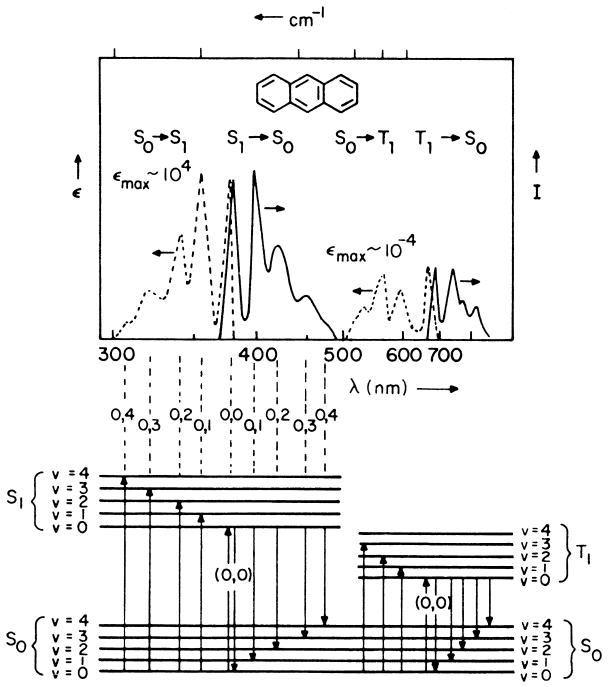



**FIGURE 1** Absorption (dashed lines) and emission (solid lines) of anthracene. The lower portion displays the electronic and vibrational assignments of the absorption and emission bands. [Reprinted with permission from Turro, N. (1978). "Modern Molecular Photochemistry," Benjamin/Cummings, Menlo Park, CA.]

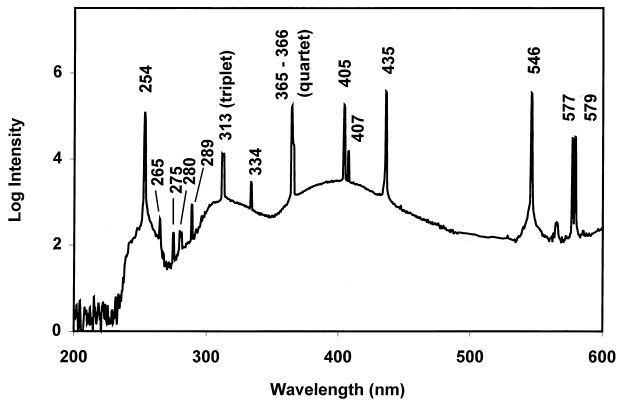



**FIGURE 2** Uncorrected emission spectrum of a low-pressure mercury vapor discharge. Wavelengths in nanometers are adjacent to each line. The intensity is on a log scale and the continuous background is from the plasma discharge. The line widths are instrumentally limited, and several of the lines are unresolved multiplets. [Data kindly supplied by W. R. Bare.]

Figure 2 shows the cathodoluminescence of atomic mercury in a low-pressure discharge. Particularly noteworthy is the exceptional narrowness of the atomic versus the molecular emissions. The characteristics and differences of these emissions are discussed in Section III. Note the logarithmic scale used to display the weaker lines. The broad weak continuum is from plasma discharge in the supporting Ar gas.

## III. EXCITED STATE TYPES

### A. Spin Multiplicity

A simplified excited state diagram is pictured in Fig. 3. Details of the quantum mechanical origins and nature of excited states are not presented here. The system is characterized by a singlet ground state, denoted by $S_o$, and singlet excited states, denoted by $S_i$ (I = 1, 2, . . .). Singlet states arise when all the electrons are spin-paired. Also shown in the excited manifold is the lowest triplet state, denoted by $T_1$. Triplets arise when there are two unpaired spins. This type of system corresponds to the vast majority of organic molecular species and occurs when the lowest energy configuration of the system is due to all of the electrons being spin-paired.

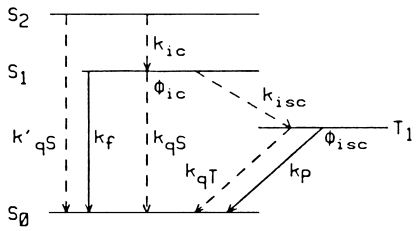



**FIGURE 3** Schematic energy level diagram, or Jablonski diagram, for a molecule showing the possible paths of energy degradation. Solid lines represent radiative emission processes, and dashed lines represent nonradiative processes. Rate constants and efficiencies of the indicated constants are denoted by *k*'s and Φ's. [Reprinted with permission from Demas, J. N. (1983). *J. Chem. Ed.* **60,** 803. Copyright 1983 Division of Chemical Education, American Chemical Society.]

Excited states of such a system generally arise when a paired electron is promoted from a filled to an unoccupied orbital. The electron can remain paired with the electron left behind to form excited singlet states, or it can undergo a "spin flip" and become unpaired; this results in a triplet state. The triplet state derived from a specific orbital promotion is of lower energy than the corresponding singlet state (Figs. 1 and 3).

Oxygen and metal ions are the common stable exceptions to this type of excited state diagram. Atomic species in flames or discharges are also frequent exceptions. Oxygen has a triplet ground state with singlet and triplet excited states. The ordering of excited singlets and triplets are inverted over those of Fig. 1, however, with the singlets being below their corresponding triplets. Metal ions can exhibit a multitude of excited state multiplicities, which can range from doublets (one unpaired electron) for $Cu^{2+}$ and Na, quartets (three unpaired electrons) for $Cr^{3+}$, and octets (seven unpaired electrons) for $Eu^{2+}$.

Regardless of the nature of the ground state, however, the excited states can have spin multiplicities that are the same as, or different from, the ground state. Spin selection rules control whether a transition between states is allowed or forbidden. Transitions between states of the same multiplicity are spin-allowed, while all others are forbidden. Forbiddenness does not mean that a transition will not occur at all, but that it will not occur as readily as an allowed one. Allowed transitions are characterized by strong absorptions, large rate constants, and short lifetimes. Spin-forbidden transitions exhibit weak absorptions, long lifetimes, and low rate constants. Compare the allowed 400-nm absorption with the forbidden 650-nm absorption of anthracene (Fig. 1), where the allowed higher-energy transition is $10^8$ times more intense.

Figure 3 is based on the assumption that spin is always a good quantum number. This assumption is not always correct, especially for species of high atomic number. Spin-orbit coupling can mix orbital and spin-angular momentum, and then the concept of electron spin fails. It is necessary to discuss the states of the system in terms of the good quantum number $J$. Pragmatically, spin-orbit coupling scrambles the singlet and triplet states and gives a large component of the other spin character to the state. Thus, the mixing of singlet character into a triplet state can greatly increase the allowedness of spin-forbidden transitions.

## B. Fluorescence and Phosphorescence

Traditionally there has been a phenomenological characterization of emission type. Short-lived emissions have been considered fluorescences and long-lived emissions, phosphorescences. One would then infer that phosphorescence arises from spin-forbidden processes and fluorescence from spin-allowed processes.

In the case of many discrete molecular systems this categorization is correct. Figure 1 demonstrates the simultaneous presence of fluorescence and phosphorescence. The short-lived emission is the fluorescence and the long-lived emission the phosphorescence.

This simple model can break down when one considers the more complex case of fluorescent lamp and television phosphors. After use, a television screen glows in a dark room for minutes to hours. In keeping with the empirical classification these glows are called phosphorescences. As it turns out, however, these very long lifetimes are generally associated with slow secondary trapping processes that have nothing at all to do with the fundamental luminescence processes. Indeed, the fundamental luminescence step in many phosphorescences is a spin-allowed process.

One also sees numerous incorrect or misleading references in the literature. The emissions of rare-earth elements (e.g., $Tb^{3+}$ and $Eu^{3+}$) and of uranyl are frequently referred to as fluorescences even though their lifetimes are hundreds of microseconds to milliseconds. Furthermore, quantum mechanically these emissions are best described as spin-forbidden processes. Thus, by the criteria of both lifetime and quantum mechanics, these emissions are actually phosphorescences. If there is doubt about the origins of an emission, it is best referred to as a luminescence.

## C. Energy Degradation Pathways (Nonradiative Pathways)

It is impossible to talk about luminescence without considering additional nonradiative processes. The anthracene emission spectrum (Fig. 1) is made up of both a fluorescence and a phosphorescence. These emissions occur with the same efficiencies regardless of whether $S_1$ or an upper singlet state is directly excited. Furthermore, because of the weakness of the $S_0 \rightarrow S_1$ absorption, it is usually extremely difficult to excite the triplet state directly. However, efficient phosphorescences on excitation into the singlet states are common. Finally, there are very few molecules that have emission efficiencies (photons emitted per photon absorbed) of close to 100%. These results imply the existence of both efficient nonradiative deactivation pathways and radiationless interconversions between states of the same, and of different, multiplicities.

Figure 3 shows a simplified representation of these additional processes. Relaxation within a manifold of the same multiplicity is called internal conversion. In condensed media, internal conversion is very fast compared with the rates of radiative emission from upper singlet states and accounts for the rarity of efficient upper-level emission. This rapidity arises because of the closeness of the lower levels, the absence of spin restrictions, and the availability of vibrational levels of the lower states that provide an efficient vibrational cascade mechanism for relaxation. In condensed media the best known example of an upper excited state emission is the $S_2 \rightarrow T_2$ fluorescence of azulenes.

Radiationless deactivation to the ground state from $S_1$ is also a special case of internal conversion. However, since it competes directly with the main emission process, it is given the separate term *quenching*. The decreased rate of quenching from $S_1$ to the ground state is attributable to the much larger energy gap between these two levels compared with the spacing between upper singlets.

Crossing between states of different multiplicities (e.g., singlet to triplets) is also possible even though the process is spin-forbidden. Conversion between states of different multiplicities is called intersystem crossing. Indeed, in some systems with small energy gaps between the singlet and triplet states, and with reduction of the forbiddenness because of spin-orbit coupling, intersystem crossing can be so fast compared with radiative coupling to the ground state that only phosphorescence is observed.

In the triplet manifold as well as in the singlet manifold internal conversion usually causes rapid relaxation to the lowest triplet state before emission occurs. The emitting triplet state is also susceptible to direct quenching to the ground state. Indeed, because of the forbiddenness of phosphorescence, the long-lived triplet state is very susceptible to quenching; room-temperature phosphorescences are relatively rare and generally not very efficient.

In the gas phase, especially at low pressures, where collisions are infrequent, internal conversion is much less rapid due to the absence of solvent or other molecular vibrations to help carry away the excess energy. This reduced efficiency of internal conversion makes upper excited state emissions much more prevalent.

## D. Atomic and Molecular Excited States

The states of discrete atoms are described by first determining the one-electron atomic orbitals, then adding the total number of electrons by filling the lowest energy orbitals with two electrons per orbital. The ground state is derived from this configuration. Excited states are then generally derived by considering the configurations arising

from promotion of an electron from an occupied to an unoccupied orbital.

For example, the ground-state configuration of atomic mercury is $[Xe](4f)^{14}(5d)^{10}(6s)^2$, where [Xe] stands for the closed-shell xenon core. If everything but the $6s$ electrons are denoted as core, the lowest excited states of atomic mercury are given by $(\text{core})(6s)^1(7s)^1$ and $(\text{core})(6s)^1(6p)^1$.

The state diagram for atomic mercury and some of the radiative transitions responsible for emissions of Fig. 2 are shown in Fig. 4. The other transitions can be derived from energy differences between states. The state designations based on the quantum numbers $S$, $L$, and $J$ are shown above each set of states. The superscript denotes the spin multiplicity of the state, $M$, and is related to the spin angular momentum quantum number $S$ by $M = 2S + 1$. The orbital symmetry of the state is determined by the orbital angular momentum quantum number $L$ and is given by the upper case letter. The $J$ quantum number, which arises from coupling of spin and orbital angular momentum and represents the total angular momentum, is the subscript. For example the 253.65-nm emission line arises from a transition from $^3P_1$ to the $^1S_0$ ground state ($S = 1 \rightarrow S = 0$; $L = 1 \rightarrow L = 0$; $J = 1 \rightarrow J = 0$).

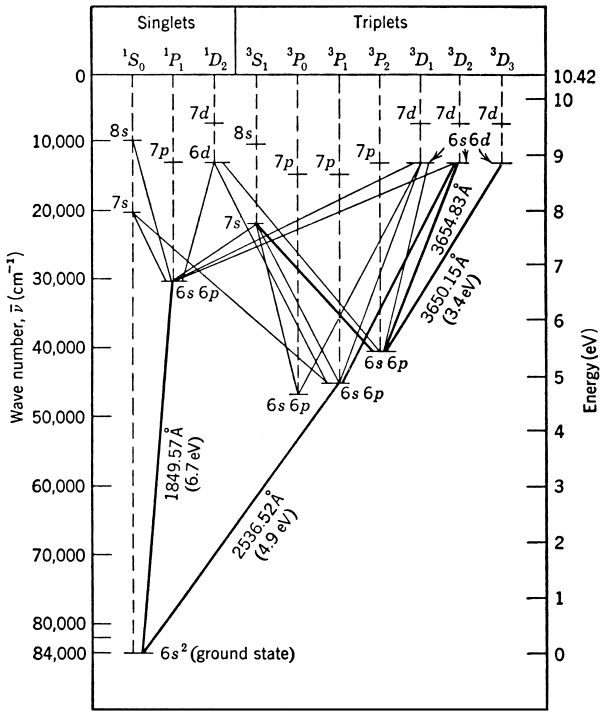



**FIGURE 4** Energy level and state diagram for atomic mercury. The term symbols for the states are indicated across the top. Some of the radiative transitions are indicated by solid lines. The orbital configuration is indicated on each state. For example, $6s\,6p$ denotes a $(6s)^1(7p)^1$ outer-shell configuration, and $7d$ denotes a $(6s)^1(7d)^1$. The core is omitted for clarity. [Reprinted with permission from Leverenz, H. W. (1950). "An Introduction to Luminescence of Solids," John Wiley & Sons, New York.]

Molecular excited states are derived in much the same way, except that the orbitals of the system are described by molecular orbital theory. The single-electron molecular orbitals are made up of combinations of atomic orbitals derived from the different atoms in the molecule. Thus, the molecular orbitals extend over the entire molecule and are not localized on a single atom. This delocalization makes for very rich bonding and spectroscopy. As with the atomic case the electrons are added to fill up the lowest energy orbitals in order to derive the ground-state configuration. Excited states usually arise from orbital promotions of electrons from occupied to unoccupied orbitals.

Excited states of molecular systems are derived from a variety of electron configurations. In organic systems the configurations responsible for the low-energy states generally involve $\pi$-$\pi^*$ and $n$-$\pi^*$ states. The $\pi$-$\pi^*$ states are derived from the promotion of an electron from a $\pi$-bonding to a $\pi$-antibonding orbital (e.g., anthracene). The $n$-$\pi^*$ excited states are derived from the promotion of an electron in a nonbonding orbital to a $\pi^*$ antibonding orbital; an example is ketones, where an electron in one of the nonbonding oxygen orbitals is promoted to the antibonding $\pi$ orbital between the carbon and oxygen atoms.

Metal complexes introduce more new states. The coordinating ligands can contribute low-lying $\pi$-$\pi^*$ or $n$-$\pi^*$ states. Splittings of the degenerate $d$ orbitals by a nonspherical ligand environment can give rise to metal-localized $d$-$d$ transitions in metal complexes with $d$ electrons. In addition, there are charge-transfer transitions derived from the promotion of an electron from a metal-localized orbital to a ligand-localized orbital or from ligand to metal orbitals.

A comparison of Figs. 1 and 2 shows a remarkable difference between the molecular and the atomic emission spectra. The atomic spectrum is incredibly sharp, while the molecular spectrum is very broad and exhibits regular progressions. The atomic states are simple because of the absence of any other vibrational or rotational states. In contrast, large molecules have a large number of vibrational and hindered rotational states superimposed on the simple energy level diagrams of Figs. 3 and 4. Furthermore, the molecule can exist in a large number of conformations in the solvent matrix, each with a characteristic absorption and emission. These factors result in a broadening of the molecular transitions.

A more complete energy level diagram is given in Fig. 1, where a dominant molecular vibration has its energy levels superimposed on each electronic state. This figure shows why the absorption and emission tend to overlap with, and be mirror images of, one another. A

well-defined vibrational progression is characteristic of systems in which there is little distortion on going from the ground to the excited state. Where distortions occur, the vibrational structure is smeared out and the emission band is broadened and red-shifted.

An interesting type of hybrid atomic molecular system is exemplified by rare-earth ions in crystal lattices or in molecular complexes. The electronic configuration of rare-earth ions is (core)$(f)^n (n = 0–14)$. The lowest excited states are derived, not by orbital promotions, but by rearrangement of the electrons within the $f$ shell. Furthermore, these $f$ electrons are so well shielded within the atom that the excited state transitions are very insensitive to the environment around the atom. Thus, the transitions of rare-earth elements look more like atomic transitions than molecular ones. Atomic-state classifications are used because of the small perturbations on the atomic transitions.

Figure 5 shows emission spectra for a neodymium(III)-doped glass at room and liquid-nitrogen temperatures. The quasi-atomic line spectra are very clear, especially at 77 K; compare these spectra with Figs. 1 and 2. Emission narrowing on cooling is common and one of the reasons why emissions are frequently studied at low temperatures. In this case the 77 K emissions are only 0.16 nm wide.

## E. Excimers and Exciplexes

Even if one fully understands ground-state chemistry, one may find surprises in the excited-state manifold, where totally unexpected species suddenly appear. A classic example of this is encountered in electrically excited mixtures of Ar and $F_2$. There are no known stable Ar compounds. However, one sees an intense $^*$ArF emission derived from

$$^*\mathrm{Ar} + \mathrm{F_2} \rightarrow {}^*\mathrm{ArF} + \mathrm{F} \qquad (1)$$

where asterisks denote excited species. The reason for the existence of $^*$ArF but not of ArF is that $^*$Ar is not the same chemical species as Ar; it has a completely different electronic configuration. Ar has a closed-shell $[\mathrm{Ne}](3s)^2(3p)^6$ electronic configuration with no free bonding electrons and so does not form ArF. The lowest excited state of $^*$Ar, however, is $[\mathrm{Ne}](3s)^2(3p)^5(4s)^1$, which has unpaired $s$ and $p$ electrons. Chemically this configuration is very similar to that of potassium metal; a free $s$ electron is bound to a singly charged core. Not surprisingly the bonding in $^*$ArF is ionic and very much like that of KF.

Rare gas chemistry can be even more complex. At high pressures the rare gas halide can react to give triatomic species

$$^*\mathrm{RgX} + \mathrm{Rg} \rightarrow {}^*\mathrm{Rg_2X} \qquad (2)$$

where Rg stands for a rare gas and X for a halogen. Figure 6 shows the emission spectra of several triatomic rare gas compounds.

Excimers and exciplexes are chemically stable excited-state species that can exist only in the excited state and do not have a corresponding ground-state form. Excimers are *ex*cited state d*imers* formed by the association of two identical subunits. Exciplexes are *ex*cited state com*plexes*

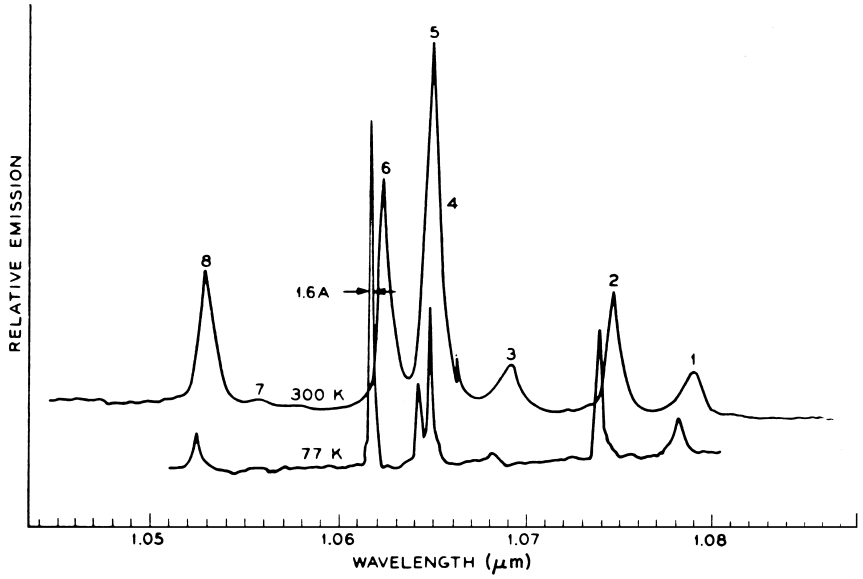



**FIGURE 5** Emission spectra of $Nd^{3+}$ in $Y_3Al_5O_{12}$ at room and liquid-nitrogen temperatures. [Reprinted with permission from Van Uitert, L. G. (1966). *In* "Luminescence of Inorganic Solids" (P. Goldberg, ed.), p. 516, Academic Press, New York.]

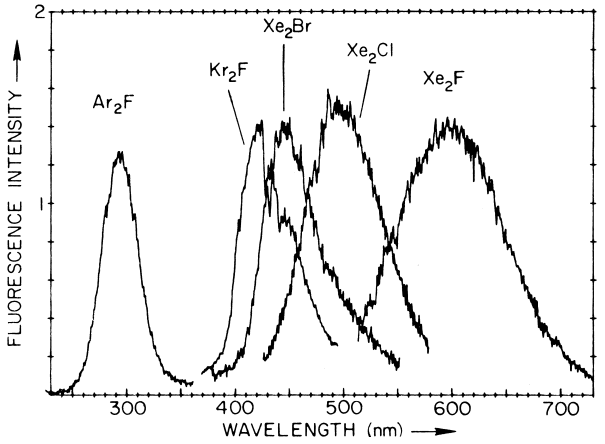



**FIGURE 6** Fluorescence spectra of rare-gas trimers in electron-beam-excited high-pressure rare gas-halogen mixtures. [Reprinted with permission from Huestis, D. L., Marowsky, G., and Tittel, F. K. (1983). *In* "Excimer Lasers—1983," AIP Conference Proceedings No. 100, Subseries on Optical Science and Engineering No. 3 (C. K. Rhodes, H. Egger, and H. Pummer, eds.), p. 240, American Institute of Physics, New York.]

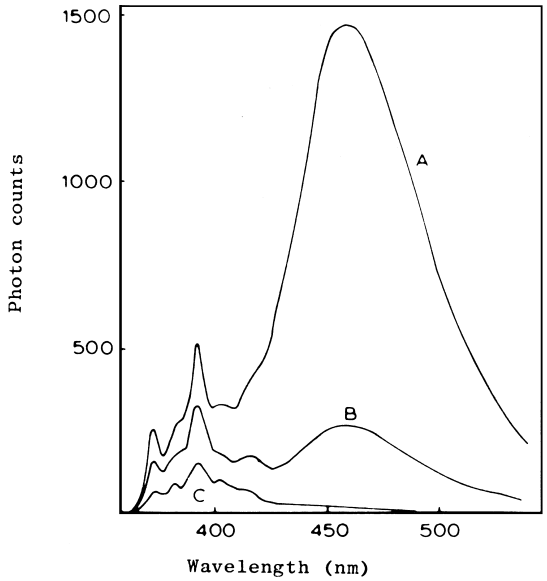



**FIGURE 7** Time-resolved fluorescence spectra of pyrene adsorbed on solid silica. The spectra are for the following delays after a short excitation pulse: (A) 7–52 nsec, (B) 108–162 nsec, and (C) 347–404 nsec. The structureless 460-nm band is the pyrene excimer, and the structured high-energy emission is the monomer. The excimer has a short lifetime, which enhances its emission at short times. [Reprinted with permission from Ware, W. R. (1983). *In* "Time-Resolved Fluorescence Spectroscopy in Biochemistry and Biology" (R. B. Cundall and R. E. Dale, eds.), p. 53, Plenum Press, New York.]

formed of two distinctly different subunits. $^*$ArF is an exciplex. If the two reactants are chemically similar, the complex is a mixed excimer.

The classic excimer, and first to be discovered, is pyrene. Pyrene exhibits no tendency to associate with itself in the ground state. At higher concentrations, however, excited-state pyrene associates strongly with a ground-state pyrene to form the pyrene excimer, which exhibits an intense emission that is shifted to the red of the monomer emission.

Figure 7 shows the exciplex emission of pyrene on silica gel. The high-energy structured emission is the pyrene monomer, while the broad low-energy emission derives from exciplexes formed from closely located adsorbed pyrenes.

An interesting dimeric emission arises in the chemiluminescent reactions of excited-state singlet oxygen. Under the chemical conditions of generation, high concentrations of $^1O_2$ exist. Dimer-like species pool energy to produce higher energy emissions:

$$^1O_2 + {}^1O_2 \rightarrow \left(^1O_2\right)_2 \rightarrow 2O_2 + h\nu \qquad (3)$$

While the emission of $^1O_2$ is in the IR, the "dimol" emission is a spectacular red. Combination bands arising from states derived by simultaneous excitation on both oxygen molecules are observed.

Although not an exciplex, a related area of excited-state behavior is acid-base reactions. Again because of the differences in electronic configurations of the ground and excited states, the ground and excited states can have greatly different p$K$ values. That is, the reaction

$$H^+ + {}^*A^- \leftarrow pK_a^* \rightarrow {}^*HA \qquad (4)$$

has a different p$K$ than the ground-state p$K_a$. This excited state p$K_a$, or p$K_a^*$, frequently differs from the ground state p$K_a$ by 5–10 p$K$ units. Thus, a strong acid may become a very weak acid in the excited state, or a weak acid may become a super acid. Bases can behave similarly. As protonated and unprotonated forms of a species can exhibit very different properties, emissions of species exhibiting excited state acid-base chemistry show remarkable and, to the uninitiated, unexpected variations in spectra with pH. Excited state acid-base reactions are now being employed in luminescence based pH sensors in biomedical and industrial applications.

The excited state reaction of Eq. (4) can involve other species than protons. Metal ion and proteins can react with the the excited-state species (or the ground-state species which is then excited) to produce a new excited state complex with different emission properties (e.g., wavelength, lifetime, quantum yield, or emission polarization). These changes can be used to quantitate the analyte. Many of the modern methods of analysis and probes of biomolecules, polymers, and surfaces are based on such changes.

### F. Band States

In strongly interacting solids, the concept of localized atomic or molecular states fails. Orbitals on adjacent atoms or molecules can interact so strongly that the molecular orbitals of the composite system must be described as extending over the entire lattice rather than being localized on a specific atom or molecule. A consequence of this is that there are no longer discrete states of the system, but a series of very closely spaced levels that make up bands. The lower levels are valence bands, and the upper unoccupied levels are conduction bands. Conduction occurs when the electrons in the valence band are promoted to the conduction band, where they are free to move through the lattice. Depending on the energy gap (forbidden band) between the filled valence band and the conduction band, the solid is an insulator, semiconductor, or conductor. A small gap permits electrons to be thermally excited to the conduction band at room temperature for conductors or semiconductors. In insulators, the gap is too large to yield any appreciable concentration of charge carriers. Luminescence seems to be restricted to semiconductors and insulators.

Excitation of an electron from the valence to the conduction band produces an excited state of the system, which can be treated as any other type of excited state and can give rise to luminescence. Electron promotion leaves behind a positively charged center or hole. Both the electron and the hole can move freely through the solid and are responsible for photoconductivity. This system is shown schematically in Fig. 8.

The electron and the hole can undergo secondary processes that influence emission. Both can be trapped at sites in the lattice. Traps may be defects in the lattice that arise from missing ion sites, interstitial ions, or replacement of normal lattice ions with impurities that may introduce additional lattice defects. Furthermore, other species may function directly as traps if they are easily oxidized or reduced. Figure 8 also shows a schematic representation of electron and hole trapping. An electron is trapped by dropping from the conduction band into a potential energy well. A hole is trapped by pulling an electron from an oxidizable site.

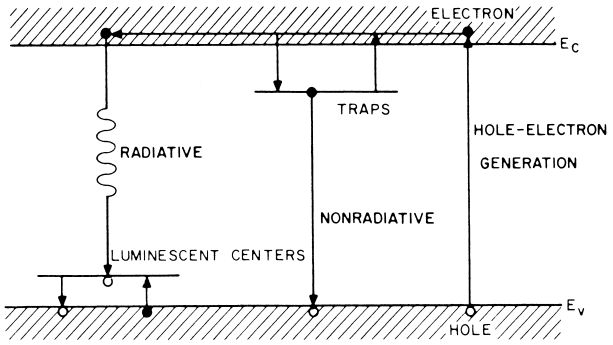



**FIGURE 8** Representation of radiative and nonradiative processes in solids. The lower striped area is the valence band, and the upper is the conduction band. [Reprinted with permission from Sze, S. M. (1981). "Physics of Semiconductor Devices," 2nd ed., Wiley, New York.]

These band states are very sensitive to the size of the semiconductor particle. Emission colors of CdS nanoparticles can vary from the blue to the IR. Such particles are currently being adapted to a variety of analytical systems.

Trapped holes or electrons can be well-defined species with their own spectroscopy, including characteristic absorption and emission spectra. Such systems are called *color centers* because of their characteristic colors. For example, an electron trapped in a halide ion vacancy in an alkaline halide lattice is called an F center. Sodium chloride has a yellow F center, potassium chloride a magenta one, and potassium bromide a blue one. These F centers can undergo reasonably efficient low-temperature emission. Doping of halide matrices with the activator can produce a number of new types of centers involving such species as $Ag^{2+}$, $Ag^{0}$, $Ag_2^{+}$, and $Ag_2^{0}$.

Emission can result from direct recombination of the conduction electron with the hole. More commonly, luminescence in band solids arises from impurities. Electrons in the conduction band can relax back to a hole close to the activator; the energy released excites the activator, which luminesces. The transition to the trap itself may be radiative. Alternatively, if the hole is trapped by oxidizing the activator, the recombination is a reduction of the center; the chemical energy released by this reaction can lead to excitation of the center. This is a form of chemiluminescence.

## IV. METHODS OF STUDYING AND CHARACTERIZING EXCITED STATES

Excited-state processes are usually studied and characterized by the following general approaches: emission and excitation spectra, luminescence efficiencies, polarization, temporal behavior, temperature effects, interactions with other species, double-resonance methods, fluorescence line narrowing, spin echoes, transient gratings, and site-selective spectroscopy. This information is correlated with absorption processes. Several of the most common approaches are discussed.

An emission spectrum is the relative intensity of emission as a function of wavelength. Data are generally acquired by scanning through the emission with a monochromator. The relative intensity is measured with an optical detector such as a photomultiplier tube or semiconductor

detector. These directly obtained data are not corrected for the transmission characteristics of the optics or for the variations in the detector's sensitivity with wavelength. Uncorrected spectra may bear little resemblance to true luminescence spectra and must be corrected, usually by calibrating the response of the system with a source of known spectral distribution such as a standard lamp.

Excitation spectra are obtained by measuring the relative emission intensity at a fixed wavelength while scanning the excitation source. For weakly absorbing solutions, the amount of light absorbed will be directly proportional to the sample absorbance. If the emission efficiency is independent of excitation wavelength, then the excitation spectrum will match the absorption spectrum. As with emission spectra the directly obtained spectra are distorted by the variations in light output of the source versus wavelength. Data are corrected by measuring the excitation source intensity as a function of wavelength.

Excitation-emission matrices (EEM) are two-dimensional plots of excitation and emission spectra. These are invaluable for characterizing complex mixtures. The matrix can provide a unique fingerprint for complex mixtures, and, as such, it is useful in identifying and tracing complex mixtures such as oil spills.

Excited state lifetime measurements are extremely useful diagnostic tools of excited-state processes. The standard method is to excite the sample with a pulse that is shorter in duration than the decay phenomena and then watch the relaxation by monitoring the luminescence. It is also possible to monitor the decay by following the excited state absorption spectrum or the electron spin resonance spectrum. Using mathematical tricks, one can also measure lifetimes appreciably shorter than the excitation. For extremely short decays, picosecond pulse probe techniques are used. Here, a sample is probed using an optical delay line where time between excitation and monitoring is set by adjusting the distance the probe pulse travels before striking the sample. Lifetimes in the low nanosecond range are readily measured using emission relaxation methods, while picosecond methods measure subpicosecond decays.

An alternative approach to lifetime determinations is the phase shift measurement where one excites the sample with a sinusoidal excitation. The emission is sinusoidal and phase-shifted from the excitation. The phase shift is related to the lifetime and the modulation frequency. A variation is to use a very short duration, high repetition source such as a mode-locked laser or a synchrotron. Such a source can be decomposed in the fundamental at the repetition frequency and the higher harmonics. The individual Fourier components of the excitation can be used to simultaneously evaluate the decay times at multiple frequencies.

In principle, multifrequency phase shift and pulse measurements provide similar information. Commercial packages are available for both types of measurements. Both types of measurements appear regularly in the literature, and the choice depends on the nature of the problem (i.e., wavelength range, lifetime, required dynamic range, funds) and the researcher's personal preference. Many workers like to see actual decay curves (the sample impulse response) rather than a set phase shift versus modulation frequency. However, with the advent of inexpensive, bright, easily modulated LEDs for frequencies into the megahertz range and sophisticated high frequency signal processing, the phase-shift measurement will certainly dominate analytical instruments based on lifetime measurements for the forseeable future.

As it turns out, many practical and fundamentally interesting systems are not characterized by a single decay time, but rather by sums of multiple exponentials or even more complex decays. This problem is mirrored in phase-shift measurements where the lifetimes determined at different frequencies differ because the decay is not a single exponential. Such complexity is the rule rather than the exception in biological systems, solid-state composite sensors, or dynamic multicomponent molecules in solution. Fortunately, the mathematical tools and the extraordinary power of inexpensive desktop computers are allowing a successful assault on these important systems.

In terms of the rate constants and paths indicated in Fig. 3, the fluorescence and phosphorescence lifetimes are given by

$$\tau_f = 1/(k_f + k_{qS} + k_{isc}) \tag{5a}$$

$$\tau_p = 1/(k_p + k_{qT}) \tag{5b}$$

where the subscripts f and p denote the fluorescence and phosphorescence processes, respectively, q denotes a quenching path, and S and T denote processes from the singlet and triplet states, respectively; $k_{isc}$ is the rate constant for intersystem crossing between $S_1$ and the triplet manifold.

Luminescence quantum efficiencies (photons emitted per photon absorbed) are given by

$$\Phi f(S_1) = k_f \tau_f \tag{6a}$$

$$\Phi p(S_1) = \Phi_{isc} k_p \tau_p \tag{6b}$$

$$\Phi p(T_1) = k_p \tau_p \tag{6c}$$

$$\Phi_{ic} = k_{ic}/(k_{ic} + k_{qS}) \tag{6d}$$

$$\Phi_{isc} = \Phi_{ic}/(S_1)/\Phi_p(T_1) \tag{6e}$$

where $\Phi_{ic}$ is the efficiency of internal conversion from the upper excited singlet to $S_1$ and $\Phi_{isc}$ is the efficiency of intersystem crossing between $S_1$ and $T_1$. The parenthetical

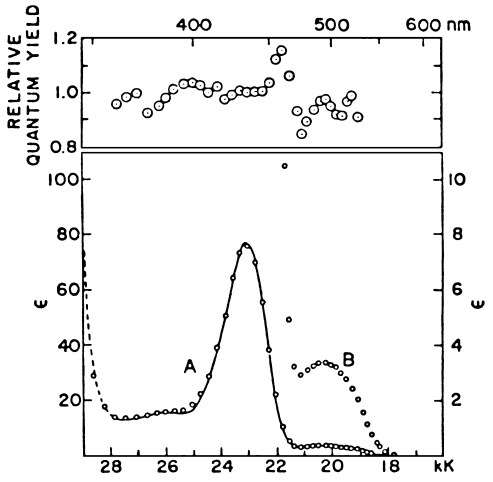



**FIGURE 9** Absorption (solid line) and excitation spectrum (○) of *trans*-$[RhBr_2py_4]^+$(py = pyridine) at 77 K. Curve A is for the scale on the left and curve B for the scale on the right. The excitation spectrum is normalized to the absorption maximum. The relative emission efficiency as a function of excitation energy is in the upper graph. [Reprinted with permission from Demas, J. N., and Crosby, G. A. (1970). *J. Am. Chem. Soc.* **92,** 7626. Copyright 1970 American Chemical Society.]

$S_1$ and $T_1$ denote the state into which the photons are absorbed.

Note that, by measuring the fluorescence efficiency on excitation into the emitting and any upper state, one can determine the internal conversion efficiency. By measuring the phosphorescence efficiency on excitation to $S_1$ and $T_1$, one can determine the intersystem crossing efficiency. Furthermore, from the luminescence yields, the lifetimes of each state, and the intersystem crossing efficiency, one can determine $k_f$, $k_p$, $k_{isc}$, and $k_q$, which largely define the dynamics of the lower excited-state process.

An example of these methods is illustrated in Fig. 9, which shows the corrected excitation and absorption spectra as well as the relative luminescence yield of *trans*-dibromotetra(pyridine)rhodium(III) bromide.

The broad band red emission exhibits a life to 500 $\mu$sec, which clearly indicates a spin-forbidden phosphorescence. There is no fluorescence; therefore, $k_{qS} + k_{isc} \gg k_f$. The relatively intense bands at 25 and 26 kK (1 K = 1 $cm^{-1}$) correspond to an $S_0 \rightarrow {}^1(d\text{-}d)$ transition, where $d$-$d$ indicates an excited state derived within the metal-localized $d$-orbitals. The much weaker 20-kK band is the spin-forbidden $S_0 \rightarrow {}^3(d\text{-}d)$ excited-state transition and is the inverse of the 15 $cm^{-1}$ emission. The intense band starting at 29 $cm^{-1}$ is another metal-localized state. If relaxation from all levels to the emitting level proceeds with 100% efficiency (i.e., $\Phi_{ic} = \Phi_{isc} = 1$), the excitation spectrum should match the absorption spectrum. In this example, the invariance with wavelength of the emission yield on the excitation into different singlet states and the emitting triplet level leaves no doubt as to the unity efficiency relaxation of all upper levels to the emitting level. Radiationless rate constants for deactivation of the emitting level are also easily calculated and can be correlated with theories.

Polarized emission spectra can be obtained in either crystals or randomly oriented samples. The direction and degree of polarization of the emission relative to the polarization of the excitation beam are recorded. From the variations of this polarization as a function of the absorption bands excited, one can frequently infer the molecular axis along which the emission originates or the absorptions arise.

An exceptionally powerful tool for unraveling the dynamics of excited-state geometry changes is time-resolved polarization anisotropy. Basically, one looks at the degree of emission polarization following excitation by a short polarized excitation pulse. If the molecules stay in a fixed orientation during the emission, the polarization will remain constant during the decay. If the molecule rotates during the emission, the degree of polarization will fall as the originally ordered system becomes randomized. From the kinetics of the depolarization one can map out the nature and the rates of such depolarization processes as energy transfer and localized or whole molecule rotations. This method is invaluable for studying the dynamics of motion of large biomolecules.

Rotational anisotropy and steady-state depolarization is also proving a powerful analytical approach. Many of the new methods of fluoroimmunoassay are based on changes in the rotational depolarization time as an analyte binds to fluorescently labeled probes. In addition, much of what we are learning about dynamics in biomolecules including proteins, DNAs, and membranes, involves luminescence measurements, especially dynamic depolarization methods.

Temperature effects on luminescence efficiencies, lifetimes, and spectral distributions are valuable diagnostic tools for finding the energies of excited states and for exploring excited state relaxation processes. For example, Fig. 10 shows the excited-state lifetime and emission yield for $[Ru(bpy)_3]^{2+}$ (bpy = 2,2-bipyridine). The odd temperature-dependence of the emission can be ascribed to the existence of three states, which are all in thermal equilibrium with one another. Each state has a characteristic radiative and radiationless lifetime as well as different emission yields. The variation in lifetime with temperature arises from the variation in Boltzmann population of the three levels. The fitting of these temperature curves permitted determination of the energy spacing of the levels and their lifetimes. State assignments were then based on these results.

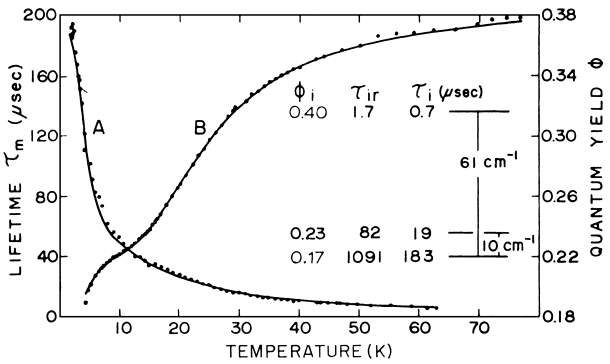



**FIGURE 10** Effects of temperature on the lifetime (A) and luminescence efficiency (B) of $[Ru(bpy)_3]^{2+}$ (2,2-bipyridine) in poly(methyl methacrylate). The solid lines are the calculated curves using the three-level energy diagram shown. For each sublevel, the efficiency as well as the radiative and nonradiative lifetimes used to fit the curves are shown. [Reprinted with permission from Harrigan, R. W., Hager, G. D., and Crosby, G. A. (1973). *Chem. Phys. Lett.* **21,** 487.]

This temperature-dependence was suggested as a cryogenic thermometer. Especially in the sub-liquid hydrogen (20 K) region, temperature measurements are difficult. A lifetime-based imaging system would allow a continuous spatial readout of the temperature over a complex object. While we are unaware of measurements at such low temperatures, an area where spatial resolution of temperature is important is in wind tunnels where temperatures may vary from ambient to 100 K. Luminescence intensities of metal complexes in conjunction with pressure-sensitive luminescence paints (PSPs) are now routinely used in temperature sensitive paints (TSPs) on models in wind tunnels.

Another interesting temperature-related effect is E-type delayed fluorescence. This is a long-lived emission with a spectrum that is indistinguishable from the prompt fluorescence. Delayed fluorescence arises by thermal backpopulation of the emitting state. Thus, the triplet functions as a storage state, and fluorescence tracks the triplet concentration.

One of the major advances in luminescence has been the advent of fluorescence microscopy (FM). It is now possible to routinely carry out luminescence measurements of objects with submicron resolution. This has revolutionized our understanding of intra- and intercellular processes, membranes, polymers, and surfaces. For example, dyes have been designed that change their luminescence on binding to specific metal ions such as calcium. One can incorporate these dyes into living cells and watch the migration and fluctuations of ion concentrations as the cells go through different processes. Fluorescent-tagged monoclonal antibodies are used in clinical diagnosis. Structural changes and the organization of membranes and monolayers can be followed by staining various domains with dyes. Dynamics on surfaces are easily followed by use of luminescent tracers. The proximity of two proteins can be judged by monitoring energy transfer between a donor and an acceptor of the different species.

Conventional FM is sufficiently sensitive in that it easy to see visually *individual* dye-stained DNA molecules. The DNAs are stained with intercalating dyes; each DNA duplex is stained with many dye molecules (about one dye per four base pairs). The large number of fluorophores per strand allows easy visualization. Further, since the DNAs are longer than the optical resolution, you can actually see the long strands. Contrary to the common view of DNA as being rigid static molecules, they actually look more like very active worms squirming around coiling and uncoiling.

Conventional FM simply replaces the traditional transmission-illuminating source with a powerful narrow band excitation source. The sample is viewed through filters that block the excitation and monitor only the sample emission. Traditionally, images were just viewed or photographed. However, low cost and/or ultrasensitive charge coupled device (CCD) image detectors have largely replaced film. The primary distortion of conventional microscopy is that the detector sees contributions from the sample lying above and below the focal plane, which smears the image.

Confocal microscopy minimizes the contribution from out of focal plane images. It uses a raster scan technique to generate the full image. The sample is excited by a tightly focused laser beam. By imaging this small volume through a spatial filter, contributions of the out-of-focus image are reduced or virtually eliminated. This ability to examine a single focal plane allows confocal microscopy to section an image into layers by varying the depth of the image plane. Then, using 3D imaging software, the scientist can slice, dice, and rotate the image in three dimensions, which allows previously impossible examination of details.

The resolution of conventional and confocal microscopy is determined by the diffraction limit of the light or about $\lambda/2$. However, resolution can be further improved by using two-photon excitation with a tightly focused beam. Materials with good two-photon cross sections can be excited at half their absorption wavelength. Since the emission intensity falls off as the square of the excitation flux, and the laser beam intensity falls off rapidly away from the maximum, the effective resolution is closer to $\lambda/3$. Figure 11 shows a beautiful example of a collection of single Rhodamine B dye molecules on a surface. Each spike represents a single fluorescent molecule. The widths of the peaks is determined by the optical resolution (about 250 nm), but the molecules are far enough apart to give clear peaks. In addition, background fluorescence from

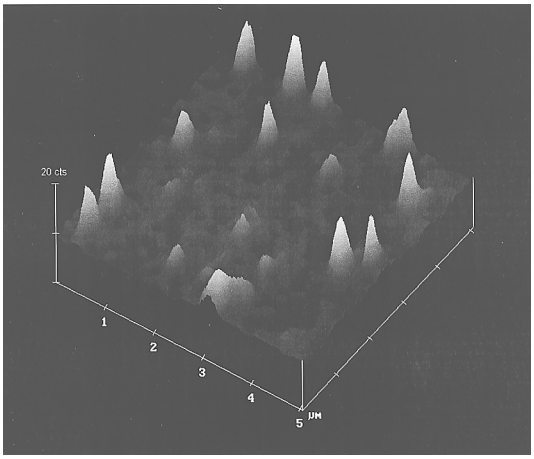



**FIGURE 11** Fluorescence image of single immobilized Rhodamine B dye molecules (5 $\mu$m $\times$ 5 $\mu$m field) dispersed on a glass substrate taken with two-photon excitation. Each peak was due to a single molecule with the fwhm being 250 nm ($<^{\lambda}/_{3}$). The femtosecond pulse train at the sample had an average power of 400 $\mu$W, center wavelength of 840, and pulse width of 180 fs. [Adapted from Sanchez, E. J., Novotny, L., Holtom, G. R., and Sunney, X. (1997). Xie, *J. Phys. Chem. A* **101,** 7020 (1997). Used with permission of the copyright holder the American Chemical Society, 1997.]

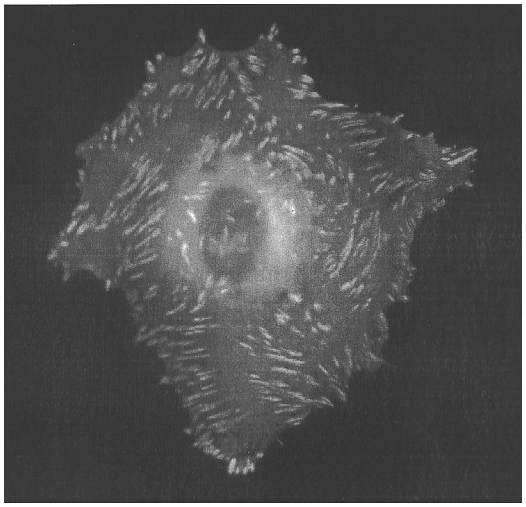


**FIGURE 12** Fluorescent image of green fluorescent protein (GFP) mutant form of Cyan fluorescent protein (CFP) labeled the Paxillin protein, which is one of the focal adhesion proteins. Cyan fluorescent protein targets Paxillin protein. The small lines all over the cell represent Paxillin protein with the nucleus at the center. The proteins are in the cytoplasmic area of the cell. The image is 128 microns $\times$ 102 microns and was taken with an Olympus IX-70 microscope with a 60$\times$ water immersion lens and a Hamamatsu Orca-2 CCD camera. [Image courtesy of A. Periasamy of the Kleck Cellular Imaging Center at the University of Virginia.]

impurities and Raleigh and Raman scattering are much less efficient at the longer wavelengths, and detection limits can actually improve in spite of the small absorption cross section of two photon processes.

Further expanding the capabilities of fluorescence microscopy is the addition of lifetime-imaging capabilities. This area is called fluorescence lifetime imaging (FLIM). One can get two- and three-dimensional imaging of species based not only on their absorption and emission characteristics, but also on their lifetimes. This allows better discrimination in complex mixtures.

Major advances in imaging are in the design of selective and sensitive probes for different analytes. Genetic engineering has proved invaluable in cloning fluorophores or reactive sites for attaching fluorophores in specific target molecules. One common tool is green fluorescent jellyfish protein (GFP) and related proteins. Green flourescent protein has been cloned into mice, which then glow a beautiful green when exposed to a black light. Figure 12 shows an example of a single cell by FM.

## V. PROCESSES AFFECTING LUMINESCENCE

### A. Intramolecular Processes

Intramolecular excited-state processes and methods of studying them have been discussed in Sections III and IV. A few additional aspects need be briefly mentioned. Intramolecular processes are very sensitive to the environment and to excited state types and ordering. For example, $n$-$\pi^*$ excited states exhibit very efficient intersystem crossing, while $\pi$-$\pi^*$ states do not. Thus, ketones such as benzophenone, the lowest excited states of which are $n$-$\pi^*$, exhibit almost pure phosphorescences, while biphenyl, the lowest excited state of which are $\pi$-$\pi^*$, exhibits predominantly fluorescence because of less efficient intersystem crossing. Similarly, molecules that phosphoresce strongly because the lowest excited state is $n$-$\pi^*$ can be converted to strongly fluorescent materials by raising the $n$-$\pi^*$ state above the fluorescent $\pi$-$\pi^*$ state. This can frequently be done by using a hydrogen-bonding solvent, which ties up the nonbonding electrons and makes their removal energetically more difficult.

Rare-earth chelate complexes with organic ligands exhibit a form of intramolecular energy transfer. Rare-earth compounds generally absorb weakly, but strongly absorbing compounds can be made by coupling a strongly absorbing ligand to the metal ion. If the energy levels on the ligand are above the emitting rare-earth level, then efficient energy transfer from the ligand-localized excited state to the metal state can occur. Indeed, in many of these systems, only metal emissions are observed, even though all of the excitation is into the organic ligand.

Another example is radiationless coupling of energy from a species to adjacent or bound solvent molecules.

$H_2O$ is a much more efficient deactivator of many molecular systems than $D_2O$, because of the higher vibrational frequencies that more easily bridge the gap between the ground and excited states. For example, crystalline $Eu^{3-}$ and $Tb^{3+}$ hydrates are more than 10 times less luminescent than the corresponding deuterates. This behavior has been used for counting the number of water molecules around metal ions in proteins or the average solvent exposure of metal complex sensitizers in organized media such as micelles.

## B. Bimolecular Processes

Excited states that persist for any appreciable time are also susceptible to quenching by other chemical species. Excited-state deactivators are called quenchers. Quenching can occur by a variety of processes

$$^*D + Q - k_2 \rightarrow D + Q + \Delta \quad (7a)$$

$$D + {}^*Q \quad (7b)$$

$$D^+ + Q^- \quad (7c)$$

$$D^- + Q^+ \quad (7d)$$

$$\text{reaction products} \quad (7e)$$

where $k_2$ is the bimolecular rate constant for deactivation of the excited state. If more than one deactivation pathway is present, $k_2$ is the sum of all processes affecting the state.

Equation (7a) denotes catalytic deactivation of the excited state. For example, external heavy atoms can increase spin-orbit coupling in a luminescent molecule during a collision and enhance intersystem crossing from a luminescent singlet to a nonemissive triplet or quench a triplet state back to the ground state.

Equation (7b) indicates energy transfer from a donor to an acceptor molecule. Efficient energy transfer quenching must generally be an exothermic reaction. The photodynamic effect is an example of bimolecular energy transfer. A dye in the presence of both light and oxygen will kill organisms. Chemically very reactive singlet oxygen is formed by energy transfer deactivation of dye triplet states. The chemically reactive singlet oxygen then kills the organism.

Electron transfer quenching denoted by Eqs. (7c) and (7d) can be either oxidative or reductive depending on whether the excited species is oxidized or reduced. Electron transfer forms the basis of a number of solar energy conversion schemes and can occur in the excited state even though the ground-state species are thermodynamically stable. This again points out the great differences in chemistry that can arise between ground- and excited-state reactions, and stresses the chemical uniqueness of the excited state relative to the ground state.

Finally, Eq. (7e) indicates any other type of chemical reaction. Examples include the addition of singlet oxygen to an olefin or the hydrogen abstraction of a proton from a protic solvent by a triplet ketone.

The kinetics of Eqs. (7) yield two Stern-Volmer equations relating emission intensity or lifetime to the concentration of the quencher,

$$(\Phi_0/\Phi) - 1 = K_{SV}[Q] \quad (8a)$$

$$(\tau_0/\tau) - 1 = K_{SV}[Q] \quad (8b)$$

$$K_{SV} = k_2\tau_2 \quad (8c)$$

where the $\Phi$'s are emission intensities and $\tau$'s are excited state lifetimes. The subscript 0 denotes the value in the absence of quencher. The $K_{SV}$'s and $k_2$'s are Stern-Volmer and bimolecular quenching constants, respectively. Thus, plots of the experimentally determined left-hand side versus [Q] provide $K_{SV}$'s and $k_2$'s if the unquenched lifetime is known. The $k_2$'s, in particular, provide a great deal of fundamental information concerning the interaction of excited states with quenchers.

Figure 13A shows a typical intensity Stern-Volmer plots for luminescence quenching by molecular oxygen of a metal complex in a polymer support. In solution, the plot would be a straight line, but in the polymer different molecules are in different environments, which leads to the characteristic downward curvature of the plot. Even though dissolved in a solid elastomer, the emission is strongly quenched by oxygen (25-fold at 1 atm of pure oxygen and 8-fold at 1 atm of pure air). This sensitivity to oxygen has led to the development of these systems as

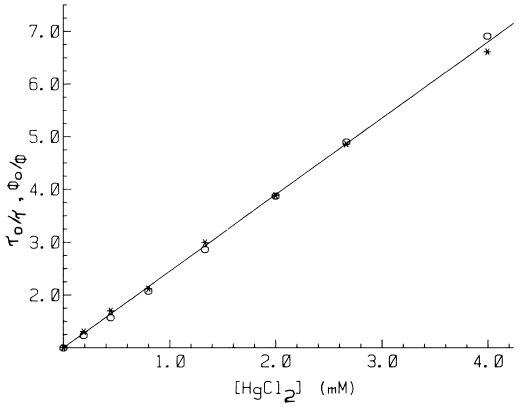



**FIGURE 13** (A) Intensity Stern-Volmer quenching plot for $[Ru(4,7\text{-}Ph_2phen)_3]^{2+}$ (4,7-$Ph_2$phen = 4,7-diphenyl-1,10-phenanthroline) in silicone rubber [Carraway, E. R., Demas, J. N., DeGraff, B. A., and Bacon, J. R. (1991). *Anal. Chem.* **63,** 337.] The solid line is the best fit for a two-site model. (B) Luminescence intensity of $Ru(4,7\text{-}Ph_2phen)_3^{2+}$ in silicone rubber while being breathed over [Bacon, J. R., and Demas, J. N. (1987). *Anal. Chem.* **59,** 2780. Copyright 1991 and 1987 American Chemical Society.]

oxygen sensors in biomedical and industrial applications where they are rapidly replacing the traditional Clark electrode sensors. The term quenchometric sensors is used to describe systems that detect an analyte by quenching of an excited state.

Figure 13B demonstrates the high sensitivity of the system to changes in oxygen concentration. The photoluminescence is being monitored while the film is being breathed on. At one point the breath is held and then breathing is resumed. The burst in emission intensity is caused by the increased exchange time in the lungs and, thus, the lower oxygen concentration and higher emission intensity. The irregularities of the following few breaths are a result of the higher blood carbon dioxide concentration.

Quenchometric oxygen sensors also function as pressure sensors if the gas composition is always the same. This led to the development of PSPs for monitoring pressure over entire aircraft or automobile models. These PSPs are largely supplanting pressure tapes, which are expensive and time-consuming to put on models.

If two partners in a collision are excited, new excited-state reactions are observed. For examples, if two excited triplets collide, one can have an energy-pooling, spin-conserving reaction that yields an excited singlet state. Also, two excited singlets can annihilate one another to form ground state species:

$$T + T \rightarrow {}^{*}S + S \tag{9a}$$

$${}^{*}S + {}^{*}S \rightarrow S + S \tag{9b}$$

Triplet-triplet annihilation is a common form of triplet decay under intense excitation conditions where high triplet concentrations exist. Because of the long lifetimes of the triplet state, triplet-triplet annihilation can also yield a $P$-type delayed fluorescence, which can persist for many microseconds after the cessation of irradiation. Singlet-singlet annihilation is less common because of the difficulty of achieving sufficiently high concentrations of the short-lived species. However, with the advent of modern high-flux lasers it is readily observed and can be a significant nonradiative pathway.

Energy transfer does not require that the molecules be in contact with one another. Resonance energy transfer can occur at distances far exceeding the physical contact distance if the emission spectrum of the donor overlaps the absorption of the acceptor and the acceptor absorption spectrum is intense or highly allowed. Dipole-dipole resonance energy transfer, which is also known as Förster transfer, can occur at distances approaching 100 nm in favorable cases, and 30- to 50-nm transfers are common. The energy collection in the photosynthetic unit generally consists of hundreds of antenna chlorophylls that collect the energy and then transfer it by a resonance mechanism to the active chlorophyll.

In crystals composed of a single component, long-range energy transfer can occur by a contact mechanism. The close proximity of the molecules permits energy to hop from one molecule to its neighbor. By a series of hops the excitation can sample a very large volume. If there are any quenchers in this volume, the energy can be transferred to the quencher, and no luminescence of the major component is observed. The classic example of this is the phosphorescence of naphthalene, where it was necessary to reduce the concentration of $\beta$-methylnaphthalene to below $10^{-7}$ mole fraction to see the host emission.

### C. Stimulated Emission

Light absorption occurs if the photon energy matches the energy gap between the ground excited states. The photon, in effect, induces a transition between the ground and excited states with the loss of a photon from the radiation field and the production of an excited state. However, the inverse process is also possible. If the excited molecule is exposed to photons the energies of which exactly match the energy gap between the excited state and a lower energy state, the photon induces a transition between the upper and lower states. The net result is the addition of a photon to the radiation field rather than a loss. This process is called stimulated emission. Not only is the photon of the same energy as the stimulating photon, but it is also of the same phasing and is traveling in the same direction as the original photon.

Stimulated emission is not generally observed because, for it to occur efficiently, a significant fraction of the molecules must be in the excited state. Furthermore, the population of the upper state must have a higher population than the terminating state (population inversion). Because of the symmetry of the absorption and the stimulated emission process, absorption will always be more efficient than stimulated emission if the terminal state concentration exceeds that of the upper state.

If stimulated emission occurs in a resonant cavity with a high degree of optical feedback, very intense highly directional monochromatic radiation results. Such systems are called lasers (light amplification by stimulated emission of radiation) and have countless practical and fundamental applications including surveying, weaponry, excited lifetime determinations, and luminescence studies.

### D. Multiphoton Processes

At the low fluxes obtained with most conventional light sources, the only absorption processes generally noted involve single photons. With the high fluxes available from

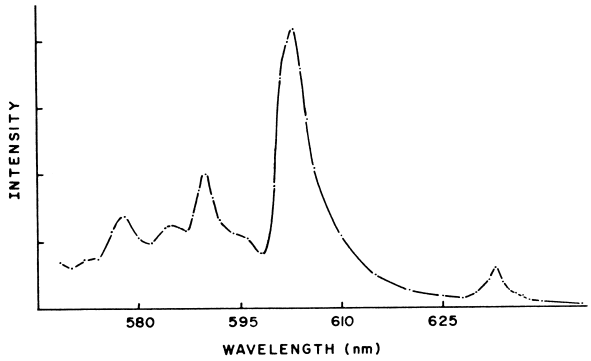



**FIGURE 14** Two-photon excitation spectrum for naphthalene. The absorption is into states that occur at half the indicated wavelength. Emission is monitored in the UV. [Reprinted with permission from Wirth, M. J., and Lytle, F. E. (1978). "New Applications of Lasers," ACS Symposium Series 85 (G. M. Hieftje, ed.) p. 24. Copyright 1978 American Chemical Society.]

lasers, however, nonlinear multiphoton processes occur. At high fluxes the absorption criterion is satisfied if the excited state transition energy equals the sum of the energy of two photons. Thus, absorption of two photons produces an excited state that is inaccessible by a single photon absorption. Indeed, efficient two-photon excitations can arise with light having energies less than the lowest excited state of the system.

Generally, because multiphoton processes have much lower cross sections than single-photon processes it is not possible to monitor the depletion of the beam as in an absorption experiment. However, luminescence from the excited species formed provides a powerful and extremely sensitive tool for monitoring the generation of excited states.

Figure 14 shows a two-photon excitation spectrum of naphthalene using visible excitation. The emission is monitored in the UV. The molecular absorptions are transitions occurring at half the indicated wavelength (twice the energy).

In addition to having different energy requirements from single-photon processes, the selection rules for two-photon absorptions also differ. Thus, multiphoton absorptions provide a valuable tool for locating and studying states otherwise invisible by conventional one-photon spectroscopy.

Multiphoton absorptions are not limited to two photons of the same wavelength. If a sample is subjected to intense irradiation by two beams of different colors, absorption can also occur if the sum of the energies of the two different photons matches the energy of a transition. Even more complex schemes can arise if three or more photons of different wavelengths are required to induce transitions between states to arrive at the final monitored state.

### E. Photon-Stimulated Emission and Quenching

Quite unrelated to simultaneous two-photon excited luminescence is a form of sequential multiphoton excited luminescence or quenching. This type of behavior is especially prevalent in solid phosphors. Initially, a high-energy photon or charged particle excites the system to a metastable state. This excited system may be a true excited state, but more often it is an electron trapped in a high-energy site. Optical excitation with different colored light can lead to a stimulated emission if the electron is ejected from the trap by light of an energy matching the gap between the trap and the conduction band (Fig. 8). Once in the conduction band the electron is free to recombine with a suitable center to produce luminescence. Generally, because of the small gaps between the traps and the conduction band, emission of this type is IR-stimulated.

Similarly, photons of suitable energy can induce transitions of the excited electron to lower nonemissive excited states. This results in quenching of subsequent emission processes. Because of the larger energy gaps to quenching levels, photon-stimulated quenching usually requires higher energy photons than does photon-stimulated emission. Indeed, it is not uncommon for the source that excites the phosphor also to be capable of quenching the luminescence.

Infrared-stimulated emission has potential application in IR detectors. The most sensitive optical detectors are photomultipliers, which do not respond to farther IR photons. However, if a suitable phosphor is pumped up and then exposed to the IR source, stimulated visible photons can be detected by a photomultiplier. A related application is in light-emitting diodes, where an IR-emitting diode is used to pump up a visible-emitting phosphor by sequential photon absorptions.

## VI. TYPES OF LUMINESCENCE

Luminescence is categorized by the method of initiation. Although each type of luminescence is described and one or more applications given, the subject is so broad that an exhaustive discussion is impossible.

### A. Photoluminescence

Photoluminescence arises when the emitting excited state is generated by the absorption of a photon. Excitation can be either a single or multiphoton process.

The applications of photoluminescence are legion. They can be as mundane as blacklight posters or as esoteric as multiphoton processes for sophisticated state analyses of molecules or ions. A particularly powerful application of

photoluminescence is in quantitative analysis. The excitation and emission spectra of materials varies greatly. For example, the characteristic excitation and emission spectra provides "fingerprints" of different crude oils and have been used to trace the sources of oil spills or illegal dumping. Furthermore, many nonluminescent materials can be made luminescent by suitable, and frequently highly selective, chemical reactions. Concentrations can then be determined from photoluminescence intensities.

A major advantage of luminescence over absorption methods of analysis is the frequently much greater sensitivity of the emission methods. This sensitivity enhancement relies in part on the nature of the measurement. Absorption measurements of weakly absorbing samples depend on being able to measure small differences in the intensities of large transmitted signals, which is an intrinsically difficult problem. On the other hand, in emission measurements one is looking for small signals on essentially zero backgrounds.

Emission measurements are so sensitive that it is possible to detect single atoms in the gas phase. Indeed, it is possible to hold, study, and cool to cryogenic temperatures a single atom by using optical tweezers. Due to the background, detection in solution has proved much more difficult.. The supporting solvent has strong scattering (Raleigh and Raman) and it is extremely difficult to get impurity emissions down to low enough levels. The solution is to reduce the sampled volume. Scattering and impurity emission scales with the volume, but the emission of a single molecule is independent of the volume. So by reducing the sampled volume to pico- or femtoliters, single molecule detection becomes possible. Single molecule detection by luminescence is now routine in solution and on surfaces.

In solution, two methods are used. The analyte may be constrained to a hydrodynamically stabilized flowing stream of a few microns. A tightly focused laser beam intersects the stream forming the analysis region of a few picoliters or less. As the molecules flow through a focused laser beam, they are excited and fluoresce multiple times. This burst of photons is collected with a high numerical aperture microscope objective, focused through a pin-hole spatial filter to remove any unwanted background regions and then focused onto a single photon-counting photomultiplier tube or avalanche photodiode. In the alternative implementation, a confocal microscope is used to image an exceptionally small solution volume (subfemtoliter). Then, as molecules diffuse in and out of the sensing volume, bursts of photons occur. One can use this strategy for determining the diffusion coefficient of the fluorophore since time spent in the vicinity of the sensing element is determined by the diffusion. Kinetics of reactions such as ligand receptor binding on proteins can be studied similarly since there is a big difference in the diffusion coefficients of the bound and unbound ligand.

Surface detection of single molecules frequently used near-field scanning optical microscopy (NSOM) to reduce the detection volume or area. The resolution of normal microscopy is limited to the diffraction limit, which is about half the wavelength of the light. However, by making the source of collection optically smaller, it is possible to excite or collect from a far smaller region (i.e., a few nanometers). The source/collector is a fiberoptic tip drawn down to sub-wavelength dimensions. If the tip is very close to the detection volume, the light does not have distance enough to broaden by diffraction. Imaging is done by raster scanning the tip over the surface much as a scanning tunneling or atomic force microscope does. Not only can single molecules be readily detected, but motion pictures of surfaces can be measured to follow such effects as surface diffusion.

The selectivity of luminescence methods is best exemplified by fluoroimmunoassay (FIA), by which specific antigens can be detected even in a messy medium such as a serum. Radioimmunoassay (RIA) has dominated the field when the ultimate in sensitivity is required (pico- to femtomoles). The problems of working with and disposing of radioactive tracers have given impetus to the use of the somewhat less sensitive FIA. Furthermore, advances in flow cytometry and laser-based detection schemes promise to make FIA fully competitive with RIA.

## B. Chemiluminescence

In chemiluminescence (CL) the energy necessary for excited-state generation is derived from the energy released in a chemical reaction. Excluding flames, probably the first man-made CL was the air oxidation of phosphorus, discovered by Brand in 1669. This discovery is the subject of a classic and beautifully detailed engraving (Fig. 15).

Three basic processes can initiate CL: (1) decomposition of a high-energy species to lower-energy ones; (2) exothermic reaction of two or more components; and (3) electron transfer reactions. These processes are represented by

$$A \rightarrow {}^{*}B \tag{10a}$$

or

$$ {}^{*}B + C$$

$$A + B \rightarrow {}^{*}C + D \tag{10b}$$

or

$$ {}^{*}C$$

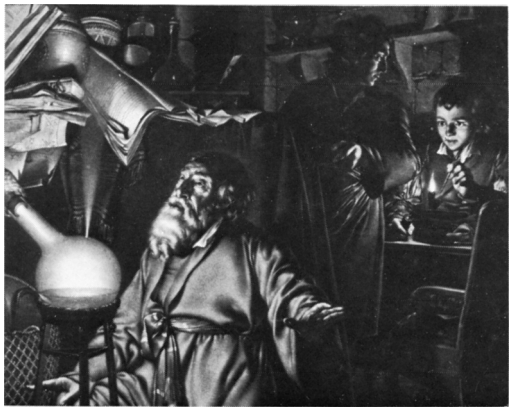


**FIGURE 15** Chemiluminescence on air oxidation of phosphorus. A portion of *The Discovery of Phosphorus* by the alchemist Brand (1669), engraving by William Pether (1775) after the painting by Joseph Wright, Fisher Collection, Pittsburgh. [Reproduction courtesy of the Fisher Collection, Fisher Scientific Co.]

$$A^+ + B^- \rightarrow {}^*A + B \qquad (10c)$$

or

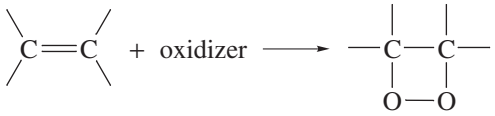
$$A + {}^*B$$

Atom transfer reactions are included in Eq. (10b). While this classification is convenient, it is frequently simplistic. Ternary reactions are possible, and electron transfer steps can also be involved in the first two types.

In all cases the energy released by the reaction must be adequate to generate the excited state. Although a small energy shortfall can be made up by drawing energy from the thermal pool, efficient CLs usually have large excess driving energies. Furthermore there must be an efficient pathway for the ground-state reactants to pass over to the surface corresponding to excited-state products.

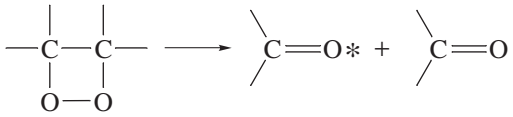
It is not necessary for one of the chemical products to be the luminescent species or even to have an available excited state. If the energy is released in a system in which a suitable acceptor is readily available, the energy can be transferred to the acceptor, which then luminesces.

Examples of Eq. (10a) CLs are the reactions of 1,2-dioxetenes. Dioxetenes can give rise to efficient CLs and are implicated in a variety of bioluminescences. Dioxetenes can be formed biochemically by oxidations of olefins or aromatic molecules:




(11)

The CL arises on cleavage to ketones, with one being excited and the other being in the ground state:




(12)

The excited ketone can then luminesce or transfer energy to another luminescent species.

The long-lived afterglow in electrical discharges in nitrogen is an example of recombination CL [Eq. (10b)],

$$N + N + M \rightarrow {}^*N_2 + M \qquad (13)$$

where M is a third body required to carry away excess energy to prevent the hot product from promptly redissociating. This reaction can give rise to pink, yellow, and blue afterglows. In spite of enormous effort, this system is far from being fully understood.

Flames are rich sources of CL. For example, the blue glow of hydrocarbon flames arises from $C_2$ as well as CN and OH radicals. At least some of these emissions are CLs.

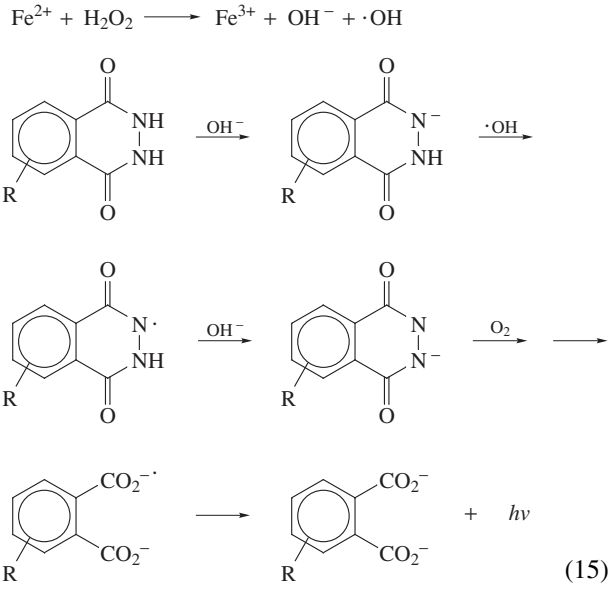
Even the luminescence of many trace elements in flames can be attributed in part to chemical excitation of the metal by such processes as

$$H + OH + M \rightarrow {}^*M + H_2O \qquad (14a)$$

$$H + H + M \rightarrow {}^*M + H_2 \qquad (14b)$$

where M represents a metal ion.

Hydrazide CL, typified by the oxidation of Luminol, is given by the following reaction sequence,




(15)

where R═$NH_2$ for Luminol. That the emissive species is the phthalate dianion is shown by the agreement between the CL and photoluminescence spectra of the product.

Thus, although paths to the excited state are quite different, the final excited state is the same.

Only a few milligrams of Luminol produce an impressive light show, but the CL efficiency (photons emitted per number of molecules reacted) is, in fact, only about 1%. This result emphasizes the impressive quantity of light that a mole of photons represents.

The third class of CLs is initiated by electron transfer processes. This type is exemplified by the one-electron transfer reduction of $[Ru(bpy)_3]^{3+}$ (bpy = 2,2′-bipyridine), which yields a highly visible orange CL with hydrazine, water, and a variety of other reductants. Indeed, with $NaBH_4$ as the reductant, the spectacular emission is highly visible in a well-lighted room.

The complexity of simple CL systems is demonstrated by the simple $[Ru(bpy)_3]^{3+}$/oxalate system:

$$[Ru(bpy)_3]^{3+} + C_2O_4^{2-} \rightarrow [Ru(bpy)_3]^{2+} + C_2O_4^{-} \quad (16a)$$

$$[Ru(bpy)_3]^{3+} + C_2O_4^{-} \rightarrow {}^*[Ru(bpy)_3]^{2+} + 2CO_2 \quad (16b)$$

The initial reduction step does not appear to be energetic enough to excite the complex, and a secondary reaction with the energetic $C_2O_4^-$ free radical is required for excitation. This type of behavior, in which the initial step is not sufficiently energetic to initiate CL, is fairly common, and thus reactive radicals are common in the actual CL step.

A special case of electron transfer CL is that in which the reductant is the ultimate reductant, an electron. $[Ru(bpy)_3]^{3+}$ exhibits an intense CL on reaction with hydrated electrons:

$$[Ru(bpy)_3]^{3+} + e^- \rightarrow [Ru(bpy)_3]^{2+} \quad (17)$$

The efficiency of excited state production approaches 100%.

Applications of CL are diverse. There are commercial emergency lights that require only the breaking of a seal to permit the mixing of solutions. The long-lived CL is bright enough to read by.

One of the simplest and most sensitive NO analyzers utilizes the very efficient CL of the reaction

$$NO + O_3 \rightarrow {}^*NO_2 + O_2 \quad (18)$$

With very simple instrumentation, part per billion levels of NO can be accurately and precisely measured.

Another useful analytical application uses metastable $^*N_2$ to transfer energy to luminescent metals or molecules. Concentrations as low as $10^6$ atoms per cubic centimeter are readily measured.

Chemical lasers use CLs. In a chemical laser the excited state population inversion is produced directly by chemical energy. One of the most efficient chemical lasers uses the hydrogen-fluorine reaction, which is initiated by the dissociation of $F_2$,

$$F_2 + h\nu \text{ (or electrical energy)} \rightarrow 2F \quad (19a)$$

$$H_2 + F \rightarrow H + {}^*HF \quad (19b)$$

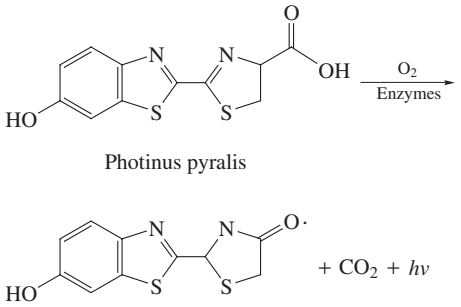
$$F_2 + H \rightarrow 2F + {}^*HF \quad (19c)$$

where $^*HF$ is vibrationally excited HF produced with population inversion. The laser transitions are near-IR emissions arising between vibrational levels of HF.

## C. Bioluminescence

Bioluminescence (BL) is a CL arising from living matter. Fireflies and the glow of disturbed microorganisms in a ship's wake are probably the best known. Although grouped separately, BLs are actually CLs in which the reactants are produced by, and organized in, living organisms.

Not surprisingly, one of the best known BLs is firefly reaction, which involves the enzymatic oxidation of a luciferin. This reaction can have an incredible efficiency, approaching 100%. The luciferin molecule and a number of synthetic analogs have been studied to elucidate the mechanism. The mechanism appears to involve a peroxide decomposition with free radical intervention.




(20)

Bioluminescence is much more pervasive than originally suspected. Low-level BLs are common to many fundamental and essential biological processes such as lipid peroxidation, intracellular redox processes, and catalase decomposition of poisonous $H_2O_2$.

Bioluminescence has had a crucial role in direct studies of cellular and biochemical processes. For example, there is a CL associated with the formation of the ultimate carcinogen benzo[*a*]pyrene-7,8-dihydrodiol 9,10-epoxide from benzo[*a*]pyrene. Also, a very sensitive CL assay for the benzo[*a*]pyrene-7,8-diol has been developed. Bioluminescence also continues to play a pivotal role in the development of the fundamental concepts of CL.

## D. Electroluminescence

Electroluminescence (EL) is luminescence occurring on electron flow in a solid-state device or electrochemical cell. The mechanisms of these two ELs are quite different.

In the electrochemical cell, electrochemical oxidations and reductions occur. It is thus not surprising that most ELs are electron transfer CLs, which are also called electrochemiluminescences (ECLs). The latter are usually produced by rapidly reversing the potential on an electrochemical cell. In this manner both oxidized and reduced species are generated near the same electrode. Electron transfer reactions between oxidizer and reductant are responsible for the ECL:

$$A + e^- \rightarrow A^- \tag{21a}$$

$$B - e^- \rightarrow B^+ \tag{21b}$$

$$A^- + B^+ \rightarrow {}^*A + B \quad \text{or} \quad A + B \tag{21c}$$

Except for chemical side reactions these systems do not run down and, thus, provide light indefinitely. Many such reactions have been studied as light sources, but light-emitting diodes have proved too simple, stable, and efficient to compete against. Among the ELs that have been studied the most thoroughly and for the longest time are the polycyclic aromatic hydrocarbons in aprotic solvents. These systems are difficult to work with, however, because of the great reactivity of the high-energy organic radicals.

Metal complexes also provide efficient ECLs. For example, $[Ru(bpy)_3]^{2+}$ is readily oxidized to the +3 and reduced to the +1 oxidation states. Reaction of the +1 and +3 species gives a very efficient CL. Because of the robustness and relatively low reactivity of the +3 and +1 states, which are stable chemical species rather than free radicals, these inorganic systems are much easier to work with than the highly reactive free radical aromatic hydrocarbons.

An ECL need not be reversible. If the solution contains an electro-inactive reactant that exhibits CL with a product of the electrode reaction, then one has an EL. The system will eventually run down when the reactant is consumed. Figure 16 shows the ECL of $[Ru(bpy)_3]^{3+}$ with oxalate. The Ru(III) is generated electrochemically at the electrode and yields a CL with the electro-inactive oxalate. The spiking is caused by the pulsing of the electrode potential to form the Ru(III) with periods of low potential to permit more oxalate to diffuse up to the electrode before the next pulse. Also shown is the photoluminescence of the Ru(II) complex. The agreement between the ECL and the photoluminescence spectra leaves no doubt that the Ru(II) complex is the active species.

Solid-state EL devices fall into two categories: (1) film-type semiconductor devices and (2) solid-state diodes.

The earliest EL was obtained by subjecting a ZnS phosphor to an AC potential. The resultant current flow in the semiconductor excited the electrons to the conduction band. Recombination of the electrons and holes either directly or at traps gave rise to the luminescence.

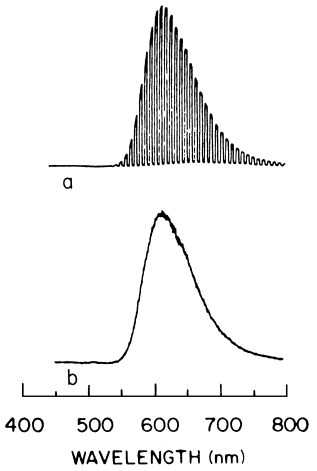



**FIGURE 16** Electrochemiluminescence (a) and photoluminescence (b) spectra of $Ru(bpy)_3]^{2+}$. The ECL is from electrogenerated $[Ru(bpy)_3]^{3+}$ in the presence of oxalate. The spiking of A is caused by pulsing of the electrode potential. [Reprinted with permission from Rubinstein, I., and Bard, A. J. (1981). *J. Am. Chem. Soc.* **103,** 512. Copyright 1981 American Chemical Society.]

Electroluminescent panels are a common source of low-level night lighting. Their efficiencies are much too low for general lighting.

Solid-state diode emitters are called light-emitting diodes (LEDs). An LED consists of a broad band-gap semiconductor diode formed by bonding a *p*-type (hole-rich) semiconductor material to an *n*-type (electron-rich) semiconductor. The resultant *pn* junction has both rectification and emitting properties.

Light-emitting diodes exhibit light emission when forward-biased, which causes current flow through the junction. Current flow arises by the flow of holes from the hole-rich *p* material to the *n* material, and by the flow of electrons from the *n*- to the *p*-type semiconductor. The electrons, for example, are now in the conduction band of the *p* semiconductor material, which has recombination sites suitable for radiative transitions from the conduction to the valence band. The light emission processes arise from direct band-gap radiative recombination with these holes or with suitable sensitizer centers. A similar fate potentially awaits the holes that traverse the junction and find themselves in an environment rich in conduction electrons. However, radiative recombination in the *p* semiconductor is more efficient, and LEDs are physically built so that emitted radiation can most freely escape from the *p* material.

Commercially available LEDs are blue/violet, green, yellow, red, and different ranges of IR. Intense, expensive blue-emitting LEDs are now available and some have reasonable intensities in the near UV. Figure 17 shows typical emission spectra for different LEDs. The emission wavelength is controlled by the band gap of the semiconductor

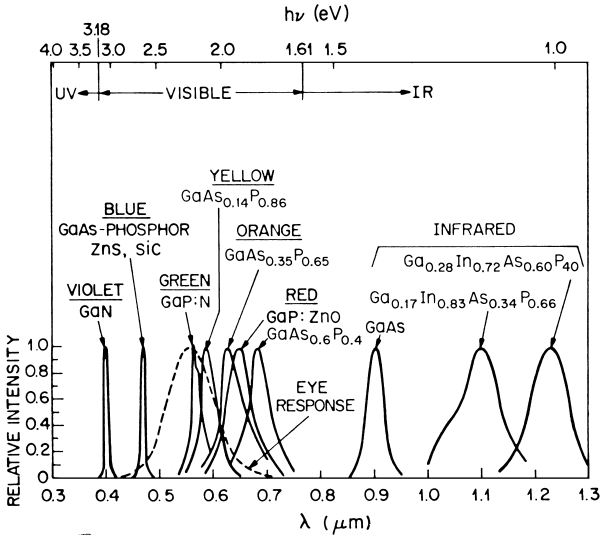



**FIGURE 17** Emission spectra of LEDs. The spectral sensitivity of the human eye is shown for comparison. [Reprinted with permission from Sze, S. M. (1981). "Physics of Semiconductor Devices," 2nd ed., John Wiley & Sons, New York.]

material. By supplying adequate optical feedback, solid-state LED lasers can be made. Very inexpensive, intense laser diodes are available down to 630 nm and at a variety of IR wavelengths. The modern laser pointers, as well as CDs, CD-Rs, CD-RWs, and DVDs, are based on these lasers.

Shorter wavelength laser diodes are only a matter of time. There is enormous commercial pressure for such development since information packing density on optical storage media scales as the inverse square of the wavelength (i.e., halving the wavelength, quadruples the storage capacity). Already LEDs and laser diodes are used in a variety of commercial scientific instrumentation. In addition, the red LEDs are so efficient and inexpensive that they are replacing incandescent bulbs in red traffic lights and the green and yellow colors will shortly follow. When blue LEDs become efficient enough, look to the replacement of fluorescent lights with banks of LEDs emitting in the three primary colors.

LEDs are used as state indicators and numeric displays on consumer and commercial electronics. For battery operated equipment, the lower power consumption of liquid crystal displays, however, has resulted in their superseding LEDs except where self-luminosity is required, and even here LCDs backed by electroluminescent panels are more common than LEDs.

Electrochemiluminescence devices are not commercialized. They do, however, have applications in some analytical work. The ECL of $[Ru(bpy)_3]^{3+}$ with oxalate is a sensitive analytical method for oxalate.

## E. Cathodoluminescence

Cathodoluminescence is luminescence arising in gas or low-pressure electrical discharges. The luminescence can arise either in the electrical discharge itself or on the electrode or target.

Luminescences from electrodes result from excitation of electrode material from the impact of electrons or charged particles on the surfaces. Electronic excitation can result directly from inelastic collisions of particles with atoms or molecules. Chemical bonds can be broken, and high-energy electrons can be ejected. These high-energy secondary electrons can produce other electronic excitations, chemistry, or ionizations. The recombination of electrons with oxidized sites can release enough energy to excite either the product of the electron transfer reaction or excitable species in the neighborhood. Chemical reactions of decomposition products can produce CL.

Gas-phase excitation arises from the same processes as surface excitation. Inelastic collisions can excite molecules or atoms, and species can be decomposed or ionized. Reactions between products can yield CL or new species that can be excited in turn giving rise to new emissions.

Gas-phase emission need not be limited to species originally in the gas phase. The energy transferred to the surface by electron or ion impact can eject or sputter surface material. This sputtering material, once introduced into the discharge, can be excited by the processes just described.

Secondary processes such as energy transfer and excited-state processes can also occur. These affect the nature and efficiency of the emission.

Cathodoluminescence forms the basis of innumerable practical devices. Color and black-and-white televisions are examples of phosphors excited by electron impact on solids. The ubiquitous hollow cathode source of elemental emission lines uses electrode sputtering to introduce the desired element into the discharge. Advertising signs and discharge lighting are very common.

The fluorescent lamp represents an interesting application of both cathodo-luminescence and photoluminescence. The discharge is in a low-pressure rare gas with a trace of mercury vapor. The 254-nm atomic mercury emission dominates the optical output (Fig. 2). The inside of the discharge tube is coated with a UV-excitable, visible-emitting phosphor, which transforms the invisible UV into usable visible light. By changing the phosphor, the emission color tint can be varied, as in the reddish greenhouse lamp. The long-wavelength blacklight works similarly except that the phosphor emits in the 300- to 400-nm region. The color TV or monitor is a beautiful example of three different color phosphors excited independently to produce an image. Close inspection of a color

screen in operation will clearly show the three color elements. The red phosphor is based on the intense red line emission of Eu(III); the ghastly flesh tones of early color TV were the result of the poor red characteristics of earlier phosphors.

## F. Radioluminescence

Radioluminescence arises from interactions between ionizing radiation and matter. This radiation may be particles, such as $\alpha$ or $\beta$ particles, and cosmic rays or photons, such as X-rays or $\gamma$ rays. Materials exhibiting radioluminescence are called scintillators, named after the scintillations, or bursts of light, that arise when radiation strikes the material.

The mechanisms of excitation by ionizing radiation are very similar to those listed under cathodoluminescence. The X- and $\gamma$-ray excitations are initiated by energy loss of the photons as they interact with matter. For higher energies, energy loss occurs by photoionization. Substrates of high atomic number absorb energy more efficiently than those of low atomic number because of the high concentration of electrons.

Charged particles and secondary electrons arising from $\gamma$ rays behave like the charged particles of gas discharges striking a surface, and the processes of excited species generation are the same. Due to the penetrating power of radiation, however, excitation can penetrate much more deeply.

An important by-product of ionizing radiation in solids is that electrons are frequently trapped in lattice sites (Fig. 8). Depending on the depth of the trap sites, the electron may either recombine rapidly or be trapped for an extended period of time, even for millions of years. Release of these electrons can then yield a delayed recombination luminescence. This delayed luminescence in thermoluminescence is discussed in Section VI.G.

Radioluminescence forms the heart of many radiation detectors. The traditional $\gamma$ detector is thallium-activated sodium iodide. The high atomic number of iodide increases the efficiency of energy absorption, and the energy is used to excite the $Tl^+$ luminescent center.

$\beta$-Particle detectors are crystals or organic molecules (e.g., anthracene) or solutions of highly fluorescent organic scintillators dissolved in solid aromatic polymers (e.g., polystyrene) or aromatic solutions. The substrate of an aromatic polymer or solvent is an essential component. The fluorescent indicator is generally present in relatively low concentrations, and much of the excitation occurs away from the scintillator. The bulk of the initial excited states formed are, thus, the easily excited bulk aromatic solvents. Due to the closeness of the aromatic species, the excitation can migrate or hop from molecule to molecule until the energy comes close enough to the scintillator molecule for efficient energy transfer. The scintillator molecule must have an excited singlet state that is below the energy levels of the solvent molecules. Thus, the scintillator traps the energy and emits.

Organic scintillators form the basis of the important liquid scintillation counter. The penetrating power of $\beta$ particles is very limited. Therefore, counting weak $\beta$ emitters such as tritium is very difficult with windowed detectors because the window or the sample itself absorbs most of the radiation. The penetration problem is solved by homogeneously dissolving the material to be counted in a "cocktail" consisting of the solvent and the scintillator. Since the particles are emitted directly in the presence of the scintillator, the penetrating power becomes irrelevant.

## G. Thermoluminescence

Thermoluminescence (TL) is luminescence that arises on gentle warming of a material and usually occurs below incandescence. Reading the sample by heating it destroys the activation, so the readout is a single-shot experiment. Most TL materials can, however, be reactivated and used repeatedly.

The mechanism of TL is shown in Fig. 18. The activation process consists of trapping an electron in a trap

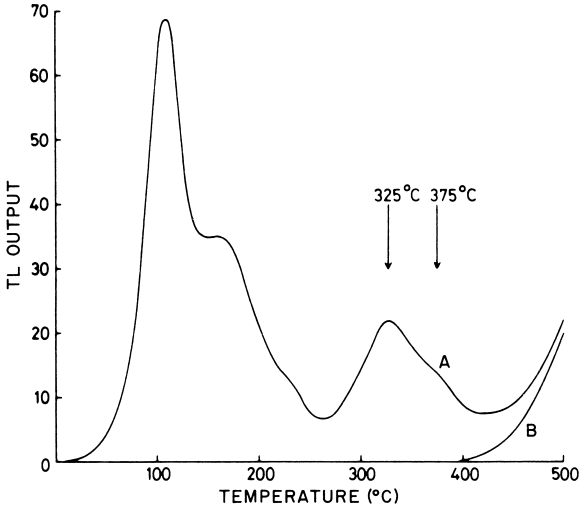



**FIGURE 18** Thermoluminescence glow curve (a) of quartz extracted from pottery for use in TL dating. The sample has been artificially irradiated with 550 rads of $\beta$ radiation as part of the calibration process. Curve (b) is the red-hot glow measured on reheating the sample. The TL is composed of both the stored and irradiated components. [Reprinted with permission from Aitken, M. J., and Fleming, S. J. (1972). *In* "Topics in Radiation Dosimetry," Supplement 1 (F. H. Attix, ed.), p. 1, Academic Press, New York.]

site. The trap site must be deep enough to prevent rapid removal of the electron at the storage temperature. Activation can be accomplished by ionizing radiation or by light. Once the sample is activated, the readout is performed by heating the sample. When the temperature becomes sufficiently high, the electrons are thermally excited into the conduction band. These conduction electrons can migrate through the lattice until they find a hole to relax into. If the hole is an emissive center, or is near one, the center is excited and emits. Alternatively, the relaxation to the hole can be emissive.

The term thermoluminescence is misleading since it does not mean thermal generation of the excited system. *Thermal-stimulated luminescence* much more accurately describes the phenomenon.

In TL measurements or development, readout is performed by heating the sample at a uniform rate and viewing the sample emission. A filter may be added over the detector to restrict the viewing wavelength. Figure 18 shows a typical TL glow curve. The development supplies information on the distribution and depth of traps. Deeper traps require higher temperatures to boil the electrons out. For example, a 375°C glow curve corresponds to a trap lifetime of several million years. Indeed, elaborate models have been developed to extract quantitative information about trap depth and distribution from glow curves.

In most examples of TL, one is examining a delayed photoluminescence. Particularly in solids subjected to radiation chemistry, the nature of the traps and of the emissive centers is modified by irradiation. Therefore, the TL emissions may look quite different from the photoluminescence of the unactivated solid.

The principle use of TL is in radiation dosimetry. The TL signal is proportional to dose over a very wide range. Modern TL dosimeters are sensitive in the microrad region. Thermoluminescence dosimeters have a number of advantages. They are inexpensive, robust, reusable, and sensitive, and they intrinsically integrate the dose. Furthermore, they require no power source or connections. At one time they were widely used for *in vivo* monitoring during therapy since they could be easily implanted or swallowed.

Lithium fluoride is one of the most popular personal dosimeters. It is sensitive down to $10^{-2}$ rads and linear to $>10^3$ rads. The average atomic number of LiF is similar to that of tissue, which gives the dosimeter a response indicative of tissue irradiation. The traps are deep enough to give long storage times without appreciable fading.

The early work on TL demonstrates the difficulty of studying emissions that intrinsically rest on impurity sites. The first attempts at TL dosimetry date back to the early 1950s, when Farrington Daniels attempted to use LiF as a dosimeter. The original work was abandoned owing to problems of sensitivity and storage times. When the work was resumed in the early 1960s, it was discovered that all the earlier TLs of LiF arose from impurity centers that were no longer present in commercially purer LiF. This led to a considerable effort to elucidate the impurity problems and eventually led to usable systems. Also, other useful systems were discovered by other groups.

Another type of sensitive, and at one time widely used, dosimeter utilized radiophotoluminescence. Irradiation of a crystal produces color centers. When photoexcited, many of these color centers emit. Since the emission intensity is directly proportional to the concentration of color centers, a dosimeter is available. An advantage of photoluminescent dosimeters is that the readout is not destructive and can be repeated.

Another important use of TL is in archeological and geological dating when radiocarbon dating is unsuitable (Fig. 18). Heating destroys any TL and initializes the phosphor. As all rocks have some degree of radioactivity from $^{238}$U, $^{235}$Th, and $^{40}$K, the TL can be used to measure the length of time that the sample was irradiated since cooling. For igneous rocks this is the time since formation. For pottery it is the time since the pottery was fired. Such dating can give results accurate to better than 10%.

## H. Flame Emissions

With the exception of hydrogen-oxygen flames, virtually all flames exhibit pronounced visual emissions. We shall discuss briefly the origin of some of these emissions. Luminescence can arise from the major components of the reaction or from trace materials. The orange glow of candle flames and oxidant-starved gas flames arises, not from luminescence, but from the incandescence of carbon particles.

Luminescence from a state is independent of how the state was populated. Many flames are hot enough and the excited state of elements and compounds low enough that a significant excited-state population can be achieved thermally. For example, the emitting state of atomic sodium is at 589 nm. For flames of various temperatures the percentages of molecules in the excited state are as follows: 2000 K, $1 \times 10^{-3}$%; 3000 K, $6 \times 10^{-2}$; and 4000 K, 0.4%. While these excited-state populations may seem very small, they are in fact very large in comparison with the concentrations that could be achieved by all but the most intense laser sources in photoluminescence experiments. This efficiency is readily seen by the intense yellow sodium fluorescence when even tap water is introduced into a relatively cool Bunsen burner flame.

An interesting aspect of flame spectroscopy is that a flame can be too hot to yield good elemental emissions. Too hot a flame can thermally ionize many of the atoms

and reduce the population of the emissive element. This propensity of luminescent species to ionize can be suppressed by increasing the concentration of free electrons in this flame. For example, ionization of calcium can be suppressed by spiking the flame with cesium or potassium, which have lower ionization potentials and increase the concentration of free electrons. Many of the products of the chemical reactions contributing to the flame are produced in excited states and emit (see Section VI.B).

While not, strictly speaking, flames, inductively coupled plasma torches produce plumes that appear to be flames. Furthermore, these plasmas behave like extremely high temperature flames, which are capable of exciting all but the most recalcitrant elements.

One of the most useful applications of flames is for elemental quantitative analysis. High-temperature flames reduce most complex matrices to their elemental components. Many elements, especially metals, then emit in the flames, and the emission can be used for analytical quantification. Even many elements that do not emit directly can still be analyzed in flames by photoluminescence or by atomic absorption.

Because of the chemical generation of many of the excited states in flames, population inversion over the ground state can result. This population inversion can form the basis of a laser. The best known chemical laser is the HF or DF laser, made by burning hydrogen or deuterium in fluorine. Visible or UV chemical lasers are still being sought.

## I. Sonoluminescence

Sonoluminescence (SL) was first observed in 1934 by H. Frenzel and H. Schultes at the University of Cologne. As an indirect result of wartime research on sonar or acoustic radar, they observed SL in an ultrasonic water bath. Very strong ultrasonic fields were found to yield clouds of chaotic flashing bubbles, which are now termed "multi-bubble sonoluminescence." Such systems were not amenable to systematic study until D. Fellipe Gaitan succeeded in trapping a single sonoluminescing bubble at the acoustic resonance in the center of a flask. This single-bubble sonoluminescence (SBSL) opened the way for an explosion of research on SL.

The majority of SL is a broad continuum, with effective black body temperatures in excess of the sun (e.g., well beyond 10,000 K). In many cases the continuum emission is still rising at the 200-nm cutoff of air. Indeed, some have suggested that temperatures suitable for fusion may be achieved this way, but this seems an extremely unlikely possibility. In addition, under some conditions, broad lines are present. Even though the typical ultrasonic transducer operates at about 25 KHz, the light pulses are much less than 50 ps and occur during the collapse phase of the bubbles. Each optical pulse can generate more then a million photons and the phenomena is readily visible to the eye as a steady star-like point.

Water is far and away the best solvent for SL and there is a strong temperature dependence. Sonoluminescence is 100-fold brighter at 0°C compared to 40°C. Dissolved rare gases appear to be critical to SL; it is the 1% Ar in air that is responsible for SL under ambient air conditions.

For a phenomenon known for over 60 years, our understanding of its precise origins is still uncertain. While there are several explanations for SL, the area is still amazingly contentious with articles appearing regularly in support of different mechanisms. The fact that the emission is a continuum with no unique molecular or atomic character exacerbates definitive modeling.

While there seems to have been no direct practical applications of SL to date, there are needs for short-lived, inexpensive, very compact light sources in such areas as luminescence lifetime measurements. Since SBSL systems can be built extremely inexpensively, SL may find use in these areas.

Although not a direct application of SL, ultrasonic-induced chemistry is a growing area with potentially significant applications. Ultrasound has been found to accelerate and control a number of chemical reactions. Some of this chemistry may be induced by the high energy photons of SL.

## J. Fracto-Emission or Triboluminescence

Fracto-emission (FE) is the emission of particles and photons before, during, and following the propagation of a crack in a stressed material. Particles include electrons, ions, and neutral atoms and molecules. Light emission under mechanical stimulation is often called triboluminescence (TL). Grinding and breaking of crystal and glasses can produce FE. However, even pulling tape off of a surface can produce FE.

Sir Francis Bacon in 1605 first reported FE by grinding sucrose, although FE could not have been missed innumerable times earlier. Amusing, if not practical, is the parlor demonstration of grinding a Wint-o-Green Life Saver® between one's teeth or with a pair of pliers in a darkened room to produce bright blue-green flashes. The Wint-o-Green Life Saver® is pressed sucrose flavored with methyl salicylate. Triboluminescence is, in many cases, due to the production of strong electric fields along fracture lines with a concomitant electrical breakdown. The observed luminescence is then a mechanically introduced cathodoluminescence. Another easily observed example of TL arises when you strike two pieces of quartz together. Bright orange flashes result that can be viewed in a dimly lighted

room. For either demonstration eye protection should be worn because of the likelihood of flying particles.

Figure 19A shows the photoluminescence of methyl salicylate, the flavoring of emission of Wint-o-Green Life Saver® which is pressed sucrose. Figure 19B is the TL from grinding pure sucrose. The UV-rich band structure can be clearly associated with the emission of nitrogen gas. Figure 19C shows the TL of a Wint-o-Green Life Saver®. The emission clearly is the nitrogen band structure augmented by a broad visible emission. This broad spectrum is the same as the PL of methyl salicylate, the flavoring.

There can be an enormous amount of chemistry going on in FE. Extensive bond breaking can occur. For example, carbon dioxide is formed during the grinding of calcite. Atomic sodium emission can arise from the breaking sodium silicate glasses.

Fracto-emission is a valuable tool for studying the physical and chemical processes that occur before, during, and following fracture. It allows examination of failure mechanisms including fatigue, microcrack initiation, and growth. Such tools are invaluable in materials research for studying the failure mechanisms of such materials as glasses, metal ceramics, polymers, and composites. In addition, FE is of interest to geologists and geophysicists for detecting fractures in mines and along geological faults.

Crystalloluminescence is emission during the growth of crystals and may arise from cleavage. It is, thus, a form of TL from internally generated breakage.

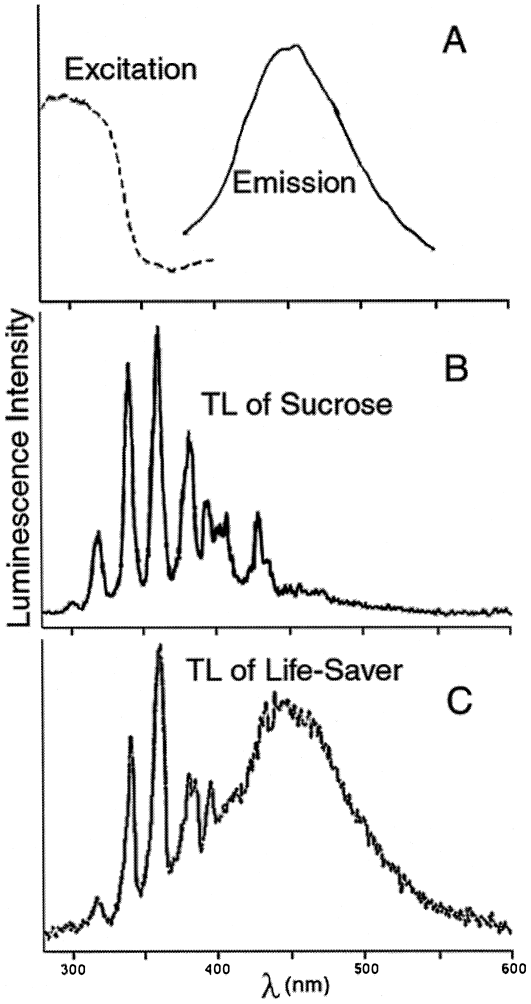



**FIGURE 19** (A) Photoluminescence and excitation spectrum of methyl salicylate, the flavoring of a Wint-o-Green Life Saver®. (B) Triboluminescence of sucrose. (C) Triboluminescence of a Wint-o-Green Life Saver®. [Adapted with permission from Linda Sweeting. Copyright 1998.]

## ACKNOWLEDGMENT

We gratefully acknowledge support of the National Science Foundation (CHE 82-06279, 86-00012, and 97-26999) and the donors of the Petroleum Research Fund, administered by the American Chemical Society.

## SEE ALSO THE FOLLOWING ARTICLES

ENERGY TRANSFER, INTRAMOLECULAR • LASERS • LASERS, DYE • MOLECULAR MICROWAVE SPECTROSCOPY • POTENTIAL ENERGY SURFACES • SONOLUMINESCENCE AND SONOCHEMISTRY • THERMOLUMINESCENCE DATING

## BIBLIOGRAPHY

Adam, W., and Cilento, G., eds. (1982). "Chemical and Biological Generation of Excited States," Academic Press, New York.

Alkemade, C. Th. J., Hollander, Th., Snellman, W., Seeger, P. J., and ter Harr, D., eds. (1982). Metal vapours in flames, *Int. Ser. Nat. Philos.* **103,** Pergamon Press, New York.

Ambrose, W. P., Goodwin, P. M., Jett, J. H., Van Orden, A., Werner, J. H., and Keller, R. (1999). Single molecule fluorescence spectroscopy at ambient temperature. *Chem. Rev.* **99,** 2929–2956.

Cundall, R. B., and Dale, R. E. (1983). "Time-Resolved Fluorescence Spectroscopy in Biochemistry and Biology," Plenum Press, New York.

Demas, J. N. (1983). "Excited State Lifetime Measurements," Academic Press, New York.

De Silva, A. P., Gunaratne, H. Q., Gunnlaugsson, T., Huxley, A. J. M., McCoy, C. P., Rademacher, J. T., and Rice, T. E. (1997). Signaling recognition events with fluorescent sensors and switches. *Chem. Rev.* **97,** 1515–1566.

Dickinson, J. T. "Fracto-Emission, Encyclopedia of Materials," Elsevier, in press.

Dunn, R. C. (1999). Near-field scanning optical microscopy. *Chem. Rev.* **99,** 2891–2927.

Fouassier, J. P., and Rabek, J. F., eds. (1990). "Lasers in Polymer Science and Technology: Applications," CRC Press, Boca Raton, FL.

Harvey, E. N. (1957). "A History of Luminescence," American Philosophical Society, Philadelphia.

Herman, B. (1998). "Fluorescence Microscopy," 2nd ed., Springer-Verlag, New York.

Horowitz, Y. S., ed. (1984). "Thermoluminescence and Thermoluminescent Dosimetry," Vols. 1–3, CRC Press, Boca Raton, FL.

Lakowicz, J. R. (1999). "Principles of Fluorescence Spectroscopy," Second Editon, Plenum Press, New York.

Pawley, J. (1995). "Handbook of Biological Confocal Microscopy," 2nd ed., Plenum Press, New York.

Periasamy, A. (2001) "Methods in Cellular Imaging," Oxford Univ. Press, New York, in preparation.

Schaefer, F. P. (1990). "Dye Lasers," 3rd ed., Springer-Verlag, New York.

Schulman, S. G., ed. (1988). "Molecular Luminescence Spectroscopy," Vol. 2, Wiley, New York.

Sweeting, L. M., Cashel, M. L., Dott, M. L., Gingerich, J. M., Guido, J. L., Kling, J. A., Pippin III, R. F., Rosenblatt, M. M., Rutter, A. M., and Spence, R. A. Spectroscopy and mechanism in triboluminescence, *Mol. Cryst. Liq. Cryst.* **211,** 389.

Sze, S. M. (1981). "Physics of Semiconductor Devices," 2nd ed., John Wiley & Sons, New York.

Weber, M. J. (1982). "CRC Handbook of Laser Science and Technology," Vols. 1–3, CRC Press, Boca Raton, FL.

Yen, W. M., and Selzer, P. M., eds. (1981). Laser spectroscopy of solids. *Top. Appl. Phys.* **49.**



# Magnetic Materials

**R. C. O'Handley**
*Massachusetts Institute of Technology*

I. Interactions and Energies in Magnetic Materials
II. Magnetic Materials, Fundamental Properties
III. Technical Properties
IV. Related Magnetic Phenomena and Applications

## GLOSSARY

**Antiferromagnetic material** A magnetic material in which the exchange interaction favors antiparallel alignment of neighboring spins. Such interactions are limited to certain crystal structures so that frustration does not occur as in a three-membered set of nearest neighbors. Antiferromagnetic interactions often occur for smaller interatomic spacings relative to $d$-orbital diameter.

**Coercivity, coercive field** The negative magnetic field required to reduce the magnetization (intrinsic coercivity, ${}_iH_c$) or flux density (${}_BH_c$) to zero after saturation in a positive field. There is no significant difference between ${}_iH_c$ and ${}_BH_c$ in soft magnetic materials.

**Curie temperature** The temperature above which long-range magnetic order vanishes. A ferromagnetic or antiferromagnetic material becomes paramagnetic above its Curie or Néel temperature, respectively.

**Demagnetizing factor** The numerical factor that indicates the extent to which a given state of magnetization in a bounded magnetic material produces an internal field that opposes the state of magnetization. The demagnetizing factor is a diagonal tensor for an ellipsoid of revolution.

**Domain** A region in a magnetic material in which all magnetic moments have essentially the same orientation. The magnetization within the domain is the saturation magnetization of the material.

**Domain wall** The surface across which the direction of magnetization rotates from that in one domain to the adjacent one. If the magnetization rotates about the normal to the domain wall, the wall is called a Bloch wall. If the magnetization rotates so that it has a component parallel to the wall normal, it is called a Néel wall.

**Exchange energy** The energy associated with the relative alignment of neighboring spins in a magnetic material. In a ferromagnet, exchange energy is more positive if the angle between the directions of neighboring spins is increased.

**Exchange interaction** The preference in a magnetic material for neighboring spins to align parallel or antiparallel to each other. This interaction is quantum mechanical in origin.

**Ferromagnetic material** A magnetic material in which the exchange interaction favors parallel alignment of neighboring spins.

**Fundamental magnetic properties** Those properties of a magnetic material that are largely independent of microstructure, defects, and processing conditions.

Examples are the saturation magnetization and Curie temperature. Fundamental properties reflect the electronic structure that in turn is a consequence of chemistry and crystallography.

**Hard magnetic material** *See* permanent magnet.

**Magnetic anisotropy** The preference for the magnetization vector in a material to point along one or more specific axes. The preferred direction(s) of magnetization can be governed by sample shape (magnetostatic energy), cyrstal structure (magnetocrystalline anisotropy energy), strain in the structure (magnetoelastic energy), as well as surface or interfacial interactions (surface or Néel energy).

**Magnetic glasses** Noncrystalline or amorphous materials that exhibit strong or useful magnetic properties. Magnetic glasses can be either oxide glasses or metallic glasses (amorphous magnetic alloys such as $Fe_{40}Ni_{40}B_{20}$).

**Magnetization** The *process* by which a magnetic material is magnetized, generally either by domain wall motion or by rotation of the magnetization within the domains. The term magnetization is also used as a noun to describe the *state* of a material in terms of the vector sum of its magnetic moments per unit volume.

**Magnetoresistance** The change in resistance of a material depending upon its state of magnetization. Like magnetostriction, it can have isotropic components that show up on heating through the Curie temperature as well as anisotropic components that vary with the relative direction of current and magnetization.

**Magnetostatic energy** The energy of a magnetic moment in a field caused by other moments. This energy, sometimes referred to as dipole energy, is particularly strong and positive inside a material magnetized normal to a thin dimension.

**Magnetostriction** The strain associated with the state of magnetization of a material. The isotropic part of this strain, that decreases sharply on heating through the Curie temperature, is called the volume magnetostriction. The anisotropic part of the strain, called the Joule magnetostriction, is a function of the direction of magnetization relative to the crystal axes.

**Nanomagnetic material** A material with variations in magnetic properties over a length scale of a few nanometers to hundreds of nanometers. Nanomagnetic materials may be manmade by thin film deposition and/or lithographic techniques or they may be natural consequences of thermodynamics and kinetics, such as the fine microstructures that can result from heat-treating certain amorphous magnetic alloys.

**Permanent magnet** A magnetic material which retains a large magnetization upon removal of an applied field after saturation. Further, permanent magnets are characterized by large coercivities so that the remanent magnetization is reduced only by a very large field. Materials with coercivities in the range of about 25–500 Oe (2–40 kA/m) are considered to be semihard magnetic materials. Greater and smaller coercivities generally characterize hard and soft magnetic materials, respectively.

**Remanence** The magnetization or flux density that remains in a sample at zero field after being exposed to a specific field.

**Soft magnetic materials** Materials that are relatively easily magnetized and demagnetized. They are typically characterized by high relative permeabilities and low coercivities (from a few tens of Oe down to as low as milli-Oe in some cases).

**Superparamagnetism** A type of magnetic behavior characterized by zero remanence and zero coercivity. It is observed in ferromagnetic particles or clusters of atoms that are so small that thermal energy overrides the tendency of the magnetization to lie in a particular direction due to crystallography or shape. Specifically, $K_u V < k_B T$.

**Technical magnetic properties** Those properties of magnetic materials that depend strongly on processing and microstructure. Examples are coercivity, remanence, and permeability.

**Zeeman energy** The potential energy of a magnetic moment in a field.

**A MATERIAL** is said to be magnetic when it possess atomic-scale magnetic moments that show long-range ordering (ferromagnets, antiferromagnets, and ferrimagnets) below a Curie temperature. In a ferromagnetic material, all magnetic moments order parallel to each other. In antiferromagnetic and ferrimagnetic materials the moments on one crystallographic site order with their moments antiparallel to those on another, crystallographically different site. In antiferromagnets, the atomic moments on the two types of sites are equal and no net moment results; in ferrimagnets, the atomic moments are not equal and a net moment results. A material might also be said to be magnetic if its atomic-scale magnetic moments show no spontaneous ordering but do respond to an applied magnetic field with an increase in its magnetic moment density (paramagnetic materials). The class of magnetic materials could as well include diamagnetic materials which need not have atomic-scale magnetic moments but do respond to an applied field by producing a weak magnetization directed opposite the applied field. Given this broad definition, it can be said that all materials are magnetic.

A more restrictive definition of magnetic materials might start with those that show ferromagnetic ordering at room temperature, that is, they have Curie temperatures which are greater than room temperature. Of all the elements, only four, Fe, Co, Ni, and Gd, are ferromagnetic at room temperature. Several other elements show antiferromagnetism (e.g., Cr and $\alpha$-Mn) or other forms of magnetic ordering such as helical or spiral moment arrangements (Tb, Sm, Ho, etc.). Yet, starting with these few *elements*, the class of ferromagnetic, antiferromagnetic, and ferrimagnetic *alloys* and compounds expands to include as significant components about two-thirds of the elements of the periodic table. At least three materials, $Cu_2MnAl$, ZrZn, and InSb, are ferromagnetic (the latter two only at cryogenic temperatures), although they contain none of the abovementioned strongly ferromagnetic elements. This article surveys elements, alloys, and compounds that are ferromagnetic or ferrimagnetic at room temperature.

Magnetic materials are said to be *soft* if they respond to relatively weak external fields by changing their state of magnetization. This change in magnetization is accompanied by a change in force on the material if the external field is inhomogeneous. Magnetic materials are said to be *hard* when they retain a state of net magnetization in the presence of a significant opposing field. Hard magnetic materials produce a field that can attract other soft or hard magnetic materials. Half a century ago, most of the interest in magnetic materials was associated either with these forces (motors, electromagnetic actuators) or with their ability to concentrate magnetic flux (shielding) or enhance flux change (inductors and transformers). Today, engineering interest in magnetic materials has expanded to include their ability to store information at high density in a nonvolatile way (the information is retained when the power to the storage device is off), as well as their varied electrical and spin transport properties.

While exciting new magnetic compositions are still being discovered, the range of properties exhibited by existing magnetic compositions is now often expanded and/or tailored to specific needs by making them in thin film, multilayer, or nanoparticle form. In these reduced-dimension structures, novel magnetic properties can be obtained because of the altered atomic and chemical structure near an interface and by arranging magnetic and sometimes nonmagnetic components at small length scales so that new magnetic interactions and transport effects can be engineered. Thus a modern overview of magnetic materials must include not only materials whose properties vary with composition, but also bond type and/or crystal structure. Reference should also be made to the exciting new developments achieved by rendering old compositions in new structures.

This article begins with an overview of some of the important manifestations of magnetism in materials (magnetic domains, magnetic anisotropy, magnetostriction) and outlines the underlying science that explains the properties of magnetic materials. Magnetic materials can be classified either by their functional properties (e.g., magnetically soft or hard) or by the nature of their bonding and structure (oxides, alloys, intermetallic compounds). Table I summarizes some of the important crystal structures and magnetic materials systems that occur in a matrix of functional classes and bonding types. In this article, the major magnetic functional classes are described and material examples are given. Also, an attempt is made to include descriptions of the major bonding and structure types of magnetic materials. Magnetic effects at surfaces, in thin films, and fine particles are also described.

The magnetic fields $B$ and $H$ are related in MKS units by $B = \mu_0 (H + M)$ where $H$ is the applied field due to *macroscopic* current densities, $J = I/\text{area}$. Ampere's law relates the $H$ field to the current:

$$\nabla \times \boldsymbol{H} = \boldsymbol{J} \quad \text{or} \quad \oint H \cdot dl = I, \tag{1}$$

thus $H$ has units $A/m$. $M$ is the magnetic moment density or magnetization in the material, $M = N \langle \mu_m \rangle / V$, where $\langle \mu_m \rangle$ is the average magnetic moment per atom or molecule in the system and $N/V$ is the number of such entities per unit volume. The magnetization, $M$, while having the same MKS units as the $H$ field, is the field due to *microscopic* currents. $B$ is the flux density, $\phi/A$ (with units of Tesla), whose time rate of change induces a voltage by Faraday's law:

$$\nabla \times E = -\partial B/\partial t \quad \text{or} \quad \oint E \cdot dl = V = \frac{-\partial \phi}{\partial t} \tag{2}$$

The microscopic currents that give rise to the magnetization are the spin and orbital angular momenta of the electrons.

The magnetic susceptibility $\chi_m$ is usually used to describe weak magnetic responses to $H$ as in *paramagnetic* and *diamagnetic* materials. The magnitude of $\chi_m$ is typically $\pm 10^{-4}$–$10^{-6}$ (dimensionless in MKS units). *Diamagnetism* is not a matter of aligning preexisting atomic magnetic moments. Rather, it is an electronic response to an applied field that creates a new component of orbital angular momentum and thus a magnetic moment. The diamagnetic response is always negative as it can be traced classically to Faraday's law or Lenz's law. A material with a paramagnetic (diamagnetic) susceptibility of $\chi = +(-)\, 10^{-5}$ would show a magnetization $M = 10$ A/m in a field of $H = 10^6$ A/m (applied field $B = \mu_0 H \approx 1.25$ Tesla). A magnetization of 10 A/m corresponds to a flux density $B = \mu_0 M$ of $1.25 \times 10^{-5}$ T which is five orders

**TABLE I Important Examples of Magnetically Soft and Hard Materials can be Found in each of the Major Bonding Types: Oxides, Metallic, and Covalent**

| Bond type | Crystal structure and examples | |
|---|---|---|
| | Soft | Hard |
| Ionic (oxides) | Spinel, TO·$Fe_2O_3$ (T = Mn, Ni. . .) | Spinel, CoO·$Fe_2O_3$ |
| | Defect spinel, $\gamma$-$Fe_2O_3$ | Hexaferrites, BaO·6$Fe_2O_3$ |
| Metallic (alloys) | | |
| 3*d* | BCC Fe, FeCo, FCC NiFe, NiCo | HCP Co-base |
| 4*f* | | HCP R metals, R-T alloys |
| Intermetallic compounds | | |
| Ordered alloys and covalent (*p-d*) | $Ni_3Fe$ | (CuAu) CoPt |
| | Heusler ($L2_1$) $Cu_2MnAl$ | $Co_5Sm$, $Co_{17}Sm_2$, $Fe_{14}Nd_2B$ |
| | Borides: $Fe_2B$, $Co_3B$ | Laves$TbFe_{24}$, $DyCo_2$ MnAl |
| Amorphous alloys | | |
| (metallic + *p-d* covalent) | $Fe_{80}B_{20}$, $Co_{70}Fe_5Si_{15}B_{10}$ $Co_{80}Gd_{20}$ | $TbFe_2$, $DyCo_2$ |

T = transition metal; R = rare-earth metal.

of magnitude less than the spontaneous magnetization of the ferromagnetic materials listed in Table I.

The reason for the weak magnetic response of paramagnetic materials to an external field is that thermal energy $k_BT$ overwhelms the energy that favors alignment of a paramagnetic moment with the field, $g\mu_B B$. The field and temperature dependence of the magnetization in paramagnets (or ferromagnets, see below) is described by the Brillouin function, $B_J(x)$:

$$\frac{M(H,T)}{M(\infty,0)} = B_J(x) = \frac{2J+1}{2J}\coth\left(\frac{2J+1}{2J}x\right) - \frac{1}{2J}\coth\left(\frac{x}{2J}\right), \qquad (3)$$

as shown in Fig. 1. In Eq. (3), $J$ is the total (spin plus orbital) angular momentum quantum number and $x = g\mu_o\mu_B JH/k_BT$, expresses the energy of the magnetic moment in the $H$ field relative to the thermal energy $k_BT$.

*Ferromagnetic* materials are characterized by a long-range ordering of their atomic moments, even in the absence of an external field. The observed field dependence of the magnetization extrapolates from its high-field values to a nonzero value, called the spontaneous magnetization (Fig. 2a). The spontaneous, long-range magnetization of a ferromagnet is observed to vanish above an ordering temperature called the Curie temperature $T_C$ (Fig. 2b). Experimental curves of $M$ versus $H$ do not always show true saturation. Thermodynamics suggests that $M(H)$ approaches saturation like $1 - H^{-1}$. For polycrystalline systems, $M(H)$ approaches saturation like $1 - H^{-2}$.

In a ferromagnetic material, the $H$ field in Eq. (3) includes not only the applied field but also the strong internal field called the Weiss molecular field, $\lambda M$: $B = \mu_o(H + \lambda M)$. The Weiss molecular field expresses the tendency for long-range magnetic order due to the exchange interactions between local moments. The constant $\lambda$ is called the molecular field constant. Because $\lambda M \gg H$, the energy of the magnetization in the exchange field, namely $g\mu_B J\mu_o\lambda M$, is comparable in magnitude to $k_BT$ and magnetic ordering occurs spontaneously in the absence of an applied field. Thus, for a ferromagnet, Eq. (3) becomes a transcendental equation that can be solved graphically or numerically to describe the results in Fig. 2b. Figure 2b also compares the experimental data for Ni with the form derived from Eq. (3) $L(x)$ for $J = \infty$).

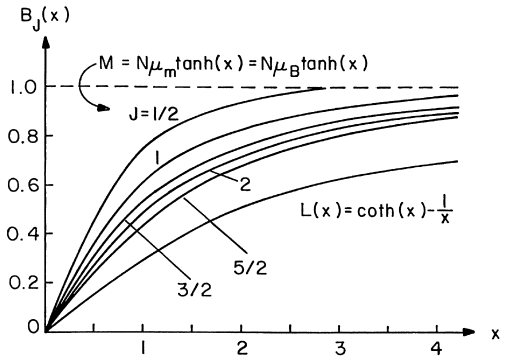



**FIGURE 1** Brillouin function versus $x = \mu_\mu B/k_B T$ for various values of $J$. The infinite-spin limit is given by the classical Langevin function, $L(x)$. [From O'Handley, R. C. (2000). "Modern Magnetic Materials, Principles and Applications." Reprinted with permission of John Wiley and Sons.]

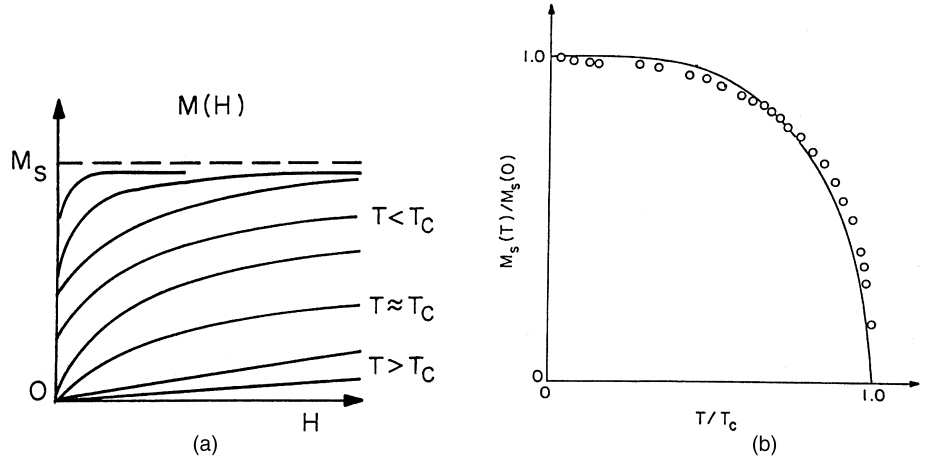



**FIGURE 2** (a) Magnetization of a ferromagnetic material versus field. (b) Reduced magnetization versus reduced temperature for nickel (open data points, from Weiss and Forrer, *Ann. Phys.* **5,** 153, 1926) and Brillouin function, $B_{1/2}(x)$ (solid line). [From O'Handley, R. C. (2000). "Modern Magnetic Materials, Principles and Applications." Reprinted with permission of John Wiley and Sons.]

Saturation magnetic moments and Curie temperatures are listed in Table II for some representative ferromagnetic materials.

The relative magnetic permeability $\mu = \mu_r/\mu_0 = B/H$ is used more often than susceptibility ($\chi_m = M/H$) to describe the response of ferromagnetic materials to $H$. This is because ferromagnets are useful in electromagnetic devices where it is $B$ rather than $H$ that is important in inducing a voltage [Eq. (2)].

# I. INTERACTIONS AND ENERGIES IN MAGNETIC MATERIALS

## A. Magnetostatics

Magnetostatics refers to the consequences of the magnetic fields that appear near the surfaces of magnetized bodies. If the magnetization of a polycrystalline ferromagnetic sample measuring, for example, $1 \times 1 \times 5$ mm is measured,

**TABLE II Fundamental Magnetic Data for Various Magnetic Materials**

| | | $\mu_0 M_s$ | $M_s$ | $\mu_0 M_s$ | $M_s$ | $n_B/FU$ | $T_C$ ($T_N$) |
|---|---|---|---|---|---|---|---|
| | | (T) | $\left(\frac{\text{emu}}{\text{cm}^3}\right)$ | (T) | $\left(\frac{\text{emu}}{\text{cm}^3}\right)$ | $= M_s/\mu_B N_v$ ($\mu_B$/FU) | |
| **Substance** | **Structure** | **(290 K)** | | **(0 K)** | | **(0 K)** | **(K)** |
| Fe | BCC | 2.1 | 1707 | 2.2 | 1707 | 2.22 | 1043 |
| Co | HCP, FCC | 1.8 | 1440 | 18.2 | 1446 | 1.72 | 1388 |
| Ni | FCC | 0.61 | 485 | 0.64 | 510 | 0.606 | 627 |
| $Ni_{80}Fe_{20}$ | FCC | 1.0 | 800 | 1.17 | 930 | 1.0 | — |
| Gd | HCP | — | | 2.6 | 2060 | 7.63 | 292 |
| Dy | HCP | — | | 3.67 | 2920 | 10.2 | 88 |
| $Ni_2MnGa$ | Heusler | 0.6 | 480 | — | | 0.6 | 373 |
| $CrO_2$ | — | 0.65 | 515 | — | | 2.03 | 386 |
| $MnOFe_2O_3$ | Spinel | 0.51 | 410 | — | | 5.0 | 573 |
| $FeOFe_2O_3$ | Spinel | 0.6 | 480 | — | | 4.1 | 858 |
| $Y_3Fe_5O_{12}$ | Garnet(YIG) | 0.16 | 130 | 0.25 | 200 | 5.0 | 560 |
| $Nd_2Fe_{14}B$ | Tetragonal | 1.6 | 1280 | — | | — | — |
| $a$-$Fe_{80}B_{20}$ | Amorphous | 1.6 | 1260 | 1.9 | 1480 | 2.0 | 650 |

The quantity $n_B$ is called the magneton number, the number of Bohr magnetons per atom or per formula unit (FU) in a material.

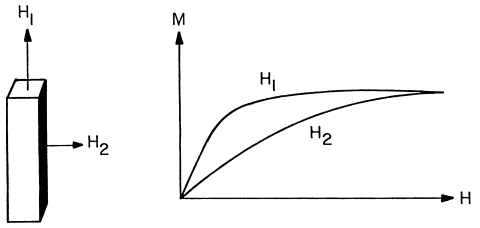



**FIGURE 3** Magnetization curves for a polycrystalline, ferromagnetic sample with field applied in different directions. [From O'Handley, R. C. (2000). "Modern Magnetic Materials, Principles and Applications." Reprinted with permission of John Wiley and Sons.]

the approach to saturation is different for application of the field along the long axis as opposed to the short direction of the sample (Fig. 3). A greater external field is needed to achieve the same degree of magnetization for fields applied in the short direction compared to the long direction.

This shearing effect on the $B$-$H$ loop is related to the magnetic poles produced at the surfaces across which the magnetization has a discontinuity in its normal component. A magnetic field, $H$, emanates from the north poles and terminates at the south poles. These magnetostatic fields are responsible for the effect of sample shape on the magnetization process. The shape and aspect ratio of the sample affect the strength of the magnetostatic field inside the sample and opposing $\boldsymbol{M}$. The magnetostatic $H$ field can be derived from a scalar magnetic potential, $\phi_m$, $H = -\nabla\phi_m$ because $\nabla \times H = 0$ in the absence of macroscopic currents.

The scalar potential results from volume as well as surface magnetic "charges" or poles, $\rho_m = \nabla \cdot \boldsymbol{M}$ and $\sigma_m = \boldsymbol{M} \cdot \boldsymbol{n}$, respectively, according to the relation

$$\phi_m = -\frac{1}{4\pi}\iiint \frac{\nabla' \cdot M(x')}{|x - x|}\, d^3x' + \frac{1}{4\pi}\iint \frac{M(x') \cdot n}{|x - x|} d^2x' \qquad (4)$$

The consequences of this result are readily seen by the exact two-dimensional solution for the field due to a line of magnetic poles. If the line of magnetic charge subtends an angle $\theta$ from the observation point and the distance to the ends of the pole distribution are $r_1$ and $r_2$ then the field components parallel and perpendicular, respectively, to the linear charge distribution are, in SI units:

$$H_{\parallel} = \frac{\sigma}{2\pi}\ln\left(\frac{r_2}{r_1}\right) \qquad H_{\perp} = \frac{\sigma}{2\pi}\theta \qquad (5)$$

Figure 4 provides an excellent pedagogical summary of the important issues in magnetostatics. This figure plots the results of micromagnetic calculations of the field distribution inside and around a uniformly magnetized bar. Only the upper right quadrant of the bar, of infinite extent out of the paper is shown. Note that the surface charges are sources for the $\boldsymbol{H}$ field inside and outside the sample. An equivalent current through the surface windings can be considered the source of the $\boldsymbol{B}$ field inside and outside the sample. Further, outside the sample the $B$ and $H$ fields are parallel to each other: $B = \mu_o H$ (SI) and $B = H$ (c.g.s.). Note also that the boundary conditions on $\boldsymbol{B}$ and $\boldsymbol{H}$ are properly satisfied in these calculated fields. $\boldsymbol{M}$ is proportional to $\boldsymbol{B} - \boldsymbol{H}$, which in this case is held uniform inside the bar. The $H$ field inside the magnetized sample in Fig. 4 opposes the state of magnetization in which the sample is held. In a sample that is not so constrained, this internal field would tend to demagnetize the sample.

The internal field due to surface magnetic poles is called the demagnetizing field $H_d$. This field is proportional to the magnetization of the sample, $H_d \propto -M$ [Eq. (5)]. In general, $H_d$ is a strong function of position inside a uniformly magnetized body. The exceptions are ellipsoids of revolution, in which case

$$H_d = -NM \qquad (6)$$

where $N$ is called the demagnetizing factor. It is a third-rank tensor with unit trace,

$$\sum_{i=1}^{3} N_i = 1 \qquad (7)$$

where $i$ indicates the direction of magnetization. The values of $N_x, N_y, N_z$ for an infinite sheet in the $x$–$y$

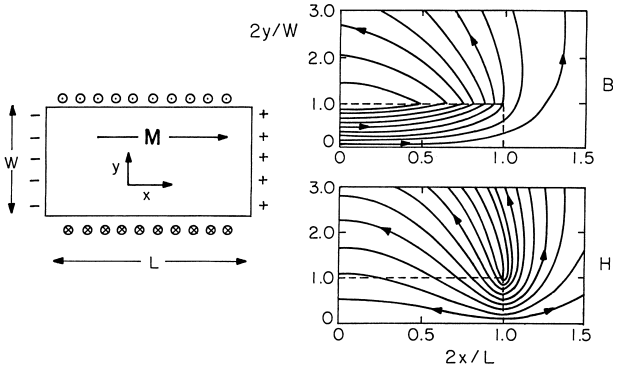



**FIGURE 4** ***B*** and ***H*** fields in and around a uniformly magnetized bar of length $L$ and width $W$ (infinite extent out of paper). At the left is sketched the magnetized bar and its surface poles (or the equivalent surface current around the bar) which are the sources for the ***H*** and ***B*** fields, respectively. Only one quarter of the bar is sketched at right because of the symmetry of the situation. [From Bertram, H. N. (1994). "Theory of Magnetic Recording," Cambridge Univ. Press, Cambridge.]

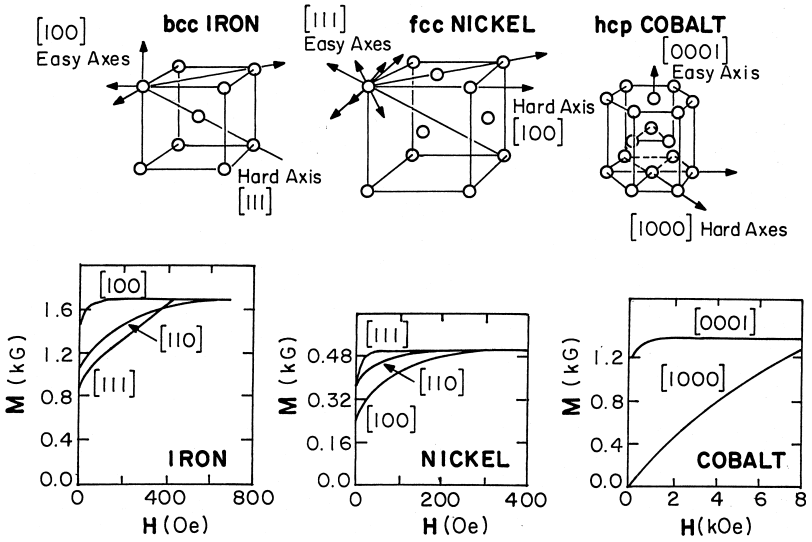



**FIGURE 5** Crystal structures for Fe, Ni, and Co showing easy and hard magnetization directions with respective magnetization curves below. [From O'Handley, R. C. (2000). "Modern Magnetic Materials, Principles and Applications." Reprinted with permission of John Wiley and Sons.]

plane are $N_x + N_y = 0$, $N_z = 1$. For a rod infinite along $z$, $N_x = N_y = 0.5$, $N_z = 0$. For a sphere $N_x = N_y = N_z = 1/3$.

## B. Magnetic and Magnetoelastic Anisotropy

The shape of a material is an *extrinsic* factor affecting the form of the $M$–$H$ curve for a field applied along different directions in a sample. Several *intrinsic* factors can also give rise to anisotropy in magnetic properties. These are (1) magnetocrystalline anisotropy (the preference for $\boldsymbol{M}$ to lie along certain high-symmetry crystallographic directions), (2) magnetoelastic anisotropy (the preference for $\boldsymbol{M}$ to lie in a direction dictated by the symmetry of an imposed anisotropic deformation of the material), and (3) field-induced or directed-pair anisotropy (the preference for magnetization along an axis determined by a field present during prior heat treatment).

### 1. Magnetocrystalline Anisotropy

The physical properties of a material must exhibit a symmetry no lower than that of the crystal structure. The symmetry of the crystal structure affects the preferred direction of magnetization at the local level. A magnetic atom senses the crystal symmetry through the Coulomb electric field of its nearest neighbors. This crystal field can be expanded in harmonic functions that reflect the symmetry about a given site. The crystal field symmetry is correlated with the orbital symmetry of the bonding valence electrons, which are characterized by their orbital angular momentum, $\boldsymbol{L}$. The spin, $S$, of the magnetic atom, couples to the electron's orbital motion via quantum mechanical spin-orbit interactions, $\xi \boldsymbol{L} \cdot \boldsymbol{S}$. Hence, the spin is coupled to the symmetry of the crystal field. Thus, for strong magnetic anisotropy, a low-symmetry crystal field and strong spin-orbit coupling are required.

Figure 5 depicts the hexagonal-close-packed crystal structure of cobalt and shows the $M$–$H$ curves for fields applied in a "hard," base-plane direction [1000] as well as along the "easy" $c$ axis, [0001]. The uniaxial anisotropy of HCP cobalt or other uniaxial magnetic materials can be described phenomenologically by the anisotropic free energy density

$$f_a^{hex} = \sum_{i=0} K_{ul} \sin^{2l} \theta. \tag{8}$$

The field energy density, $-\mu_o \boldsymbol{M} \cdot \boldsymbol{H}$, must be added to Eq. (8) to give the total free energy

$$f_{Tot} = -\mu_0 M_s H \sin\theta + K_{\mu 1} \sin^2 \theta \tag{9}$$

where $\theta$ is the angle between $M_s$ and the $c$ axis. The zero-torque condition, namely $-\partial f / \partial \theta = 0$, gives for the first-order, uniaxial hard-axis (base plane) magnetization process:

$$M = M_s \sin\theta = \frac{\mu_o M_s^2 H}{2K_{u1}} \tag{10}$$

Fitting this form to the cobalt magnetization data gives $K_{u1} = 4.1 \times 10^5$ J/m$^3$. Inclusion of the second-order anisotropy, $K_{u1} \sin^4 \theta$, describes the negative curvature of the hard axis $M$–$H$ curve near the approach to saturation. Fitting experimental curves gives $K_{u2} \approx 1.5 \times 10^5$ J/m$^3$.

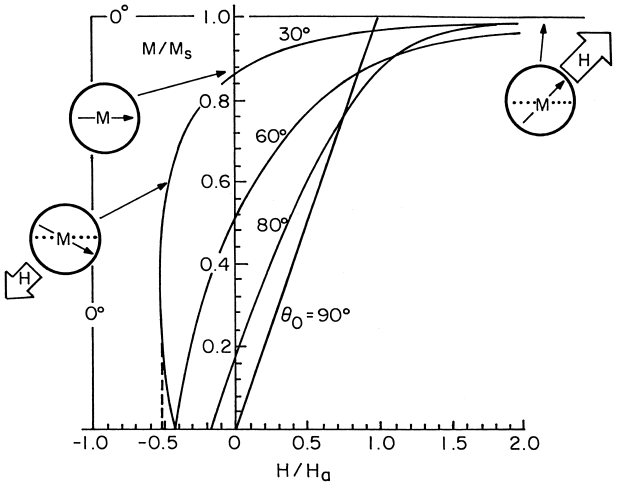



**FIGURE 6** Reduced magnetization versus reduced field applied at an angle $\theta_o$ to the easy axis. The linear *m-h* curve is for $\theta_o = 90°$, and the other curves of increasing remanence are for $\theta_o = 80°$, 60°, and 30°. Possible magnetization distributions are shown as inserts for nucleation-inhibited, single-domain particles. [From O'Handley, R. C. (2000). "Modern Magnetic Materials, Principles and Applications." Reprinted with permission of John Wiley and Sons.]

For fields applied at arbitrary angles $\theta_0$ relative to the easy axis of a uniaxial material, the $M$–$H$ loops vary in a richer way between the square easy-axis ($\theta_0 = 0$) limit and the hard-axis ($\theta_0 = 90°$) limit (Fig. 6).

The cubic crystal structures of $\alpha$-Fe (BCC) and $\gamma$-Ni (FCC) require inclusion of $\phi$ dependence in the harmonic expansion of the anisotropy energy:

$$f_a^{cubic} = K_0^c + K_1^c\left(\alpha_1^2\alpha_2^2 + cycl\right) + K_2^c\alpha_1^2\alpha_2^2\alpha_3^2 + \cdots \tag{11}$$

Values of $K_1^\alpha$ and $K_2^\alpha$ ($\alpha = u$ or $c$) are listed in Table III for some common magnetic materials. Positive (negative) $K_1$ implies ⟨100⟩ (⟨111⟩) easy axes if $K_2 = 0$.

Figure 7 shows the variation of $K_1$ with composition in the FCC Fe-Ni alloys. Compositions exist near 80% Ni for which $K_1 = 0$, that is, the magnetization is equally stable along ⟨100⟩ and ⟨111⟩ directions. These materials are generally easily magnetized and hence make excellent soft magnetic materials.

## 2. Magnetoelastic Interactions

The most common manifestation of magnetoelastic (ME) interactions is magnetostriction, $\lambda = \Delta l/l$, the strain in a material that accompanies a change in its direction of magnetization. Figure 8 shows the field dependence of the magnetostrictive strain measured parallel and perpendicular to the applied field direction, starting from a randomly demagnetized, isotropic sample. Below saturation, the strain is generally quadratic in the field-induced magnetization; above the anisotropy field, the strain saturates.

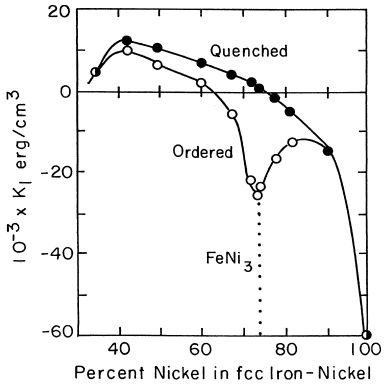



**FIGURE 7** First-order anisotropy constant for FCC Ni-Fe alloys. [From Bozorth, R. M. (1993). "Ferromagnetic Materials," IEEE Press, New York.]

The difference between the parallel and perpendicular strain curves is 3/2 of the saturation magnetostriction constant. When the strain parallel to the field direction is positive, the magnetostriction is positive.

Because magnetocrystalline anisotropy reflects the crystal field symmetry, a distortion of the crystal field results in a new, strain-induced anisotropy. For example, a uniaxial strain applied to a cubic crystal causes a uniaxial ME anisotropy that adds to the cubic anisotropy [cf. Eq. (11)]:

$$f_{me}^c = B_1\left[\varepsilon_1\left(\alpha_1^2 - \tfrac{1}{3}\right) + cycl\right] + B_2[\varepsilon_{12}\alpha_1\alpha_2 + cycl] + \cdots \tag{12}$$

Here $B_1$ and $B_2$ are axial and shear magnetic stress coefficients that, when multiplied by the appropriate strain component, give the ME anisotropy energy density. Thus, if $B_1 > 0$, an extensional strain in the $z$ direction ($\varepsilon_{zz} > 0$) increases the energy for magnetization in that direction ($\alpha_z = 1$) relative to magnetization in the $x$–$y$ plane. Even though the anisotropic strain may be small, e.g., $|\varepsilon_{ii}| \bullet 10^{-3}$, the ME anisotropy may be comparable to or exceed

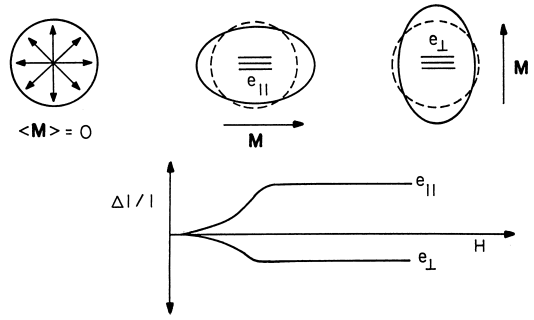



**FIGURE 8** Representations of a demagnetized sample (left) and two states of magnetization with the sense of magnetostrictive strain shown for positive magnetostriction constant: $(\Delta l/l)_{||} > 0$. At the bottom is illustrated the field dependence of strain parallel and perpendicular to the magnetization direction. [From O'Handley, R. C. (2000). "Modern Magnetic Materials, Principles and Applications." Reprinted with permission of John Wiley and Sons.]

**TABLE III Magnetic Anisotropy Constants $K_1$ and $K_2$ for Selected Materials**

| | $T = 4.2$ K | | RT | |
|---|---|---|---|---|
| | $K_1\,(-k_2)$ | $K_2\,(-k_4)$ | $K_1$ | $K_2$ |
| 3*d* metals | | | | |
| Fe | $5.2 \times 10^4$ | $-1.8 \times 10^4$ | $4.8 \times 10^4$ | $-1.0 \times 10^4$ |
| Co$^u$ | $7.0 \times 10^5$ | $1.8 \times 10^5$ | $4.1 \times 10^5$ | $1.5 \times 10^5$ |
| Ni | $-12 \times 10^4$ | $3.0 \times 10^4$ | $-4.5 \times 10^3$ | $-2.3 \times 10^3$ |
| $Ni_{80}Fe_{20}$ | — | — | $-3 \times 10^2$ | — |
| $Fe_{50}Co_{50}$ | | | $-1.5 \times 10^4$ [a] | |
| 4*f* metals | | | | |
| Gd$^u$ | $-1.2 \times 10^5$ | $+8.0 \times 10^4$ | $+1.3 \times 10^4$ | — |
| Tb$^u$ | $-5.65 \times 10^7$ | $-4.6 \times 10^6$ | — | — |
| Spinel ferrites | | | | |
| $Fe_3O_4$ | $-2 \times 10^4$ | — | $-0.9 \times 10^4$ | — |
| $CoFe_2O_4$ | $+10^6$ | — | $2.6 \times 10^5$ | — |
| Garnets | | | | |
| YIG | $-2.5 \times 10^3$ | — | $1 \times 10^3$ | — |
| Hard magnets | | | | |
| $BaO{\cdot}6Fe_2O_3^u$ | $4.4 \times 10^5$ | — | $3.2 \times 10^5$ | — |
| $Sm^uCo_5$ | $7 \times 10^6$ | — | $1.1$–$2.0 \times 10^7$ | — |
| $Fe_{14}Nd_2B^u$ | $-1.25 \times 10^7$ [b] | — | $5 \times 10^6$ | — |

[a] Disordered; $K_1 \approx 0$ for ordered phase.

[b] Uniaxial materials are designated with a superscript "$u$" and their values $K_1^u$ and $K_2^u$ are listed under $K_1$ and $K_2$, respectively. The sign convention for the uniaxial materials is based on the $\sin^2\theta$ notation of Eq. (8): $K_1 > 0$ implies easy axis. Units are J/m$^3$; multiply these values by 10 to get erg/cm$^3$.

the magnitude of the cubic anisotropy. For example, in Ni, $B_1 \approx 6.2 \times 10^6$ N/m$^2$, so for $\varepsilon \approx 10^{-3}$, $|f_{\rm me}^{\rm c}| \approx 6.2 \times 10^3$ J/m$^2$ which is comparable to $|K_1| \approx 4.8 \times 10^3$ J/m$^3$.

Some values of $B$s and $\lambda$s are listed in Table IV for representative materials.

The ME free energy in Eq. (12) also implies that if the direction of magnetization changes (i.e., the $\alpha_i$s change) then the material may change its equilibrium state of strain. This is the phenomenon of anisotropic magnetostriction shown in Fig. 8. The dependence of the magnetostrictive strain on the direction of magnetization may be derived by adding to Eq. (12) the form of the elastic energy for a cubic crystal

$$f_{el} = \tfrac{1}{2}C_{11}\left(e_{11}^2 + cycl\right) + C_{12}(e_{11}e_{22} + cycl) + \tfrac{1}{2}C_{44}\left(e_{12}^2 + cycl\right). \quad (13)$$

Minimization of Eqs. (12) and (13) with respect to the $\varepsilon_{ij}$'s leads to the form of the magnetostrictive strains $\lambda_{100}$ and $\lambda_{111}$:

$$\lambda_{100} = -\frac{2}{3}\frac{B_1}{C_{11} - C_{12}} \quad (14)$$

and

$$\lambda_{111} = -\frac{1}{3}\frac{B_2}{C_{44}} \quad (15)$$

Eqs. (14) and (15) show that the magnetostrictive strains $\lambda_{ijk}$ are related to the ME stress coefficients $B_i$ by the elastic constants in a way that parallels the mechanical stress strain relations, $\varepsilon_{ij} = C_{jkl}\sigma_{kl}$, (except for the

**TABLE IV ME Coupling Coefficients or Stresses (MPa) and Magnetostriction Constants $\lambda_{100}$ and $\lambda_{111}$ in PPM at Room Temperature for Several Materials**

| | $\lambda_{100}$ ($\lambda^{\gamma,2}$) | $B_1$ (MPa) | $\lambda_{111}$ ($\lambda^{\varepsilon,2}$) | $B_2$ (MPa) | Polycrystal, $\lambda_s$ |
|---|---|---|---|---|---|
| 3*d* metals | | | | | |
| BCC-Fe | 21 | −2.9 | −21 | +2.9 | −7 |
| HCP-Co$^u$ | (−140) | +6 | (50) | +13 | (−62) |
| FCC-Ni | −46 | +6.2 | −24 | +4.3 | −34 |
| *a*-$Fe_{80}B_{20}$ | — | | — | | +32 |
| 4*f* metals | | | | | |
| $TbFe_2$ | | | 2600 | | 1753 |
| Oxides | | | | | |
| $Fe_3O_4$ | −15 | | 56 | | +40 |
| $CoFe_2O_4$ | −670 | | 120 | | −110 |
| Yttrium-iron garnet | −1.4 | | −1.6 | | −2 |

See Eqs. (14) and (15). Some polycrystalline magnetostriction values are also listed. The prefix "a-" designates an amorphous material. For uniaxial materials (superscript "$u$") where $\lambda^{\gamma,2}$ or $\lambda^{\varepsilon,2}$ was reported, their values are given in parentheses in the $\lambda_{100}$ and $\lambda_{111}$ column, respectively.

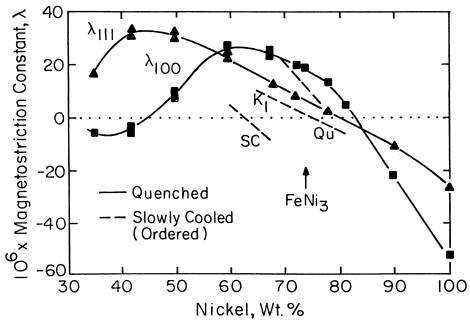



**FIGURE 9** Room temperature magnetostriction constants for FCC Fe-Ni alloys. [From Bozorth, R. M. and Wakiyama, T. (1969). *J. Phys. Soc. Jpn.* **17,** 1669; Hall, R. C. (1960). *J. Appl. Phys.* **31S,** 5157.] Crossover behavior of anisotropy constant $K_1$ for slow-cooled (labeled SC) and quenched (labeled Qu) Fe-Ni alloys are also shown for reference.

minus sign which results from the definitions of the $B_i$'s).

Figure 9 shows the variation of $\lambda_{100}$ and $\lambda_{111}$ with composition in the FCC-Ni-Fe alloys. Note that the magnetization coefficients vanish near 80% Ni, close to the composition at which $K_1 = 0$ in these alloys (Fig. 7). This further establishes the soft magnetic character of these alloys.

## C. Exchange Energy, Magnetic Domain Walls, Domains

The exchange interaction, which couples the directions of the spins in a material to each other, is a quantum-mechanical phenomenon based on the Coulomb interaction between electrons. The result is that spins behave as if there were an interaction of the form

$$\mathcal{H}_{Heisenberg} = -2\sum_{i<J} \mathcal{J}_{ij} S_i \cdot S_j. \tag{16}$$

This form of exchange coupling is called the Heisenberg interaction and $\mathcal{J}$ is the exchange integral over the wave functions responsible for transferring the spin information.

On a macroscopic, continuum level, Eq. (16) can be expressed as an exchange energy density

$$u_{exch} = \frac{U_{exch}}{V} = A\left(\frac{\partial\theta(x)}{\partial x}\right)^2 = A\left(\frac{\nabla \boldsymbol{M}}{M}\right)^2. \tag{17}$$

Here, $A = \mathcal{J}S^2a^2/(2\Omega)$ is the exchange stiffness constant (with $a$ being the lattice constant and $\Omega$, the atomic volume) and $\theta(x)$ is the position-dependent orientation of magnetization in the material. Eq. (17) expresses the fact that there is an energy cost (proportional to $A$) associated with local departures from uniform magnetization orientation.

The magnetostatic energy (Section II.A) can be reduced by the formation of magnetic domains (each domain is uniformly magnetized to saturation along one of the anisotropy easy axes). These magnetic domains are separated by magnetic *domain walls*, which are surfaces over which the magnetization direction gradually changes from one easy direction to another. The equilibrium domain wall thickness is determined by minimizing the integral over the wall of its position-dependent magnetic energy. This energy is made up of local magnetic anisotropy and exchange energy densities. The result is that a 180-degree domain wall in a uniaxial material has domain wall thickness and energy density given, respectively, by

$$\delta_{dw} = \pi\sqrt{\frac{A}{K_u}} \tag{18}$$

and

$$\sigma_{dw} = 4\sqrt{AK_u}. \tag{19}$$

These parameters for Ni and Fe are $\delta_{dw} = 72$ and 30 nm, respectively, and $\sigma_{dw} = 0.7$ and 3.0 mJ/m$^2$, respectively. The thickness and energy density of domain walls are the fundamental parameters that determine the mobility of a domain wall or, on the other hand, the ability of various defects to pin or impede domain wall motion.

When the energy of a domain wall fluctuates *gradually* with position in a material, its motion can be impeded because a gradient in surface energy is equivalent to a pressure on the wall. The Zeeman energy difference across the domain wall is also a pressure working against the wall energy gradient. Balancing these pressures leads to an expression for the coercivity due to *defects that are large relative to the wall width* ($D > \delta_{dw}$) and across which the wall energy varies gradually:

$$H_c \approx \frac{2H_a}{\pi}\frac{\delta_{dw}}{D}\left[\frac{\Delta A}{A} + \frac{\Delta K}{K} + \frac{3}{2}\lambda_s\frac{\Delta\sigma}{K} + \cdots\right]. \tag{20}$$

Eq. (20) indicates that the upper limit to the coercivity is governed by the anisotropy energy density (expressed here through the anisotropy field, $H_a = 2K_u/M_s$) and the ratio of domain wall thickness to defect size. The extent to which $H_c$ approaches this limit is proportional to the magnitude of the relative fluctuation in the strength of parameter, $X$, over a defect size $D$. A more general theory of coercivity treats defects whose properties differ from those of the host material in a *step-like* fashion. This model predicts that *in the small defect regime* ($D < \delta_{dw}$), the coercivity increases linearly with defect size *and for* $D > \delta_{dw}$, the coercivity is independent of defect size [rather than dropping off with the inverse of the defect dimension as for a gradual-fluctuation defect, Eq. (20)]. The

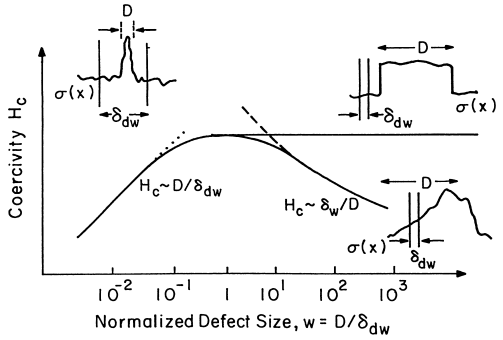



**FIGURE 10** Schematic variation of coercivity with normalized defect size spanning two regions, small defects and large defects relative to wall thickness. The predicted behavior in each case is shown. [From O'Handley, R. C. (2000). "Modern Magnetic Materials, Principles and Applications." Reprinted with permission of John Wiley and Sons.]

predictions of these models are compared schematically in Fig. 10.

## II. MAGNETIC MATERIALS, FUNDAMENTAL PROPERTIES

### A. Electronic Structure of Magnetic Oxides and Metals

In oxides, the electronic wave functions are atomic-like; quantum numbers can be used to describe the energies and spatial distributions of the electronic states. Oxygen (Fig. 11a) has a valence electronic structure of $2p^4$; the $2s^2$ states are at much lower energy. For transition (T) metal atoms (Fig. 11b) the valence electrons include those in the $3d^n$ states ($0 < n < 10$) as well as the lower-energy $4s^2$ states. Figure 11c depicts what happens to these atomic energy levels (T atom, oxygen atom) when transition metal atoms bond with oxygen to form stable ionic compounds (T-oxide, c) or to form a transition metal solid (T-metal, d,e).

When $3d$ T atoms interact with oxygen (whose $2p^4$ electrons are more electronegative than the $3d^n$ states) to form a *T-metal oxide*, bonding and antibonding *s-p-(d)* orbitals are formed; $4s$ (and possibly some $3d$) electrons are transferred to the oxygen atom to completely fill the bonding *s-p-(d)* orbial, which is localized at the oxygen site. The remaining $3d^{n-x}$ electrons on the T ion assume lower energies than before the charge transfer and are localized more strongly there because the core potential now exceeds the valence electronic charge. This electronic stabilization and the Coulomb attraction between the $O^{2-}$ and $T^{x+}$ ions are responsible for the binding energy of the T-metal oxide. The $3d^{n-x}$ states lie in the energy gap between the bonding and antibonding *p-d* states (Fig. 11c). These localized *d* states may be split by the symmetry of the crystalline electric field of the surrounding oxygen ions. The crystal field and exchange splittings of these states determine the magnetic moment per T ion, $n_T = (n_d^{\uparrow} - n_d^{\downarrow})\,\mu_B$, as well as other physical properties of the oxide such as color and electrical conductivity.

When transition metal atoms are brought together to form a *solid T metal*, the *s* states ($l = 0$) of different atoms begin to interact first as interatomic distance decreases and the *d* states ($l = 2$) interact at smaller atom separations. This interaction of the electronic states over many atoms spreads their energies over a range (energy band) due to bonding and antibonding interactions among the variously spaced atoms (Fig. 11d). The energy bands of the $4s$ and $3d$ states of the approaching atoms can overlap so that the available valence electrons ($4s^2\ 3d^n$) can be redistributed over the lowest energy states giving

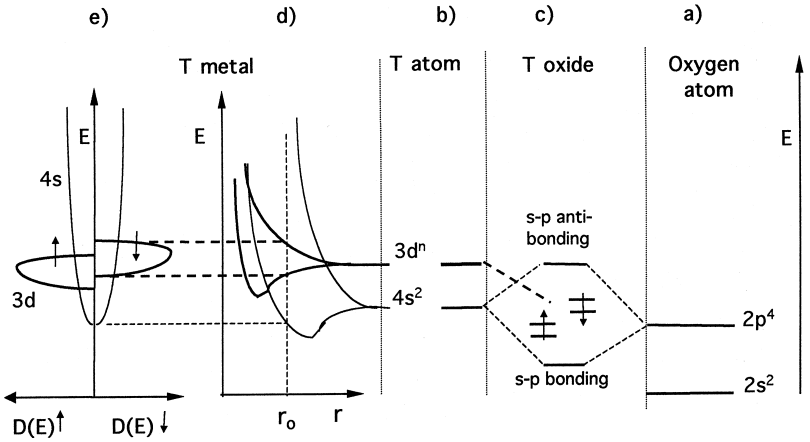



**FIGURE 11** Schematic valence electronic structures of atomic oxygen (a), T atom (b), and how they combine to form T-metal oxides (c), and T metals (d, e).

$4s^{2-y}3d^{n+y}$ per atom. This decreased electronic energy accounts for the stability of the metal at a particular interatomic spacing $r_o$ (Fig. 11d). At $r_o$ the energy overlap of the $d$ and $s$ bands can be plotted versus the density of states (states/energy/atom) as shown in Fig. 11e. The partial filling of the delocalized 4s band accounts for the electrical conductivity, the unequal filling of the exchange-split, narrow $3d^{\uparrow}$ and $3d^{\downarrow}$ bands accounts for the metallic magnetic moment per atom, $n_B = (n_d^{\uparrow} - n_d^{\downarrow})$.

The bonding in *intermetallic compounds* generally has both metallic and covalent character. At the metallic end of the spectrum one finds magnetically soft or hard ordered compounds such as $Ni_3Fe$ ($Cu_3Au$ structure) and CoPt (tetragonal CuAu structure). With increasing covalent bond character (i.e., more directional $p$-$d$ or $p$-$f$ bonds) the transition metal borides and the Heusler alloys based on $Cu_2MnAl$ are found. A mixture of metallic and covalent bonding is found also in *amorphous magnetic alloys* such as $Fe_{80}B_{18}Si_2$ or $Fe_{40}Ni_{40}B_{20}$. Here, the introduction of $p$ character from glass-forming elements (such as boron or phosphorus) adds directional character to the bonding and tends to stabilize the liquid state, forming a eutectic in the phase diagram, in composition ranges between stable compounds (e.g., between BCC Fe and $Fe_2B$).

## B. Selected Fundamental Magnetic Properties

The magnetic properties at a given site in a material are a function of the number, type, distance, and symmetry of the nearest neighbors. This dependence is expressed in terms of the Stoner criterion for magnetic moment formation

$$\mathcal{J}(E_F)\mathcal{D}(E_F) > 1. \tag{21}$$

Here $\mathcal{J}$ is the strength of the *intra-atomic* exchange energy and $\mathcal{D}$ is the density of electronic states, both evaluated at the Fermi level; they are both functions of the local environment. Bonding tends to reduced $\mathcal{D}(E_F)$ and weaken magnetism. The Curie temperature can also be expressed as a function of the local environment by the mean-field expression,

$$T_C = 2z\mathcal{J}\, s(s+1)/k_B \tag{22}$$

In this case, $\mathcal{J}$ is the *interatomic* exchange energy, $z$ is the coordination number, and $s$ is the spin quantum number.

### 1. Spinel Ferrites

Oxides (including the magnetically hard magnetoplumbite, hexaferrites) are very stable and mechanically brittle; they are typically used in polycrystalline (sintered ceramic) form. The most commonly used soft magnetic oxides are the ferrites of the spinel structure, having

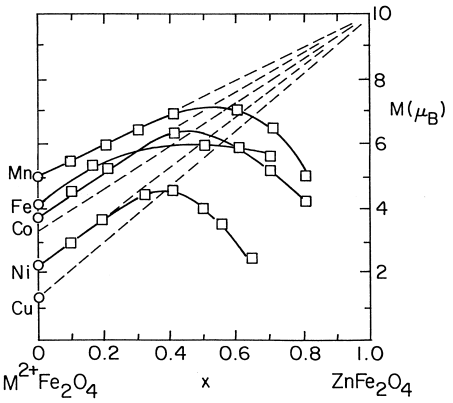



**FIGURE 12** Magnetic moments in transition metal-zinc ferrites as T = Cu, Ni, Co... are substituted for divalent iron. [From Guillaud, P. C. (1951). *J. Phys. Rad.* **12,** 239; Gorter, E. W. (1954). *Phillip. Res. Rpt.* **9,** 295.]

the formula $TO{\cdot}Fe_2O_3$ where T is a divalent transition ion. In this structure the $Fe^{+3}$ and $T^{+2}$ ions can occupy two types of sites having different oxygen coordination: tetragonal "A" sites and octahedral "B" sites. Because of the dominant negative superexchange interaction between the moments on A and B sites, these materials are typically ferrimagnetic. For the $TO{\cdot}Fe_2O_3$ series with $T = Mn^{2+}$ ($3d^5$, $\mu_m = 5\,\mu_B$), $Fe^{2+}$ ($3d^6$, $\mu_m = 4\,\mu_B$), ... $Cu^{2+}$ ($3d^9$, $\mu_m = 1\,\mu_B$), the two $Fe^{3+}$ ions occupy different types of sites so their moments (5 $\mu_B$ each) cancel. The net magnetization is given by the moment on the $T^{2+}$ species (see Fig. 12). As $Zn^{2+}$ is substituted for the $T^{2+}$ species, the Zn ion displaces the $Fe^{3+}$ ions from A to B sites because, among the most common ions, $Zn^{2+}$ has the strongest chemical affinity for the A site. As a result, the net moment per formula unit (FU) of the compounds of the series $T_{1-x}Zn_xO{\cdot}Fe_2O_3$ increases with $x$ even though $Zn^{2+}$ bears no moment. At higher Zn concentrations, the moment on the A sites becomes small enough to allow the antiferromgnetic exchange interaction that exists among the moments on the B sites to dominate, thus reducing the net moment per FU. Of the oxides in these series, Ni-Zn ferrites and Mn-Zn ferrites are widely used for soft magnetic applications because of their relatively large magnetization densities (few ferrites exhibit $B_s > 0.4$ T at room temperature), high electrical resistivity, and small values of $K$ and $\lambda$.

The temperature dependence of the specific magnetization in ferrites is generally more complicated than that of a single-sublattice ferromagnet (cf. Fig. 2). In ferrimagnetic materials the moments on the two sublattices may show different temperature dependences leading the net moment to show compensation temperatures and $dM/dT > 0$ over some temperature ranges (schematics at right in Fig. 13).

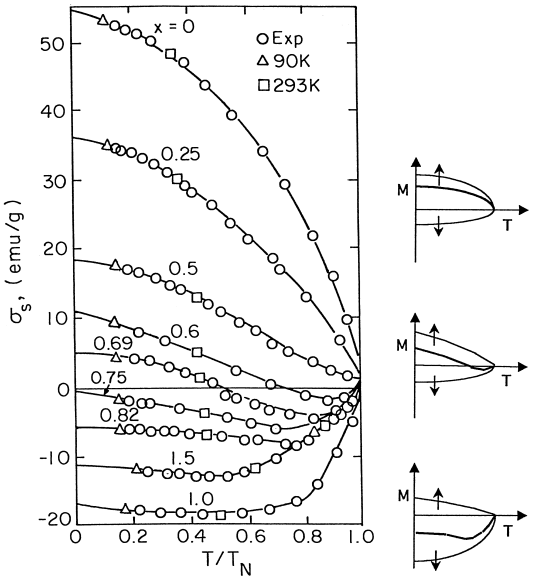



**FIGURE 13** Temperature dependence of the magnetization in $NiFe_{2-x}V_xO_4$. [From Blasse, G., and Gorter, E. W. (1962). *J. Phys. Soc. Jpn.* **16,** Suppl. B-1, 176.] The schematic *M–T* curves at right illustrate how the magnetization of two sublattices having different temperature dependences combine to give the unique features observed in the data.

## 2. Transition Metals and Alloys

In 3*d* transition metal alloys, the saturation magnetic moment $\mu_B$ (per average T species) varies as indicated in Fig. 14, which is called the Slater-Pauling curve. For alloys to the right of the peak moment, the Fermi level cuts across only the spin-down *d* band. As the *d* electron concentration, $n_d = n_v - n_s \approx n_V - 1$, increases from 7.5 e/atom oward 10 e/atom, the number of holes per atom (unpaired spins), and hence the moment per atom, decreases linearly toward zero at $Ni_{40}Cu_{60}$. The systematic behavior on the left-hand side of the curve involves holes in both spin subbands and is not as easily explained in a simple model. Band structure calculations (Diederichs. P. H. *et al.*, 1991) are able to explain most of the features in Fig. 14.

Figure 14 could lead one to expect that every atom, for example, each Fe and Ni atom in a $Ni_{50}Fe_{50}$ alloy, has the same magnetic moment, namely $\mu_{T\text{-ave}} \approx 1.6\ \mu_B$. This is, in fact, *not* the case. Spin-polarized neutron scattering shows that there are different local magnetic moments on the different T species with their sum given by the data in Fig. 14. For example, $Ni_{50}Fe_{50}$ shows $\mu_{Fe} \approx 2.5\ \mu_B$, and $\mu_{Ni} \approx 0.7\ \mu_B$.

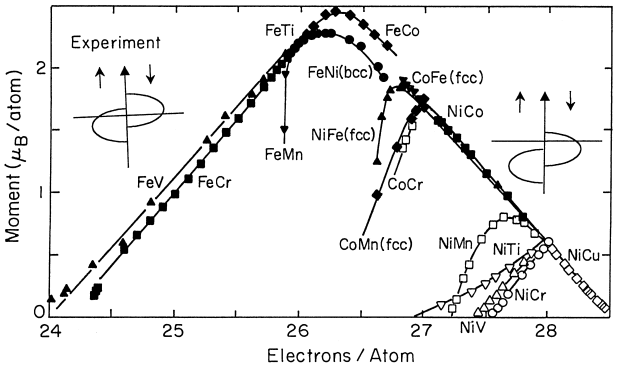



**FIGURE 14** The Slater-Pauling curve showing moment per atom (in Bohr magnetons) for metallic 3*d* alloys as a function of valence electron concentration or alloy composition. [From Diederichs, P. H. *et al.* (1991). *In* "Magnetism in the Nineties." (Freeman, A. J. and Gschneider, K. eds.), North-Holland.] The inset schematic band structures illustrate the main difference between alloys to the left and right of the peak in the curve.

## 3. Intermetallic Compounds and Amorphous Alloys

The dependence of magnetic moment on the local environment (number, type, distance, and symmetry of nearest neighbors) is shown clearly for a series of Mn-containing *intermetallic compounds*. The magnetic moment per Mn atom shows a linear decrease with increasing Pauling valence (Fig. 15). The Pauling valence is a function of interatomic distance and coordination and is a measure of covalent (as opposed to ionic) bonding. Increased covalent bonding hinders magnetic moment formation because covalent bonds have spin-paired electrons. A similar model (Corb, B. W. *et al.*, 1983) has been shown to apply to 3*d* metalloid alloys; it also shows that increasing the number of bonding orbitals per T atom decreases its moment.

The variation of magnetic moment with valence electron concentration in *crystalline* transition-metal borides TB and $T_2B$, shows a characteristic variation that appears similar to that of the Slater-Pauling curve but with the peak moment suppressed and shifted more toward lower

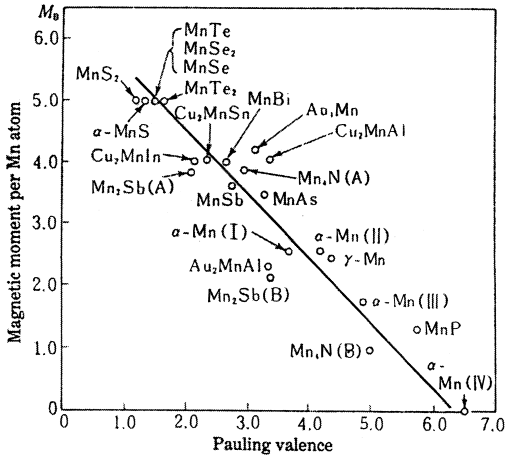



**FIGURE 15** Variation of Mn magnetic moment with Pauling valence. [From Mori, N., and Mitsui, T. (1968). *J. Phys. Soc. Jpn.* **25,** 82.]

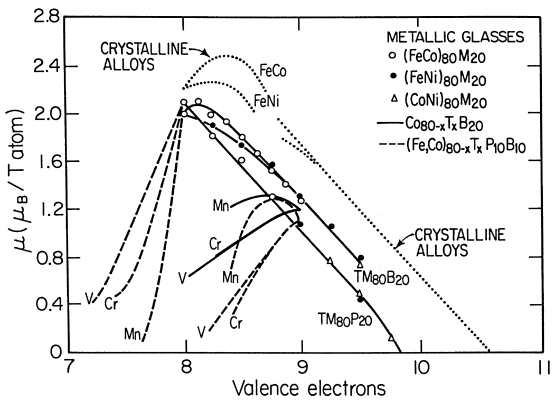



**FIGURE 16** Variation of magnetic moment per transition metal atom in crystalline and amorphous alloys as a function of number of valence electrons, $n_V$. $n_V = 8, 9$, and 10 corresponding to Fe, Co, (or $Fe_{0.5}Ni_{0.5}$) and Ni, respectively. The data for crystalline materials are based on Fig. 14. [From O'Handley, R. C. (2000). "Modern Magnetic Materials, Principles and Applications." Reprinted with permission of John Wiley and Sons.]

T-valence with increasing boron content (Cadeville, M. C. *et al.*, 1966). This shift is not due to the effect of the presence of boron on the valence electron concentration (the shift would require boron to increase the valence electron concentration). The shift is more subtle than that: these compounds show an entirely different electronic structure than the $3d$ metallic alloys.

*Amorphous magnetic alloys* combine some of the bonding characteristics of both $3d$ alloys and intermetallic compounds. Figure 16 shows the variation of the low-temperature saturation moment per transition metal atom as a function of T content for amorphous alloys based on boron, $T_{80}B_{20}$, and on phosphorus, $T_{80}P_{20}$. The variation of magnetic moment in crystalline alloys (cf. Fig. 12) is shown as a dotted line for reference. Amorphous $T_{80}B_{20}$ alloys show magnetic moments that are shifted relative to the Slater-Pauling curve in a way that is consistent with data for crystalline TB and $T_2B$ compounds and alloys.

In *rare-earth-transition metal intermetallics* (R-T), the magnetic moments of transition metals couple ferromagnetically with light rare-earth moments ($\boldsymbol{J}_R \cdot \boldsymbol{s}_T > 0$) and antiferromagnetically with heavy rare-earth moments ($\boldsymbol{J}_R \cdot \boldsymbol{s}_T < 0$). Thus the spin-spin coupling between R and T species is always antiferromagnetic (Fig. 17, left). This coupling can be ascribed to the $5d$ conduction electrons of the rare-earth (whose spin is always parallel to that of the $4f$ electrons) and their interaction with the symmetry-compatible $3d^n$ electrons of the transition metal. Exchange between these two sets of $d$ states is invariably antiferromagnetic with respect to the $d$ electrons involved; there are only minority-spin holes in

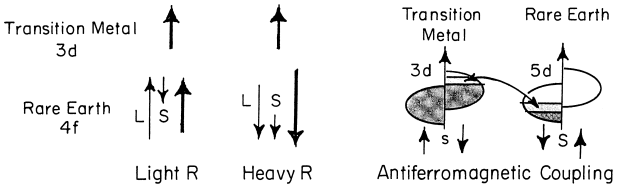



**FIGURE 17** Simplified schematic representation of spin and angular momentum coupling at rare-earth site and antiferromagnetic exchange coupling between R and TM spins. Right, schematic band structure that accounts for antiferromagnetic R-TM spin coupling. [From O'Handley, R. C. (2000). "Modern Magnetic Materials, Principles and Applications." Reprinted with permission of John Wiley and Sons.]

the TM $3d$ orbitals, and spin is conserved in the dominant hopping process (Fig. 17, right). The antiferromagnetic R-$5d$–T-$3d$ interaction explains the *ferromagnetic* coupling of T moments to light ($J = L - S$) R moments and *antiferromagnetic* coupling to heavy ($J = L + S$) R species. The net result of R-T exchange coupling is generally larger magnetic moments for intermetallics of light rare-earths and transition metals.

## III. TECHNICAL PROPERTIES

### A. *B-H* Loops and Magnetic Domains

Consider a magnetic material in the demagnetized state ($B = 0$, $H = 0$ in Fig. 18). Even though the local magnetic moments show long-range order (typically over several microns), the demagnetized state can be achieved by the formation of magnetic domains. Domains are regions of homogeneous magnetization separated by domain walls, surfaces over which the orientation of the atomic moments

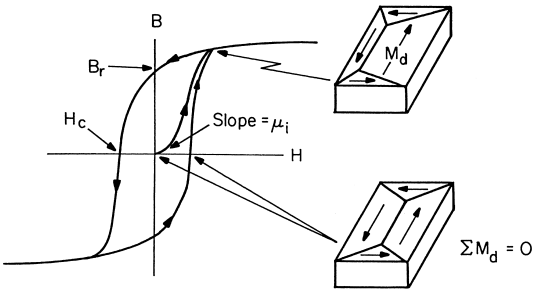



**FIGURE 18** Hysteresis loop of a magnetic material showing the variation of $B$ with changing $H$. Initial magnetization curve from the demagnetized state is shown with the initial permeability $\mu_i$, indicated. The remanence $B_r$ and coercive field, $H_c$, are indicated. The approximate domain structures are indicated at right for demagnetized state and for approach to saturation. [From O'Handley, R. C. (2000). "Modern Magnetic Materials, Principles and Applications." Reprinted with permission of John Wiley and Sons.]

changes relatively abruptly (Section II.C). The vector magnetizations of the domains sum to zero in the demagnetized state. Application of a field to a demagnetized sample results in the motion of domain walls so as to expand the volume of those domains having the largest component of $\boldsymbol{M}$ along $\boldsymbol{H}$. The initial flux density, $B = \mu_o\,(H + M) = \phi/\text{area}$, produced in response to a small field, $H$, defines the initial permeability, $\mu_i = B/H]_{H\approx 0}$. At stronger fields, the permeability increases to its maximum value $\mu_{max}$. The relative initial permeability, $\mu_r$, can be as great as $10^5$ or $\mu_r\mu_o \approx 10^{-1}$ in some materials. When most domain wall motion has been completed, there often remain domains with nonzero components of magnetization at right angles to the applied field direction. The magnetization in these domains must be rotated into the field direction to minimize the potential energy—$\boldsymbol{M}\cdot\boldsymbol{B}$. This process generally costs more energy than wall motion.

Upon decreasing the magnitude of the applied field from saturation, the magnetization rotates back toward its "easy" directions, generally without hysteresis. As the applied field decreases further, new domains may nucleate and domain walls begin moving back across the sample. Because energy is lost when a domain wall jumps abruptly from one local energy minimum to the next (Barkhausen jumps), wall motion is an hysteretic or lossy process. The flux density and magnetization remaining in the sample when the applied field is zero are called the residual flux density, $B_r$, and remanence, $M_r$, respectively. The reverse field needed to restore $B$ to zero is called the coercivity, $H_c$. (The field needed to restore $M$ to zero is called the intrinsic coercivity, ${}_iH_c$. The distinction between $H_c$ and ${}_iH_c$ is important only in permanent magnets, because in a soft magnetic material $H_c \ll M$ so that $M = 0$ for essentially the same field that gives $B = 0$). $H_c$ is a good measure of the ease or difficulty of magnetizing a material.

In some soft magnetic materials, domain walls can be moved with fields of order 0.1 A/m. In general, defects such as grain boundaries and precipitates cause the wall energy to depend on position, so in most soft magnetic materials, higher fields (of order 10–1000 A/m) are required to move domain walls or rotate the magnetization vector.

## B. Soft Magnetic Materials

A soft magnetic material is one for which $H_c$ is less than or equal to about $5 \times 10^3$ A/m. Pure iron is the prototypical soft magnetic material. It has a very high saturation flux density, $B_s = \mu_o M_s = 2.2$ T, and its cubic crystal structure leaves it with a small magnetocrystalline anisotropy, $K_1 = +4.8 \times 10^4$ J/m$^3$, and small magnetostriction, $\lambda_{100} = +21 \times 10^{-6}$. Domain images from a (100) iron −3% Si single crystal taken by scanning electron microscopy with spin polarization analysis (SEMPA) are shown in Fig. 19. The magnetization within each domain follows the easy ⟨100⟩ direction (dictated by $K_1 > 0$) leading to 90° and 180° domain walls. Addition of small amounts of silicon improves the usefulness of iron for a number of soft magnetic applications at modest frequencies (typically for 50–60 Hz transformers).

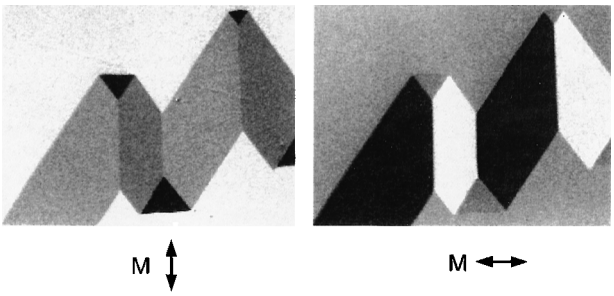



**FIGURE 19** Magnetic domains at the surface of a 3% Si-Fe crystal taken by scanning electron microscopy with spin polarization analysis (SEMPA). Crystallographic ⟨100⟩ axes lie in the image plane along the horizontal and vertical directions. Left panel shows magnetic contrast when the instrument is sensitive to the horizontal component of magnetization: dark is magnetized to the left, light to the right. In the right panel, the contrast is sensitive to vertical component of magnetization: dark is magnetized down, light is magnetized up. (Courtesy of Celotta, R. J. *et al.*, unpublished.)

There are three major Fe-Ni or permalloy compositions of technical interest:

1. % nickel permalloys (e.g., Supermalloy®, Mumetal®, Hi-mu 80®). The 80% nickel permalloys are very important because the magnetostriction and magnetocrystalline anisotropy both pass through zero near this composition (see Figs. 7 and 9). They show a saturation flux density of about 1 T. These alloys are used where the highest initial permeability is required. This includes inductors for power supplies and circuits as well as magnetic recording read and write heads.
2. 65% nickel permalloys (e.g., A Alloy®, 1040 Alloy®). The 65% nickel permalloys show a strong response to field annealing while maintaining small anisotropy.
3. 50% nickel permalloy (e.g., Deltamax®). The 50% nickel permalloys are important because of their higher flux density ($B_s = 1.6$ T) as well as their responsiveness to field annealing to give a very square loop.

The equiatomic BCC FeCo alloys (called Permendurs, Fig. 20a) have very high saturation flux density ($B_s \approx 2.4$ T) as well as relatively low magnetic anisotropy. While the magnetocrystalline anisotropy (as well as stress-induced-anisotropy) limits the soft magnetic properties

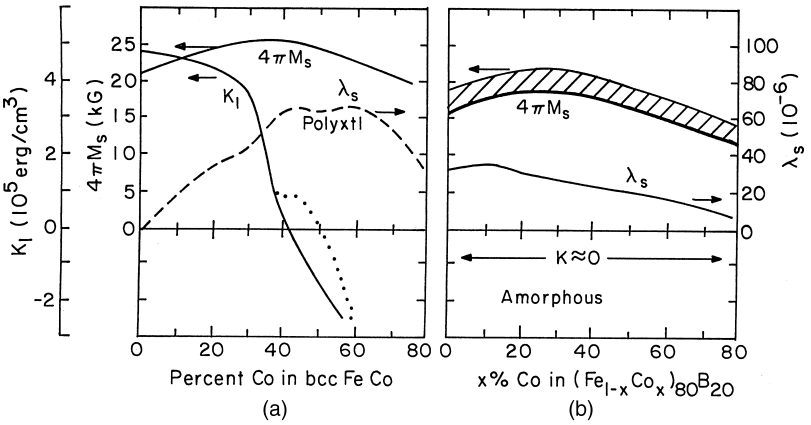



**FIGURE 20** (a) Magnetic properties of BCC Fe-Co alloys. Anisotropy and magnetostriction. [After Hall, R. C. (1959). *J. Appl. Phys.* **30,** 816; dotted lines are for CsCl-ordered phase.] (b) Magnetic properties of amorphous Fe-Co-B alloys for which the magnetocrystalline anisotropy is essentially zero over the entire composition range. [From O'Handley, R. C. *et al.*, (1979). *J. Appl. Phys.* **50,** 3603; O'Handley, R. C. (2000). "Modern Magnetic Materials, Principles and Applications." Reprinted with permission of John Wiley and Sons.]

of these alloys, grain size and texture are the primary factors determining the technical magnetic properties actually attained.

*Amorphous metallic alloys* are materials that are rapidly quenched from the melt or deposited from the vapor states as thin films so that their atomic structure lacks the long-range order of a crystalline solid. Without *long-range* order, amorphous alloys have no magnetocrystalline anisotropy. Thus amorphous metallic alloys based on transition metals can show a very easy magnetization process. Magnetoelastic anisotropy remains as the major impediment to easy magnetization in amorphous alloys. Figure 21 shows that the coercivity of amorphous $Co_{80-x}Fe_xB_{20}$ alloys is minimized near the composition $x = 4$ for which $\lambda \approx 0$.

The presence in amorphous magnetic alloys of significant concentrations of non-magnetic, glass-forming species reduces the saturation magnetization but often has other beneficial effects (increased resistivity, decreased magnetostriction). The high electrical resistivity of amorphous alloys (120–150 $\mu\Omega$cm) compared to Si-Fe (30−50 $\mu\Omega$cm) and iron-nickel alloys (20 $\mu\Omega$cm) makes them attractive for high-frequency operation. Reasonably strong magnetization can be realized in a variety of amorphous alloys based on iron, cobalt, and/or nickel (cf. Figs. 16 and 20b for Fe-Co-B amorphous alloys).

The most widely used *soft magnetic ferrites* are based on manganese-zinc ferrite and nickel-zinc ferrite. These materials have the spinel structure of $Fe_3O_4$. Figure 22 shows the dependence of magnetostriction and permeability on $Fe_2O_3$ content in $[(MnO)_{0.7}(ZnO)_{0.3}]_{1-x}\cdot(Fe_2O_3)_x$. In these compositions, $K_1$ is small so the permeability peaks near the composition at which $\lambda_s = 0$ (cf. Fig. 21). The insulating character of oxide magnets makes them useful at frequencies in excess of the MHz range.

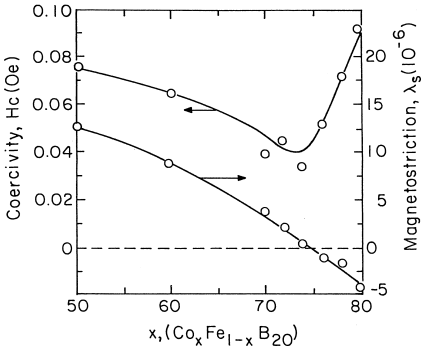



**FIGURE 21** Variation of coercivity (left scale) and magnetostriction (right scale) with Fe/Co ratio in amorphous $(CoFe)_{80}B_{20}$ alloys. [From O'Handley, R. C. *et al.* (1976). *IEEE Trans.* **MAG-12,** 924.]

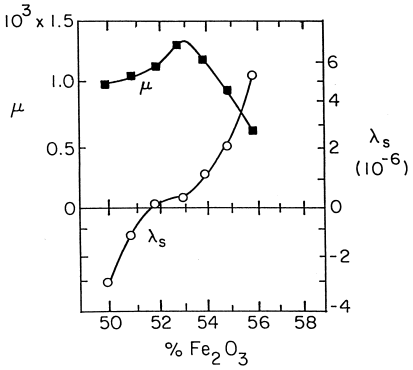



**FIGURE 22** Variation of permeability and magnetostriction with iron oxide content in MnZn ferrites. [After Guillaud, P. C. (1957). *Proc. IEEE* **104B,** 165.]

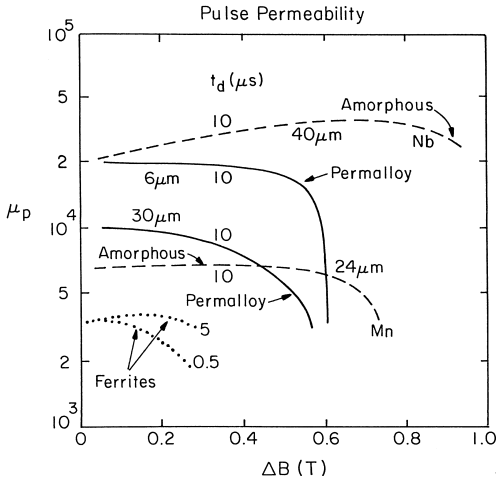



**FIGURE 23** Pulse permeability versus maximum flux swing for three classes of soft magnetic materials: amorphous metallic alloys (Co-Fe-Nb-B-Si and Co-Mn-Fe-Mo-b-Si), crystalline Mo-permalloy (Ni-Fe-Cu-Mo), and ceramic Mn-Zn ferrites. Data for two thicknesses are given in each case. [From Boll, R., and Hiltzinger, H. R. (1983). *IEEE Trans.* **MAG-19,** 1946.]

Figure 23 compares the AC performance of representative soft magnetic materials in a plot of pulse permeability versus flux density with thickness (or particle size for ferrites) given as a parameter. While ferrites generally have an advantage at higher frequencies, in pulse power applications, where large $dB/dt$ is important, the other classes of materials show advantage.

## C. Nanomagnetic Materials

Composite materials in which one of the component microstructures has one, two, or three nanoscale dimensions allow new properties and functions to be realized that may not be achievable in simpler structures or by changing composition in a single-phase alloy. Controlling structure and feature sizes at the nanometer scale in magnetic materials is particularly effective because many of the important magnetic length scales that govern magnetic properties fall in the nanometer range.

1. The *domain wall thickness*, Eq. (18), can range from 5 nm to several hundred nanometers on going from hard to soft magnetic materials.
2. The *exchange length*, $l_{ex} = \sqrt{2A/(\mu_o M_s^2)}$, is the minimum distance over which the local moment direction can change to reduce magnetostatic energy. Exchange lengths are typically less than 10 nm.
3. The critical size below which magnetic particles are unable to support a domain wall, the single-domain limit, is typically in the range of 20–40 nm.
4. Below a certain particle size called the *superparamagnetic* limit, thermal energy, $k_B T$ is sufficient to demagnetize a particle that would otherwise retain a set direction of magnetization by virtue of its anisotropy. This size depends exponentially on the time scale over which thermal demagnetization occurs, but is typically less than 10 nm.

*Nanostructured soft magnetic materials* can be engineered by creating single-domain crystalline particles in an amorphous magnetic matrix. The prototype of this class of materials is the nanocrystalline magnet Fe-Si-B-Nb-Cu ($\alpha$-$Fe_3Si$ particles in a matrix of residual amorphous phase (Yoshizawa, Y. *et al.*, 1988). The Cu and Nb are added to what is otherwise a common glass-forming composition in order to enhance nanocrystal nucleation and to retard growth of those nuclei, respectively. In these materials, the properties can vary widely depending upon the size of the nanocrystalline particles and the dimensions and magnetic properties of the intervening amorphous matrix.

The effect of the intergranular amorphous phase as an exchange-coupling medium between the single-domain particles is illustrated by experiments shown in Fig. 24. Above $T_C$ of the Nb-rich, amorphous matrix, the material is an assembly of noninteracting, single-domain, $Fe_3Si$ particles; as such, it shows low coercivity. Below $T_C$ of the intergranular phase, the single domain particles are exchange coupled to each other so that they switch coherently in an applied field. Here too, the coercivity is relatively small. A peak in coercivity appears near the decoupling temperature. It is due to the independent switching of adjacent grains, creating domain walls between them. The presence of domain walls leads to discontinuous response and hysteresis in the system.

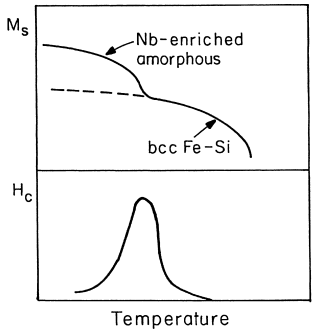



**FIGURE 24** Schematic summary of the results of measurements on nanocrystalline Fe-B-Si-Nb-C alloys having different Nb contents. [From Skorvanek, I. *et al.* (1995). *J. Magn. Magn. Mater.* **140–144,** 467.]

Starting with amorphous Co-Nb-B alloys, the coercivity increases by nearly four orders of magnitude through various stages of devitrification corresponding to various nanostructure length scales. The coercivity initially increases consistent with the $D^6$ power law derived theoretically (Herzer, 1993). Above a peak, $H_c$ drops off more gradually, consistent with the $1/D$ behavior predicted for defect sizes greater than the domain wall width (Fig. 10).

Iron-rich amorphous and nanocrystalline alloys generally show larger magnetization than those based on Co or Ni. However, iron-rich amorphous alloys generally have fairly large magnetostriction, limiting their permeability (Fig. 20b). Formation of a nanocrystalline iron-rich alloy can lead to a dramatic reduction in magnetostriction, thus favoring easy magnetization (Hasegawa, N. *et al.*, 1993).

Most magnetically soft nanocrystalline systems are based on $DO_3$ crystallites (e.g., $Fe_3Si$) in an amorphous matrix. However, the presence of 25 atom% Si significantly reduces the saturation flux density of the alloy (Hasegawa, N. *et al.*, 1991). Spinodal decomposition in metastable amorphous transition metal-carbon alloys can be used to form nanocrystalline alloys of the general formula $(Fe, Co)_{81}Ta_9C_{10}$. Annealing at 550°C for 20 min results in primary crystallization of $\alpha$–Fe (or $\alpha-$FeCo) particles measuring 5–10 nm in diameter and dispersed transition metal carbide nanocrystals (generally at triple junctions). The primary nanocrystals share grain boundaries making grain-to-grain exchange coupling stronger. The softest magnetic properties are obtained for the smallest nanocrystalline grain sizes. In these alloys the stable carbide grain boundary phase inhibits grain growth just as Nb does in FeBSiCuNb.

Other magnetic nanocrystalline materials show potential for increased saturation flux density. Alloys of the type $Fe_{88}Zr_7B_4Cu_1$ (Suzuki, K. *et al.*, 1991) and $Fe_{44}Co_{44}Zr_7B_4Cu_1$ (Willard, M. A. *et al.*, 1998) crystallize to high-magnetization $\alpha$-Fe nanostructures. The latter show saturation flux densities in excess of 2 T.

## D. Hard Magnetic Materials

Permanent magnets are used to produce strong fields without having to apply a current to a coil. Hence they should exhibit a strong net magnetization. It is also important that the magnetization be stable in the presence of external fields. These two conditions indicate that the *B-H* loop should have large values of remanent induction, $B_r$, and coercivity $H_c$, respectively. Permanent magnets have coercivities in the range of $10^4$–$10^6$ A/m.

The shapes of the *M-H* loops in the second quadrant (which determines demagnetization behavior) are compared in Fig. 25 for some common permanent magnets. The permanent magnets with the highest energy products are based on $Fe_{14}Nd_2B_1$. Figure 25b shows the evolution of the maximum energy product in permanent magnets over recent decades.

The earliest permanent magnets were the natural forms of magnetite, $Fe_3O_4$. More recently, the high-anisotropy hexagonal ferrites (barium or strontium hexaferrite) (Kojima, H. *et al.*, 1982) and the magnets of the alnico family were developed (McCurrie, R. A., 1982). The hexagonal ferrites derive their magnetic hardness from their very large magnetocrystalline anisotropy (Table I) and relatively weak magnetization. The alnico magnets achieve high-coercivity by the formation of a high-aspect-ratio structure consisting of columns of nonmagnetic $Ni_3Al$ in a matrix of $\alpha$-Fe by spinodal decomposition from a Heulser composition such as $Fe_2NiAl$. Alnico magnets can achieve high remanence by processing in a mgentic field to achieve strong orientation of the columnar microstructure.

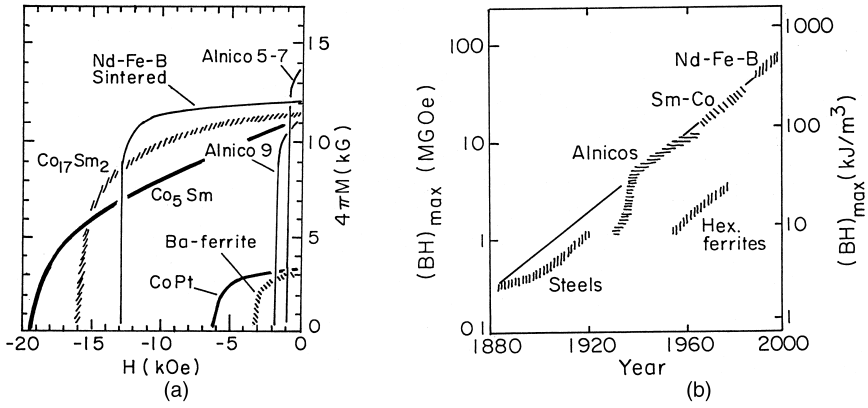



**FIGURE 25** (a) Second quadrant *M-H* loops of some common permanent magnets. (b) Increase in $(BH)_{max}$ of permanent magnets over recent decades. [From O'Handley, R. C. (2000). "Modern Magnetic Materials, Principles and Applications." Reprinted with permission of John Wiley and Sons.]

Hard magnets of the type $SmCo_5$ boast the highest uniaxial anisotropies of any class of magnets, $K_u \approx 10^7$ J/m$^3$. On the other hand, phases in the class $Sm_2Co_{17}$ exhibit higher flux density and Curie temperature. Most *Co-R magnets* are, in fact, multiphase composites of these two structures and sometimes other phases.

Because of the large magnetic anisotropy of $SmCo_5$, a 180° Bloch wall in this material should have a width of only 3.1 nm. Further, the domain wall energy density is 40 mJ/m$^2$ (40 erg/cm$^2$), 100 times that of a soft material. Such walls are not easily nucleated and thus the magnetization process in single-phase $RCo_5$ intermetallics is limited by reversal domain nucleation. Once nucleation occurs, domain walls move relatively easily until they reach a grain boundary or other defect. Hence, initial efforts to produce cobalt rare-earth magnets focused on the fabrication of single-domain $SmCo_5$ particles. Small substitutions of Cu for Co would lead to the precipitation of a nonmagnetic phase that increases the coercivity. They showed that heat treatment of $R(CoCu)_5$ magnets results in precipitation of a dispersion of fine ($d \approx 10$ nm) second-phase, Cu-rich particles in a $R_2Co_{17}$ matrix having a grain size of order 10 $\mu$m (Fig. 26). The coercivity mechanism becomes domain wall pinning on the small nonmagnetic $SmCu_5$ particles.

The more-recently-developed magnets based on $Fe_{14}Nd_2B$ exhibit the highest energy products achieved so far in permanent magnets. Their development came as a result of the cost and limited world supply of cobalt. Commercial *Fe-Nd-B magnets* based on sintering and melt spinning are available.

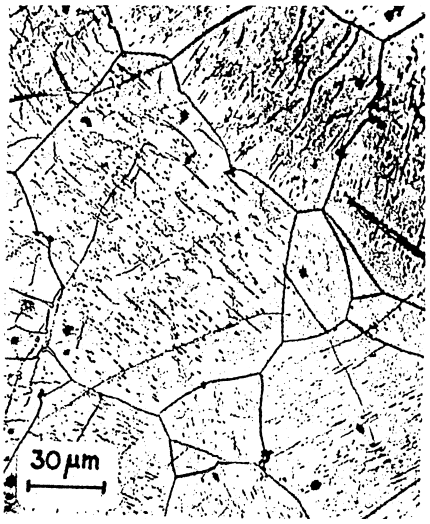



**FIGURE 26** Precipitation microstructures of 1–5 type R-Co magnets, bulk-hardened with copper: $SmCo_{3.5}Cu_{1.0}Fe_{0.5}$ homogenized at 1100°C for 3 hr, quenched and aged at 525°C. The fine (10-nm) dark precipitates are platelets of Cu-rich, $Sm(CuCo, Fe)_5$. [After Strnat, K. (1988). *In* "Ferromagnetic Materials," Vol. 4 (Wohlfarth, E. P. and Buschow, K. H. J., eds.), Elsevier Press, North Holland, Amsterdam.]

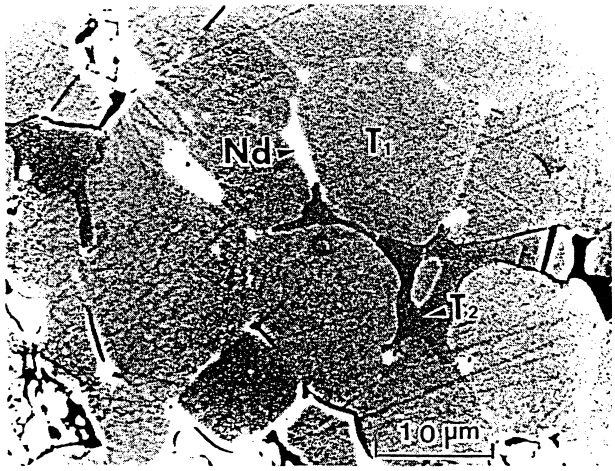



**FIGURE 27** X-ray composition micrograph of a sintered $Nd_{0.15}Fe_{0.77}B_{0.08}$ magnet; $T_1$, $T_2$, and Nd denote $Nd_2Fe_{14}B$, $Nd_{1+e}Fe_4B_4$, and an Nd-rich phase, r espectively. [From Sagawa, *et al.* (1987). *Jpn. J. Appl. Phys.* **26,** 785.]

The attractive permanent magnet properties of $Fe_{14}Nd_2B_1$ magnets arise from several factors: (1) the large uniaxial magnetic anisotropy ($K_u = +5 \times 10^6$ J/m$^3$) of this tetragonal phase, (2) the large magnetization ($B_s = 1.6$ T) owing to the ferromagnetic coupling between the Fe and Nd moments; and (3) the stability of the 14-2-1 phase which allows development of a composite microstructure characterized by 14-2-1 grains separated by nonmagnetic B- and Nd-rich phases (Fig. 27) which tend to decouple the magnetic grains. Figure 28 shows

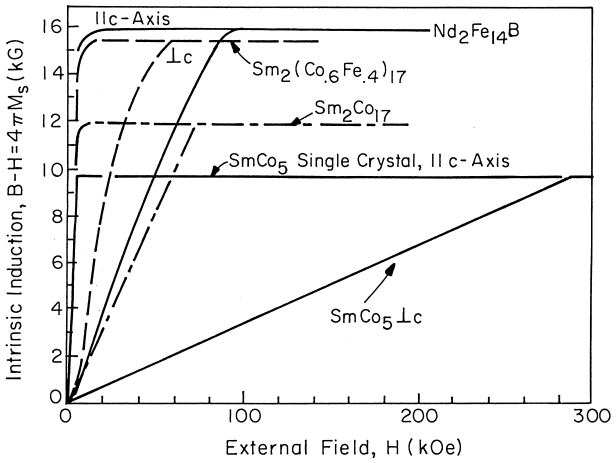



**FIGURE 28** The easy and hard axis magnetization curves for the two principal Co-Sm magnets and $Fe_{14}Nd_2B$. [After Strnat, K. (1988). *In* "Ferromagnetic Materials," Vol. 4 (Wohlfarth, E. P. and Buschow, K. H. J., eds.), p. 131. Elsevier Press, North Holland, Amsterdam.]

**TABLE V Comparison of Some Magnetic Properties for $SmCo_5$, $Sm_2(CoFe)_{17}$, and $Fe_{14}Nd_2B$ Permanent Magnets at 25°C**

| | $\mu_o M_s$ (T) | $T_C$ (°C) | $K_u$ (MJ/M$^3$) | $_iH_c$ (MA/m) | | $(BH)_{max}$ (MGOe) | |
|---|---|---|---|---|---|---|---|
| | | | | Isotr. | Aligned | 2-D | 3-D |
| $SmCo_5$ | 1.0 | 685–700 | 10 | 0.8–1 | 2.9 | 14–16 | 18–24 |
| $Sm_2(CoFe)_{17}$ | 1.2–1.5 | 810–970 | 3.3 | 1–1.3 | 2.4 | 16–20 | 24–30 |
| $Fe_{14}Nd_2B$ | 1.6 | 312 | 5 | — | 1.2–1.6 | 34–45 | |

the easy and hard axis magnetization curves for the two principal Co-Sm magnets and $Fe_{14}Nd_2B$ (see also Table V).

Many permanent magnets are essentially nanostructured materials consisting of high-anisotropy, single-domain particles (or multidomain particles with pinned domain walls) that are magnetically decoupled from each other.

## IV. RELATED MAGNETIC PHENOMENA AND APPLICATIONS

### A. Thin Film and Surface Magnetism

Research in magnetic thin films has grown in sophistication as advances in ultrahigh vacuum technology and surface characterization techniques kept pace with improvements in theoretical understanding and computational accuracy.

At a clean surface, the atoms of a material have reduced nearest-neighbor coordination, reduced symmetry and generally different bond lengths compared to interior atoms. These changes generally result in narrower $d$ bands and increased density of states, which in turn favor increased saturation magnetic moment and lower Curie temperature. First principles band structure calculations at surfaces bear this out. It is very difficult experimentally to isolate the magnetic moment of surface atoms sufficiently to detect a 10% increase in moment.

Near a surface, the narrower, more atomic-like $d$ bands give rise to increased angular momentum; the reduced symmetry near a surface generally leads to stronger magnetic anisotropy. The ratio $L_z^2/(L_x^2 + L_y^2)$ is increased because $\boldsymbol{L} = \boldsymbol{r} \times \boldsymbol{p}$ and momentum perpendicular ot the surface must be reduced. The surface magnetic anisotropy can be expressed as an energy density in powers of the direction cosines of $\boldsymbol{M}$ [cf. Eq. (11)] having the appropriate uniaxial symmetry:

$$\sigma = K_1^s \alpha_3^2 + K_2^s \alpha_3^2 + K_3^2 \alpha_1^2 \alpha_2^2 + \cdots \qquad (23)$$

With this convention, $K_1^s > 0$ ($<0$) favors in-plane (perpendicular) magnetization near the surface. The first term in Eq. (23) is most often measured in thin film systems and is clearly a function of the nature of the material across the interface from the magnetic species. Compared to enhanced surface magnetic moments, surface anisotropy is more readily determined experimentally but is harder to calculate from first principles. The literature abounds with observations of surface anisotropy effects in thin films and at surfaces.

The first term in Eq. (23) is often expressed as: $K^{\text{eff}} \sin^2\theta$, with the contributions to the effective anisotropy defined by

$$K^{eff} = -2\pi M_s^2 + 2B_1 \varepsilon_{xx} + 2\frac{K^s}{t}. \qquad (24)$$

Here $B_1$ is the magnetoelastic coupling coefficient, $\varepsilon_{xx}$ is the in-plane film strain, $K^s$ is the surface anisotropy [$-K^s$ in Eq. (23)]. Measured values of $K^s$ are typically of order 0.1–0.5 mJ/m$^2$.

Figure 29 shows the measured effective, first-order uniaxial anisotropy for Cu/Ni/Cu(001) epitaxial films. The positive values of $K^{eff}$ over a relatively wide thickness range, 2 nm $< t_{\text{Ni}} <$ 140 nm, imply perpendicular magnetization there. Here the perpendicular magnetization results from small magnetostatic energy and large positive magnetoelastic energy.

There has been considerable interest in the switching of the easy axis of magnetization in ultrathin, epitaxial Fe films ($t_{\text{Fe}} < 1$ nm). This phenomenon has been studied as a function of film thickness and temperature and the results point to the origin of perpendicular magnetization in surface anisotropy.

Co thin films layered with Au (den Broeder, F. J. A. *et al.*, 1988) or with Pd (Engle, B. *et al.*, 1991) can show strong perpendicular magnetic anisotropy up to 2 nm in thickness. In these systems both magnetoelastic energy and interface anisotropy appear to be important for perpendicular magnetization.

### B. Electronic Transport in Magnetic Materials

The electrical transport properties of magnetic materials (called galvanomagnetic effects) are interesting because the response of the conduction electrons can depend on

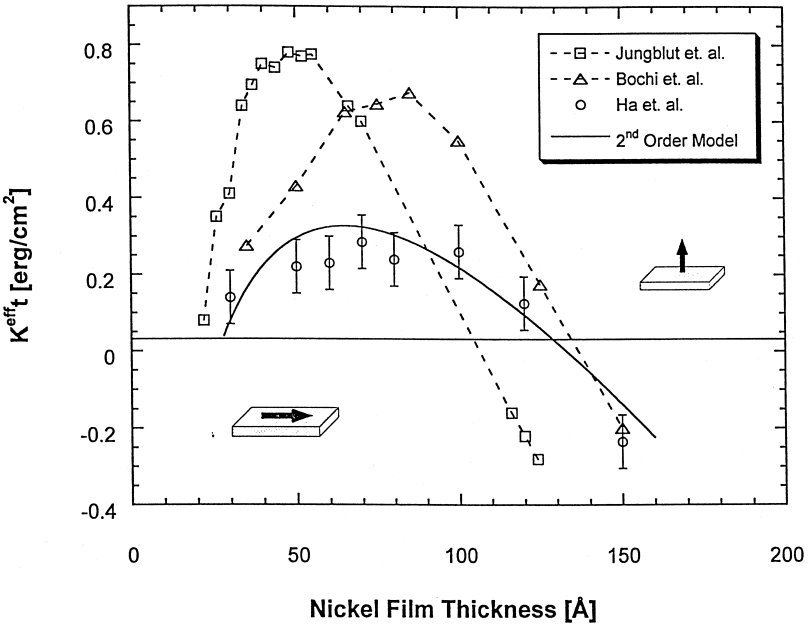



**FIGURE 29** Variation of effective anisotropy times Ni thickness ($K^{eff} \times t$) versus Ni thickness in Cu/Ni/Cu(001) sandwiches [after Jungblut, R. *et al.* (1994). *J. Appl. Phys.* **75,** 6424; Bochi, G. *et al.* (1995). *Phys. Rev. B* **52,** 7311; and Ha, K. *et al.* (1999). *J. Appl. Phys.* **85,** 5282], using *in situ* magneto-optic Kerr effect, *ex situ* vibrating-sample magnetometry, and *ex situ* torque magnetometry, respectively, to determine the effective anisotropy. $K^{eff} \times t > 0$ implies perpendicular magnetization in the uniaxial approximation. [From O'Handley, R. C. (2000). "Modern Magnetic Materials, Principles and Applications." Reprinted with permission of John Wiley and Sons.]

their spin direction. The unique features result from the presence of both $s$ and $d$ electrons at the Fermi energy (Fig. 11e) as well as differences in the density of $3d\uparrow$ and $3d\downarrow$ states at $E_F$. In magnetic transition metals, the two equally-populated sub-bands of $s$ electrons ($\uparrow$ and $\downarrow$) carry most of the current in two parallel spin channels. In each spin channel there is a small resistivity associated with $s$ electron scattering and a larger resistivity associated with scattering of $s$ electrons into localized $d$ states, $s$-$d$ scattering. To a first approximation, there is no mixing between the two spin channels. For iron, the Fermi energy cuts across both spin-up and spin-down $d$ bands, hence there is a large $s$-$d$ contribution to the resistivity in each channel. In Ni, the Fermi energy is above the top of the majority-spin $d$ band and hence the resistivity in that band is less than in the minority-spin sub-band where $s$-$d$ scattering can occur.

Magnetotransport effects in a given magnetic material can arise from two types of magnetic interactions—exchange coupling and spin-orbit interaction. A conduction electron generally has a greater scattering cross section at sites having spin opposite its own. This is referred to as exchange scattering or spin disorder scattering.

Conduction electrons can also interact with a scattering site by the spin-orbit interaction. The total angular momentum must be conserved during the scattering process. Thus, an electron of spin-up before scattering can only scatter into a spin-up state (no change in spin-angular momentum), if it does not change its orbital angular momentum; it can scatter into a spin-down state if it increases its orbital angular momentum. These two spin-dependent scattering mechanisms can therefore serve to mix the carriers in the two spin sub-bands. Opening a path between two parallel conduction paths always lowers the net resistivity.

The "ordinary" *galvanomagnetic effects*, i.e., the *Hall effect*, $\boldsymbol{E} = R_H(\boldsymbol{J} \times \mu_0 \boldsymbol{H})$ and *magnetoresistance*, are observed in most materials. These effects are classical and arise from the Lorentz force, $\boldsymbol{F} = q(\boldsymbol{v} \times \boldsymbol{B})$, on the charge carriers. In ferromagnetic materials, the ordinary effects are present but are generally overshadowed by phenomena with similar symmetries. The galvanomagnetic effects unique to ferromagnets are called "extraordinary," "spontaneos," or "anomalous" because of their greater strength relative to the ordinary effect. The extraordinary galvanomagnetic effects derive their strength from the fact that the role of the external field is replaced by the internal field, which is proportional to the magnetization and hence is much stronger than an applied field. The mechanism by which $\boldsymbol{M}$ couples to the electron trajectory in ferromagnets is the spin-orbit interaction between the current carrier (orbit) and the magnetization (spin).

In ferromagnetic materials the *Hall resistivity* may be written

$$\rho_H = \frac{E_H}{J} = \rho_{oH} + \rho_{sH} = R_o B + R_s \mu_o M \qquad (25)$$

where the first term is the ordinary Hall resistivity proportional to the external field and the second term is the spontaneous effect, proportional to the magnetization. We can write these two terms as a simple sum because the vector symmetry of the spin-orbit energy, $\boldsymbol{L} \cdot \boldsymbol{s}$, responsible for the spontaneous Hall effect, is compatible with the energy of the classical Lorentz force, $\boldsymbol{r} \cdot \boldsymbol{F} \propto \boldsymbol{r} \cdot (\boldsymbol{v} \times \boldsymbol{B}) = (\boldsymbol{r} \times \boldsymbol{v}) \cdot \boldsymbol{B} \propto \boldsymbol{L} \cdot \boldsymbol{M} \propto \boldsymbol{L} \cdot \boldsymbol{s}$.

The *anisotropic magnetoresistance* (AMR) may be determined by extrapolation of high field MR (ordinary MR) data to $H = 0$ (see Fig. 30). The fractional change in resistance with field due to anisotropic magnetoresistance, while only a few percent in the best cases ($Ni_{90}Fe_{10}$), is used in numerous sensor applications. While the Hall effects are linear in $B = \mu_o(H + M)$, the MR effects are quadratic in $B$.

There is a more-recently-discovered galvanomagnetic effect that is unique to electronic transport in thin films. Called *giant magnetoresistance* (GMR), it is of a different physical origin than AMR. GMR arises from spin-spin exchange scattering and requires small dimensions in the components of a ferromagnetic/noble metal/ferromagnetic sandwich or multilayer. In 1988, Baibich, M. N. *et al.* (1988) reported an MR ratio of order 50% at 4.2 K in multilayers of the Fe-Cr alloys. This magnetoresistance was approximately an order of magnitude greater than the highest values known to that time. The Fe layers in these experiments were typically 30–60 Å thick and separated by Cr layers from 9 to 60 Å. The iron layers are strongly coupled antiferromagnetically through the Cr layers and hence are difficult to saturate. If electrons bearing the polarization of one ferromagnetic layer drift into the other (through a thin 2–3 nm of noble metal) the scattering probability is greater or less depending on whether the polarizations of the two layers are antiparallel or parallel. In order for the difference in scattering probability to show up as a large $\Delta R/R$, two general conditions should be satisfied.

1. The thicknesses of the three layers must be small enough that a large fraction of the charge carriers from one ferromagnetic component diffuse into the other before experiencing spin-dependent scattering
2. The magnetization in the two components should be able to be controlled independently either by having different coercivities or different anisotropy fields

A *spin valve* is a simple embodiment of the GMR effect in which there are only two magnetic layers separated by a nonmagnetic conductor. The magnetic layers are uncoupled or *weakly* coupled in contrast to the generally *strong* AF exchange operating in Fe/Cr-like multilayer systems. The magnetoresistance can be made to change in fields of a few tens of Oe rather than tens of kOe. One of the layers is magnetically soft and the other is magnetically hard. Thus, a modest field can cause a change in the angle between the moments of these two magnetic layers.

In operation of the spin valve, cycling the field causes $M_1$ and $M_2$ to lie antiparallel or parallel to each other. The resulting $M$-$H$ loops are shown schematically in Fig. 31. The sharp magnetization reversal near $H = 2$ Oe is due to the switching of the soft NiFe layer 2 in the presence of its weak coupling to layer 1. The more rounded magnetization reversal near 100 Oe is the switching of the hard layer. The relative orientations of layers 1 and 2 are indicated by the pairs of arrows in each region of the $M$-$H$ curve. In

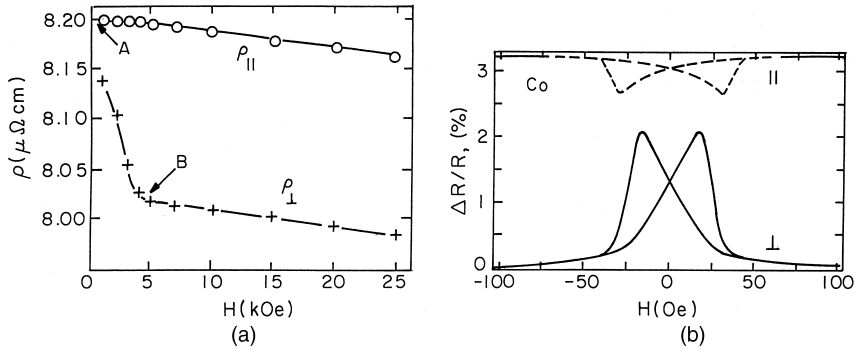



**FIGURE 30** (a) Resistivity of $Ni_{0.9942}Co_{0.0058}$ at room temperature versus applied field (McGuire, T. 1975). (b) Low-field magnetoresistance for cobalt thin film showing even field symmetry and hysteresis. [After Parkin, S. S. P. (1994). *In* "Ultrathin Magnetic Structures," Vol II (Heinrich, B. and Bland, J. A. C. eds.) Springer-Verlag, Berlin; O'Handley, R. C. (2000). "Modern Magnetic Materials, Principles and Applications." Reprinted with permission of John Wiley and Sons.]

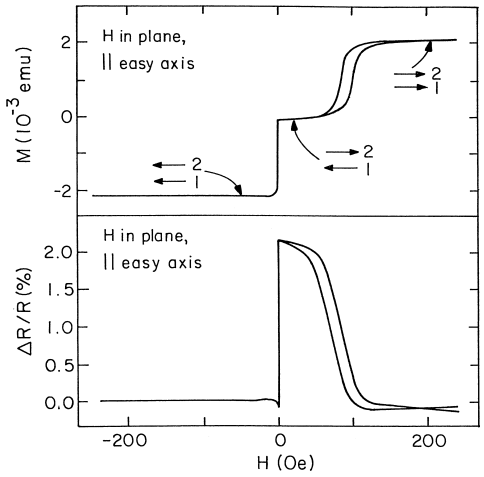



**FIGURE 31** Room temperature magnetization and relative change in resistance for Si/(NiFe 150 Å)/(Cu 26 Å)/(NiFe 150 Å)/(FeMn 100 Å)/(Ag 20 Å). Current is perpendicular to the easy axis determined by the FeMn film which is exchange coupled to the adjacent permalloy layer. [After Dieny, B. *et al.* (1991). *Phys. Rev. B* **43,** 1297.]

the lower panel, the change in resistance during the same magnetization cycling is shown. The GMR resistance is larger for antiparallel alignment of the two magnetic layers, whereas the classic AMR of single magnetic layer generally shows $\rho_{\parallel} > \rho_{\perp}$

*Spin tunnel junctions* have some similarities to spin valves and spin switches in their structure and field dependence. However, in a tunnel junction, the nonmagnetic spacer layer is an insulator.

A tunnel junction is conveniently formed from a crossed pair of metal film stripes. The first deposited stripe, e.g., aluminum, may be oxidized partially to form a barrier before deposition of the second electrode. A voltage applied across such a junction can result in a current if there are occupied states in one electrode at the same energy as unoccupied states in the other, plus or minus $k_B T$.

Moodera *et al.* (1996) have shown that by using thin magnetic films of different coercivities, the ferromagnet-insulator-ferromagnet (F-I-F) tunnel junction can form a sensitive magnetic field probe. Figure 32 shows the fractional resistance change ($\Delta R$ normalized to the high-field value of resistance) in FeCo-$Al_2O_3$-Co junctions. Also shown is the AMR measured in each individual electrode. These AMR measurements show that the small value of the AMR effect contributes little to the tunneling MR effect. They also clearly indicate the coercivities of the two uncoupled ferromagnetic layers. The field-dependence of the tunneling MR ratio then appears much like that of a spin valve or a spin switch with higher resistance when the two ferromagnetic electrodes are magnetized antiparallel to each other.

Very large magnetic-field-induced changes in resistivity have been observed in the doped perovskite, lanthanum-strontium manganate, $(La_{1-x}Sr_x)MnO_3$ (Jin *et al.*, 1994). This so-called *colossal magnetoresistance* (CMR) occurs in a region where a metal-insulator transition coincides with the ferromagnetic-paramagnetic transition. The resistivity increases with increasing temperature in the metallic magnetic phase. Application of a field expands the ferromagnetic phase, displacing to higher temperatures the metal-insulator transition and hence displacing the sharp increase in metallic resistivity. Thus the MR ratio is proportional to the temperature derivative of the $R(T)$ curve times the field derivative of the metal-insulator transition by a derivative chain rule. That is, the sharper the resistivity transition and the stronger the field dependence of that transition, the greater will be the CMR ratio.

Although CMR is a different physical effect than GMR, it has a similar formal dependence on magnetization orientation; aligning the moments on adjacent cation sites (AF to F) causes the resistance to decrease. This is due to the increase in hopping conductivity of the cation $e_g$ electrons for parallel spins. Fields of tens of kOe are needed to saturate the effect because they are working against thermal energy, which overcomes the long-range ferromagnetic coupling above the Curie temperature.

## C. Magnetic Recording

A variety of magnetic materials, transport effects, thin film, and nanostructured materials find applications in magnetic data storage systems.

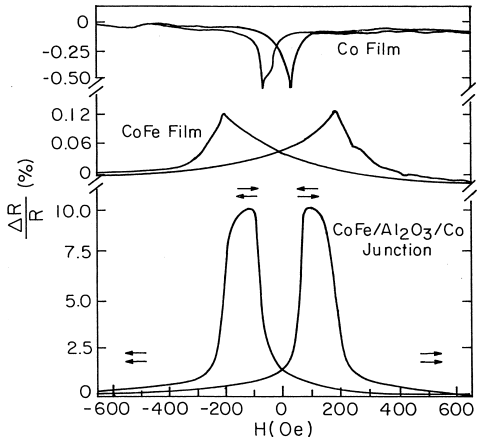



**FIGURE 32** Above, anisotropic magnetoresistance in each individual CoFe and Co electrode, and below, junction magnetoresistance in CoFe/$Al_2O_3$/Co spin tunnel junction versus applied field. Measuements done at room temperature and arrows indicate the relative directions of magnetization in the two magnetic layers. [After Moodera, J. *et al.* (1996). *Appl. Phys. Lett.* **69,** 708.] Note similarity with Fig. 30.

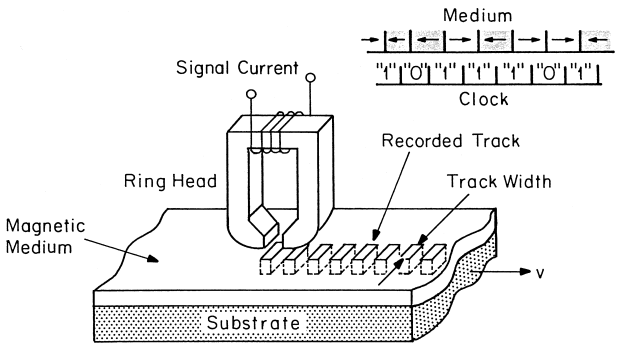



**FIGURE 33** Schematic representation of longitudinal, digital magnetic recording write processes. Insert, upper right, sequence of transitions constitute the bits which are read as binary information. [From O'Handley, R. C. (2000). "Modern Magnetic Materials, Principles and Applications." Reprinted with permission of John Wiley and Sons.]

In the digital recording process (Fig. 33), the magnetic recording medium (tape or disk) moves relative to an electromagnetic transducer that is essentially a magnetic circuit with a gap. When a current passes through windings about the head, the head is magnetized and a fringe field appears in the gap. In the write process, the fringe field in the gap magnetizes the medium alternately in one direction or the other as the drive current changes polarity. Because the head and the medium move relative to each other, information can be described in the head reference frame in terms of the variable $\omega t$ (e.g., $e^{-i\omega t}$) or in the medium reference frame by the variable $kx$ where $k = 2\pi/\lambda$. The sequence of binary states has digital information significance (Fig. 33, inset). A clock sets the system frequency, indicating when or where a transition might occur. The presence or absence of a transition at expected intervals (called bits) is read as a "one" or a "zero" to represent binary coded information.

The *write head* must have adequate magnetic permeability at high frequency so that it can be driven to saturation with minimal current. The write head must have a high enough saturation magnetization so that its fringe field exceeds the coercivity of the medium, typically 40–250 kA/m (500–3000 Oe). $Ni_{81}Fe_{19}$ ($\mu_o M_s \approx 1.0$ T) is generally used in thin-film write heads but higher induction permalloys, such as $Ni_{50}Fe_{50}$, and iron nitrides based on $Fe_{16}N_2$ ($\mu_o M_s \approx 3$T), are beginning to be used. The film thickness in the write head is typically 2–3 $\mu$m and the air gap between the head and medium is of order 100 nm for high density recording.

A *read head*, on the other hand, operates from its quiescent or demagnetized state. A read head should respond to the fringe field of the medium by magnetization rotation rather than wall motion (wall motion generates noise). Thus, the read head material should be able to be field annealed to develop a weak, cross-track uniaxial anisotropy in order to define the demagnetized domain state. The read head must have low coercivity, low noise, and extremely high permeability in order to respond with a substantial change in flux to the weak fringe field above the medium. Near-zero-magnetostriction permalloy is generally used in thin-film read heads. The read and write functions can be filled by the same inductive head but there are advantages to separating these functions.

In 1975, Thompson *et al.* (1975) described the use of the MR effect in magnetic recording heads. The resistance versus field for the anisotropic MR effect follows the general form shown in Fig. 30a: $\Delta\rho(H)/\rho = \Delta\rho/\rho](\cos\theta - 1/2)$. A bias field is needed to allow the head to operate on the steep, nearly linear portion of the curve. Shield layers on either side of the MR element were found to increase its sensitivity and reduce signal pickup from adjacent transitions.

Spin valves are also used as magnetic read heads. They show increased sensitivity compared to AMR heads and allow for higher recording density.

*Magnetic recording media* are ideally comprised of a regular array of isolated single-domain magnetic elements. These elements should be capable of being magnetized using a reasonable field strength. Further, they should be bistable, i.e., when the field is removed, the elements should have a large remanent magnetization. The medium should be comprised of small, independent magnetic entities (grains or single-domain particles) which can retain their direction of magnetization across a sharp transition. A bit ideally should be comprised of a single-domain, isolated magnetic particle. Because this is generally not practical, approximately $N = 10^3$ particles should constitute a bit in order to insure a sharp transition.

Currently used particulate magnetic media include chromic oxide, $Cr_2O_3$, so-called metal particles (iron with an unavoidable oxide surface layer), variants of $\gamma$-$Fe_2O_3$ usually doped with cobalt, as well as Ba or Sr hexaferrite ($BaO{\cdot}6Fe_2O_3$).

The most widely used magnetic recording media today are the thin-film media found in hard disk drives. Longitudinal thin-film media are based on CoCr with Pt an Ta additions. Pt is used to increase the magnetic anisotropy of the cobalt-rich film. It also improves the epitaxial relation between the cobalt film and the Cr underlayer. Cr plays a role in isolating the magnetic grains. The use of Ta as an alloying addition in CoCr longitudinal media is found to enhance segregation of Cr to the grain boundaries as well as improving epitaxy to the Cr underlayer.

One way to achieve low-noise, high-density media is to make each bit consist of a single piece or grain of magnetic

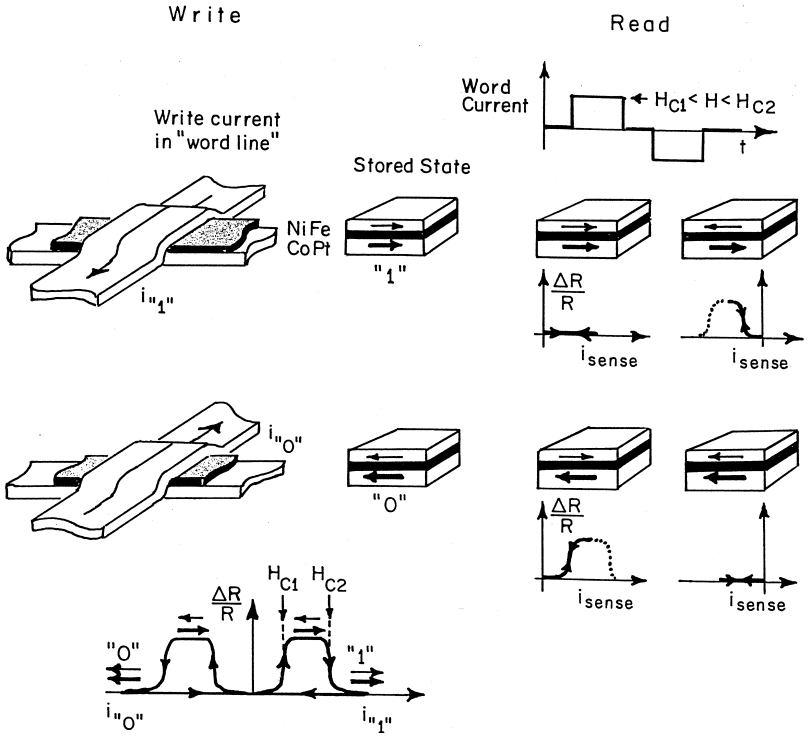



**FIGURE 34** Schematic of the read and write processes in a pseudo-spin-valve random access memory. The write process involves current pulses through both the word line and the sense line such that the field at the PSV exceeds $H_{c2}$. The read process involves a field pulse that takes the device to the high-resistance state without switching the semihard layer ($H_{c1} < H < H_{c2}$). [From O'Handley, R. C. (2000). "Modern Magnetic Materials, Principles and Applications." Reprinted with permission of John Wiley and Sons.]

material. Such bits should be arranged periodically in order to be synchronized with the signal channel. This can be achieved using high resolution lithography. The term *patterned media* is used to refer to media for which each bit consists of a single, lithographically defined grain. Such a recording medium eliminates the random $\sqrt{N}$ noise associated with multigrain bits. It also eliminates the noise associated with irregular or saw-tooth transitions that cause noise in thin-film media. Patterned media will allow for high bit densities because the superparamagnetic limit will thus apply to a single bit, not to each of the many grains in a multigrain bit. Finally, the patterning process defines a sharper transition between bits, and dispersion of easy axes can be minimized relative to that in thin-film media. Thus patterned media have relaxed conditions on coercivity and $M_r t$ product.

Considerable current interest is focused on a class of storage devices called *magnetic random access memories* (MRAMs). They have several advantages over hard disk drives: they have no moving parts, need no read or write heads, and offer the ability to access information at an arbitrary sequence of addresses (random access) as opposed to sequential access as in tape and disk storage. An MRAM is basically an array of spin valves, pseudo-spin valves or spin tunnel junctions (Fig. 34), each of which can be set to a given bistable state by a relatively large write field; that state can be read without destroying it by using a smaller read field.

The write process consists of magnetizing *both* the free and semihard layers in one direction or another by an appropriately directed word current pulse (and simultaneous sense-line current). After the write process, the two layers are in their remanent states and the resistance of either state has the same minimum value (Fig. 34).

The MRAM read process consists of applying a bipolar current pulse to the word line. This pulse produces a field sufficient to switch the soft layer but not the hard layer: $H_{c_1} < H_{word} < H_{c_2}$. Thus, depending upon the state of the element, "0" or "1," the resistance in the sense line changes in phase or out-of-phase, respectively, with the word current pulse. After application of the read pulse, the MR element reverts to its original remanent state (↑↑ or ↓↓). This is only possible if the two magnetic layers are ferromagnetically exchange coupled through the Cu spacer, or if the read pulse is followed by a small reset pulse of opposite polarity.

The materials requirements for MRAMs are similar to those of spin-valves and spin-tunnel junctions plus the additional resrtrictions imposed by nanofabrication and multimaterial compatibility.

### D. Postscript

The recent history of magnetic materials research and development is characterized by frequent reinvention of the field. The decade of the 1960s was dominated by work on permalloys, ferrites, transport phenomena, and magnetic resonance in bulk and thin film samples. In the1970s, magnetic "bubble films" (perpendicularly magnetized domains in which information could be stored and manipulated) and rapid developments in amorphous magnetic alloys attracted considerable attention. The 1980s saw continued improvements in amorphous magnetic alloys, rapidly accelerating activity in magnetic thin films and surfaces, and the magnetic recording developments those activities supported, as well as the emergence of Fe-Nd-B permanent magnets. The 1990s witnessed enormous growth in magnetic recording research, development of nanocrystalline magnetic materials, and extensive basic and applied work on a host of new magnetotransport phenomena.

The immediate future of magnetic materials will involve greater use of nanofabrication (thin films, phase separation) to engineer new properties and devices. Many of these new magnetic materials technologies will converge in the emerging field of spin-tronics, where spin-dependent transport phenomena hold the potential to supplant semiconductors in some microelectronic applications.

## SEE ALSO THE FOLLOWING ARTICLES

FERROMAGNETISM • GEOMAGNETISM • MAGNETIC FIELDS IN ASTROPHYSICS • MAGNETIC RECORDING • THIN FILM TRANSISTORS

## BIBLIOGRAPHY

Baibich, M. N. *et al.* (1988). *Phys. Rev. Lett.* **61,** 2742.
Bozorth, R. M. (1955, 1993). "Ferromagnetic Materials," Van Nostrand, New York; IEEE Press, New York.
Cadeville, M. *et al.* (1996). *J. Phys. (Paris)* **27,** 29.
Chikazumi, S. (1997). "Physics of Ferromagnetism," Oxford University Press, Oxford.
Corb, B. W. *et al.* (1983). *Phys. Rev. B* **27,** 636.
Diedrichs, P. H. *et al.* (1991). *In* "Magnetism in the Nineties," (Freeman, A. J. and Gschneider, K. eds.), North-Holland, Amsterdam.
den Broeder, F. J. A. *et al.* (1988). *Phys. Rev. Lett.* **60,** 2769.
Engle, B. *et al.* (1991). *Phys. Rev. Lett.* **67,** 1910.
Hasegawa, H. *et al.* (1993). *J. Mater. Eng. Perf.* **2,** 181.
Hasegawa, H. *et al.* (1991). *J. Appl. Phys.* **70,** 6253.
Herzer, G. (1993). *J. Mater. Eng. and Perf.* **2,** 193.
Jin, S. *et al.* (1994). *Science* **264,** 413.
Livingston, J. D. (1996). "Driving Force: The Natural Magic of Magnets," Harvard Univ. Press, Cambridge MA.
Moodera, J. *et al.* (1996). *Appl. Phys. Lett.* **69,** 708.
O'Handley, R. C. (2000). "Modern Magnetic Materials; Principles and Applications," John Wiley and Sons, New York.
Willard, M. A. *et al.* (1998). *J. Appl. Phys* **84,** 1.
Yoshizawa, Y. *et al.* (1988). *J. Appl. Phys.* **64,** 6044.



# Permittivity of Liquids

**J. Barthel**
**R. Buchner**
*University of Regensburg*

I. Phenomenological Aspects
II. Analysis of Complex Permittivity Spectra of Liquids
III. Experimental Methods
IV. Molecular Interpretation of Relaxation Modes

## GLOSSARY

**Dielectric relaxation** Delayed response of the electric polarization of a material system to a perturbation of the electric field.

**Dispersion** Frequency dependence of a material property, here the relative permittivity or specific conductivity of a solution.

**Kinetic depolarization** Reduction of polarization of the dipolar solvent molecules of electrolyte solutions (and in turn the reduction of ionic mobilities) resulting from the torque produced by an ion in its adjacent solvent molecules, counteracting the force of the external electric field.

**Libration** Partial reorientation (small-angle oscillation) of molecular dipoles or the translational (linear) oscillations of ions in the cage produced by the adjacent molecules.

**Loss angle** Phase angle between the polarization and electric field vectors of a dissipative system, characterizing the energy absorption of the sample.

**Relaxation time** Time constant of a relaxation process; inverse of the rate for approach to equilibrium.

**Rotational diffusion** Random (Brownian) movement of a molecular probe vector, the dipole moment in the case of dielectric relaxation, referred to an initial state.

**Step response function** Characteristic time-dependent function controlling the development of polarization toward equilibrium after a jump of the electric field.

**THE APPLICATION** of an electric field $\boldsymbol{E}$ ($\boldsymbol{E}$, electric field strength) to a liquid polarizes the molecules of the fluid and produces a charge transport if the investigated sample is an electrolyte solution.

The polarization of the molecules results from an alignment of their permanent dipole moments $\boldsymbol{\mu}$ against thermal motion (orientation effect, orientational polarization $\boldsymbol{P}_\mu$) and the induced dipole moments, $\boldsymbol{\mu}_{\text{ind}} = \alpha \boldsymbol{E}$, because of the action of the electric field on their polarizability $\alpha$ (deformation effect, induced polarization $\boldsymbol{P}_\alpha$).

The total polarization $\boldsymbol{P}$ of the sample relates the macroscopically measurable relative permittivity $\varepsilon$ ($\varepsilon_0$, permittivity of the vacuum) to the microscopically defined molecular quantities $\boldsymbol{\mu}$ and $\alpha$, owing to its definitions at both macroscopic and microscopic levels.

macroscopic: $\varepsilon_0(\varepsilon - 1)\boldsymbol{E} = \boldsymbol{P} = \boldsymbol{P}_\mu + \boldsymbol{P}_\alpha$: microscopic (1)

The flow of electric charges in electrolyte solutions is caused by the gradient of the galvanic potential $\psi$, $\boldsymbol{E} = -\text{grad}\,\psi$. It satisfies Ohm's law, $\boldsymbol{j} = \kappa\boldsymbol{E}$ ($\boldsymbol{j}$, electric current density; $\kappa$, specific conductivity).

The material properties $\varepsilon$ and $\kappa$ depend on the frequency $\nu$ ($\nu$, linear frequency; $\omega = 2\pi\nu$, circular frequency) of the applied electric field $\boldsymbol{E}$. Separate measurements under static and quasi-static (very low frequency) conditions yield the static values of permittivity $\varepsilon$ and specific conductivity $\kappa$; but note that for $\kappa > 0$ the electric current prevents the direct determination of $\varepsilon$.

Static permittivity $\varepsilon$ is related to the equilibrium polarization of the sample. In this case, no energy is dissipated in the system. With increasing frequency, situations are reached in which the polarity change of the electric field causes significant variation of the electric field strength within periods that are characteristic of molecular motions (dipole rotation, ionic mobility, etc.) or reactions (formation of ion pairs or hydrogen bonds, etc.). Then the polarization lags behind the electric field, and energy is dissipated in the system. This effect is commonly called dielectric dispersion or dielectric relaxation. The energy dissipation, expressed as dielectric loss, is practically used for dielectric heating devices, such as the common microwave oven. At optical frequencies, the relaxation processes are paralyzed, and only the resonance processes (intramolecular atomic and electronic movements) underlying induced polarization are observable.

The ultimate aim of dielectric relaxation spectroscopy (DRS) is the deconvolution of the observed dielectric relaxation behavior into individual contributions which can be interpreted on a molecular scale. Such an analysis yields a wealth of information about molecular and cooperative motions, kinetic processes, and liquid structure which is often not available by other methods. In principle, processes ranging in time scale from several hundred femtoseconds to hours are susceptible to dielectric relaxation. Figure 1 gives a survey of processes and corresponding frequencies relevant for the dielectric relaxation of common solvents and electrolytes around ambient temperature. In Section IV, a brief introduction into some relaxation mechanisms will be given.

Note that for conducting samples there is always a contribution to energy dissipation from Ohm's law. The characteristic time constant of conductivity is the relaxation time for the reestablishment of electroneutrality.

## I. PHENOMENOLOGICAL ASPECTS

The phenomenological description of the interactions between material systems and electromagnetic fields is based

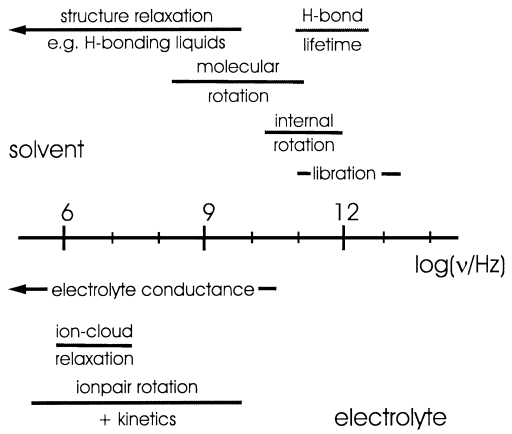



**FIGURE 1** Frequency regions of relaxation processes of the solvent and of the electrolyte which may contribute to the dielectric permittivity and relaxation of liquids and solutions around room temperature.

on the Maxwell equations

$$\textbf{curl}\,\boldsymbol{H} = \boldsymbol{j} + \frac{\partial \boldsymbol{D}}{\partial t} \tag{2}$$

$$\textbf{curl}\,\boldsymbol{E} = -\frac{\partial \boldsymbol{B}}{\partial t} \tag{3}$$

$$\text{div}\,\boldsymbol{D} = \rho \tag{4}$$

$$\text{div}\,\boldsymbol{E} = 0 \tag{5}$$

and the relations

$$\boldsymbol{D} = \varepsilon_0\varepsilon\boldsymbol{E} \tag{6}$$

$$\boldsymbol{B} = \mu_0\mu\boldsymbol{H} \tag{7}$$

$$\boldsymbol{j} = \kappa\boldsymbol{E} \tag{8}$$

which define the material properties $\varepsilon$ (relative permittivity; $\varepsilon_0$, permittivity of the vacuum), $\mu$ (relative permeability; $\mu_0$, permeability of the vacuum), and $\kappa$ (specific conductivity), using the electric vectors $\boldsymbol{E}$ (electric field strength), $\boldsymbol{D}$ (dielectric displacement), and $\boldsymbol{j}$ (electric current density) and the magnetic vectors $\boldsymbol{H}$ (magnetic field strength) and $\boldsymbol{B}$ (magnetic flux density); $\rho$ is the charge density of the electric field. Polarization $\boldsymbol{P}$ is related to the electric field strength $\boldsymbol{E}$ via Eq. (1). The contributions $\boldsymbol{P}_\mu$ and $\boldsymbol{P}_\alpha$ to the total polarization, Eq. (1), are considered to be linearly independent. This allows us to express $\boldsymbol{P}_\alpha$ with the help of the "infinite frequency" permittivity $\varepsilon_\infty$, separating relaxation and resonance processes so that the total polarization $\boldsymbol{P}$ can be split into two parts

$$\boldsymbol{P}_\alpha = \varepsilon_0(\varepsilon_\infty - 1)\boldsymbol{E} \tag{9}$$

$$\boldsymbol{P}_\mu = \varepsilon_0(\varepsilon - \varepsilon_\infty)\boldsymbol{E} \tag{10}$$

A time domain (TD) experiment, Fig. 2(a), shows the response of $\boldsymbol{P}_\alpha$ and $\boldsymbol{P}_\mu$ when the static field $\boldsymbol{E}$ applied

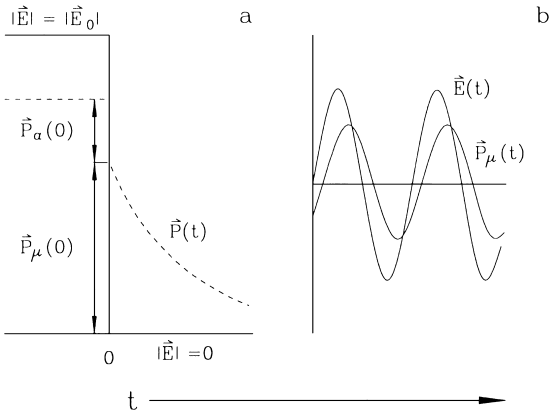



**FIGURE 2** Methods for the determination of complex permittivity: (a) time domain method: application of a jump in the field strength $|\boldsymbol{E}|$ and measurement of the response $\boldsymbol{P}(t) = \boldsymbol{P}^{\mathrm{eq}} \cdot F_{\mathrm{p}}(t)$; (b) frequency domain method: application of a harmonic field $\boldsymbol{E}(t) = \boldsymbol{E}_0 \cdot \exp(i\omega t)$ and measurement of amplitude and phase of the response $\boldsymbol{P}_\mu(t, \omega) = \varepsilon_0(\varepsilon - \varepsilon_\infty)\boldsymbol{E}(t)$.

initially to a liquid sample is switched off at time $t = 0$. $\boldsymbol{P}_\alpha$ breaks down without a time lag, that is, it is always in equilibrium with $\boldsymbol{E}$, whereas $\boldsymbol{P}_\mu$ decreases monotonically with time to its final value $\boldsymbol{P}_\mu(\infty) = 0$. Formally, this can be expressed by the relationship

$$\boldsymbol{P}_\mu(t) = \boldsymbol{P}_\mu^{\mathrm{eq}} \cdot \boldsymbol{F}_{\mathrm{p}}^{\mathrm{or}}(t) \tag{11}$$

with the static (equilibrium) orientational polarization $\boldsymbol{P}_\mu^{\mathrm{eq}} = \boldsymbol{P}_\mu(0)$ and the step response function (time autocorrelation function) of the orientational polarization

$$F_{\mathrm{p}}^{\mathrm{or}}(t) = \frac{\langle \boldsymbol{P}_\mu(t) \cdot \boldsymbol{P}_\mu(0) \rangle}{\langle \boldsymbol{P}_\mu(0) \cdot \boldsymbol{P}_\mu(0) \rangle} = \frac{\langle \boldsymbol{M}_\mu(t) \cdot \boldsymbol{M}_\mu(0) \rangle}{\langle \boldsymbol{M}_\mu(0) \cdot \boldsymbol{M}_\mu(0) \rangle} \tag{12}$$

which can be shown to be identical with the autocorrelation function of the macroscopic dipole moment of the sample, $M = \sum_i \boldsymbol{\mu}_i(t)$. $\boldsymbol{M}$ is defined as the vector sum of all constituting molecular dipole moments, $\boldsymbol{\mu}_i(t)$.

Conversion to the frequency domain (FD) is made by the consideration that an arbitrary time dependence of the field strength $\boldsymbol{E}$ can be expressed as a sequence of infinitely small time steps. In the case of a harmonically changing electric field, $\boldsymbol{E}(t) = \boldsymbol{E}_0 \exp(i\omega t)$, that is, a monochromatic electromagnetic wave of circular frequency $\omega$, the relation

$$\begin{aligned} \boldsymbol{P}_\mu(\omega, t) &= \varepsilon_0(\varepsilon - \varepsilon_\infty)\boldsymbol{E}(t) \\ &\quad \times \int_0^\infty \left( -\frac{\partial F_{\mathrm{p}}^{\mathrm{or}}(t')}{\partial t'} \right) \exp(i\omega t')\, \mathrm{d}t' \\ &= \varepsilon_0(\varepsilon - \varepsilon_\infty)\tilde{F}(\nu)\, \boldsymbol{E}(t) \end{aligned} \tag{13}$$

is obtained, which expresses the phase shift and amplitude modulation of $\boldsymbol{P}_\mu(\omega, t)$ relative to $\boldsymbol{E}(\omega, t)$, Fig. 2(b). The relaxation function $\tilde{F}(\nu)$ is the frequency domain analogue of the autocorrelation function $F_{\mathrm{p}}^{\mathrm{or}}(t)$; note that $\tilde{F}(\nu)$ is the Fourier transform of the time derivative of $F_{\mathrm{p}}^{\mathrm{or}}(t)$, called the step response function $f_{\mathrm{p}}^{\mathrm{or}}(t) = -(\partial F_{\mathrm{p}}^{\mathrm{or}}(t)/\partial t)$. To incorporate the observed energy absorption, the treatment of dissipative systems exposed to electromagnetic waves of circular frequencies $\omega$ may be advantageously based on the use of electric and magnetic vectors of the type $\boldsymbol{A}(t) = \boldsymbol{A}_0 \cdot \exp(i\omega t)$ in the Maxwell equations ($\boldsymbol{A} = \boldsymbol{E}, \boldsymbol{D}, \boldsymbol{P}, \boldsymbol{j}, \boldsymbol{H}, \boldsymbol{B}$) and complex material properties such as $\hat{\varepsilon} = \varepsilon' - i\varepsilon''$, $\hat{\kappa} = \kappa' - \kappa''$, and $\hat{\mu} = \mu' - i\mu''$. Note that permeability can usually be set equal to unity in liquids. The comparison of Eqs. (9, 10) and (1), after transformation to complex notation, then yields the frequency dependence of permittivity:

$$\begin{aligned} \hat{\varepsilon}(\omega) &= \varepsilon'(\omega) - i\varepsilon''(\omega) \\ &= \varepsilon_\infty + (\varepsilon - \varepsilon_\infty)\tilde{F}(\nu) \end{aligned} \tag{14}$$

On the frequency scale, the quantity $\varepsilon_\infty$ is the permittivity corresponding to a nonpolar liquid (zero dipole moment) between zero frequency and the infrared (IR) range (before the onset of intramolecular vibrations) and to a polar liquid at IR frequencies only where $\boldsymbol{P}_\mu$ equals zero. The fluctuation of the induced polarization, $\boldsymbol{P}_\alpha$, is very rapid and is affected only by resonant (quantum mechanical) transitions in the IR and ultraviolet (UV)/visible regions.

Processes linked to structural rearrangements take place on the nanosecond to subpicosecond time scale for most liquids at room temperature, corresponding to frequencies in the megahertz to terahertz region on the frequency scale; they contribute only to $F_{\mathrm{p}}^{\mathrm{or}}$ and hence to $\boldsymbol{P}_\mu$. Such processes are dipole reorientation, breaking and reforming of hydrogen bonds, or returning the system to chemical equilibrium (for example, ion pair or complex formation), and reestablishement of electroneutrality in electrolyte solutions (see Fig. 1). Figure 3 shows schematically the typical behavior of the frequency dependence of the real, $\varepsilon'$, and imaginary, $\varepsilon''$, parts of the complex permittivity spectrum of a nonconducting liquid.

The real part $\varepsilon'(\omega)$, the permittivity spectrum, is a measure of the polarization at frequency $\nu$; the imaginary part $\varepsilon''(\omega)$, the dielectric loss or absorption curve, characterizes energy dissipation in the system and is the relevant quantity for the assessment of dielectric heating effects. The phase shift between electric field and polarization, $\delta = \arctan[\varepsilon''(\omega)/\varepsilon'(\omega)]$, is called the loss angle; the dissipated energy per unit volume and time is $\dot{W} = E_0^2 \omega \varepsilon_0 \varepsilon''(\omega)/2$.

Electrically conducting systems require special consideration. The appropriate combination of Maxwell equations yields the wave equation

$$\mathbf{grad}\, \mathrm{div}\boldsymbol{A} + \hat{k}^2 \boldsymbol{A} = 0 \tag{15}$$

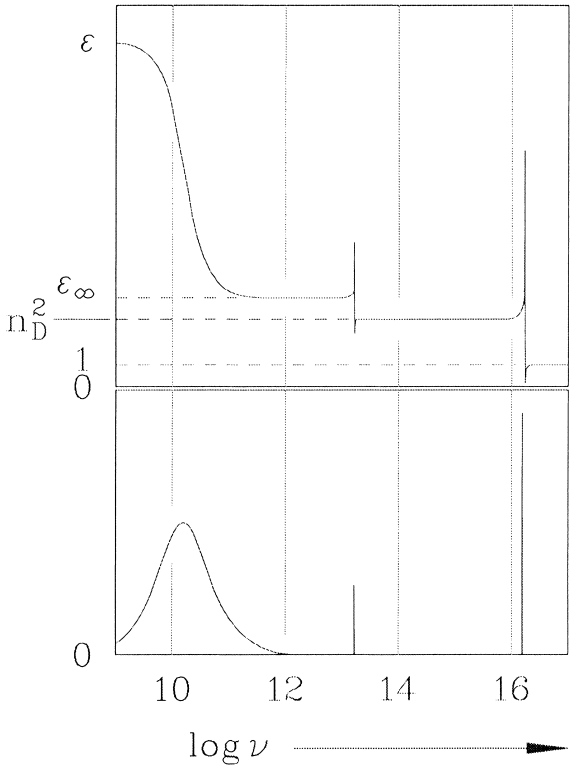



**FIGURE 3** Dielectric permittivity, $\varepsilon'(\nu)$, and loss, $\varepsilon''(\nu)$, spectra for a polar liquid with a single Debye relaxation process in the microwave region and two resonant transitions in the IR and UV/visible range; $n_D$ is the refractive index in the visible spectral range.

where the vector $\boldsymbol{A}$ is either the electric, $\boldsymbol{E}$, or the magnetic field strength, $\boldsymbol{H}$, both propagating perpendicularly to one another in the medium which is characterized by the complex propagation coefficient $\hat{k}$

$$\hat{k}^2 = k_0^2\left(\hat{\varepsilon}(\nu) + \frac{\hat{\kappa}(\nu)}{i\omega\varepsilon_0}\right) = k_0^2\hat{\eta}(\nu) \tag{16}$$

where $k_0$ is the propagation coefficient of the vacuum. The dissipated energy per unit volume and time is $\dot{W} = E_0^2\omega\varepsilon_0\eta''(\omega)/2$. Equation (16) clearly shows that the only measurable quantity is the generalized permittivity $\hat{\eta}(\nu)$, which is reduced to $\hat{\varepsilon}(\nu)$ for nonconducting samples where $\boldsymbol{j} = 0$. It is known from experiments that the dispersion of conductance, the so-called Debye–Falkenhagen effect, due to the relaxation of the diffuse ion cloud surrounding each ion, is rather small for electrolyte solutions and may be neglected at high frequencies, that is, $\kappa''(\nu) = 0, \kappa'(\nu) = \kappa$. Then follows

$$\eta'(\nu) = \varepsilon'(\nu) \tag{17}$$

$$\eta''(\nu) = \varepsilon''(\nu) + \frac{\kappa}{\omega\varepsilon_0} \tag{18}$$

from which $\varepsilon''(\nu)$ can be calculated using the measured quantity $\eta''(\nu)$ when $\kappa$ is known from conductance measurements under quasi-static conditions. For systems in which interfacial charges are important (dispersions of charged polymers or colloids, ionic micelles), the consideration of the frequency dependence of conductivity is always required.

## II. ANALYSIS OF COMPLEX PERMITTIVITY SPECTRA OF LIQUIDS

The complex permittivity spectrum of ethanol, Fig. 4, is typical for the dielectric response of liquids in the microwave region. The aim of DRS is the interpretation of the observed relaxation behavior on the molecular scale. Prerequisite of such an endeavor is the fitting of $\hat{\varepsilon}(\nu)$ by an appropriate model to extract the independent relaxation modes contributing to dielectric permittivity. Generally, this analysis is performed in the frequency domain. Due to the often limited frequency range and the large bandwidth associated with each process this is a nontrivial task. Although a satisfactory fitting of the spectra, that is, with a superposition of dispersion steps, is always possible, the

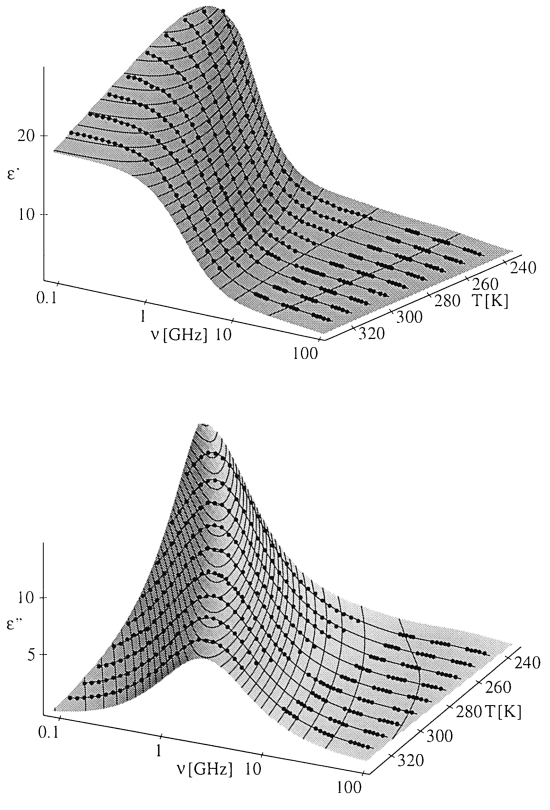



**FIGURE 4** Dielectric permittivity, $\varepsilon'(\nu)$, and loss, $\varepsilon''(\nu)$, spectra of ethanol in the temperature range $-25 \leq \vartheta/^\circ\mathrm{C} \leq 55$.

obtained result is not necessarily correlated with physical processes. If possible, only extended series of measurements, for example, as a function of temperature or electrolyte concentration, should be considerd to check the self-consistency of the applied model. To minimize the impact of systematic errors, which affect $\varepsilon'$ and $\varepsilon''$ differently, permittivity and loss spectra should be fitted simultaneously.

The simplest case of response function, generally found for the dipole rotation of small symmetric molecules like acetonitrile, is a first-order exponential decay of orientational polarization with response function

$$F_{\mathrm{p}}^{\mathrm{or}} = \exp(-t/\tau) \tag{19}$$

governed by the relaxation time $\tau$. In the frequency domain this corresponds to the Debye equation

$$\hat{\varepsilon}(\nu) = \frac{\varepsilon - \varepsilon_\infty}{1 + i2\pi\nu\tau} + \varepsilon_\infty \tag{20}$$

with the dispersion of permittivity

$$\varepsilon'(\omega) = \frac{\varepsilon - \varepsilon_\infty}{1 + \omega^2\tau^2} + \varepsilon_\infty \tag{21}$$

between $\varepsilon = \lim_{\nu\to 0}\varepsilon'$ and $\varepsilon_\infty = \lim_{\nu\to\infty}\varepsilon'$ (a centrosymmetric curve with respect to the critical frequency $\nu_{\mathrm{c}} = 1/(2\pi\tau)$), and a Lorentzian bandshape of the dielectric loss

$$\varepsilon''(\omega) = \frac{(\varepsilon - \varepsilon_\infty)\omega\tau}{1 + \omega^2\tau^2} \tag{22}$$

which peaks at $\nu_{\mathrm{c}}$.

Generally, however, even simple molecular liquids exhibit a more complex relaxation behavior and require more complex bandshape models and/or superposition of $n$ individual relaxation processes $j$ of amplitude (relaxation strength) $S_j = \varepsilon_j - \varepsilon_{j\infty}$

$$\hat{\varepsilon}(\nu) = \sum_{j=1}^{n} S_j \tilde{F}_j(\nu) + \varepsilon_\infty \tag{23}$$

where

$$\varepsilon = \varepsilon_1 = \sum_{j=1}^{n} S_j + \varepsilon_\infty; \quad \varepsilon_{j\infty} = \varepsilon_{j+1}; \quad \varepsilon_{n\infty} = \varepsilon_\infty \tag{24}$$

The relaxation functions $\tilde{F}_j$ of the individual dispersion steps may generally be represented by modifications of the Havriliak–Negami equation

$$\tilde{F}_j(\nu) = \left[1 + (i2\pi\nu\tau_j)^{1-\alpha_j}\right]^{-\beta_j} \tag{25}$$

each one with relaxation time $\tau_j$ and relaxation time distribution parameters, $0 \le \alpha_j < 1$ and $0 < \beta_j \le 1$. Special cases of Eq. (25) are the asymmetric Davidson–Cole relaxation time distribution, $\alpha_j = 0$, and the symmetrically broadened Cole–Cole distribution, $\beta_j = 1$. The limiting case of $\alpha_j = 0$ and $\beta_j = 1$ is the Debye equation. Note that for $\kappa \neq 0$ only an analysis based on Eq. (23) permits the determination of the static permittivity $\varepsilon$, a quantity important for many thermodynamic properties of electrolyte solutions.

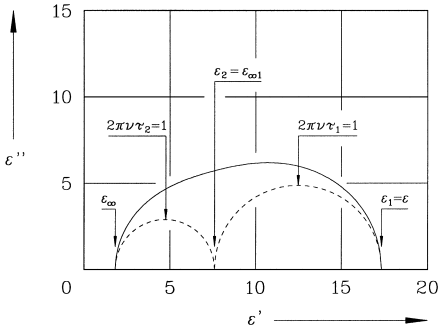



**FIGURE 5** Propylene carbonate–1,2-dimethoxyethane (PC-DME) mixture (mole fraction $x_{\mathrm{PC}} = 0.2$; 25°C) as an example for two superposed Debye processes with characteristic parameters $\varepsilon = \varepsilon_1$, $\varepsilon_{1\infty} = \varepsilon_2$, $\varepsilon_{2\infty} = \varepsilon_\infty$, $\tau_1$ and $\tau_2$. The contribution of PC is characterized by the dispersion amplitude $S_1 = \varepsilon_1 - \varepsilon_{1\infty} = 9.74$ and relaxation time $\tau_1 = 22.0$ ps; that of DME by $S_2 = \varepsilon_2 - \varepsilon_\infty = 5.79$ and $\tau_2 = 4.7$ ps.

Examples of spectra with a superposition of dispersion steps where the individual contributions to $\hat{\varepsilon}$ are indicated are given in Figs. 5–8. The significance of the quantities $\varepsilon_i$, $\varepsilon_{i\infty}$, and $S_j$ becomes obvious in a plot of $\varepsilon''(\nu)$ versus $\varepsilon'(\nu)$, the so-called Argand diagram or Cole–Cole plot, Fig. 5. The example is typical for a mixture of dipolar aprotic liquids without specific intermolecular interactions. For such systems, dominated by dipole–dipole interactions and packing requirements, the contributions of the individual components are detectable over the entire mixture range, with their dynamics reflecting the smooth change from only A:A interactions in the pure liquid A to A:B interactions in dilute solutions of A in B.

A disadvantage of the Argand diagram is the lacking frequency coordinate so that assignment to the dynamics of the system is not immediately obvious. Such information is especially brought out by the loss spectra, $\varepsilon'' = f(\nu)$, since the relaxation times are immediately related to the peak frequencies of the individual contributions. Figure 6 shows the complex permittivity spectrum of an associating electrolyte in water. The two observed dispersion steps, which are of Debye type and well separated in the frequency scale, can be attributed to the cooperative relaxation of the water's hydrogen bond network, $\tau_{\mathrm{H_2O}} \approx 8.3$ ps, and to the tumbling motion of long-lived La[Fe(CN)$_6$] ion pairs formed by this electrolyte, $290 \le \tau_{\mathrm{IP}}/\mathrm{ps} \le 630$. Note that pure water, Fig. 7, and most aqueous electrolytes show an additional fast process, $\tau_{\mathrm{H_2O},2} \approx 1$ ps, which is not resolved for the La[Fe(CN)$_6$] solutions. The spectrum of ethanol, Fig. 4, is a superpostion of three dispersion

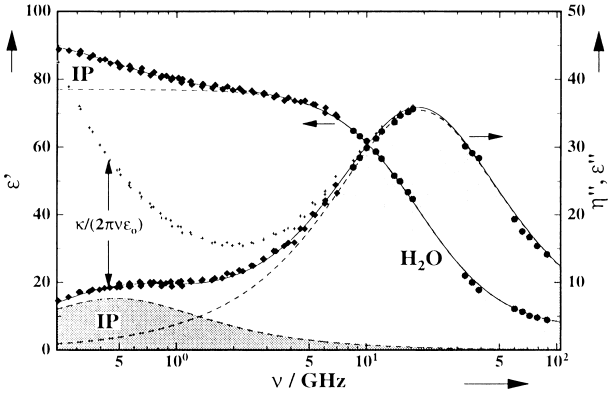



**FIGURE 6** Dielectric permittivity, $\varepsilon'(\nu)$, and loss, $\varepsilon''(\nu)$, of 0.0345 mol dm$^{-3}$ La[Fe(CN)$_6$] in water at 25°C. Experimental data (FD: ●, TDR: ♦) are fitted to a superposition of two Debye processes (solid lines) attributed to the ion pair (IP) and to the solvent ($H_2O$). Also indicated is the total loss, $\eta''(\nu)$, of the solution.

steps with relaxation times in the order of $\tau_1 = 164$ ps, $\tau_2 = 7.9$ ps, and $\tau_3 = 1.5$ ps at room temperature.

Figure 8 shows the spectrum of the monohydrate of trifluoromethanesulfonic acid which is of potential interest as a solvent in fuel cells. Its spectrum is rather featureless, and the extraction of four Debye processes pushes the band fitting procedure with a model based on Eq. (23) to the numerical limits. For such spectra, a physically meaningful deconvolution into molecular-level processes is only possible with information from other methods. Spectra of similar shape are usually observed for liquids of flexible molecules with high molecular weight like polymer melts. Here a broad distribution of relaxation times must be assumed, which significantly complicates a detailed analysis.

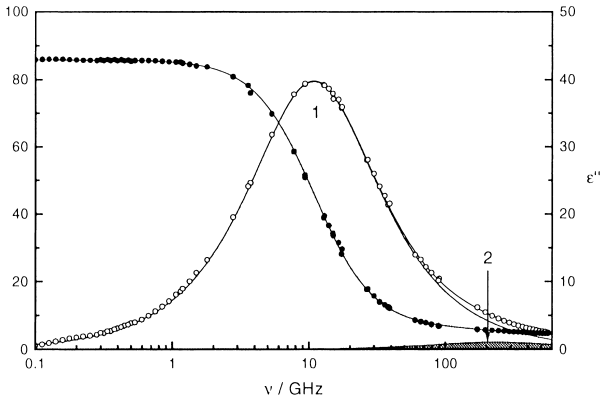



**FIGURE 7** Dielectric permittivity, $\varepsilon'(\nu)$ (●), and loss, $\varepsilon''(\nu)$ (○), of water at 5°C. Experimental data are fitted to a superposition of two Debye processes (solid lines) attributed to the cooperative relaxation of the hydrogen-bond network (1) and to the rotation of mobile $H_2O$ molecules (2).

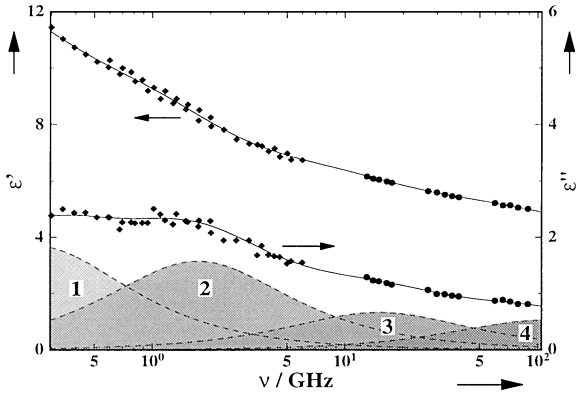



**FIGURE 8** Dielectric permittivity, $\varepsilon'(\nu)$, and loss, $\varepsilon''(\nu)$, of the monohydrate of trifluoromethanesulfonic acid at 40°C. Experimental data (FD: ●, TDR: ♦) are fitted to a superposition of four Debye processes (solid lines).

## III. EXPERIMENTAL METHODS

For electrolyte solutions of common solvents around room temperature the time scale of dielectric relaxation processes is in the order of 0.1 ps to 10 ns, meaning that ideal experiments should span from megahertz to terahertz (far-infrared, FIR) frequencies to obtain the full information on the dynamics of the investigated system. The relevant equation for the construction of measurement devices is Eq. (15). However, in the broad frequency range, to cover the ratio of the characteristic dimension, $l$, of the measurement cell to the wavelength $\lambda$ of the applied electromagnetic radiation changes considerably. At low frequencies, $l/\lambda \ll 1$, broadband coaxial transmission lines can be applied. Broadband experiments are again possible with free-space methods from the FIR region upward where $l/\lambda \gg 1$, but for the intermediate microwave range, where $l/\lambda \approx 1$, narrow-band waveguide equipment is necessary, which makes experiments cumbersome and expensive.

Below 50 MHz, impedance bridges with the sample enclosed between the electrodes of a parallel plate condenser are used for the determination of resistance and capacitance of the sample, to yield the complex permittivity $\hat{\eta}$. Limitations of the method result from electrode polarization and from stray fields at high frequencies. Low loss samples permit the use of heterodyne beat measurements linking the capacitance of the measuring cell to the frequency shift of a megahertz oscillator.

Such resonance circuits are also applied above 50 MHz. However, due to the recent progress in electronic instrumentation coaxial transmission lines probing amplitude attenuation and phase shift of a transmitted or (more convenient) reflected wave have become far more attractive. Signal generation and detection can now be performed in the FD with the help of accurate vector network analyzers

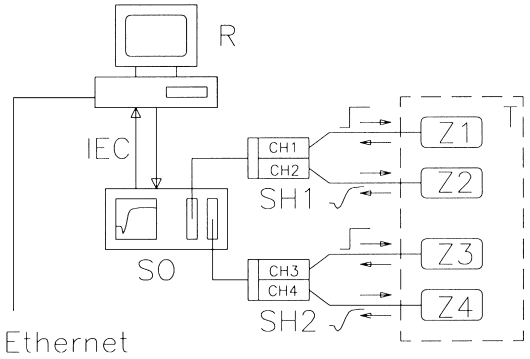



**FIGURE 9** Schematic diagram of a coaxial-line time domain reflectometer (TDR). SO: digital sampling scope (20 GHz bandwidth); SH1, SH2: sampling heads; Z: matched pairs of reflection cells; T: precision thermostat; R: personal computer with access to a work station for data analysis.

(VNA). Alternatively, in a TD experiment, a fast rising voltage pulse can be applied to the sample-filled cell, and $\hat{\eta}(\nu)$ is obtained from the Fourier transform of the reflected [time domain reflectometer (TDR)] or transmitted pulse. A typical TDR instrument is shown in Fig. 9. Stringent geometric requirements for cell construction currently limit the maximum frequency for coaxial line techniques to 20 GHz with VNA and to somewhat less with TDR.

Above about 10 GHz the wavelength of the electromagnetic wave is comparable to the dimensions of the measuring equipment, making the use of waveguides unavoidable. Resonator techniques as well as methods based on the transmission or reflection of propagating waves are in use. For the investigation of lossy liquids, transmission line techniques yield superior results, but resonators are more easily adapted to measurements at high temperatures and pressures. The drawback with waveguides is their limited frequency band. Several setups are needed to bridge the gap between coaxial-line techniques and free-space methods. For instance, in the authors' laboratory four interferometers of the type sketched in Fig. 10 are used to cover the range $8.5 \leq \nu/\mathrm{GHz} \leq 89$.

Above 100 GHz, the small wavelengths ($\leq$3 mm in the vacuum) permit free-space propagation with optical lenses and mirrors, as commonly used in conventional IR spectroscopy. The technique of Fourier transform spectrometers based on an asymmetric Michelson interferometer is rather mature. In this arrangement the sample is placed in one of the active arms of the interferometer, which allows the simultaneous determination of the refractive index and the absorption coefficient, both necessary for the calculation of $\hat{\varepsilon}$. However, until now measurements in the FIR region, which are of high potential interest as a source of information on the short-time dynamics (0.1–1 ps) of the sample, are rather limited due to the low intensities of conventional FIR sources. This situation is changing now with the rapid development of femtosecond-laser pumped terahertz emitters and detectors.

## IV. MOLECULAR INTERPRETATION OF RELAXATION MODES

The step response function $F_{\mathrm{p}}^{\mathrm{or}}$ underlying the interpretation of the complex permittivity spectrum is a macroscopic property related to the fluctuations of the macroscopic dipole moment $\boldsymbol{M} = \sum_i \boldsymbol{\mu}_i(t)$. The relation to molecular dynamics is made obvious by the rearrangement of its correlation function as

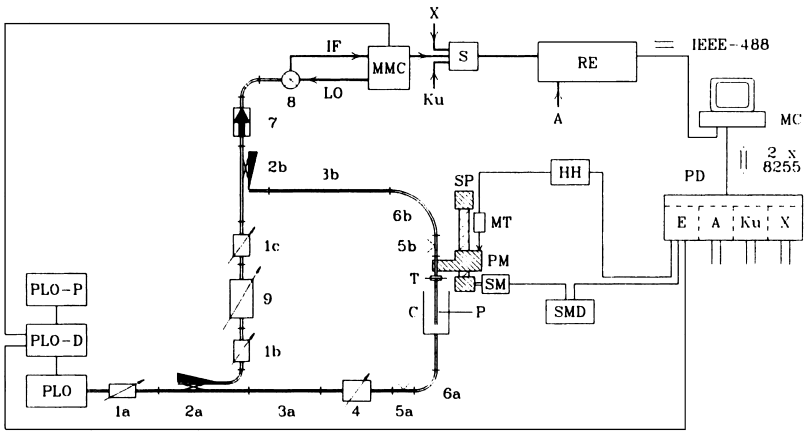



**FIGURE 10** Waveguide apparatus for the determination of $\hat{\varepsilon}(\nu)$ in the E-band (60–90 GHz) range with transmission measurements. 1–9: waveguide interferometer with cell C and movable probe P; PLO, PLO-D, PLO-P: microwave signal source and control unit; 8, MMC, S, RE: signal detection unit; HH, MT, SMD, SM, PM, SP: probe position control unit; PD: interface enabling the control of four interferometers (E-, A-, Ku-, X-band) in the frequency range 8.5 to 90 GHz; MC: personal computer.

$$\langle \boldsymbol{M}(t) \cdot \boldsymbol{M}(0) \rangle = \sum_i \langle \boldsymbol{\mu}_i(t) \cdot \boldsymbol{\mu}_i(0) \rangle + \left\langle \sum_i \boldsymbol{\mu}_i(t) \cdot \sum_{j \neq i} \boldsymbol{\mu}_j(0) \right\rangle \quad (26)$$

self

distinct

The *self* term gives the orientational correlation function, that is, the dynamics of individual molecules in a continuous "bath" of fluctuating frictional forces. It can be compared with results from NMR relaxation or IR and Raman bandshapes. The *distinct* term explicitly expresses the coupling of the motion of molecule $i$ with the structure and dynamics of surrounding molecules, that is, it probes cooperative effects. Such terms also contribute to Rayleigh scattering, Kerr-effect relaxation, and quasi-elastic neutron scattering data. In combination with models and computer simulations this allows the deduction of intermolecular interaction potentials.

In liquids lacking specific long-range interactions, like hydrogen bonds, the reorientation of the molecular dipole moment can generally be described as rotational diffusion. The molecular relaxation time deduced from $\langle \boldsymbol{\mu}_i(t) \cdot \boldsymbol{\mu}_i(0) \rangle$ is proportional to molecular size and viscosity. For simple dipolar fluids of highly symmetric rigid molecules, like acetonitrile, the long-time behavior in the microwave region is reproduced by an exponential correlation function, and the influence of the *distinct* term can be expressed by a static correlation factor of orientational correlations, the well-known Kirkwood factor $g$, and a dynamic correlation factor $\dot{g}$. The short-time dynamics available from FIR spectra, however, deviates because of the predominance of libration and inertial effects. Reduction of molecular symmetry, or possible intramolecular rotations of polar groups, leads to the emergence of additional Debye-type relaxation processes. Examples are the observed second relaxation process when going from $C_n$ symmetry (acetonitrile) to $C_{2V}$ symmetry (benzonitrile) or the high-frequency relaxation process of butylene carbonate due to fast rotation of the ethyl side chain. As can be seen in Fig. 11 for $N,N$-dimethylformamide and $N,N$-dimethylacetamide, for this class of dipolar aprotic solvents the influence of dissolved electrolytes on the static permittivity is moderate (correctly, on the relaxation strength, but usually $\varepsilon_\infty(c) \approx \varepsilon_\infty(0)$). Additionally, the relaxation time parallels the observed increase of solution viscosity with electrolyte concentration.

The *distinct* term attains a predominant role in hydrogen-bonding liquids. Here, the intermolecular dynamics dominates the relaxation behavior with a slow cooperative process of large amplitude (relaxation strength), and the topology of the hydogen-bond system is reflected in the number of observed relaxation processes. This can be nicely exemplified with the series $N,N$-dimethylformamide (no H bonds), $N$-methylformamide (chains), and formamide (network). When winding chains are formed, as with alcohols or $N$-methylformamide, three relaxation processes are found, $\tau_1 > \tau_2 > \tau_3$. The magnitude of the cooperative relaxation time, $\tau_1 = 164$ ps for ethanol at 25°C, exceeds the relaxation time expected for molecular rotation (which is roughly equal to the observed $\tau_2 = 7.92$ ps) at least by a factor of five. $\tau_1$ and the amplitude $S_1$ are very sensitive to added solutes, especially electrolytes, as can be seen in Fig. 11. This process essentially probes the interchain dynamics. $\tau_2$ probes the reorientation of monomers and that of molecules with a single H bond, and the fast process with relaxation time $\tau_3 = 1.54$ ps is indicative for partial reorientation within the H-bonded chains. For molecules able to form hydrogen-bond networks, like water and formamide, only two relaxation processes are found (polyols show additional intramolecular dynamics). Compared to the value expected for rotational diffusion from molecular size and viscosity, the relaxation time $\tau_1$ of the dominating slow process ($S_1/(\varepsilon - \varepsilon_\infty > 0.95)$ is increased by a factor of 3–5. For water (Fig. 7), data suggest that $\tau_1$ (8.33 ps at 25°C) is a measure for the rate with which $H_2O$ molecules are released from the three-dimensional network before they can rapidly reorient with time constant $\tau_2 \approx 1$ ps. Experiments with dissolved electrolytes and nonelectrolytes reveal that up to intermediate concentrations, $c < 1$ mol dm$^{-3}$, such hydrogen-bond networks can accommodate solutes without much disturbance of the bulk solvent structure and dynamics (see Fig. 11).

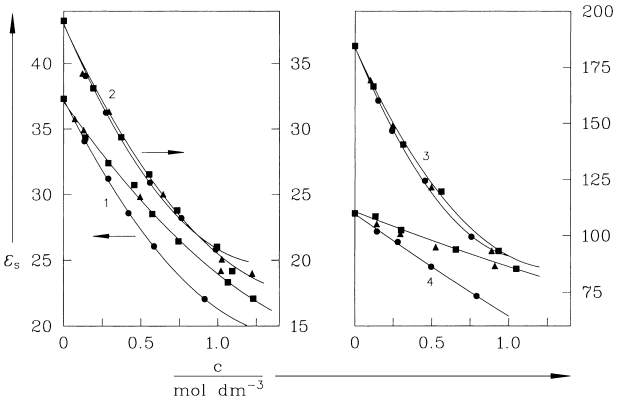



**FIGURE 11** Solvent permittivity $\varepsilon_S$ as a function of $LiClO_4$ (■), $NaClO_4$ (▲), and $Bu_4NClO_4$ (●) concentration in $N,N$-dimethylformamide (1), $N,N$-dimethylacetamide (2), $N$-methylformamide (3), and formamide (4) at 25 °C.

It is well known that especially liquids formed by large, asymmetric and flexible molecules tend to supercool and

to form a glass instead of crystallizing. It is thought that glass formation is closely connected with the cross-over from molecular reorientation which is dominated by the *self* term at high temperatures to cooperative relaxation close to the glass transition temperature, $T_g$. Currently such processes are a 'hot topic' in liquid state research. Investigations were mainly motivated by the emergence of the so-called *mode coupling theory* which predicts a universal scaling law for all dynamical variables in the system, linking thus viscosity, mechanical relaxation and dielectric relaxation.

The influence of dissolved nonelectrolytes in polar solvents is mainly controlled by volume effects. The solution of electrolytes, however, yields a decrease of the static solvent permittivity that significantly surpasses the volume requirements of the dissolved ions (Fig. 11). Two additional effects occur in electrolyte solutions: ion solvation and kinetic depolarization. Kinetic depolarization, a dynamic feature, arises from the interaction of the solvent molecules and the solvated ions moving in the external field. The moving ion creates a torque in the surrounding solvent dipoles opposite to the force of the external field, leading to a reduction of the orientational polarization of the solvent molecules and the mobility of the ions. The continuum theory of Hubbard and Onsager approximately accounts for this contribution. Ion solvation, a static effect, is caused by the high field strength at an ion surface aligning the solvent molecule dipoles in its vicinity so that they cannot contribute to the solvent relaxation. This permits one to deduce effective solvation numbers, $Z_{\mathrm{IB}}$, of ions from the solvent permittivity. As expected, the higher the charge density at the ionic surface, the greater the solvation effect ($Li^+ > Na^+ > K^+ > Rb^+ \approx Cs^+$). It should be noted that the thus-d etermined $Z_{\mathrm{IB}}$ may differ from first-shell coordination numbers deduced from scattering experiments or computer simulations, as the ability of the ions to align solvent dipoles may be rather small ($ClO_4^-$) or extend beyond the first solvation shell ($Mg^{2+}$).

Another effect characteristic of electrolyte solutions is ion association (Fig. 12), which may lead from encounter complexes (2SIP) where both ions keep their primary solvation shell via subsequent desolvation steps to solvent-shared (SSIP) and contact ion pairs (CIP) depending on the relative balance of ion–ion and ion–solvent interactions. Such speciation processes are important for many electrolyte systems of biological, geochemical, and technological interest. For chemical reactions involving ions, generally the CIP state has to be reached before the product can be formed. Despite of the importance of ion association, up to now many systems are only ill characterized because thermodynamic methods yield only the overall equilibrium constant, $K_{\mathrm{A}}$, whereas spectroscopic

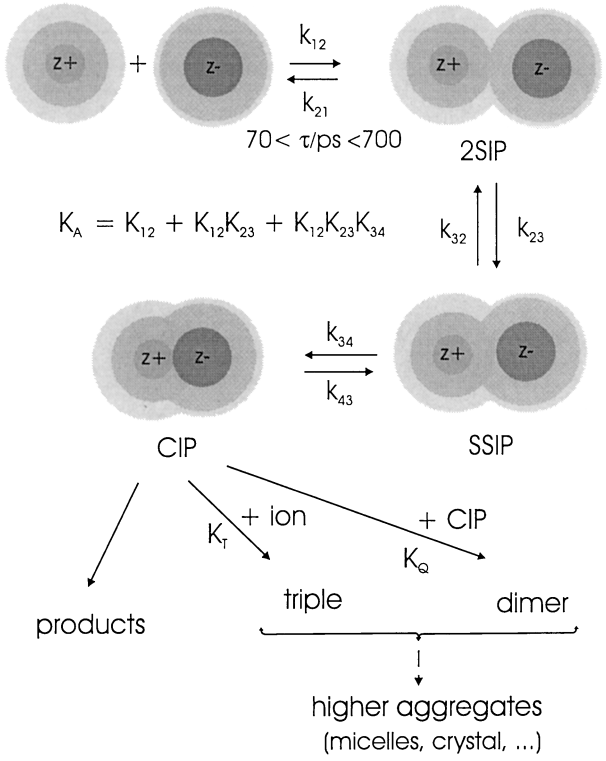



**FIGURE 12** Scenario of possible ion association equilibria in solution involving free ions, doubly solvent-separated (2SIP), solvent-shared (SSIP), and contact ion pairs (CIP), as well as possible higher aggregates. The $K_{ij}$ are equilibrium constants and the $k_{ij}$ rate constants relating different steps $i$ and $j$.

techniques are generally only sensitive to CIP. As 2SIP, SSIP, and CIP have permanent dipole moments, DRS is sensitive to all ion pair types, provided the lifetime of the species is at least comparable to their rotational correlation time. An example for ion pair relaxation can be seen in Fig. 6. In favorable cases dielectric studies, eventually combined with other techniques, not only permit the determination of the concentrations of the formed species but also allow one to estimate the rate constants of ion pair formation and decay from the concentration dependence of the ion pair relaxation time(s). Dielectric investigations of the dynamics of micelles fall into the same category.

In biological systems, both hydrogen bonding and ion–dipole interactions are of crucial importance. Dielectric relaxation studies can be carried out to study the interaction of water with biomolecules and its modulation by electrolytes. For instance, information on the flexibility of proteins or nucleic acids is available. Differences in solvent mobilities allow the identification of several binding states for water on biomolecules with the help of permittivity studies. DRS also becomes increasingly important in

material analysis and the characterization of pharmaceutical systems.

## SEE ALSO THE FOLLOWING ARTICLES

ELECTROLYTE SOLUTIONS, THERMODYNAMICS • ELECTROLYTE SOLUTIONS, TRANSPORT PROPERTIES • HYDROGEN BOND • INFRARED SPECTROSCOPY • LIQUIDS, STRUCTURE AND DYNAMICS • MICROWAVE COMMUNICATIONS

## BIBLIOGRAPHY

Bagchi, B., and Chandra, A. (1991). "Collective orientational relaxation in dense dipolar liquids," *Adv. Chem. Phys.* **80,** 1–126.

Barthel, J., Buchner, R., Eberspächer, P. N., *et al.* (1998). "Dielectric relaxation in electrolyte solutions. Recent developments and prospects," *J. Mol. Liq.* **78,** 82–109.

Barthel, J., and Buchner, R. (2000). "Relative permittivities of electrolytes," *In* "Experimental Thermodynamics" (Goodwin, A. R. H., Marsh, K. N., and Wakeham, W. A., eds.), Vol.6, Ch.9c, Blackwell, Oxford.

Buchner, R., and Barthel, J. (1994). "Dielectric relaxation in solutions," *Annu. Rep. Progr. Chem., Sect. C* **91,** 71–106.

Craig, D. Q. M. (1995). "Dielectric Analysis of Pharmaceutical Systems," Taylor & Francis, London.

Davis, J. L. (1990). "Wave Propagation in Electromagnetic Media," Springer, Berlin.

Kaatze, U. (1997). "The dielectric properties of water in its different states of interaction," *J. Solution Chem.* **26,** 1049–1112.

Madden, P., and Kivelson, D. (1984). "A consistent molecular treatment of dielectric phenomena," *Adv. Chem. Phys.* **56,** 467–566.

Scaife, B. K. P. (1989). "Principles of Dielectrics," Clarendon, Oxford.



# Perovskites

**C. N. R. Rao**
*Nehru Center for Advanced Scientific Research*

I. Crystal Chemistry
II. Electrical and Magnetic Properties of Perovskite Oxides
III. Magnetic Properties of Perovskite Fluorides
IV. Metal–Insulator Transitions
V. Oxides of $K_2NiF_4$ Structure
VI. Ferroics
VII. High-Temperature Superconductors
VIII. Colossal Magnetoresistance
IX. Why Are Perovskites Special?

## GLOSSARY

**Ferroics** Materials possessing two or more orientation states or domains that can be switched from one to another through the application of one or more appropriate forces.

**Intergrowths** Structures in which unit cells of two related materials occur randomly or recurrently.

**Magnetoresistance** Phenomenon whereby the resistance of a solid varies by the application of a magnetic field.

**Polytypism** Phenomenon that may be regarded as polymorphism in one dimension, exhibited by solids with close-packed and layered structures where the primary coordination around an atom is satisfied in more than one way (for example, cubic versus hexagonal close packing).

**Superconductivity** Phenomenon whereby the electrical resistance of a material vanishes below a critical temperature, accompanied by exclusion of the magnetic field.

**Transfer energy** Measures the strength of interaction between two orbitals of neighboring atoms and is proportional to the orbital overlap.

**PEROVSKITES** constitute one of the most fascinating classes of materials exhibiting diverse properties anywhere between ferroelectricity and superconductivity. Oxides and fluorides are the most commonly found materials of perovskite structure with the general composition $ABX_3$. A variety of other structures, especially in complex metal oxides, contain the perovskite unit. Structure and properties of perovskites constitute an excellent case study of the chemistry and physics of materials.

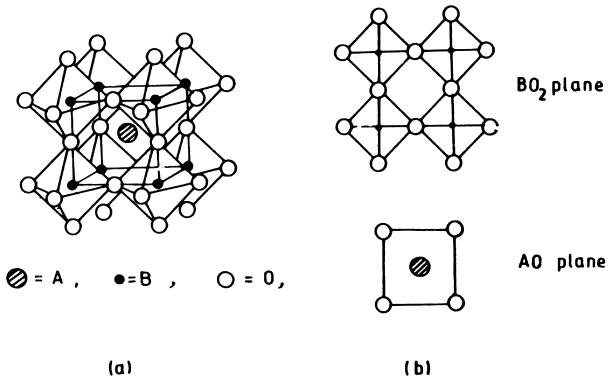



**FIGURE 1** (a) The $ABO_3$ perovskite structure. Without the large A atom in the body center position, the structure becomes that of cubic $ReO_3$; (b) layer sequence in the perovskite structure parallel to (001).

## I. CRYSTAL CHEMISTRY

Perovskites of the general formula $ABX_3$ may be regarded as derived from the $ReO_3$ structure as shown in Fig. 1. The $BX_3$ framework in the perovskite is similar to that in $ReO_3$ structure consisting of corner-shared $BX_6$ octahedra. The large A cation occupies the body center, 12-coordinate position. In an ideal cubic perovskite structure, where the atoms are just touching one another, the B–X distance is equal to $a/2$ and the A–X distance is $\sqrt{2}(a/2)$, where $a$ is the cube unit cell length and the following relation between radii of ions holds: $R_A + R_X = \sqrt{2}(R_B + R_X)$. Goldschmidt found that the perovskite structure is retained in $ABX_3$ compounds even when this relation is not exactly obeyed and defined a tolerance factor, $t$, as

$$t = \frac{R_A + R_X}{\sqrt{2}(R_B + R_X)}$$

For the ideal perovskite structure, $t$ is unity. The perovskite structure is, however, found for lower values of $t(\sim 0.75 < t \leq 1.0)$, also. In such cases, the structure distorts to tetragonal, rhombohedral, or orthorhombic symmetry. This distortion arises from the smaller size of the A ion, which causes a tilting of the $BX_6$ octahedra in order to optimize A–X bonding. Perovskite oxides, $ABO_3$, can be thought of as consisting of alternating $BO_2$ and AO layers stacked one over the other in the [001] direction. An alternative description of the $ABO_3$ structure in terms of close packing of A and O ions is one where close-packed $AO_3$ layers [Fig. 2(a)] are stacked one over the other with the B cations occupying octahedral holes surrounded by oxygen.

Several $ABO_3$ oxides, where A is a large cation such as Ba and B is a small cation of the $d$-transition series, are known to exhibit polytypism. The stacking of an $AO_3$ layer in the structure may be cubic (c) or hexagonal (h) with respect to its two adjacent layers depending on whether it is in the middle of the ABA or ABC sequence. If the stacking is entirely cubic, the B-cation octahedra share only corners in three dimensions to form the perovskite (3C) structure (Fig. 2). If the stacking is all hexagonal, the

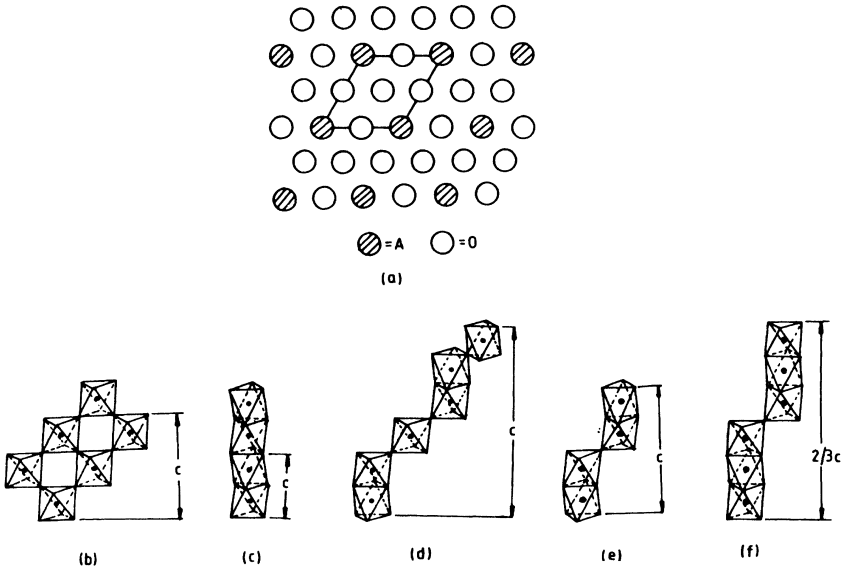



**FIGURE 2** (a) Close-packed $AO_3$ layer in perovskites. (b)–(f) $BO_6$ octahedra in different perovskite polytypes: (b) 3C, (c) 2H, (d) 6H, (e) 4H, and (f) 9R.

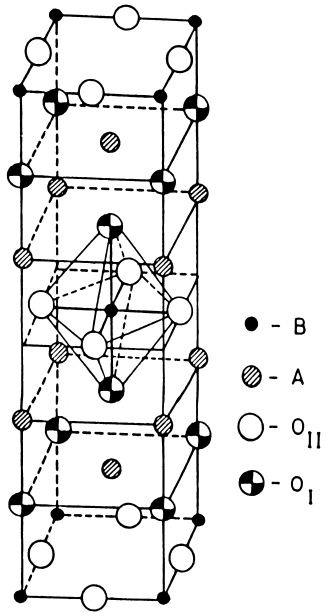



**FIGURE 3** The $K_2NiF_4$ structure of oxides, $A_2BO_4$.

B-cation octahedra share opposite faces, forming chains along the *c*-axis as in $BaNiO_3$ (2H). In between the two extremes, there can be several polytypic structures consisting of mixed cubic and hexagonal stacking of $AO_3$ layers; for example, the 6H and 4H polytypes have the stacking sequences cch cch and chch, respectively. Typical $ABO_3$ oxides showing polytypism are $BaCrO_3$, $BaMnO_3$, and $BaRuO_3$.

Oxides of the general formula $A_2BO_4$ crystallize in the $K_2NiF_4$ structure, which is closely related to the perovskite structure. The tetragonal structure of $K_2NiF_4$ (Fig. 3) can be regarded as consisting of $KNiF_3$ perovskite slabs of one unit cell thick, which are stacked one over the other along the *c*-direction. The adjacent slabs are displaced relative to one another by $\frac{1}{2}\,\frac{1}{2}\,\frac{1}{2}$, such that the *c*-axis of the tetragonal structure is roughly equal to three times the cell edge of the cubic perovskite. The structure is two-dimensional in the sense that only the equatorial anions of the $NiF_6$ octahedra are linked through corners. Tolerance factors for the $K_2NiF_4$ structure can be worked out just as for perovskites, and oxides of this structure often show orthorhombic distortion. Oxides of the $K_2NiF_4$ structure (for example, $La_2CuO_4$, $LaNiO_4$) have been investigated extensively.

The perovskite structure can tolerate vacancies at the A or X sites giving rise to nonstoichiometric compositions, $A_{(1-x)}BX_3$ and $ABX_{3-x}$. B-site vacancies are energetically not favored unless there are compensating factors such as B–B interaction. Typical examples of A-site vacancies are the tungsten bronzes, $A_xWO_3$ and $Cu_{0.5}TaO_3$; brownmillerite, $CaFeO_{2.5}$, is an example of an anion-deficient perovskite. Perovskite-type oxides also show anion excess nonstoichiometry as in the case of $LaMnO_{3+x}$, where the apparent anion excess probably arises from La vacancies. Examples of B-site vacancy hexagonal perovskites are $Ba_3Re_2O_9$ and $Ba_5Nb_4O_{15}$.

There are many other interesting oxide systems possessing perovskite units. There is a family of oxides, first described by Aurivillius, of the general formula $Bi_2A_{n-1}B_nO_{3n+3}$ containing $(Bi_2O_2)^{2+}$ layers and $(A_{n-1}B_nO_{3n+1})^{2-}$ perovskite layers. Typical members of this family are $Bi_4Ti_3O_{12}(n=3)$ and $BaBi_4Ti_4O_{15}(n=4)$. These oxides form disordered as well as ordered intergrowth structures. There are other intergrowth structures in oxides derived from the perovskite structure. Thus, the $A_nB_nO_{3n+2}$ family consists of slabs of $A_{n-1}B_nO_{3n+2}$ obtained by cutting the perovskite structure parallel to the (110) planes; a series of oxides with *n* between 4 and 4.5 is known in the Na–Ca–Nb–O system. The $A_{n+1}B_nO_{3n+1}$ family (for example, in Sr–Ti–O and La–Ni–O systems) is generated by cutting the perovskite structure into slabs along the (100) planes.

## II. ELECTRICAL AND MAGNETIC PROPERTIES OF PEROVSKITE OXIDES

Perovskite oxides exhibit a variety of electronic properties. Thus, $BaTiO_3$ if ferroelectric, $SrRuO_3$ is ferromagnetic, $LaFeO_3$ is weakly ferromagnetic, $BaPb_{1-x}Bi_xO_3$ is superconducting, and $LaCoO_3$ shows an insulator—metal transition. Properties of known perovskites have been compiled by Goodenough and Longo as well as Nomura. Several perovskite oxides exhibit metallic conductivity, typical examples being $ReO_3$, $A_xWO_3$, $LaTiO_3$, $AMoO_3$ (A = Ca, Sr, Ba), $SrVO_3$, and $LaNiO_3$. Metallic conductivity in perovskite oxides is caused by strong cation–anion–cation interaction.

We have listed important perovskite oxides containing B-site transition metal atoms in Fig. 4, where oxides with the same *d*-electron configuration are grouped together in the columns. Entries in each column are arranged in the decreasing order of B cation–anion transfer energy *b* (B–O covalency) from top to bottom. Covalent mixing parameters $\lambda_\sigma$ and $\lambda_\pi$ (and hence transfer energies $b_\sigma$ and $b_\pi$) increase with the increasing valence state of the B cation; for the same valence state, mixing varies as $5d > 4d > 3d$. The influence of A cations on the covalency of the B–O bond is indirect. Acidic A cations decrease B–O covalency; $\lambda_\sigma > \lambda_\pi$ in all the compounds.

The dotted lines in Fig. 4 representing $b_\pi = b_m$ ($b_m$ is the critical value for spontaneous magnetism), $b_\pi = b_c$,

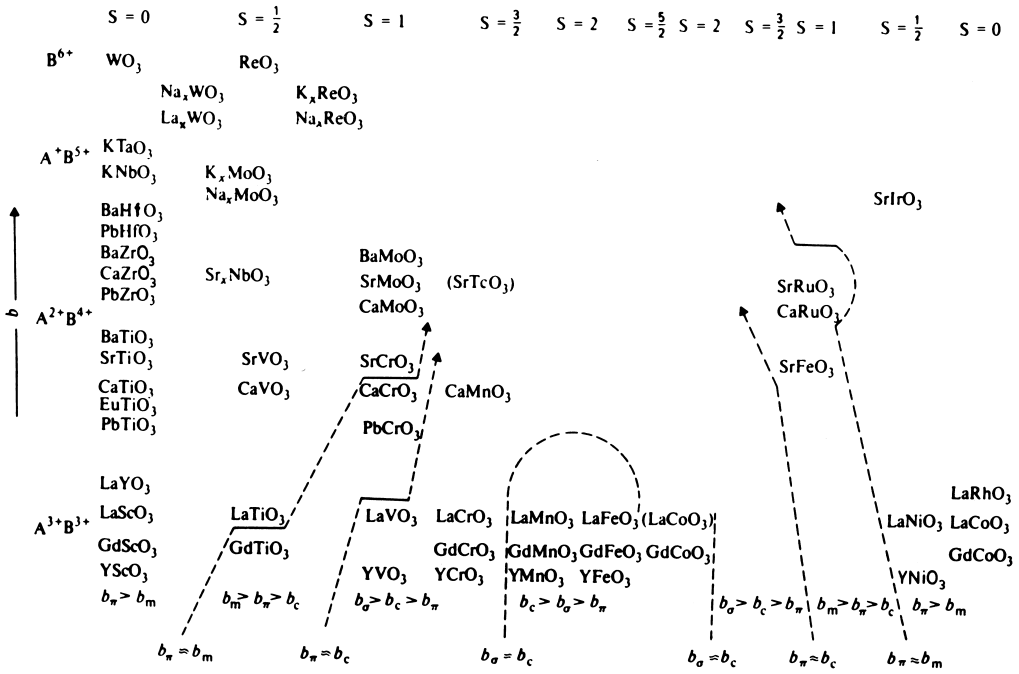



**FIGURE 4** Periodic table of perovskites. [From Goodenough, J. B. (1974) *In* "Solid State Chemistry" (C. N. R. Rao, ed.), Dekker, New York.]

and $b_\sigma = b_c$ ($b_c$ is the critical value of the transfer energy) separate oxides exhibiting localized electron behavior from those with collective electron properties. Compounds in column 1 are insulators because the B cations are of $d^\circ$ electron configuration. Most of the compounds in column 2 (spin $S = \frac{1}{2}$) are metallic and Pauli paramagnetic; the line $b_\pi = b_m$ separates $LaTiO_3$ from $GdTiO_3$ because $GdTiO_3$ is a semiconductor with a ferromagnetic Curie temperature ($T_c$) of 21 K. $AMoO_3$ (A = Ca, Sr, Ba) and $SrCrO_3$ in the third column ($S = 1$) are metallic and Pauli paramagnetic. Other compounds in this column are semiconducting and antiferromagnetic. The line $b_\pi = b_m$ separates metallic and Pauli paramagnetic $SrCrO_3$ from the antiferromagnetic semimetal $CaCrO_3$. The line $b_\pi = b_c$ separates $PbCrO_3$ from $LaVO_3$ because the latter exhibits a crystallographic transition at a temperature lower than the Neel temperature ($T_N$) characteristic of localized electrons. The region $b_m > b_\pi > b_c$ appears to be quite narrow as revealed by electrical, magnetic, and associated properties. Pressure experiments are valuable in the study of this region; thus, $dT_N/dP < 0$ in $CaCrO_3$, while $dT_N/dP > 0$ in $YCrO_3$ and $CaMnO_3$. Since increasing pressure increases $b_\pi$ (by decreasing lattice dimensions), $dT_N/dP > 0$ for $b_\pi < b_c$ (localized behavior) and $dT_N/dP < 0$ for $b_m > b_x > b_c$ (collective behavior). Compounds in columns 4, 5, and 6 are antiferromagnetic insulators. Since the intraatomic exchange $\approx S(S+1)$ decreases the covalent mixing, it is natural that maxima in the curves $b_\pi = b_c$ and $b_\sigma = b_c$ corresponding to smallest values of $b_\pi$ and $b_\sigma$ occur in the middle of the columns with $S = \frac{5}{2}$. Rare earth orthoferrites with $S = \frac{5}{2}$ are antiferromagnetic insulators and exhibit parasitic ferromagnetism. The important contributions here are: (a) $Fe^{3+}$ spins canted in a common direction either by cooperative buckling of oxygen octahedra or by anisotropic superexchange and (b) canting of an antiferromagnetic rare earth sublattice because of interaction between two sublattices.

$LaCoO_3$ is shown twice in Fig. 4, both in the $S = 2$ and $S = 0$ columns at the end, since $Co^{3+}$ in this solid can have either the low-spin or the high-spin configuration. The compound exhibits a transition from a localized electron state to a collective electron state (metal-insulator transition). In the ninth column of Fig. 4, perovskites containing $d^4$ cations are placed. Of the three compounds in this column, $SrRuO_3$ is a ferromagnetic metal ($T_c = 160$ K); $CaRuO_3$ is antiferromagnetic ($T_N = 110$ K) with a weak ferromagnetism. Since both the compounds have the same $RuO_3$ array, the change from ferromagnetic to antiferromagnetic coupling is of significance. $SrFeO_3$ is placed in the same column on the assumption that $Fe^{4+}(3d^4)$ is in the low-spin state, but recent work suggests that $Fe^{4+}$ in this oxide is in the high-spin state down to 4 K. $CaFeO_3$, on the other hand, shows disproportionation of $Fe^{4+}$ to $Fe^{3+}$ and

$Fe^{5+}$ below 290 K. In the last but one column containing $S = \frac{1}{2}$B cations, metallic and Pauli paramagnetic $LaNiO_3$ should be separated from antiferromagnetic $YNiO_3$ and $LuNiO_3$ indicating that in $LaNiO_3$ $b_\sigma > b_m$ and in $YNiO_3$ $b_\sigma < b_m$. Similarly, in the last column, $LaCoO_3$ should be separated from $LaRhO_3$ because the latter is a narrow gap semiconductor with a filled $t_{2g}(\pi^*)$ band and an empty $e_g(\sigma^*)$ band.

## III. MAGNETIC PROPERTIES OF PEROVSKITE FLUORIDES

In oxides, ferrimagnetism is common in spinel, garnet, and magnetoplumbite structures because of the occupation of tetrahedral, octahedral, or dodecahedral sites by magnetic ions. In fluorides, the transition metal cations invariably occupy octahedral sites. Ferrimagnetism therefore results from the manner in which the octahedra are linked. In fluorides of hexagonal $BaTiO_3$ structure, for example, one of the sublattices consists of octahedra linked by corners and the other octahedra linked by faces; one third of the metal ions are present in corner-shared octahedra and two-thirds in face-shared octahera. Both the sublattices have different magnetic moments that are coupled antiferromagnetically leading to ferrimagnetism. $CsFeF_3$ is a typical material possessing this structure; Fe–Fe interaction between face-shared octahedra is ferromagnetic and that between corner-shared octahedra is antiferromagnetic, giving rise to ferrimagnetism with a $T_c$ of 60 K. Isostructural $RbMnF_3$ is, however, antiferromagnetic since all interactions between the neighboring octahedra are antiferromagnetic in this compound.

$KCrF_3$ is an antiferromagnetic solid adopting an antiferro-distortive structure, the magnetic structure being similar to that of $LaMnO_3$ (A type). Superexchange interaction $d^0_{x^2-y^2} - p - d^1_{z^2}$ leads to ferromagnetic layers that are antiferromagnetically coupled through the empty $d_{x^2-y^2}$ orbitals. $KCuF_3$ is a 1D antiferromagnet (A type) with the spins lying in the *ab* plane. The magnetic behavior is again a consequence of antiferro-distortive ordering of distorted octahedra. Interaction between two half-filled $d_{x^2-y^2}$ orbitals occurs along the *c*-axis, while interchain coupling is through filled $d_{z^2}$–half-filled $d_{x^2-y^2}$ interaction in the *ab* plane. There are two forms of tetragonal $KCuF_3$, one with I4/*mcm* symmetry and the other P4/*mbm* symmetry; the 1D character is more pronounced in the latter. $CsNiF_3$ is the only fluoride crystallizing in the hexagonal 2H perovskite structure, where infinite chains of face-sharing $NiF_6$ octahedra parallel to the *c*-axis exist. It exhibits 1D ferromagnetism at high temperatures ($70 < T < 300$ K) and 3D antiferromagnetism at low temperatures. Neutron diffraction and specific heat measurements show that the 3D transition occurs at 2.65 K. The 3D magnetic structure consists of ferromagnetic planes parallel to the *c*-axis that are coupled antiferromagnetically.

Among the fluorides with layered structures, the $K_2NiF_4$ family has been widely investigated. $K_2NiF_4$ is a two-dimensional Heisenberg antiferromagnet with $T_N = 97$ K and $J/k = -50$ K. The isostructural $K_2CoF_4$ behaves as a 2D, $S = \frac{1}{2}$, Ising system ($T_N = 107.8$ K, $J/k = -97$ K). $K_2CuF_4$ (and its rubidium and cesium analogs) crystallizing in an orthorhombic-distorted $K_2NiF_4$ structure are 2D Heisenberg ferromagnets. The distortion of the structure and ferromagnetic properties arise from the ordering of the elongation axis of the $CuF_6$ octahedra alternately in the *a* and *b* directions. The intralayer exchange constant $J/k$ is ~11 K and the value for interlayer coupling is ~0.03 K.

## IV. METAL–INSULATOR TRANSITIONS

$LaCoO_3$, which is an insulator at ordinary temperatures, becomes metallic at high temperatures ($\gtrsim 1200$ K). More interesting are the transitions from the metallic to the insulating state brought about by compositional changes. Oxides of the type $La_{1-x}Sr_xMO_3$ (M = V or Co) show metal–insulator transitions with an increase in $x$. Thus, $La_{1-x}Sr_xCoO_3$ becomes metallic at all temperatures when $x > 0.3$, while $LaCoO_3(x = 0)$ is an insulator at room temperature. When M = Mn or Co, the oxide becomes ferromagnetic at the same composition when the *d*-electrons become itinerant. Another interesting system showing compositionally controlled metal–insulator transitions is $LaNi_{1-x}M_xO_3$, where M = Cr, Mn, Fe, or Co. In this system, the metallic resistivity of the $x = 0$ oxide gives way to a semiconducting or an insulating behavior above a particular value of $x$. Such a change–over occurs essentially at a universal value of resistivity (2000 $\mu\Omega$ cm) in all these oxide systems, the value corresponding closely to Mott's minimum metallic conductivity. A metal–insulator transition is brought about in the $La_{4-x}Ba_{1+x}Cu_5O_{13+\delta}$ system by a change in the La:Ba ratio or oxygen stoichiometry.

## V. OXIDES OF $K_2NIF_4$ STRUCTURE

Quasi-two-dimensional oxides of the type $A_2BO_4$ possessing the $K_2NiF_4$ structure contain the $ABO_3$ perovskite layers in between the rock-salt AO layers with B–O–B interaction occurring only in the *ab* plane. Electrical and magnetic properties of the $A_2BO_4$ oxides are considerably different from those of the $ABO_3$ perovskites.

Accordingly, $LaNiO_3$ is metallic and Pauli paramagnetic, while $La_2NiO_4$ exhibits two-dimensional antiferromagnetic ordering around 200 K and a semiconductor–metal transition around 600 K. In the $LaO(LaNiO_3)_n$ family, the electrical conductivity decreases with an increase in $n$, becoming essentially metallic when $n = 3$. $SrRuO_3$ is a metallic ferromagnet, but $Sr_2RuO_4$ is a paramagnetic insulator. A more interesting comparison is provided by $LaCuO_3$ and $La_2CuO_4$. The former is a metal and the latter is an antiferromagnetic insulator. Oxygen-excess $La_2CuO_4$ and $La_{2-x}Sr_x(Ba_x)CuO_4$ are superconductors, ($T_c \sim 30$–40 K), although in the normal state, they are marginally metallic. A strict comparison of the properties of three- and two-dimensional oxides can be made only when the *d*-electron configuration of the transition metal ion B is the same. A comparative study of the two systems has been made with respect to their electrical and magnetic properties. For example, members of the $La_{1-x}Sr_{1+x}CoO_4$ are all semiconductors with a high activation energy for conduction unlike $La_{1-x}Sr_xCoO_3 (x \geq 0.3)$ which is metallic; the latter oxides are ferromagnetic. $La_{0.5}Sr_{1.5}CoO_4$ shows a magnetization of $0.5\mu_B$ at 0 K (compared to $1.5\mu_B$ of $La_{0.5}Sr_{0.5}CoO_3$), but the high-temperature susceptibilities of the two systems are comparable. In $SrO(La_{0.5}Sr_{0.5}MnO_3)_n$, both magnetization and electrical conductivity increase with an increase in $n$ approaching the value of the perovskite, $La_{0.5}Sr_{0.5}MnO_3$. $LaSrMn_{0.5}Ni_{0.5}(Co_{0.5})O_4$ shows no evidence of long-range ferromagnetic ordering, unlike the perovskite $LaMn_{0.5}Ni_{0.5}(Co_{0.5})O_3$; high-temperature susceptibility behavior of these two insulating systems is, however, similar. $LaSr_{1-x}Ba_xNiO_4$ exhibits high electrical resistivity with the resistivity increasing proportionately with the magnetic susceptibility (note that $LaNiO_3$ is a Pauli paramagnetic metal). High-temperature susceptibilities of $LaSrNiO_4$ and $LaNiO_3$ are comparable. Susceptibility measurements show no evidence for long-range ordering in $LaSrFe_{1-x}Ni_xO_4$, unlike that in $LaFe_{1-x}Ni_xO_3$ ($x \leq 0.35$), and the electrical resistivity of the former is considerably higher.

## VI. FERROICS

Ferroics are materials possessing two or more orientation states or domains that can be switched from one to another through the application of one or more appropriate forces. In a ferromagnet, the orientation state of magnetization in domains is switched by the application of a magnetic field. In a ferroelastic, the direction of spontaneous strain in a domain is switched by the application of mechanical stress. In a ferroelectric, spontaneous electric polarization is altered by the application of an electric field. These three ferroics are primary ferroics since they are governed by switchability of the properties. Perovskites provide many examples of ferroics. $BaTiO_3$, $Bi_4Ti_3O_{12}$, and $KNbO_3$ are well-known ferroelectrics, while $PbZrO_3$ and $NaNbO_3$ are antiferroelectric. Some of the perovskites exhibit paired properties: ferroelectric-ferroelastic, $KNbO_3$; ferroelectric-antiferromagnetic, $HoMnO_3$; ferroelectric-superconducting $SrTiO_3$; and antiferroelectric-antiferromagnetic, $BiFeO_3$.

There are several secondary ferroic properties that occur as induced quantities, and the orientation states differ in derivative quantities that characterize the induced effects (for example, induced electric polarization by dielectric susceptibility). Thus, $SrTiO_3$ is a secondary ferroic showing ferrobielectricity.

Materials such as $Pb(Mg_{1/3}Nb_{2/3})O_3$ are relaxor ferroelectrics. Materials such as $Pb(Zr_{1-x}Ti_x)O_3$ or PZT are electro-optic materials. Aurivillius oxides of the formula $Bi_2A_{n-1}B_nO_{3n+1}$ are high-$T_c$ ferroelectrics.

## VII. HIGH-TEMPERATURE SUPERCONDUCTORS

Superconductivity in perovskite oxides has been known for some time, the highest $T_c$ observed until 1987 being ~13 K in $Ba(Bi, Pb)O_3$. The oxide system where high $T_c$ was first reported in the 30–40-K region, $La_{2-x}Ba_x(Sr_x)CuO_4$, has the tetragonal $K_2NiF_4$ structure at ordinary temperatures, but becomes orthorhombic around 180 K. Oxygen-excess $La_2CuO_4$ also shows superconductivity in the 30-K region. Superconducting oxides of the 123 type ($T_c \sim 90$ K) with the general formula $LnBa_2Cu_3O_{7-\delta}$ (Ln = Y, La, Nd, Sm, Eu, Gd, etc.) are defect perovskites containing Cu–O sheets as well as Cu–O chains, the latter imparting the orthorhombic structure. The 123 oxides are the $x = 1$ members of the $Ln_{3-x}Ba_{3+x}Cu_6O_{14+\delta}$ defect perovskites. The bismuth and thallium cuprate superconductors conforming to the formulas $Bi_2(Ca, Sr)_{n+1}Cu_nO_{2n+4}$, $Tl_2Ca_{n+1}Ba_2Cu_nO_{2n+4}$, $TlCa_{n-1}Ba_2Cu_nO_{2n+3}$, with $T_cs$ reaching 125 K contain defect perovskite layers and rock-salt-type layers.

In Fig. 5, the schematic structures of $n = 2$ cuprates of different families are shown in order to illustrate how they arise from the intergrowth of perovskite and rock salt layers. The family of mercury cuprates is also related to these intergrowths. The highest $T_c$ to date is in a Hg cuprate of the formula $HgBa_2Ca_2Cu_3O_8$ which becomes superconducting at 165 K under pressure. It may be noted that the highest $T_c$ found in a copper-free oxide material

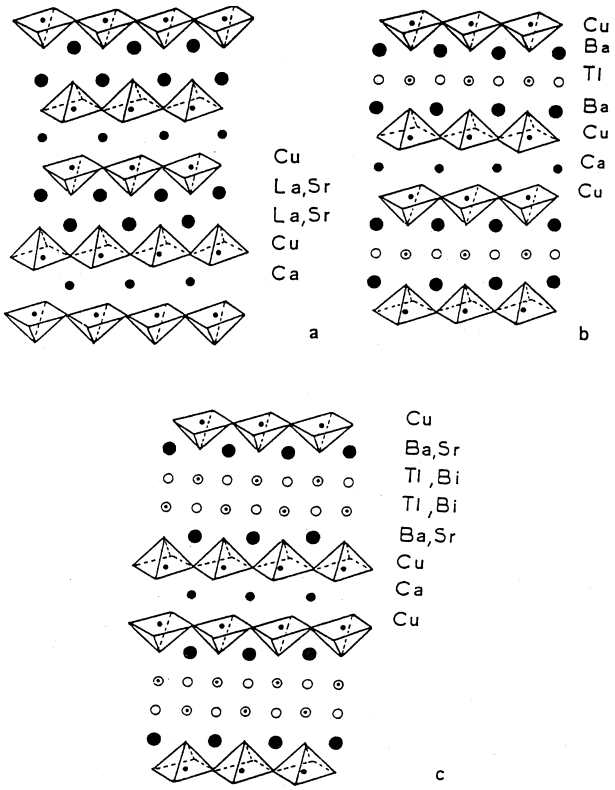



**FIGURE 5** Schematic structure of $n=2$ cuprates $(ACuO_{3-x})_m(AO)_n$: (a) $LaSrCaCu_2O_6$, (b) $TlBa_2CaCu_2O_7$, and (c) $Tl_2Ba_2CaCu_2O_8$ or $Bi_2Sr_2CaCu_2O_8$.

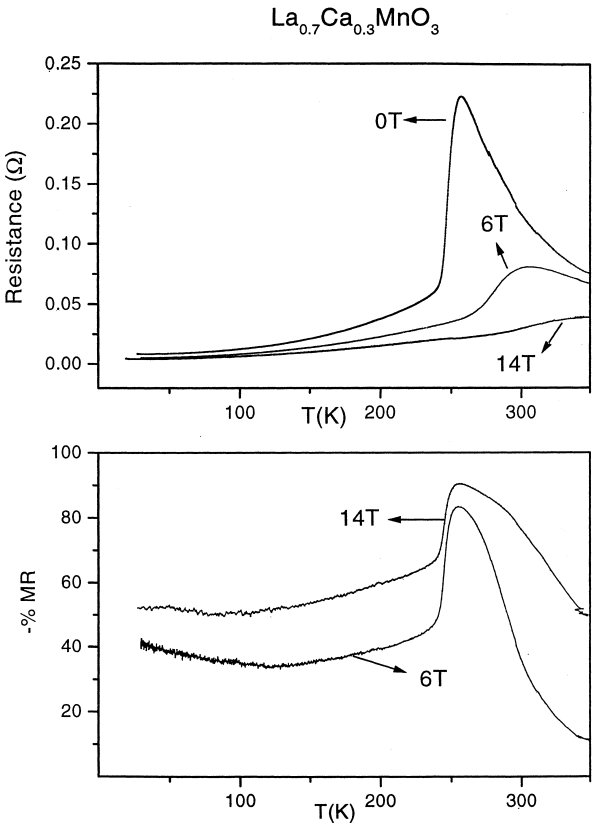



**FIGURE 6** Electrical resistivity data of $La_{0.7}Ca_{0.3}MnO_3$ in the absence and presence of magnetic field. Temperature variation of magnetoresistance (MR) is also shown.

is ~30 K in $Ba_{1-x}K_xBiO_3$ which is a three-dimensional perovskite.

## VIII. COLOSSAL MAGNETORESISTANCE

Rare earth manganates of the formula $Ln_{1-x}A_xMnO_3$ (Ln = rare earth, A = alkaline earth) with the perovskite structure show many interesting properties. Because of the double-exchange mechanism of electron hopping between $Mn^{3+}$ and $Mn^{4+}$ ions, these materials exhibit ferromagnetism and an insulator-metal transition at the ferromagnetic transition temperature, $T_c$. Application of a moderate or high magnetic field (1–6T) causes a large decrease in the resistivity, particularly around $T_c$. The negative magnetoresistance can be as high as 100%, and hence the term colossal magnetoresistance. In Fig. 6, typical magnetoresistance data are shown in the case of $La_{0.7}Ca_{0.3}MnO_3$. The $T_c$ in these manganates is extremely sensitive to the average size as well as the size mismatch of the A-site cations. Charge-ordering exhibited by some of the manganate compositions is also sensitive to the average size of the A-site cations. The competing interactions present in the manganates such as double-exchange and Jahn–Teller distortion (and also charge-ordering) are responsible for the fascinating phenomena and properties of these materials.

## IX. WHY ARE PEROVSKITES SPECIAL?

There is hardly any other class of solids that exhibits the variety of fascinating properties as the perovskites. Thus, perovskite oxides show high-temperature superconductivity, colossal magnetoresistance, and a variety of ferroic properties. This is because of the unique structure of perovskites wherein the B–O interaction, B–O–B angle, BB transfer integral, and other important factors can be sensitively varied by changing the A-site or B-site cations. We should remember that there is no B–B interaction in perovskites. It is by and large the B–O–B interaction and the nature of the $BO_6$ octahedra that determine the properties. The sensitivity of the properties of perovskites to the cations in the A- and B-sites illustrates why perovskites are special. Thus, varying the A-site cations can affect the

properties through changes in (a) the tolerance factor and lattice distortion, (b) B–O–B angle and related parameters, (c) disorder due to A-site ion size mismatch, and (d) B-O bonding through competitive interaction, involving $\sigma$ or $\pi$ bonds. Varying the B-site cations can change the properties because of changes in (a) the tolerance factor, (b) spin and electronic configuration, and (c) disorder and size effects.

## SEE ALSO THE FOLLOWING ARTICLES

CRYSTALLOGRAPHY • FERROMAGNETISM • MATERIALS CHEMISTRY • SOLID-STATE CHEMISTRY • SUPERCONDUCTIVITY • SUPERCONDUCTORS, HIGH TEMPERATURE

## BIBLIOGRAPHY

Goodenough, J. B. (1974). *In* "Solid State Chemistry" (C. N. R. Rao, ed.). Dekker, New York.

Goodenough, J. B., and Longo, J. M. (1970). "Landbolt-Bornstein Tabellen, New Series III/4a," Springer-Verlag, Berlin and New York.

Nomura, S. (1978). "Landbolt-Bornstein Tabellen, New Series III/12a," Springer-Verlag, Berlin and New York.

Rao, C. N. R., and Gopalakrishnan, J. (1997). "New Directions in Solid State Chemistry," Cambridge Univ. Press, London and New York, Second Ed.

Rao, C. N. R., and Raveau, B. (1998). "Transition Metal Oxides," VCH-Wiley, New York, Second Ed.

Rao, C. N. R., and Raveau, B. (1999). "Colossal Magnetoresistance, Charge-Order and Related Properties of Manganese Oxides," World Scientific, Singapore.



# Positron Microscopy

**Arthur Rich, deceased**
**James Van House**
*University of Michigan*

I. Production of Low-Energy Positron Beams
II. Overview of Positron Microscopy
III. Theoretical Basis
IV. The Transmission Positron Microscope
V. Positron Reemission Microscopy

## GLOSSARY

**Brightness** Particle intensity of a beam per unit solid angle per unit energy.
**Brightness enhancement** Increase in the brightness of a positron beam by use of one or more stages of moderation. The nonconservative moderation process circumvents the restrictions described by Liouville's theorem.
**Liouville's theorem** States that an ensemble of particles occupies a constant volume in phase space when acted on by conservative forces.
**Moderator** Material with a negative surface affinity for positrons that is used to stop a flux of high-energy positrons. Some fraction of the incident flux is reemitted at an energy, typically 1 eV, determined by the negative surface work function.
**Positronium** Hydrogenlike bound state of a positron and an electron.

**THE POSSIBILITY** of obtaining images of the microscopic world with resolutions of the order of atomic dimensions ($\sim$1 Å) was made possible by the introduction of the transmission electron microscope (TEM) over 50 years ago. Since that time, a variety of instruments have been constructed to improve on this resolution or to obtain different contrasts by using a particle other than the electron as a microscopic probe of the target being investigated. Because of the positron's positive charge and other antimatter properties, such as annihilation with an electron to produce $\gamma$-rays, positron imaging promises to reveal features of materials not previously seen with electrons.

## I. PRODUCTION OF LOW-ENERGY POSITRON BEAMS

All the positron microscopes exploit the now well-developed technology of slow positron beams. These beams are typically generated from a radioactive source (usually $^{22}$Na or $^{58}$Co) by a process called moderation. In the moderation process, the initially high-energy positrons generated by the decay of radioactive nuclei are slowed from their initial energy of several hundred kiloelectron

volts to thermal energies ($\sim\frac{1}{40}$ eV) in a material, called a moderator, whose surface has a negative work function for positron emission. About 1 in $10^3$ of the incident high-energy positrons diffuse to the surface where the negative work function for positrons causes the spontaneous emission of the positrons into the vacuum. The emission kinetic energy is usually about 1 eV, and the emission energy spread is about 0.1 eV. These reemitted slow positrons are then accelerated and focused into a beam. The moderation process has the virtue that the initial energy spread (typically 500 keV) of the positrons emitted by the radioactive source is concentrated into an energy band of 0.1 eV in the reemitted positron beam, resulting in a 5000-fold increase in the brightness of the positron flux emitted from the radioactive source. Brightness, which is defined as the intensity of particles per unit solid angle per unit energy, is a measure not only of the beam intensity, but of its angular divergence as well. Increases in beam brightness for systems acted on by conservative processes are prohibited by Liouville's theorem, which states that an ensemble of particles occupies a constant phase space volume when acted on by conservative forces. The moderation of positrons, however, is not a conservative process, thus Liouville's restriction does not apply. Furthermore, as originally proposed by A. P. Mills, Jr., one could generate a positron beam with enhanced brightness by accelerating and focusing the beam onto a small spot on a crystal. That crystal, the remoderator, then reemits low-energy positrons from a smaller area than the original moderator diameter, and thus has enhanced brightness. Through successive stages of this brightness enhancement, one loses intensity (typically less than half the incident positrons are reemitted from the remoderator for 1000-V incident energy), but one gains a factor of 10–100 in brightness per stage.

## II. OVERVIEW OF POSITRON MICROSCOPY

The first positron microscope image (Fig. 1) was taken at the University of Michigan and appeared in the January 18, 1988, issue of *Physical Review Letters*. It was taken by the positron analog of the TEM, and is called the transmission positron microscope (TPM). While the initial TPM resolution of 5 $\mu$m falls significantly short of TEM standards, improvements in resolution to below 1 nm should be quite feasible, as discussed later. Using the TPM in direct conjunction with a TEM, the TPM and TEM images can be compared. The differences observed in such a comparison, caused by the opposite sign of the coulomb interaction for positrons versus electrons, will contain information on atomic form factors, as well as on the material composition of the sample being studied.

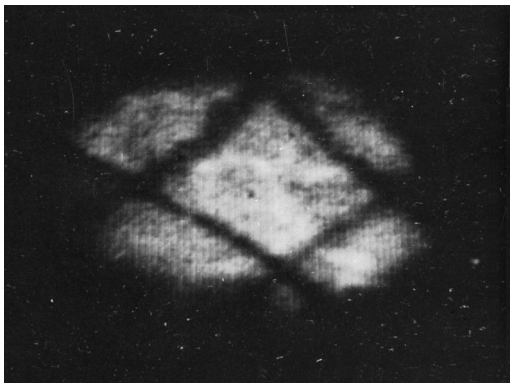


**FIGURE 1** The first positron microscope image. The image, discussed in detail in Section II, is of a V.Y.N.S. foil. The image magnification is 29× and the resolution is 4.5 mm. [From Van House, J., and Rich, A. (1988). *Phys. Rev. Lett.* **60,** 169–172.]

In addition, a small fraction of the positrons will annihilate, during passage through the sample, into two or three $\gamma$-rays. Detection of these $\gamma$-rays could provide information on the momentum space distribution of the electrons bound in the sample.

The potential for developing the positron counterpart to the scanning electron microscope (SEM) was demonstrated at Brandeis University. As discussed in the February 1988 edition of the *Review of Scientific Instruments*, using a brightness enhanced beam, the Brandeis group, in collaboration with Bell Laboratories, have developed a scanning positron microprobe that utilizes the annihilation of a positron with an electron to give $\gamma$-rays as a signal. This instrument was used to produce a one-dimensional $\gamma$-ray scan of a target with a resolution of 20 $\mu$m. Such $\gamma$-ray images will be sensitive to large-scale defects because positrons are readily trapped in sample defects. While the demonstrated resolution is not as good as the resolution of the prototype TPM, and falls far short of electron SEM standards, resolutions of 100 nm (limited by positron diffusion) should eventually be possible using higher intensity radioactive sources and a sufficient number of brightness enhancement stages.

The potentially most useful positron microscope exploits the phenomenon of spontaneous positron reemission by the negative surface work function. The first image from such a positron reemission microscope (PRM) was taken by the University of Michigan group. This first PRM operated in a reflection geometry and had a resolution of 2 $\mu$m. Independently, the Brandeis group, together with Bell Labs, operated a PRM using a transmission geometry with a resolution of 0.3 $\mu$m—the limit of optical microscopes—and with a magnification of 1150. Both groups feel that further improvements can give the PRM a resolution below 1 nm.

These new positron reemission microscopes are variants of a proposal first suggested by L. Hulett, J. M. Dale, and S. Pendyala of Oak Ridge National Laboratory. In their original conception, a small sample was to be placed on a 100-nm-thick substrate, and the positron beam was to illuminate the back side of the substrate. The positrons would slow to thermal energies by inelastic scattering in the substrate and would be reemitted from the front side. They would image the sample lying on the substrate by a shadowing effect. The versions of the PRM actually built differ from the original proposal in that the sample serves as the reemission surface, greatly enhancing the prospects of observing positron specific contrasts, as discussed later.

## III. THEORETICAL BASIS

Since the number of low-energy (∼1 eV) positrons directly emitted by a radioactive source that may be used to form a positron beam of the quality needed for imaging is essentially zero, the successful demonstration of positron microscopes is partially based on the fact that the effective brightness of a radioactive source for positron emission at low energies may be increased enormously by the moderation process discussed previously. The $e^+$ beam so formed is then focused onto a target. The $e^+$ transmitted through, or reemitted from the target, are then imaged by an objective lens through a contrast aperture. They are then imaged by a projector lens onto a channel electron multiplier array (CEMA) with a phosphor screen anode. The CEMA/phosphor combination converts each individual positron into a spot of light that can be detected by a camera. The camera output is then fed into a computer, which digitizes the image and adds it to events previously stored in computer memory, to eventually form a digitized image. This signal averaging is crucial to successful positron microscopy since it allows an image to be built up from the signal events registered by the CEMA at rates as low as 20–200 Hz. Such rates are a factor of $10^4$ lower than the lowest intensities typically used in an electron microscope.

The optimum resolution of a positron microscope occurs when the detector resolution and lens resolution are matched. The detector resolution is determined by the image magnification, as well as the single particle image size produced by the detector. The magnification is, in turn, determined by the desired particle current density at the detector. This detector current density is determined by the initial current density of the beam, the mechanical construction of the instrument, and the acceptance parameters of the focusing elements.

The optimum resolution determined by this matching of lens and detector resolutions is given by

$$R = \left\{ \frac{C_c}{6}\left(\frac{\Delta E}{E}\right)\left(\frac{E_m}{E}\right)^{1/2} \sin\theta_m \left(\frac{\rho_d}{\rho_m}\right)^{1/2} r_d \times \left[1 + \sqrt{1 + \frac{9C_s r_d \sin\theta_m((\rho_d/\rho_m)(E_m/E))^{1/2}}{(C_c(\Delta E/E))^2}}\right]\right\}^{1/2}. \quad (1)$$

Here, $C_s$ and $C_c$ are the spherical and chromatic aberration coefficients of the imaging lenses (primarily the objective lens), $E$ is the energy of the beam at the anode of the objective lens, and $\Delta E$ is the variation in $E(\Delta E \ll E)$. The quantities $\rho_d$ and $\rho_m$ are the beam current densities at the detector ($\rho_d = I_d/A_d$, where $A_d$ and $I_d$ are the beam area and the current at the detector, respectively) and the current density at the moderator ($\rho_m = I_m/A_m$, where $A_m$ is the moderator area and $I_m$ the positron current at the moderator). The angle $\theta_m$ is the spread in emission angle of the positrons at the moderator, $E_m$ is the emission energy, and $r_d$ is the size of the single particle image produced by the detector.

In the case where spherical aberrations dominate $R$, which commonly occurs in transmission microscopy, Eq. (1) simplifies to

$$R = \left(\frac{C_s}{4}\right)^{1/4}\left(\frac{\rho_d}{\rho_m}\frac{E_m}{E}\right)^{3/8}(r_d \sin\theta_m)^{3/4}. \quad (2)$$

We note that these expressions are valid for both the TPM and the PRM, as well as other possible types of microscopes that are based on acceleration and focusing of charged particles using electromagnetic lenses to produce a final image. These expressions are also valid for positron sensitive detectors other than the CEMA/phosphor detector, for example, a charge coupled device (CCD). Using Eqs. (1) and (2), developed to describe the optimum resolution, we now discuss two of the positron microscopes.

## IV. THE TRANSMISSION POSITRON MICROSCOPE

### A. Description and Operation

We will discuss the operation of the TPM in some detail to illustrate some of the features unique to positrons. The first transmission positron microscope is shown in Fig. 2. Positrons from a 40-mCi, 5-mm diam, 15-mg/cm$^2$-thick $^{22}$Na source are incident on a set of W vanes annealed at 2200°C. After losses from source self-absorption, approximately $4 \times 10^{-4}$ of the positrons emitted by the source are reemitted at an energy of 2 V, accelerated and then focused into a beam of $7 \times 10^5$ positrons/s. This beam is transported to a bending magnet that turns it through 90° and is subsequently focused into a parallel beam that is then incident on a low aberration condenser lens. The condenser

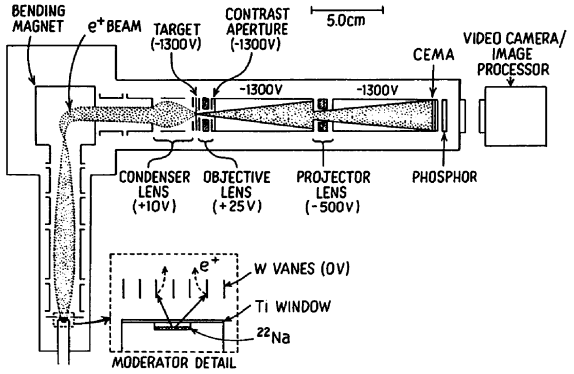



**FIGURE 2** The transmission positron microscope. [From Van House, J., and Rich, A. (1988). *Phys. Rev. Lett.* **60,** 169–172.]

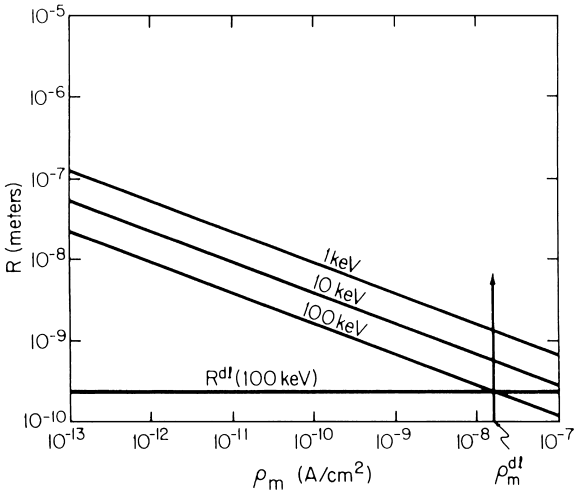



**FIGURE 3** The resolution of the TPM as a function of the positron current density at the moderator. The diffraction limit on resolution is indicated by the arrow. [From Van House, J., and Rich, A. (1988). *Phys. Rev. Lett.* **60,** 169–172.]

lens focuses the beam down to a 1.7-mm-diam spot on the target at an energy of 1.3 keV. The $e^+$ transmitted through the target are imaged by an objective lens and then by a projector lens onto a three-plate CEMA with a phosphor screen anode. The CEMA/phosphor combination converts each $e^+$ into a 200-$\mu$m-diam spot of light that is detected by an image analysis system (Fig. 2) The system adds the event to the appropriate memory location in a $384 \times 384$ array, resulting in a digital signal averaging that allows an image to be built up at rates as low as 20–200 Hz. The actual rate used at the detector depends upon the detector noise level and the desired signal averaging time, with shorter signal averaging times requiring higher rates. The magnification and resolution predicted by Eq. (1) for this prototype TPM are $M = 125\times$, and $R = 2.2\,\mu$m. In practice, transmission losses in the target must be taken into account. These losses result in lower magnifications, and thus resolutions other than given by Eq. (1). In the prototype TPM, it was found that the typical magnifications were around $M = 55\times$, where the predicted resolution is $R = 4\ \mu$m.

The images that were obtained using the TPM were of polyvinylchlorylacetate copolymer (V.Y.N.S.) foils less than 800 Å thick supported on a 100 line, 82% transmitting copper mesh. The first TPM image is shown in Fig. 1. It required 4 h of signal averaging to accumulate. The magnification was calibrated from the known 250-$\mu$m grid wire spacing. Gaussian fits to a histogram of one of the grid wires in Fig. 1, both with and without a V.Y.N.S. film, yielded, respectively (errors are statistical), $R_m = (9 \pm 1)\,\mu$m and $R_m = 4.2 \pm 0.5\,\mu$m. The first measurement shows that the effects of chromatic aberrations because of inelastic scattering in the target degrades the resolution by only a factor or two. Such an effect is not serious enough to prevent the use of the TPM at energies down to 1 keV, where the largest new contrasts are expected to occur, as discussed later. The second measurement shows that the predictions of Eq. (1) are verified to within experimental error.

The potential values of resolution that can be achieved with the TPM as a function of the current density of positrons emitted from the moderator are shown in Fig. 3. The diffraction limit on resolution is reached at current densities of $1.6 \times 10^{-8}$ A/cm$^2$. Such current densities are well within the reach of the current technological development of slow positron beams, if such beams make use of the technique of brightness enhancement discussed in Section I.

## B. Applications of the TPM

Because of the opposite sign of the coulomb interaction, a number of well-known differences exist between positrons and electrons in their scattering interactions with matter. Comparison of the resulting contrast differences between TEM and TPM images taken under otherwise identical conditions could help isolate the effects of particular terms in the scattering cross section on image formation in a manner that cannot be done using the TEM alone. In particular, the screening of the nucleus is more effective for positrons, resulting in substantially reduced small-angle scattering. Because of this, a strongly $Z$-dependent ($Z$ is the atomic number) fractional difference in the amplitude contrast, ranging at an energy of 50 keV, from 10% for $Z = 8$ to 130% for $Z = 80$, exists between the TPM and TEM. These differences should be 2–3 times larger at lower energies. Comparison of the contrast differences (once they are calibrated) between TPM and TEM images may provide information on atomic form factors because

of the screening differences, and could eventually, because of the strong $Z$ dependence, prove a sensitive microanalysis technique. The repulsion of the positrons by the nuclear charge will also alter the phase shifts of the partial waves compared to those of electrons. This will result in differences in the phase contrast that could be as large as those calculated previously for the amplitude contrast.

A number of smaller signals should also prove to yield interesting information. Among these are several new signals that have no analog in the TEM, but which will appear in the TPM because of positron annihilation in flight. These signals contain information on the binding energies and momentum space distribution of the target electrons. Exploitation of these signals will require a scanning technique to establish a spatial correlation between the signal and the target.

## V. POSITRON REEMISSION MICROSCOPY

### A. Description and Analysis of Resolution

The first PRM is shown in Fig. 4, and its operation is described in the legend. The PRM differs from the TPM in that the nature of the slow positron emission process and the short instrumental depth of information combine to make the PRM an extremely surface-sensitive instrument, sampling positron interactions that occur at an energy of 1–3 eV. By comparison, in the TPM, contrast formation occurs for positron interactions at energies of 1 keV and above, sampling primarily bulk properties of thin targets. The areas of application of the PRM are thus radically different from the TPM.

The Michigan PRM was used to conclusively demonstrate the unique defect sensitivity of the PRM, as well as the feasibility of directly imaging biological samples and semiconductor devices. An image demonstrating the defect sensitivity is shown in Fig. 5. For this image, a W foil target was masked with a 100 line, 82% transmitting Cu grid and sputtered at normal incidence with 2-keV argon ions to a fluence of $\simeq 10^{16}/\text{cm}^2$. The resulting image, after removal of the masks, unambiguously demonstrates a defect map of the surface. This defect sensitivity will be of particular interest in the study of surface atom diffusion and surface diffusion of defects in a variety of materials, including semiconductors.

Working independently, a group at Brandeis University has developed a transmission geometry PRM, which closely follows the original Oak Ridge concept. In an image taken at 0.3-$\mu$m resolution with their positron microscope, the Brandeis group identified what they describe as a finger of greatly attenuated emissivity in the middle of a region of relatively high positron counts. They suggest that the positrons in the finger region may have been trapped by boundaries at large tilt angles.

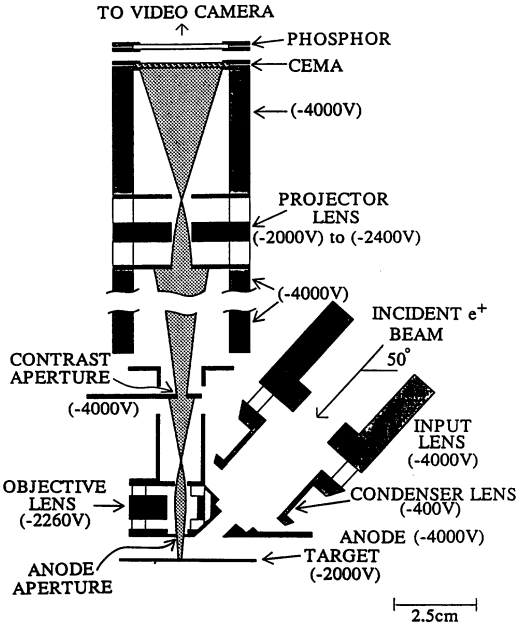



**FIGURE 4** The reflection positron reemission microscope. The incident slow positron beam of $7 \times 10^5$ positrons/s is generated by a 40-mCi $^{22}$Na source and W vane combination and transported by standard electron optics to the entrance of the PRM. A low aberration condenser lens focuses the beam into a 3000-V/cm electric field in a geometry designed to deflect the beam to the target (positioned on the axis of the imaging optics) with minimum distortion of the incident beam shape. The reemitted positrons are accelerated by the electric field and are focused by the anode aperture and then by the objective lens through a variable diameter contrast aperture to the projector lens. The projector lens focuses the positrons that constitute the final image onto a channel electron multiplier array with a phosphor screen anode that generates a 100-mm-diam single particle image for each incident positron. The single particle image is detected and signal averaged by an image analysis system. [From Van House, J., and Rich, A. (1988). *Phys. Rev. Lett.* **61,** 488–491.]

The optimum resolution of either the transmission or reflection PRM can be deduced from Eq. (1), with two important modifications:

1. In the TPM, the resolution was determined based upon the properties of the moderator used to generate the slow positron beam. In the PRM, the target serves as the moderator for the subsequent image beam. Consequently, all quantities that relate to the moderator in Eq. (1) must now be replaced with the corresponding target properties.
2. The acceleration of the reemitted slow positrons by the electric field between the target and anode results in the formation of a virtual target at a distance $l^* \approx 2l$ behind the real specimen. The dependence of $l^*$ upon the emission energy $V_e$ gives rise to nonnegligible spherical

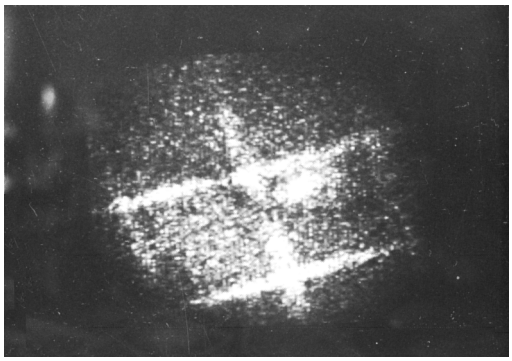


**FIGURE 5** A demonstration of the near-surface defect sensitivity of the PRM. The preparation of the damaged W foil is described in Section III. The dark sputtered regions have low positron reemission because of positron trapping in defects. The bright, high-reemission areas were masked from sputter damage. The image is an argon sputtered W foil and was taken at a magnification of 17×, with a resolution of 4.6 mm. [From Van House, J., and Rich, A. (1988). *Phys. Rev. Lett.* **61,** 488–491.]

and chromatic aberrations during the acceleration of the reemitted positrons from the target to the anode. The complete spherical and chromatic aberration coefficients, $C_s$ and $C_c$, consist of a linear sum of the spherical and chromatic aberration coefficients of the objective lens referred back to the object space of the virtual specimen, $C_s^0$ and $C_c^0$, to the spherical and chromatic aberrations of the accelerating field, yielding

$$C_s = \frac{2l}{27}\sqrt{\frac{V_a}{V_e}} + C_s^0$$
$$C_c = \frac{l\sqrt{V_e V_a}}{\Delta V_e}\left(1 - \sqrt{1 + \frac{\Delta V_e}{V_e}}\right) + C_c^0. \tag{3}$$

Here, $l$ is the target–anode separation, $V_a$ is the accelerating potential of the anode (all voltages are given with respect to the target potential), $V_e$ is the emission energy, and $\Delta V_e$ is the variation in $V_e$.

Using parameters similar to those discussed in Section II for the TPM, it can be deduced from Eqs. (1) and (3) that the diffraction limit on the resolution for the transmission PRM operating at 100 keV is 6 Å, whereas for the reflecton PRM operating at 5 keV (limited by the geometry), the diffraction limit on the resolution is about 15 Å. Obtaining such resolutions will require positron current densities at the target of greater than $10^{-5}$ A/cm$^2$, about 3 orders of magnitude higher than the $10^{-8}$ A/cm$^2$ diffraction limit current density required for the TPM. This higher current density is required primarily to compensate for higher spherical and chromatic aberration coefficients given by Eq. (3).

## B. Contrast Formation in the PRM and PRM Applications

The mechanisms responsible for the contrast expected to occur in PRM images can be loosely grouped into two categories: topographical contrast and materials contrast. It is in the second category that new types of contrast, and thus new applications, unique to positrons are expected to appear in the PRM.

The topographical contrast in the PRM should be identical to that which appears in the photo-electron emission microscope (PEM), since contrast caused by the effects of the specimen relief on the accelerating field are identical for both positrons and electrons. Heights of 30 Å are detectable experimentally in the PEM, and comparable heights should ultimately be detectable in the PRM. The analysis of Hulett indicates that resolutions approaching atomic dimensions may be detectable in the transmission PRM under conditions involving tip and edge effects of objects placed on a single crystal source of reemitted positrons. Additional sources of topographical contrast include shadowing caused by the incident beam striking a protrusion at an angle. In addition, as discussed by the Bell–Brandeis group, positron reemission microscopy may lend itself to a form of holographic imaging, in which one records the interference between the positrons emitted from a substrate and those scattered from atoms within the molecule lying on the surface. The multiple scattering would make this technique extremely complex, but it has the possibility of circumventing the 0.6-nm limit on resolution imposed by diffraction.

The material contrast is related to the positron reemission yield of a substance, $Y(E)$, which is defined as the number of slow positrons reemitted per incident positron. The yield decreases as the incident positron energy $E$ increases and has a maximum value $Y_0$ at $E = 0$. The magnitude of $Y(E)$ depends on the various interactions that occur as the incident positrons slow down to near thermal energies and then diffuse through the bulk to the emission surface, as well as interactions that occur at the point of emission.

The manner in which this information will appear in the PRM image is dependent on the nature of the imaging system and on positron diffusion in the bulk of the target. The electron optics form a Gaussian image that is sharply defined at the emitting surface (hence the high topographical resolution discussed previously). Consequently, the surface information will display the approximately 5-Å resolution given by Eq. (1), whereas information at depths below the instrumental depth of field (given by similar relations for the PEM) will appear as a diffuse background. The long positron diffusion length will also result in a randomization of information at depths below 100 Å.

Consideration of these two factors leads to the conclusion that, in the PRM, the most important new contrasts will be due to phenomena that occur near the surface and near the point of emission of the slow positron, that is, the PRM will be an extremely surface sensitive instrument.

Slow positron emission has a demonstrated sensitivity to the density-of surface and near surface defects in metals (Fig. 5). Areas of application include studies of defect formation at material interfaces and around the edges of islands and terraces of submonolayer thin-film overlayers on metal substrates. If a similar defect sensitivity is obtained in insulators and semiconductors, applications to studies of high-density integrated chips and color centers in phosphors are indicated. Assuming that atomic resolutions at edges can be achieved, imaging of single atom defects may be possible. The defect sensitivity of the PRM will be of particular interest in the study of surface atom diffusion and surface diffusion of defects, and may also be useful in understanding the role of surface defects in crack formation.

The sensitivity of $Y(E)$ to the surface work function will give rise to an orientational contrast that should allow studies of polycrystalline materials, such as metal alloys and geological samples. The demonstrated sensitivity of this work function to surface conditions on metals will allow determination of surface cleanliness and should aid in the study of the effects of monolayer and submonolayer coverages of adsorbates on single crystal substrates.

Possible biological applications of the PRM may include elimination of the need for the staining agents required to give sufficient contrast. Such staining agents often produce sample distortion—the sample is no longer in its natural state. The Michigan group has already shown that the PRM can directly image the surface of biological specimens without the use of staining agents, and hence, distortion can be prevented. Another application of the PRM may be in the study of polar surface sites on biological cells. Special labeling compounds attached to site specific antibodies (which then attach to the polar site on the cell) are typically used in light microscopy to link the cell's biological function to its observable topography. Unfortunately, only polar sites specific to the particular antibodies used can be indirectly studied through the labeling compound. The PRM may, however, be directly sensitive to broad classes of polar surface phenomena through two possible mechanisms: positron trapping at negatively charged sites and positron emission after positronium breakup at the polar site, that is, positronium chemistry. Both mechanisms would directly affect the PRM image contrast at the polar site without the use of any labeling or attachment compounds.

## SEE ALSO THE FOLLOWING ARTICLES

MICROSCOPY • SCANNING ELECTRON MICROSCOPY • SCANNING PROBE MICROSCOPY • TRANSMISSION ELECTRON MICROSCOPY

## BIBLIOGRAPHY

Brandeis, G. R. (1991). *Phys. Rev. B* **43**(10), 111.
Brandeis, G. R., Canter, K. F., Horsky, T. N., Lippel, P. H., and Mills, A. P., Jr. (1988). *Rev. Sci. Instrum.* **59,** 228–232.
Brandeis, G. R., Canter, K. F., Horsky, T. N., and Mills, A. P., Jr. (1988). *Appl. Phys.* **A46,** 335–337.
Brandeis, G. R., Canter, K. F., Horsky, T. N., and Mills, A. P., Jr. (1988). *Phys. Rev. Lett.* **61,** 492–495.
Brandeis, G. R., Canter, K. F., and Mills, A. P., Jr. (1991). *Phys. Rev. B* **43**(10), 103.
Charlton, M., and Humberson, J. W. (2000). "Positron Physics," Cambridge Univ. Press, Cambridge, UK.
Rich, A., and Van House, J. (1988). *J. Electron Miscrosc. Tech.* **9,** 209.
Rich, A., and Van House, J. (1989). U.S. Patent #4,864,131.
Schultz, A. (1997). "Slow-Positron Beam Techniques for Solids and Surfaces," Springer-Verlag, Berlin/New York.
Van House, J., and Rich, A. (1988). *Phys. Rev. Lett.* **60,** 169–172.
Van House, J., and Rich, A. (1988). *Phys. Rev. Lett.* **61,** 488–491.
Xixiang, Z., Madison, D. H., McEachran, R. (1996). "Computational Atomic Physics: Electron and Positron Collisions with Atoms and Ions," Springer-Verlag, Berlin/New York.



# Radiation Physics

**John H. Hubbell**
*National Institute of Standards and Technology*

I. Description of Radiation Physics
II. Radiation Sources
III. Quantifying Radiation
IV. Interaction with Matter
V. Useful Radiation Data

## GLOSSARY

**Activity** The activity, A, of an amount of radioactive nuclide in a particular energy state at a given time is the quotient $A = dN/dt$, in which dN is the expectation value of the number of spontaneous nuclear transitions from that energy state in the time interval dt. The unit for activity is the Becquerel (Bq), in which $1\ Bq = 1\ s^{-1}$. The activity unit in the older literature is the curie (Ci), in which $1\ Ci = 3.7 \times 10^{10}\ s^{-1}$ (exactly) $= 37.0$ GBq.

**Alpha ray ($\alpha$ ray)** Radiation in the form of particles equivalent to helium nuclei, ejected from radioisotopes undergoing nuclear transitions according to this mode.

**Attenuation coefficient ($\mu$)** Sometimes called absorption coefficient. Coefficient in Lambert's law $I = I_0 e^{-\mu t}$ which describes the attenuation to intensity I of a narrow beam of radiation of incident intensity $I_0$, after penetrating a thickness t of material. Units of $\mu$: Inverse of units of t (e.g., $cm^{-1}$). Data tabulations are usually in terms of the **mass attenuation coefficient ($\mu/\rho$)**, where $\rho$ is the density of the material. Units of $\mu/\rho$: $cm^2\ g^{-1}$. Primarily applicable to photon radiation (x rays, gamma rays, bremsstrahlung).

**Beta ray ($\beta$ ray)** Radiation in the form of electrons ejected from an radioisotopic nuclei undergoing nuclear transitions according to this mode.

**Bremsstrahlung** Photon produced when charged particles (e.g., electrons or protons) are slowed by interactions with atoms in passing through matter. Bremsstrahlung (German, braking radiation) is sometime called white radiation, or continuous spectrum, to distinguish it from fluorescence (x-ray) line spectra characteristic of each element.

**Cross section ($\sigma$)** Effective area of a target particle (e.g., atom, electron, nucleus, etc.) for intercepting a photon or a unit of particle radiation, resulting in absorption or deflection of the incident radiation. Units of $\sigma$: $cm^2$ or b (barns), where $1\ b = 10^{-28}\ m^2 = 10^{-24}\ cm^2$. The total cross section $\sigma_{tot}$ (probability for any interaction with the target particle) is related to the mass attenuation coefficient $\mu/\rho$ according to the relation $\sigma_{tot} = (\mu/\rho) \cdot (A_r/N_A)$ where $A_r$ is the relative atomic mass (atomic weight) and $N_A$ is the Avogadro constant.

**Dose (D)** Used broadly for energy deposited in matter from radiation. Used in dosimetry for the energy absorbed per unit mass of material, usually by ionization processes. Units are the rad and the Gray (Gy), which

are equivalent, respectively, to 100 ergs/g and 1 J/kg. Therefore, 1 rad = 1/100 Gray or 1 cGy.

**Exposure (X)** Exposure, X, related to the air ionization properties of photon radiation, is the quotient dQ/dm of the amount of charge dQ of the ions of one sign produced by the electrons (negatrons and positrons) liberated by photons in a volume element of air having mass dm are completely stopped in air. X has the units C/kg. In the older literature, and in the calibrations of many existing instruments, one finds the special unit *roentgen* (R) in which $1\ \mathrm{R} = 2.58 \times 10^{-4}$ C/kg (exactly).

**Fluence** Time-integrated flux of particles or photons. Unit: $\mathrm{cm}^{-2}$.

**Flux** Number of particles or photons passing through some defined zone per unit time. For parallel beams, this is a unit area; for omnidirectional radiation, the zone chosen is usually a sphere with cross section of 1 $\mathrm{cm}^2$. In both cases, the unit is $\mathrm{cm}^{-2}\ \mathrm{sec}^{-1}$.

**Gamma-ray ($\gamma$-ray)** Photon resulting from a transition in an atomic nucleus, either from natural decay of a radioisotope, or from an induced nuclear transition.

**Gray** Radiation absorbed dose unit of the Systeme Internationale (SI), of value 1 J $\mathrm{kg}^{-1}$ and equal to 100 rad.

**Mean-free-path (mfp)** For photons in an attenuating medium, distance over which the primary (unscattered) beam intensity is reduced by a factor 1/e and is equal to the reciprocal, $1/\mu$, of the attenuation coefficient $\mu$.

**Photon** Quantum of electromagnetic radiation. Can be from any region of the electromagnetic spectrum including radio waves and visible light, but in this article referring to quanta in the energy (or wavelength) region of x rays, gamma rays, or bremsstrahlung.

**X ray (x-ray when used as an adjective)** Specifically referring to characteristic (line-spectra: fluorescence) photons resulting from atomic (extra-nuclear) transitions, but often used more broadly, for photons from any source, over the energy range from tens of eV (electron volts) through the GeV region.

**RADIATION PHYSICS** ties together a variety of otherwise separate and compartmentalized scientific, medical, engineering disciplines, all involving aspects of radiation including radiation sources, radiation transport (penetration), radiation detection and radiation effects. These disciplines include, for example:

1. Atomic and nuclear physics, theory and measurements
2. Medical radiation physics: imaging, therapy
3. Environmental radiation dosimetry
4. Nuclear power engineering, shielding, radiation transport theory, Monte Carlo
5. X-ray crystallography
6. Industrial radiation processing, radiometric gauging, on-stream monitoring and control
7. Fluorescence XRS, XRF materials analysis, radiation archeometry, dating
8. X-ray, $\gamma$-ray astronomy, astrophysics, space vehicle dosimetry
9. Radiation damage to electronic circuitry

This cross-disciplinary feature of radiation physics has resulted in the formation in 1985 of the International Radiation Physics Society (IRPS) which meets triennially to bring researchers from the above and other radiation-related disciplines together to share their experiences and results, bound together by the common thread of radiation physics. Information on this Society and its triennial International Symposia may be obtained from the IRPS Secretariat, c/o Department of Physics, University of Pittsburgh, Pittsburgh, PA 15260.

## I. DESCRIPTION OF RADIATION PHYSICS

The field of radiation physics is distinct from the fields of atomic physics, nuclear physics, and particle physics and other material sciences, all of which focus on the nature of tangible matter, with radiation playing the role of the probe. In radiation physics, the roles are reversed, with the focus on the radiation, and matter playing the role of the probe.

The radiation in question arises from both natural and artificial sources. Natural sources include, for example, space radiation and radium and uranium in rocks, as well as trace radionuclides in the human body and in other organisms. Artificial sources include, for example, concentrations of radioisotopes such as in cobalt-60 plaque irradiators, accelerators such as x-ray and electron-beam machines, and nuclear chain reactions in reactors and weapons.

The technologies of radiation physics are used in academic research, industrial technology, aerospace technology, and medicine. In some machines, materials and devices have to withstand high doses of radiation. In the treatment of tumors, beams of radiation have to be generated and controlled with great accuracy. Thus, while the effects of radiation on living tissue fall into the field of radiobiology, the task of irradiating tissue is often carried out by experts in radiation physics.

The field of radiation effects is an important subtopic of radiation physics. Radiation deposits energy in matter. This energy is then distributed among atoms and molecules in a great variety of ways. The energy excites atoms to higher energy states or moves them about in the

material. The scientific disciplines employed in radiation physics center around the interactions of photons or particles with solids, liquids, and gases. The primary transfer of energy is complex but well understood; but the energy transferred then dissipates via secondary interactions with the target material. Secondary processes are complex and not well understood except in a few cases, such as the various results of the displacements of atoms in silicon by an electron beam.

The technological aspects of radiation physics are very wide. Instruments for measuring radiation (dosimeters and counters) are widely used. The prevention of degradation in electronic and optical materials is important. Furthermore, in some cases, controlled irradiation can improve commercial products. For example, food preservation and toughening of plastics are sometimes most economically achieved by irradiation with noncontaminating radiation (electrons or $\gamma$ rays). Infectious materials such as sewage can be sterilized by radiation, and in-package sterilization of medical supplies using radiation is now done routinely.

An equally important subtopic of radiation physics is noninvasive interrogation of systems. In addition to medical imaging by conventional film and by the increasing number of tomographic modalities, hot-rolling of steel now employs noncontact radiometric monitoring of the thickness, and control of the rollers, and voids (bubbles) in liquids flowing inside closed pipes can be monitored by radiometric transmission and backscattering arrangements, to name a few examples. The design of equipment for airport x-ray inspection of luggage utilizes radiation physics to produce useful images with minimal damage to camera film and other sensitive items.

## II. RADIATION SOURCES

### A. Overview

With reference to radiation effects, Fig. 1 shows a survey of the broad range of energy values that have to be considered under the term radiation. Most of the radiations of interest have energies above 1 keV ($10^3$ electron volts), but neutrons with much lower energies are still described as particle radiation. Ultraviolet photons also have sufficient energy to cause the same chemical effects as we see with x rays and very high energy photons. The shaded areas in Fig. 1 show that different fundamental effects in matter occur at different radiation energies.

### B. Photons

Photons that have energies in the keV and MeV (megavolt) range (x rays and $\gamma$ rays) have the ability to penetrate

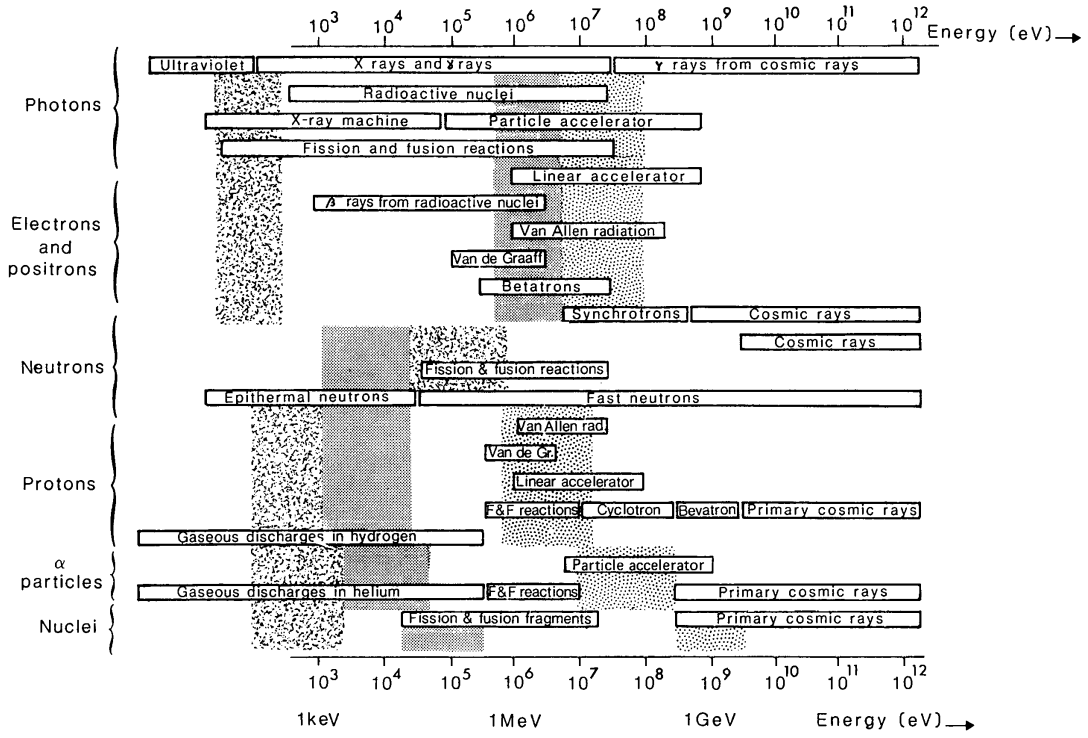



**FIGURE 1** Energy values for various radiation environments and threshold energies. Approximate threshold for producing ionization ▒, for atomic displacements in solids ▒, and for nuclear transmutations ▒. Note: F & F = fission and fusion. (Courtesy of Hughes Aircraft Co.)

matter deeply and, when absorbed, to produce strong effects. X rays are generated when an electron beam strikes matter; an x-ray generator consists of a powerful electron gun and a metal target in which the photons are generated. For x-ray crystallography applications, the photons will be in the energy range from 5 to 30 keV, including line-energies characteristic of the atomic number of the metal in the target, superimposed on a bremsstrahlung continuum spectrum. For imaging and irradiation applications, the energies commonly range up to 3 MeV, mostly in the form of bremsstrahlung. For research applications, electron accelerators such as synchrotrons and linacs (linear accelerators) produce photons up to the GeV ($10^9$ electron volts) region, but for imaging and irradiation purposes the energy is usually kept below 5 MeV to avoid producing radioactivities in the sample due to the photonuclear effect which has a resonance peak in the region 5 to 40 MeV.

$\gamma$ Rays are high-energy monoenergetic photons. The term is used specifically for photons created during the disintegration of atomic nuclei. A well-known example is the pair of photons created in the spontaneous disintegration of the cobalt-60 atomic nucleus. These photons have energies of 1.1732 and 1.3325 MeV, and cobalt-60 ($^{60}$Co) is frequently used in plaque and other geometrical configurations in irradiation facilities. $\gamma$ Rays are created by nuclear chain reactions such as those that occur in a nuclear reactor core or a nuclear explosion. The isotopes contained in nuclear fuel represent concentrated sources of $\gamma$ rays. These can be used for experimental irradiation in spent-fuel ponds. Table I presents data on $^{60}$Co and some of the other radioisotopes useful in medical therapy and industrial irradiation applications, also on $^{90}$Sr which can be important as an environmental hazard.

In recent decades, synchrotron radiation has become a major high-flux source of photons for research and analytical applications. This radiation is produced in high-energy accelerators from the bending of electron orbital trajectories in the confining magnetic field, sometimes by magnetic "wigglers and undulators" interposed in the electron

**TABLE I Radioisotopes Important in Medical Therapy and Industrial Irradiation Applications, also as Environmental Hazards ($^{90}$Sr)[a]**

| | | | Photons | | Particles | |
|---|---|---|---|---|---|---|
| Nuclide | Half-life (year) | Type of decay | Energy (MeV) | Percentage emitted (%) | Energy (MeV) | Transition probability (%) |
| $^{60}$Co | 5.27 | | 1.173 | 99.86 | 0.318 | 99.9 |
| | | $\beta^-$ | 1.333 | 99.98 | 1.491 | 0.1 |
| | | | (av 1.25) | | | |
| $^{192}$Ir | 0.526 | | 0.296 | 29.6 | | |
| | (192 d) | | 0.308 | 30.7 | | |
| | | | 0.316 | 82.7 | 0.530 | 42.6 |
| | | $\beta^-$ | 0.468 | 47.0 | 0.670 | 47.2 |
| | | | 0.604 | 8.2 | | |
| | | | 0.612 | 5.3 | | |
| $^{137}$Cs | | | 0.662 | 85.1 | | |
| | | | | | 0.512 | 94.6 |
| | 30 | | 0.032 | | 1.174 | 5.4 |
| | | $\beta^-$ | −0.038 | 8 | | |
| | | | ($^{137}$Ba) | | Plus internal conversion electrons 0.65 MeV | |
| | | | K X rays | | | |
| $^{90}$Sr | 28 | $\beta^-$ | 0.54 | 100 | — | — |
| + daughter | | | | | | |
| $^{90}$Y | 0.176 | $\beta^-$ | 2.27 | 100 | — | — |
| $^{85}$Kr | 10.6 | $\beta^-$ | 0.15 | 0.7 | 0.51 | 0.7 |
| | | $\beta^-$ | 0.67 | 99.3 | — | — |
| $^{252}$Cf | 2.65 | Spontaneous fission | — | — | Neutrons, 2 MeV | |
| | | | | | $\gamma$s, 5.9–6.1 MeV | |
| | | | | | Fission fragments, 80 and 104 MeV | |

[a] Main emission energies.

path. The photon energies thus produced range from tens of eV from accelerators in the hundreds of MeV range, to above 100 keV for electron accelerators in the multi-GeV range. Another recently developed source of photons in the $\gamma$-ray energy range is by inverse Compton scattering. In such devices, intense laser beams in the visible or ultraviolet (eV range) are collided with GeV-range electrons in an accelerator, boosting the eV laser photons up to MeV energies.

### C. Electrons

Electron beams for research and for irradiation purposes are produced by a variety of orbital and linear accelerators, with output energies from the keV to the GeV regions, and beam currents from microamperes to kiloamperes. Linear accelerators (linacs) consist of an electron gun injector at one end, followed by a series of accelerating cavities containing rf (radio frequency) power. Natural sources of high-energy electrons include $\beta$ rays from radioisotope decay, and also from electrons in space which have been accelerated in magnetic fields. There is a particularly high concentration of electrons trapped around planets that have high magnetic fields, such as Earth and Jupiter. These trapped-electron concentrations, called Van Allen belts, offer a radiation damage hazard to components in unmanned space vehicles traversing these altitudes, and are avoided as much as possible in the case of manned flights.

### D. Positrons

The emissions from decay of radionuclides can also include electrons of the opposite charge sign, from that of the orbital atomic electrons and $\beta$'s ($\beta^-$), which are called positrons ($e^+$ or $\beta^+$), and these can also be accelerated to provide positron beams. Positrons are also produced in the process of pair production by photons of energy higher than that equivalent to the rest-mass of two electrons (1.022 MeV total). The signature of positrons is the annihilation radiation resulting from the collision between an electron and a positron, usually, if the positron has come to rest, consisting of two photons in opposite directions, each with one rest-mass energy of 0.511 MeV. This radiation is important in medical diagnostic imaging by PET (positron emission tomography). Variants of this method include SPECT (single-photon emission computed tomography).

### E. Protons

Proton beams can also be produced by accelerators. One especially well-known form of proton accelerator is the cyclotron, which uses radio-frequency energy. Nuclear reactions also produce protons in a material sample. Protons are emitted from the sun, and protons are also found trapped in planetary magnetic fields. High-energy protons are emitted from the sun in bursts associated with solar flares. A less energetic, steady stream of protons emitted from the sun is called the solar wind.

### F. Ions

$\alpha$ Particles (which are ions) consist of high-energy helium nuclei; these too are emitted by radioisotopes and stars and are thus found in interplanetary space. In addition to helium ions, very energetic ions of all atomic masses are found in space. These are called cosmic rays. Ion beams can be generated in accelerators. One use for ion beams is the ion implantation of solids to modify their properties. High-current ion implanters are available for industrial use. These beams cause large amounts of radiation damage in the solids so treated.

### G. Neutrons

The main sources of neutrons are nuclear fission and nuclear fusion reactions. Of these, the fission of uranium is the most common. The primary product of fission is fast neutrons having an energy distribution described as a fission spectrum. This spectrum has a large content with energies above 1 MeV. This is the spectrum that would be observed near a nuclear explosion. In a nuclear reactor, interaction with surrounding material, most efficiently with light nuclei, reduces the neutron energy. Thermal neutrons or cold neutrons of much lower energy are produced. On collision with matter, fast neutrons produce much damage while thermal neutrons produce radioactivation. Neutrons reduced to cryogenic temperatures, with their long effective wavelength, are useful in surface physics studies. Nuclear fusion reactions produce neutrons having a much higher energy than fission. For example, one commercial generator of fusion neutrons produces a beam of D-T neutrons of a single energy of 14 MeV.

## III. QUANTIFYING RADIATION

### A. Flux and Fluence

A parallel beam of radiation passing through free space can be quantified by quoting the number of particles passing through a unit area. See flux and fluence in this article's Glossary. If the particles come from a variety of directions (as in outer space) then we quote the number intersecting a sphere of unit cross-sectional area.

### B. Energy Spectrum

A curve plotting the population of particles in a given energy range is called the energy spectrum of the particle.

### C. Exposure

In radiation physics, we are often concerned with the end-result of the absorption of radiation-borne energy in solids. In order to calculate this energy, we first need to quantify the fluence and energy of the particles that impinge on the solid. Even if we cannot do this, we can at least specify an observable effect in a familiar medium such as air. These forms of a statement of a quantity of radiation do not describe the exact energy absorbed in a given material. They are, however, a useful measure of radiation quantity and are called "exposure" units.

We express exposure in two ways:

1. We note the fluxes impinging on the solid of interest and the energies of the particles or photons concerned, such as $10^{15}$ cm$^{-2}$ of 1 MeV electrons or $10^{12}$ cm$^{-2}$ of fission-spectrum neutrons.
2. We measure the quantity of ionization produced by such a flux in a standard medium, usually air.

The units used in (1) are often used in the radiation testing of silicon devices with radiation beams such as electrons. The units used in (2) are often used in medicine, where x-ray generators and isotopes are the most common sources of radiation, and air ionization chambers are the most common measuring instruments. Calculation methods are available to convert units of exposure into units of absorbed energy, namely, dose and kerma, which are discussed later.

### D. Dose and Kerma

The term dose is a useful general description of the energy per unit mass that has been deposited in a material by a high-energy particle on its way through. Given a value for dose, we can calculate biological and nonbiological radiation effects. For crystalline solids such as silicon and metals, some further complications arise. It may be necessary to divide the energy deposition into two fractions: ionization and atomic displacement. A quantity defined to assist with these distinctions is *kerma*, meaning kinetic energy released in a material. To distinguish between the two forms of energy deposition, we can speak, for example, of ionization kerma when referring to the fraction of energy going into ionization.

Radiation dose and kerma are both measures of energy deposited. The SI unit for either is the Gray (Gy), which represents energy deposited per unit mass of 1 J/kg. Many authoritative publications still employ the older practical unit, the rad, representing 100 ergs/g. One roentgen represents 86.9 ergs per gram in air. A dose of 1 Gy thus equals 100 rad.

The term *equivalent dose* is a useful description of the biological effects of different kinds of radiation on different organs of human tissue. The equivalent dose is defined as the absorbed dose multiplied by an appropriate *radiation weighting factor*, $w_r$. For x rays and gamma rays the weighting factor is about unity. The SI unit for the equivalent dose is the Sievert (Sv), where $1\ \text{Sv} = 1\ \text{J/kg}$. The older unit is the rem, representing 100 ergs per gram. An equivalent dose of 1 Sv thus equals 100 rems.

### E. Range–Energy Relations

Photons and particles, when passing through a slab of material, are attenuated by collisions and other interactions. We can plot the intensity of the emergent radiation as a function of the slab thickness $d$. In many cases, the attenuation can be expressed by an exponential law, of the form

$$I/I_0 = e^{-\mu d},$$

where $I_0$ is the incident radiation intensity, $I$ is the emergent intensity, and $\mu$ is the attenuation coefficient in units reciprocal to those of the thickness $d$ of the slab.

For photons (x- and $\gamma$-ray), this law is followed exactly for an idealized "narrow beam" geometry in which secondary radiations produced in the absorber are not seen by the detector. For electrons, the law is followed over the early part of the curve but, after a certain distance, the practical range, no electrons emerge. For neutrons, the law is followed approximately, although the interactions with the atoms of the material are very different from those for electrons and photons.

For electrons and other charged particles there is a minimum slab thickness $W$ that stops all the particles. This called the "stopping range." We can plot curves of $W$ versus energy. These are useful for calculations of shielding. Depth-dose curves as a function of penetration depth, such as shown in Fig. 2, are useful in designing radiation cancer-therapy treatment plans.

## IV. INTERACTION WITH MATTER

### A. Phenomena Observed during Absorption of Radiation

#### 1. General

The laws by which radiation is attenuated by and absorbed in matter are derived from the several competing types of

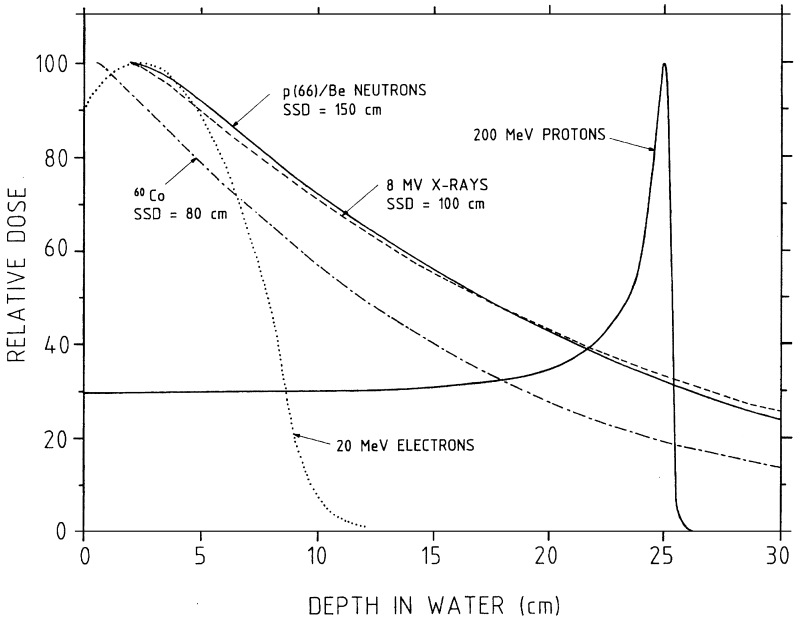



**FIGURE 2** Central axis depth dose curves for some typical photon, electron, neutron and proton radiotherapy treatment beams. [Reproduced courtesy of D. T. L. Jones "Present status and future trends of heavy particle raadiotherapy," pp. 13–20 in Cyclotrons and their Applications 1998 (Proceedings of the 15th International Conference on Cyclotrons and their Applications, Caen, France, pp. 14–19 June 1998) (E. Baron and M. Lieuvin, eds), Copyright Institute of Physics, Bristol 1998.]

interactions of photons or particles with the atoms of the material. This section describe the primary processes, for the various types of radiation.

### 2. Photons (e.g., X Rays, $\gamma$ Rays, Bremsstrahlung)

Photons are electromagnetic wave-trains, differing from visible light only by having a shorter wavelength and a higher energy, with the same velocity $c = 2.99792458 \cdot 10^8$ m $s^{-1}$ in a vacuum. Photons can also be treated as particles, with energies E (usually in multiples of eV) inversely related to the wavelength $\lambda$ (usually in angstroms [Å] where 1 Å $= 10^{-10}$ m $= 0.1$ nm) according to:

$$E(eV) = 12398.42/\lambda(Å).$$

As a particle, a photon can transfer momentum to a target particle such as an electron, in a variety of scattering and absorption collision processes. Compared with particles such electrons and protons, the probability of collision is low, so that photons in the energy range here considered are regarded as penetration radiation. For each element, the probability of interaction with incident photons is a function of the energy of the photon and the atomic number Z of the material. Interaction is primarily by four processes: the photoelectric process, the Compton (inelastic) scattering process, the Rayleigh (elastic) scattering process, and the pair (and triplet) production process. Among the other less-probable processes is photonuclear interaction which can induce radioactivity.

*a. Photoelectric absorption ($\tau$).* In this interaction a photon is completely absorbed by at atom, and an electron is ejected. For a given atomic electron shell or subshell to participate in this process, the photon energy $h\nu$ must be greater than the binding energy $B$ of electrons in that shell or subshell. The ejected electron has a kinetic energy of $h\nu - B$. The highest value of $B$ for a given element is for the two (except one for hydrogen) innermost and most tightly bound electrons, in the K shell, ranging from 13.598 eV for hydrogen ($Z = 1$) up to 88.005 keV for a high-Z element such as lead ($Z = 82$). Progressing outward from the nucleus, the L shell has three subshells $L_I$, $L_{II}$, and $L_{III}$, each with a slightly different binding energy $B$. These threshold energy values result in a characteristic "sawtooth" shape of a plot of the photoelectric effect cross section as a function of incident photon energy.

Although the regions at photon energies just above these "sawtooth" absorption edges (thresholds) can exhibit considerable fine structure, with oscillations of 10% or more, from both matrix and atomic (outer unfilled shells, etc.) effects, this fine structure is generally ignored in

general-purpose compilations of x-ray attenuation coefficients for medical and shielding applications. However, EXAFS (x-ray absorption fine structure) is a widely used and important analytical tool.

*i. Fluorescence (characteristic) x rays.* In addition to the ejection of a photoelectron, this process results in the emission of one or more fluorescence x rays, due to an outer-shell electron falling into the inner-shell vacancy created by the departure of the ejected photoelectron. The fluorescence x-ray energy is equal to the difference in binding energy between the participating inner and outer electron shells or subshells, hence a fluorescence x-ray spectrum consists of a number of discrete lines, each corresponding to one of the many transition possibilities.

*ii. Auger effect.* The fluorescence x rays do not necessarily emerge from the collided atom, but can instead dislodge outer electrons, similar in effect to a secondary photoelectric absorption. This additional emission of electrons is called the Auger effect, and the emitted electrons are called Auger electrons. The probability that the primary fluorescence x ray will escape from the atom without undergoing the Auger process is called the fluorescence yield, $\omega_i$, in which i refers to the electron shell or subshell from which the primary photoelectron was dislodged.

The photoelectric effect ($\tau$) is the dominant interaction process for low photon energies.

*b. Compton scattering ($\sigma_{inc}$).* In this interaction, also called incoherent or inelastic scattering, only part of the energy of the photon is transferred to an electron. Both the electron and photon are scattered from the collision, with energies and directions related as determined by conservation of momentum and energy between the deflected photon and the recoiling electron. The energy E′ of the deflected photon is reduced from that of the incident photon energy E according to the relation

$$E' = E/\{1 + (E/mc^2)(1 - \cos\theta)\}$$

in which $mc^2 = 0.5110$ MeV is the rest mass energy of an electron (or positron) and $\theta$ is the deflection angle of the scattered photon. This equation can also be written in terms of the shift in wavelength of the photon

$$\lambda' - \lambda = 1 - \cos\theta$$

in which the wavelengths $\lambda'$ and $\lambda$ of the deflected and incident photons are in Compton units

$$\lambda = mc^2/E = 0.5110/E[\text{MeV}].$$

It can thus be seen that the maximum shift in photon wavelength is two Compton units, at the photon backscattered angle of $\theta = 180^\circ$, at which angle, no matter how high the incident photon energy E, the backscattered photon energy E′ will never exceed $mc^2/2 = 0.2555$ MeV. In collimated gamma-ray sources, this energy can show up in a spectrum of the primary beam, in addition to high-energy photons from the source itself, due to 180° Compton scattering from material behind the source.

Over most of the region where the Compton cross section is a major part of the total cross section, the target electron can be assumed to be free and at rest, in which case the elegant Klein-Nishina equations apply. At lower energies, where electron motion and binding energies are a significant fraction of the incident photon energy, the Klein-Nishina theoretical cross section can be modified by use of an incoherent scattering function S(x, Z) in which x is a momentum transfer variable related to the incident photon energy E and its subsequent deflection angle $\theta$, and Z is the atomic number of the target atom. Calculations of S(x, Z) require knowledge of atomic wave functions, and values of S(x, Z) are usually taken from available tables of the incoherent scattering function, an example of which is provided in the bibliography.

Compton scattering ($\sigma_{inc}$) is the dominant interaction process in the intermediate energy range from a few tens or hundreds of keV up to a few MeV, this range being the broadest for the lowest-Z elements.

*c. Rayleigh scattering ($\sigma_{coh}$).* In this interaction, also called coherent or elastic scattering, a photon is scattered by the atomic electron cloud as a whole, with the entire atom, including the heavy nucleus (compared to the electrons) taking up the recoil. Thus, from energy and momentum conservation considerations, the deflected photon undergoes negligible loss of energy or change in wavelength.

Since the angular distribution Rayleigh scattering is sharply peaked in the forward direction for photon energies in the gamma-ray region, and the energy loss of the deflected photon is negligible, the contribution of this process is sometimes omitted in radiation transport calculations. However, at the lower photon energies (4 to 30 keV) utilized in x-ray crystallography, the photon wavelength is comparable to the lattice spacing between atoms in crystals, and this effect becomes of prime importance. Due to the coherence between the incident and deflected photon, interferences from waves scattered from successive crystalline layers result in reflections only into sharply discrete directions, making possible studies not only of inorganic crystalline materials, but also of biological structures such as DNA.

Rayleigh (coherent) scattering ($\sigma_{coh}$) is never the dominant photon interaction process, but for heavy elements in the energy region just below the photoelectric K-shell

absorption edge (threshold) it can account for up to 10% of the total attenuation coefficient.

*d. Pair and triplet production ($\kappa_n$ and $\kappa_e$).* In this type of interaction, which can happen only for photons with energy in excess of two electron rest-mass energies ($2mc^2 = 1.022$ MeV), the photon interacts with the electrostatic field of a charged particle, such that the photon disappears and in its place is created an electron–positron pair ($e^-, e^+$). If this process happens in the field of the positively charged atomic nucleus ($\kappa_n$), with negligible recoil due to its effectively infinite mass in comparison with the electron or positron, only these two particles emerge from the atom as the products of the primary interaction. Any photon energy in excess of 1.022 MeV appears as kinetic energy, divided between the electron and the positron.

*i. Triplet production.* If the target charged particle for this process is an atomic electron ($\kappa_e$), the target electron recoils in the forward direction, along with the created electron-positron pair, with all three equal-mass particles ($2e^-, e^+$) sharing as kinetic energy the excess photon energy. In the early cloud-chamber observations of this process, the three particles appeared as a three-pronged trident, or "triplet." Due to the kinematics, with the recoiling electron target, the threshold energy for triplet production ($\kappa_e$) is four electron rest-mass energies ($4mc^2 = 2.044$ MeV).

*ii. Annihilation radiation.* The produced positron ($e^+$), as was mentioned earlier, eventually collides catastrophically with an electron ($e^-$) and both are annihilated, reverting usually to a pair of photons, each of energy of one electron rest-mass energy ($mc^2 = 0.511$ MeV), higher if the positron annihilates while still in flight, and (rarely) to a single or to three photons.

The probability for pair (or triplet) production is approximately proportional to the square of the charge of the target particle. Hence, the cross section for pair production in the field of the nucleus ($\kappa_n$) varies as $Z^2$, whereas for triplet production ($\kappa_e$) the cross section varies simply as Z, the number of unit-charge electron target particles per atom. Thus for hydrogen (Z = 1), at photon energies well above the thresholds, the cross sections for $\kappa_e$ and $\kappa_n$ are approximately equal, whereas for higher-Z elements $\kappa_e$ is smaller approximately as

$$\kappa_e/\kappa_n \approx Z/Z^2 = 1/Z.$$

Pair and triplet production ($\kappa_n$ and $\kappa_e$) are the dominant processes for photon interactions with atoms at high energies, from a few MeV upwards.

*e. Photonuclear absorption ($\sigma_{ph.n.}$).* This type of interaction, with a threshold of the order of 5 MeV or higher, is somewhat the analog of the atomic photoelectric effect, but with the photon absorbed by the atomic nucleus, rather than by an electron in the shells surrounding the nucleus. The most likely result of such an interaction is the emission of a single neutron, in which case we can express the interaction as

$$N_i(\gamma, n)N_f$$

in which $N_i$ is the target nucleus and $N_f$ is the final nucleus, with one less neutron, and may be a radioactive isotope of the target nucleus $N_i$. Besides single neutron emission, other possibilities include the emission of charged particles, gamma rays, or more than one neutron. Emission of charged particles, such as protons, changes the atomic number Z of the target element, as well as its mass.

The most characteristic feature of the photonuclear absorption cross section is the "giant resonance." This is a broad peak in the absorption cross section centered at about 24 MeV for light nuclei, decreasing in energy with increasing mass number to about 12 MeV for the heaviest stable nuclei. The width "Γ" (energy difference between the points at which the cross section drops to one half its maximum value) varies from about 3 MeV to 9 MeV depending on the detailed properties of individual nuclei.

The magnitude of the photonuclear cross section, even at the resonance peak energy, is small in comparison with the sum of the above "electronic" cross sections, and in no case contributes more than 10% to the total cross section. However, photonuclear absorption can be important in shielding technologies since the emitted neutrons are usually far more penetration than the incident photons, and obviously important in irradiation technology due to the induced radioactivities in the target materials.

*f. Other photon interactions.* A number of other less-probable things can happen to photons as they traverse and encounter tangible matter, but the interactions listed above are at present the only ones of practical interest in medical and technological applications. One of these less-probable but interesting interactions is Delbrück scattering, which is scattering of photons in the Coulomb (electrostatic) field of nuclei as a consequence of vacuum polarization. Delbrück scattering is considered to be "scattering of light by light," and has been observed for photon energies from a few hundred keV up to a few GeV. Other small effects include, for example, resonance scattering and Thomson scattering by the nucleus, Compton scattering by nucleons, meson production (at a few hundred MeV), resonance scattering associated

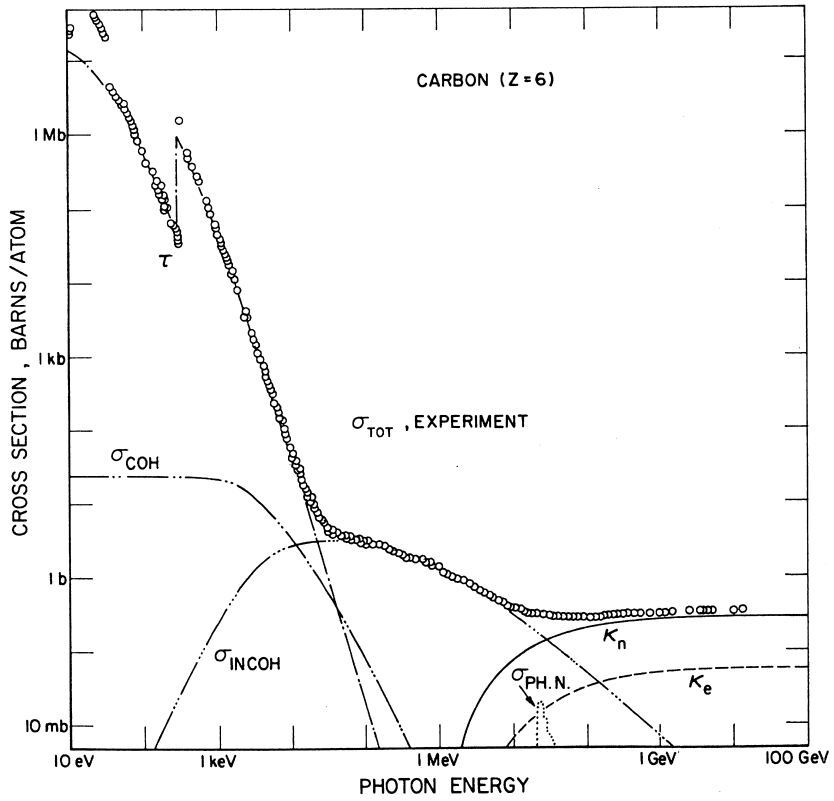



**FIGURE 3** Contributions of (a) atomic photoeffect, $\tau$, (b) incoherent (Compton) scattering, $\sigma_{inc}$, (c) coherent (Rayleigh) scattering, $\sigma_{coh}$, (d) nuclear-field pair production, $\kappa_n$, (e) electron-field pair production (triplet), $\kappa_e$, and (f) nuclear photoabsorption, $\sigma_{ph.n.}$, to the total measured cross section, $\sigma_{tot}$ (circles) in carbon over the photon energy range 10 eV to 100 GeV. The measured $\sigma_{tot}$ points, taken from 90 independent literature references, are not all shown in regions of high measurement density.

with meson production, and nucleon–antinucleon pair production. Nucleon–antinucleon pair production, of interest in cosmogenic modeling, has a threshold which varies from 3.75 GeV (4 nucleon masses) in the field of a proton at rest down to 3.0 GeV (3.2 nucleon masses) for protons moving in a complex nucleus.

*g. Total cross section ($\sigma_{tot}$) and the mass attenuation coefficient ($\mu/\rho$).* The total photon interaction cross section $\sigma_{tot}$, for a given material and a given photon energy is given as the sum over the photoelectric effect cross section $\tau$, the incoherent (Compton) and coherent (Rayleigh) scattering cross sections $\sigma_{inc}$ and $\sigma_{coh}$, the pair and triplet production cross sections $\kappa_n$ and $\kappa_e$, and the photonuclear absorption cross section $\sigma_{ph.n.}$:

$$\sigma_{tot} = \tau + \sigma_{inc} + \sigma_{coh} + \kappa_n + \kappa_e + \sigma_{ph.n.}$$

The relative importance of all six of these different types of interactions of photons with atoms, in the different regions of photon incident energy, can be seen in Figs. 3 and 4 for carbon (Z = 6) and for lead (Z = 82).

The photonuclear cross section $\sigma_{ph.n.}$ is at present not amenable to systematic compilation and tabulations, due to its complex dependence on the irregular variations in nuclear structure as a function of the atomic number Z and of the mass numbers A of the variable isotopic mixes of the stable elements. Hence, despite its importance (production of radioactivities) in biological and technological applications involving photon energies above 5 or 10 MeV, this cross section is still omitted from systematic general-use tabulations of x-ray attenuation coefficient and energy-absorption coefficient data. Hence, in currently available tabulations of the mass attenuation coefficient, $\mu/\rho$, the composition will be found to be

$$\mu/\rho = (\tau + \sigma_{inc} + \sigma_{coh} + \kappa_n + \kappa_e)/(u \cdot A)$$

with $\mu/\rho$ in units of $cm^2/g$ if the five component cross sections are in units of $cm^2$/atom, where u ($= 1.660\ 538\ 73(13) \cdot 10^{-24}$ g) is the atomic mass unit

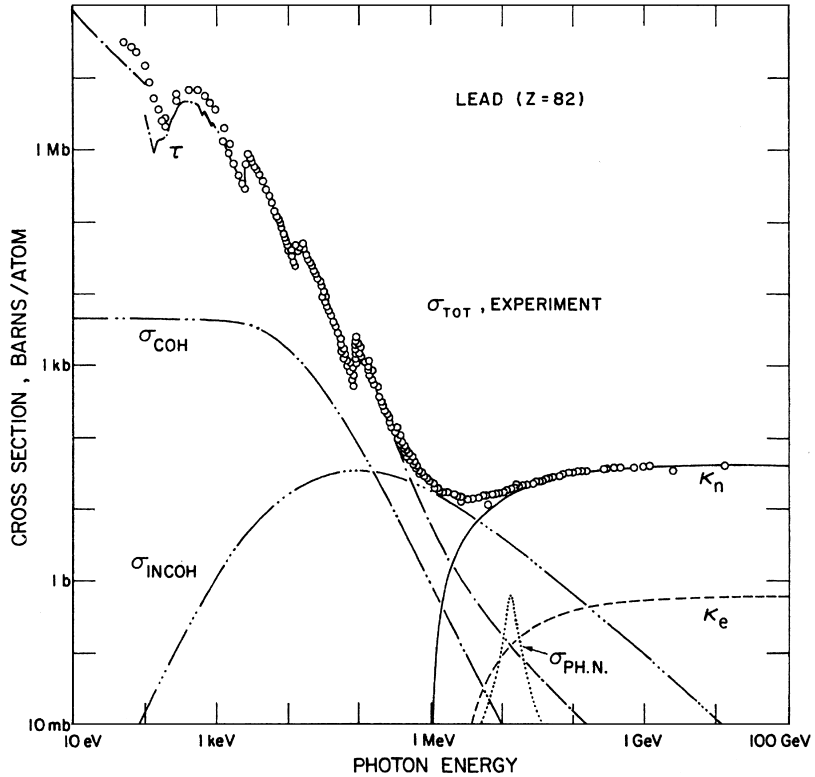



**FIGURE 4** Contributions of (a) atomic photoeffect, $\tau$,(b) incoherent (Compton) scattering, $\sigma_{\text{inc}}$, (c) coherent (Rayleigh) scattering, $\sigma_{\text{coh}}$, (d) nuclear-field pair production, $\kappa_{\text{n}}$, (e) electron-field pair production (triplet), $\kappa_{\text{e}}$, and (f) nuclear photoabsorption, $\sigma_{\text{ph.n.}}$, to the total measured cross section, $\sigma_{\text{tot}}$ (circles) in lead over the photon energy range 10 eV to 100 GeV. The measured $\sigma_{\text{tot}}$ points, taken from 121 independent literature references, are not all shown in regions of high measurement density.

(1/12 of the mass of an atom of the nuclide $^{12}$C) and A is the relative atomic mass of the target element.

*i. Additivity.* For compounds and mixtures, the mass attenuation coefficient $\mu/\rho$ at a particular photon energy E can be obtained according to simple additivity:

$$\mu/\rho = \sum_{\text{i}} w_{\text{i}}(\mu/\rho)_{\text{i}}$$

in which $w_{\text{i}}$ is the fraction by weight of the $i$th elemental constituent and $(\mu/\rho)_{\text{i}}$ is the mass attenuation coefficient for that element for photons of energy E.

*ii. Narrow-beam attenuation.* For an idealized pencil beam of photons and a detector shielded by a collimator such that secondary radiations (e.g., scattered and fluorescence photons) from the interposed target material cannot be seen by the detector, the exponential attenuation law mentioned earlier holds exactly:

$$\text{I}/\text{I}_0 = \text{e}^{-(\mu/\rho)\text{x}}$$

in which x is in units of g/cm$^2$ if $\mu/\rho$ is in units of cm$^2$/g as mentioned above.

*h. Other factors governing photon attenuation.* Although the idealized pencil beam (collimated source, collimated detector) geometry is useful for performing basic measurements of the mass attenuation coefficient $\mu/\rho$ in support of theoretical modeling of the individual types of interaction cross sections and their sum $\sigma_{\text{tot}}$, in practical situations two other factors modify the exponential attenuation law. One of these is the inverse square law, in which the radiation flux at various radial distances $r$ from a point isotropic (PTI) source of photons varies according to $1/r^2$. For a point isotropic source embedded in an absorbing and scattering medium, the flux detected by an uncollimated detector will also be modified by a buildup factor $\text{B}_{\text{PTI}}(E_0, Z, r)$, due to scattering and other secondary radiations reaching the detector. In this representation of the buildup factor, $E_0$ is the photon source energy, $Z$ represents the element or mixture comprising the medium, and $r$ is the distance from the source, usually in units of mfp (mean-free-path), in which 1 mfp $= 1/\{(\mu/\rho)[E_0, Z]\}$. For applications in radiation field modeling for extended sources, $\text{B}_{\text{PTI}}(E_0, Z, r)$ has been parametrized as a function of $r$ in a variety of

empirical and semi-empirical analytical formulations, as will be discussed in a later section on radiation fields from point and extended sources. However, to a large extent Monte Carlo calculations have replaced the use of buildup factors in radiation transport predictions except for some benchmark limiting cases.

*i. Mass energy-transfer coefficient $\mu_{tr}/\rho$ and mass energy-absorption coefficient $\mu_{en}/\rho$.* For computing the energy deposited at a given point within an absorbing medium, subject to a photon flux traversing that point, the mass energy-transfer coefficient $\mu_{tr}/\rho$ and mass energy-absorption coefficient $\mu_{en}/\rho$ are useful quantities. These coefficients account for the energy loss from a collision site of the energy of secondary photon radiation which departs from the collision site, instead of the entire photon energy being deposited at the collision site in the form of charged-particle kinetic energy. The difference between these quantities, and how they differ also from the total mass attenuation coefficient $\mu/\rho$, is illustrated pictorially in Fig. 5, in which the upward branching arrows represent the energy of the departing secondary photon radiations.The mass energy-transfer coefficient, used for computing kerma, is defined as

$$\mu_{tr}/\rho = (f_{\sigma inc}\sigma_{inc} + f_{\tau}\tau + f_{\kappa n}\kappa_n + f_{\kappa e}\kappa_e)/(u \cdot A)$$

in which the factors $f_i$ are the average fractions of the incident photon energy, for each type of interaction, left at the collision site in the form of ionization, excitation and kinetic energy of charged particles, the remainder departing, as shown, the collision site in the form of Compton (incoherent) scattered photons, fluorescence photons from photoelectric absorption or Compton vacancies, and annihilation radiation from the catastrophic encounters of pair- and triplet-produced positrons with electrons. For Rayleigh (coherent) scattering, the deflected photon loses negligible energy, so that $\sigma_{coh}$ does not contribute to the mass energy-transfer coefficient $\mu_{tr}/\rho$. A more-detailed version of this energy-deposition quantity, as shown in Fig. 5, is the mass energy-absorption coefficient $\mu_{en}/\rho$ which takes into account the bremsstrahlung (photon) radiation produced by the energetic charge particles, including positrons, which can depart from the region around the original collision site without depositing energy. A further refinement takes into account the fact that the combined energies of the annihilation radiation photons from positrons annihilating in flight is greater than $2mc^2$. The mass energy-absorption coefficient $\mu_{en}/\rho$ is expressed as

$$\mu_{en}/\rho = (1-g)\mu_{tr}/\rho$$

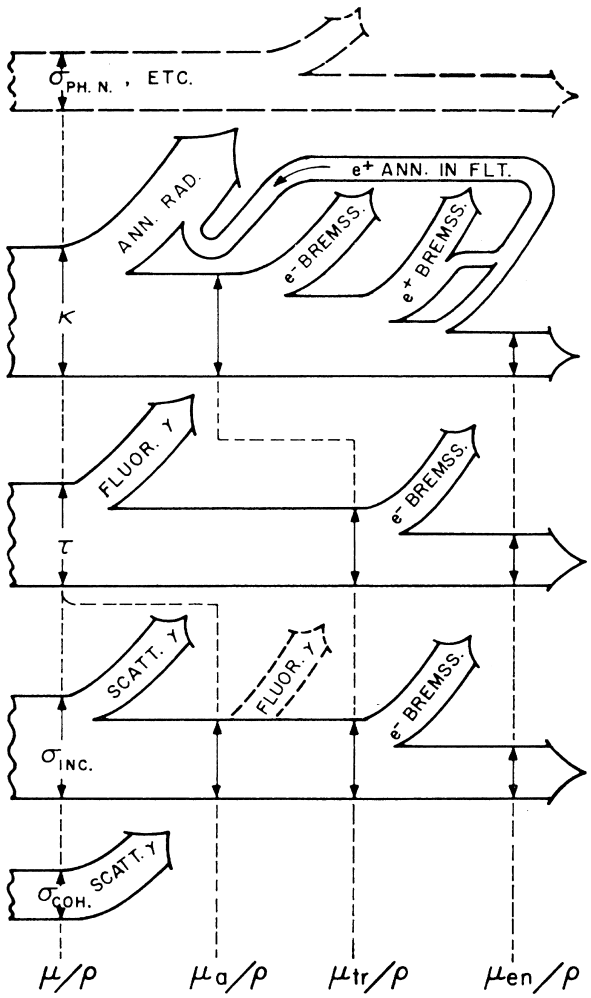



**FIGURE 5** Schematic representation of the mass attenuation coefficient $\mu/\rho$, the mass absorption coefficient $\mu_a/\rho$, the mass energy-transfer coefficient $\mu_t/\rho$, and the mass energy-absorption coefficient $\mu_{en}/\rho$ in terms of the cross sections for coherent ($\sigma_{coh}$) and incoherent ($\sigma_{incoh}$) scattering, atomic photoeffect ($\tau$), pair production ($\kappa$), and photonuclear reactions ($\sigma_{ph.n.}$). The upward-branching arrows represent the fraction, of the incident photon energy, lost to the volume of interest in the form of secondary photons such as positron annihilation radiation (ANN. RAD.), bremsstrahlung ($e^-$, $e^+$ BREMSS.), fluorescence x rays (FLUOR. $\gamma$) and scattered photons (SCATT. $\gamma$). The enhancement of annihilation photon energies due to positron annihilation in flight ($e^+$ ANN. IN FLT.) at the expense of positron bremsstrahlung and energy deposition is also indicated.

in which g is the average bremsstrahlung yield (fraction) for all of the charged particles produced in the original collision in the form of photoelectrons (and Auger electrons), Compton-ejected electrons, and electron-positron pairs (and triplets) from pair production events. A further refinement could include the effects of secondary photons and charged particles from photonuclear absorption $\sigma_{ph.n.}$, but insufficient systematic information is

available to include this process in currently available data tales.

## 3. Neutron Absorption Mechanisms

Neutrons are uncharged particles that can interact easily with a positively charge atomic nucleus where various capture processes can occur. The probability of such an interaction is low, and hence neutrons are regarded as highly penetrating radiation. In elastic scattering the neutron loses part of its energy by displacing atoms. It can be shown that the energy transferable during collisions is greatest for atoms of low atomic mass, especially hydrogen. Therefore, hydrogenous materials are useful for neutron shielding and detection. Energetic protons are created during elastic collisions with hydrogenous materials such as polymers and water.

After a number of collisions in a material at room temperature, the neutrons become thermalized, that is, their kinetic energy is of the order of 0.025 eV, the average vibrational energy of the atom in a solid at room temperature. If the solid is cooled below room temperature, then neutrons of even lower energies are produced. These are called cold neutrons and have uses in materials physics. The National Institute of Standards and Technology, from its reactor, produces neutrons at cryogenic temperatures, valuable in surface studies due to the effective long wavelength of the neutrons.

Thermal neutrons can be captured by atomic nuclei. This usually leads to the emission of a $\gamma$ photon from the excited nucleus and the creation of a new isotope, a process known as transmutation. The new isotope may or may not be radioactive. If it is, then the process is termed radioactivation. If the neutron capture results in the emission of more neutrons, and a moderator material is present, a chain reaction can occur. In particular, in the case of heavy elements such as uranium which can undergo fission from neutron capture, large amounts of energy can be released in either controlled or uncontrolled modes.

## 4. Electron ($\beta$-Ray) Absorption

The two principal mechanisms by which electrons shed their kinetic energy are by collision, which yields ionized or excited atoms, and the production of bremsstrahlung radiation. A third mode of interaction of electrons with matter is the Cherenkov effect which plays a minor role in energy deposition, but is quite visually dramatic in refractive transparent media.

*a. Collisions.* As they pass through matter, electrons gradually lose energy by means of collisions with atoms, along a wandering path. The collision energy is then usually taken up by other electrons; these ejected electrons are termed secondary electrons. Occasionally, atoms are displaced from their lattice positions by the collision, and a defect is left in the solid. Secondary electrons, after further collisions, generate electron-hole pairs or electron-ion pairs, depending on the medium in question. In a liquid or gas, a complex track of electron-ion pairs is produced, which has a core of intense ionization and small spurs called delta rays, where secondary electrons have departed radially from the original track.

*b. Electromagnetic (bremsstrahlung) radiation.* High-energy electrons, when deflected or slowed in Coulomb (electrostatic) fields of nuclei, produce photons known as bremsstrahlung ("braking" radiation). Electron–electron bremsstrahlung is also possible, but this is a less-probable process. Thin-target bremsstrahlung has energy and angular distributions well described mathematically according to equations developed by Schiff, and this broad continuum is the major photon output of x-ray machines for imaging and for therapy, particularly in the case of high-energy machines using electron linacs or other high-energy accelerators as sources of electrons to impinge on heavy-element bremsstrahlung converter targets.

*c. Cherenkov radiation.* The speed of light has a constant value of c in a vacuum. However, in a transparent material with a refractive index $n > 1$, the velocity of light in the material is reduced to c/n. Thus for a high-energy particle traveling at velocity $v = \beta c$, if $\beta > 1/n$, we can say that the particle is "traveling faster than light" in the material. As first observed experimentally by Cherenkov in 1934 and explained by Frank and Tamm in terms of Maxwell's theory, when a charged particle traverses matter at a velocity in excess of the speed of light in the medium, radiation in the optical-wavelength region is emitted. This light, called "Cherenkov radiation," is subject to constructive interference such that it is confined to a cone whose axis coincides with the direction of propagation of the charged particle, in a manner analogous to the production of a shock wave from a supersonic aircraft. The aperture half-angle $\theta$ of this propagating cone of light is given by

$$\cos\theta = 1/(\beta n)$$

in which $\beta$ is the ratio of the particle's velocity to that of light in a vacuum and n is the refractive index of the medium. In water moderated and shielded research reactors, where the fuel elements are visible, the eerie blue glow in the water surrounding the elements is Cherenkov radiation, mostly from Compton electrons induced by high-energy $\gamma$ rays.

## 5. Proton Radiation Absorption

Protons interact primarily with atomic electrons, hence their effects, including biological, are similar to those of x rays. However, as seen in Fig. 2, their depth-dose pattern is quite different. The proton depth-dose curve, here shown for 200-MeV protons in water, is characterized by a relatively low entrance dose plateau region, followed by a sharp dose peak (the Bragg peak) near the end of the range. Beyond the Bragg peak is a very sharp cut-off, with no protons and dose deposition reaching the areas deeper into the material. Protons scatter very little, hence the lateral dose fall-off is also very sharp. Thus, in cancer therapy involving tumors adjacent to critical structures which need to be spared, use of protons can avoid the use of isocentric treatments in which the source-beam is rotated around the patient. If the tumor is larger than the typically 1–2 cm width of the Bragg peak, the peak can be spread by rotating variable-thickness absorbers intercepting the incident beam.

## 6. Charged Nuclei

Besides protons (hydrogen nuclei, $Z = 1$), nuclei of heavier atoms moving at high energy can cause trails of intense ionization. Examples of such particles include $\alpha$ particles from radon and its short-lived daughter products (5, 6, and 7.7 MeV), cosmic ray particles with energies extending up to the TeV region, ions in accelerator beams, and fission fragments. Light atoms produce a track of significant length. For example, a neon ($Z = 10$) ion at 40 MeV has a track length in polymer plastics approaching 100 $\mu$m. Over a core region having a diameter of a few nanometers, the dose imparted to the material in the track is $10^6$ rad. Over a penumbra region having a diameter of about 10 $\mu$m the dose falls off to about 100 rad. Thus, the details of the effects of different energetic atoms on a solid vary strongly according to atomic weight. Heavy atoms in motion are encountered in cosmic rays and in fission fragments. During the fission of an atomic nucleus, a large amount of momentum is given to the two fragments of the nucleus. If the fission occurs in a solid, then the fragments are stopped rapidly. The resulting bulk damage to the solid is heavy and is localized in a small volume.

## 7. Neutrinos, $\nu$

It is beyond the scope of this article to list and classify the entire "zoo" of known charged and uncharged particles and anti-particles which could be considered to be within the domain of "radiation physics." However, the neutrino is here singled out for mention since it plays a role in the nuclear fusion processes energizing stars including our Sun as well as in supernovae and other astrophysical events. The neutrino was first postulated by W. Pauli in 1930 to avoid an apparent non-conservation of energy and linear momentum in radioactive $\beta$-decay. This elusive "massless" uncharged particle interacts with matter principally through the weak nuclear force, hence the cross section is so small that it can pass through the Earth with only a small probability of collision. Detectors consist of large tanks of water with surface studded with inward-looking photomultiplier tubes to observe Cherenkov or other radiations from secondary products of neutrino collisions.

The three known types of neutrinos, associated with electrons $\nu_e$, muons $\nu_\mu$, and with tau particles $\nu_\tau$, although the latter has not been observed directly.* The Sun's power is thought to be generated principally by hydrogen nuclei (protons, p) undergoing fusion into helium (He, $Z = 2$) and two positrons ($e^+$), according to

$$4p \rightarrow {}^4He + 2e^+ + 2\nu_e + 26.7\,\text{MeV}$$

from which the neutrinos $\nu_e$, even those produced deep in the core of the Sun, can easily escape to enable terrestrial detection and analysis. Currently, the observed Solar neutrino flux is less than the theoretical predictions.

A rare (but observed in geochemical experiments) example of the neutrino participation in radioactivity is double $\beta$-decay, which would be forbidden except for the neutrino production is:

$$\text{Nucleus}(Z, A) \rightarrow \text{Nucleus}(Z + 2, A) + 2e^- + 2\nu_e$$

in which Z is the atomic number and A is the atomic mass.

## 8. Radiation Fields from Point and Extended Sources

This section will apply mostly to photon sources, but for sources embedded in a medium with negligible scattering or absorption, also negligible magnetic or electrostatic fields in the case of charged particles, the analytical geometrical treatments here will also be applicable.

*a. Point source.* For a point isotropic (PTI) source of monoenergetic photons embedded in a homogeneous absorbing and scattering medium, as schematized in Fig. 6, the total response D of an isotropic detector at distance r from the source is

$$D = D^\circ + D^s$$

in which $D^\circ$ represents the response of the detector to photons arriving from the source without suffering a scattering interaction, and $D^s$ the response of the detector to

*See, however, B. Schwarzschild (2000). "The tau neutrino has finally been seen," *Physics Today* **53**(10), 17–19.

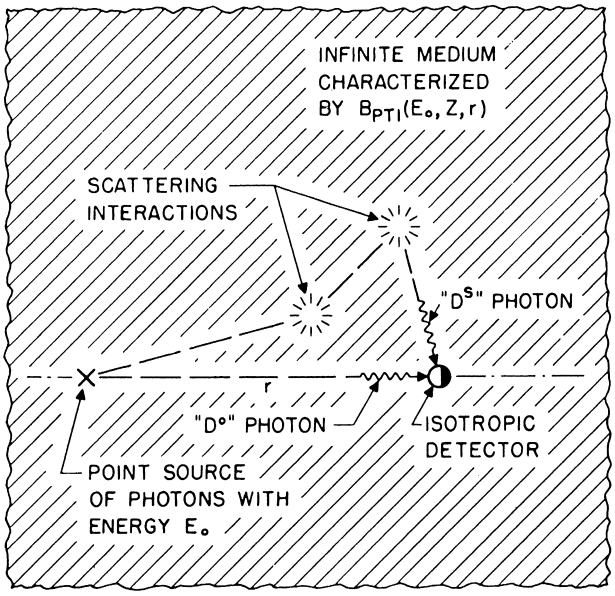



**FIGURE 6** Schematization of a point isotropic monoenergetic radiation source, and an isotropic radiation detector, embedded in a medium characterized by an attenuation coefficient $\mu$ and a buildup factor $B_{PTI}$ which accounts for scattered and other secondary radiations "seen" by the detector.

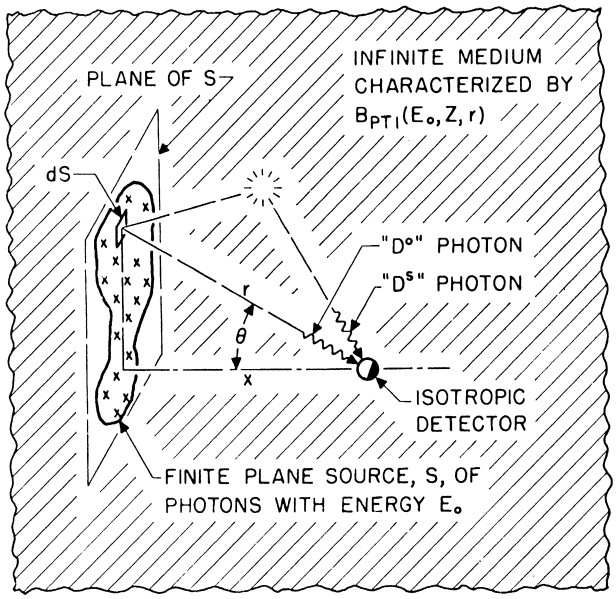



**FIGURE 7** Finite plane isotropic source S uniformly covered with isotropically radiating material.

photons arriving indirectly, via scattering or other modes of secondary radiations. The unscattered response can be calculated according to

$$D^\circ = (\sigma/4\pi)[\exp(-\mu r)]/r^2$$

in which the source-strength constant $(\sigma/4\pi)$ is the detector response at unit-distance from a unit-strength point source in the absence of attenuation [e.g., $\sigma/4\pi =$ 1.27 rads/hr in tissue 1 m from a 1 Ci $^{60}$Co source, from Table VI. C]. Defining the buildup factor $B_{PTI}$ for a given photon energy $E_o$, material Z and distance r, as

$$B_{PTI} = D/D^\circ = (D^\circ + D^s)/D^\circ$$

we have

$$D = D^\circ \times B_{PTI} = (\sigma/4\pi)B_{PTI}[\exp(-\mu r)]/r^2.$$

The build-up factor $B_{PTI}$ has been expressed in a variety of parametrized formulations suitable for different applications, most of them listed and compared in the articles by Hubbell (1963) and by Harima (1993) in the bibliography.

*b. Finite plane isotropic source with exponential and inverse-square-law attenuation, and buildup factor.* The detector response D to a finite plane isotropic source S such as shown in Fig. 7 can be expressed as the integral

$$D = \frac{\sigma}{4\pi}\int_s \frac{\exp(-\mu r)}{r^2} \times B_{PTI}(\mu r)\, dS.$$

For regular geometries such as the rectangular plaque source shown in Fig. 8 found in irradiation facilities, a convenient formulation for the buildup factor, to separate the

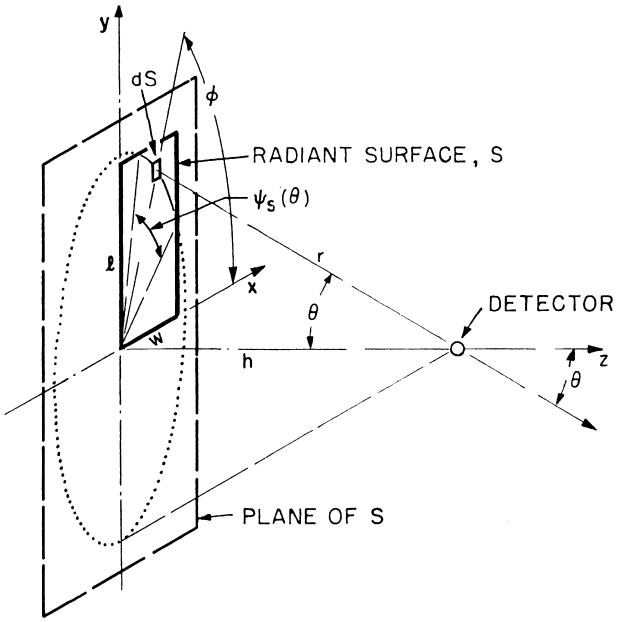



**FIGURE 8** Schematization showing the geometry parameters for a rectangular plaque source, S, such as used in irradiation processing, in which the detector is at a perpendicular distance h from a corner of the rectangle of width w and length l.

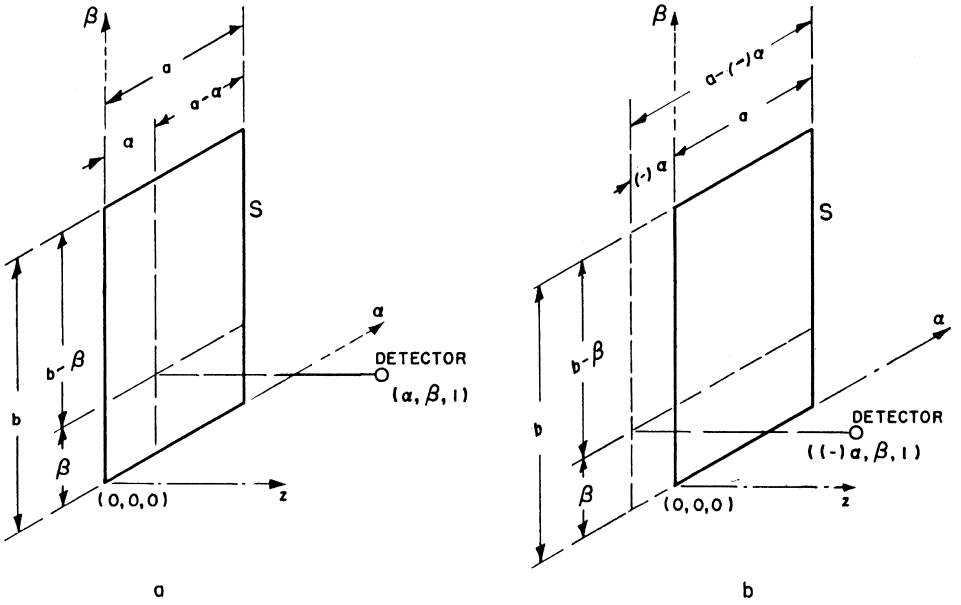



**FIGURE 9** Schematization for obtaining the detector response when the detector is not over a corner, but is displaced by distances $\alpha$ and $\beta$, respectively, from the corner of a plaque source with width a and length b.

shape (geometry) variables from the material-dependent penetration variables, is

$$B_{PTI}(\mu r) = \exp(+\mu r) \sum_{n=0}^{N \text{ or } \infty} b_n(-\mu r)^n/n!$$

in which the $b_n$ coefficients are tabulated in Hubbell (1963) which also provides transforms for obtaining $b_n$'s from other formulations given in termsof polynomials or exponential combinations. The detector response from the extended source S can then be computed according to

$$D = (\sigma/4\pi) \sum_{n=0}^{N \text{ or } \infty} b_n \times q_n(\text{geom}) \times (\mu x)^n$$

in which x is a fixed distance for a given geometry. The geometry (shape) coefficients $q_n$ can be calculated analytically for a regular shape such as the rectangular source in Fig. 8 according to

$$q_n(\text{geom}) = \int_S (-r/x)^n \, (1/r^2 n!) \, dS$$

or in polar coordinates

$$q_n(\text{geom}) = \int_{\Omega_S} [(-\sec\theta)^{n+1}/n!] \, d\Omega_S$$

in which $\Omega_S$ is the solid angle subtended by the source S with respect to the detector, and $\theta$ is the angle between r and x as indicated in Fig. 7 or between r and h as in Fig. 8.

*c. Rectangular plaque source.* For a detector opposite a corner of a rectangular plaque source, analytical expressions and tabulations of the geometry coefficients $q_n$ are available in Hubbell *et al.* (1960). For the detector not opposite a corner, but displaced by distances $\alpha$ and $\beta$ from the corner of a rectangular source of width a and length b, in units of the detector distance from the source-plane (h = 1), as shown in Fig. 9, corner-position data can be combined according to

$$D(a, b; \alpha, \beta) = D(\alpha, \beta) + D(a - \alpha, b - \beta) + D(\alpha, b - \beta) + D(a - \alpha, \beta).$$

For a "bare" rectangular plaque source (embedded in a medium with negligible absorption and scattering), the corner-position (h = 1) detector response is given by the zeroth term of the above series

$$D(a, b) = (\sigma/4\pi) \times q_0(a, b).$$

in which

$$q_0(a, b) = \int_0^b \frac{1}{\sqrt{\beta^2 + 1}} \tan^{-1} \frac{a}{\sqrt{\beta^2 + 1}} \, d\beta.$$

Although the rectangular source geometry coefficients $q_n(a, b)$ are available in closed form (Hubbell *et al.*, 1960) for all $n \geq 1$, the above integral for n = 0 is not soluble in closed form. However, a series solution which is rapidly convergent for all values of a and b, i.e., for $0 \leq a \leq b \leq \infty$, given in Hubbell et al. (1960) is

$$q_0(a,b) = \frac{\pi}{2}\sinh^{-1} a - \frac{b}{a}\sum_{i=0}^{\infty}\frac{1}{2i+1}\left(\frac{a^2}{a^2+1}\right)^{i+1} \times \left\{\frac{\sqrt{a^2+1}}{b}\tan^{-1}\frac{\sqrt{a^2+1}}{b} - \sum_{j=0}^{i-1}\frac{2^{2j}(j!)^2}{(2j+1)!}\left(\frac{a^2+1}{a^2+b^2+1}\right)^{j+1}\right\}$$

For all a ≤ b, only terms in the outer sum i = 0 to i = 6 are required for 0.01% accuracy, since the inner finite summation is actually the leading terms of the $\tan^{-1}$ (and factor) expression just before it, hence the very rapid convergence of the outer infinite summation.

*Infinite strip "bare" source*. For large b ($b \to \infty$) the summation vanishes, leaving only

$$q_0(a, \infty) = (\pi/2)\sinh^{-1} a.$$

Hence the response of a detector at unit height centered above an infinite strip source of width 2a and surface activity $\sigma$, in a nonattenuating medium, is

$$D = 4(\sigma/4\pi)(\pi/2)\sinh^{-1} a = (\sigma/2)\sinh^{-1} a$$

*d. Circular disk plane isotropic source, off-axis.* Similarly, for a unit-radius (r = 1) circular disk plane isotropic source as shown in Fig. 10, the response of a detector at height h (in disk radii) above the plane of the source and at distance $\rho$ (in disk radii) off-axis can also be computed using disk source (off-axis) geometry coefficients $q_n(\rho, h)$ given analytically and in tables in Hubbell *et al*. (1961)

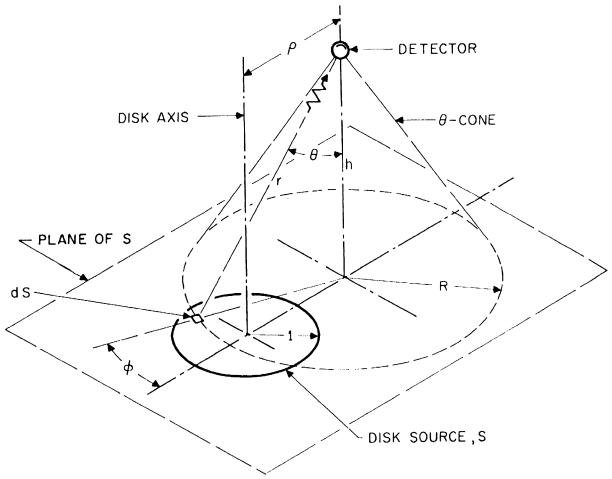



**FIGURE 10** Source-detector geometry for a circular disk source, with the detector at an off-axis position, showing the relevant parameters $\rho$, h, $\phi$, r and R, in which the linear dimensions are measured in radii of the source disk.

Analogous to the detector response coefficient for a bare rectangular source $q_0(a, b)$, the corresponding coefficient $q_0(\rho, h)$ for a detector at height h (in disk radii) from the plane of a disk source of unit radius, displaced from the disk axis by distance $\rho$ (in disk radii), can be expressed in the closed form

$$q_0(\rho, h) = \pi \ln\left((1 + h^2 - \rho^2 + \sqrt{(1+h^2-\rho^2)^2 + 4\rho^2 h^2})/2h^2\right)$$

Formulas and tabulations of higher terms through n = 9 of $q_n(\rho, h)$ are given in Hubbell (1961), also formulas and data for treating anisotropic angular source distributions.

### 9. Monte Carlo Simulations of Radiation Transport

Although analytical methods of mapping radiation fields, such as in the examples above, are sometimes useful as benchmarks for simple or limiting situations, the availability of modern high-speed and large-memory computers has resulted in the wide use of Monte Carlo simulation. In this method, trajectories of photons and/or particles are determined by random numbers at each collision point in the medium, weighting the azimuthal and deflection angles by the probabilities, or differential cross sections, for the interactions with target atoms. For more extensive information on Monte Carlo simulations, see the reviews by Morin (1988) and by Jenkins *et al*. (1988).

## V. USEFUL RADIATION DATA

### A. Radiation Units

$$1\,\text{rad} = 100\,\text{erg g}^{-1} = 6.25 \times 10^{13}\,\text{eV g}^{-1} = 10^{-2}\,\text{Gy}$$

$$1\,\text{Mrad} = 6.25 \times 10^{19}\,\text{eV g}^{-1} = 10\,\text{kGy}$$

$$1\,\text{Gy} = 1\,\text{J kg}^{-1} = 100\,\text{rad}$$

$$1\,\text{kGy} = 10^5\,\text{rad} = 100\,\text{krad}$$

$$1\,\text{MGy} = 10^8\,\text{rad} = 100\,\text{Mrad}$$

$$10^{20}\,\text{eV g}^{-1} = 1.6\,\text{Mrad} = 16\,\text{kGy}$$

1 roentgen (R) = 86.9 erg $g^{-1}$(air) = $2.58 \times 10^{-4}$ Coulombs $kg^{-1}$(air)
1 R of 1-MeV photons ≡ $1.95 \times 10^9$ photons $cm^{-2}$.

This fluence deposits 0.869 rad in air
0.965 rad in water
0.865 rad in silicon
0.995 rad in polyethylene
0.804 rad in LiF
0.862 rad in Pyrex glass
(80% $SiO_2$).

One curie (Ci) of radioactive material produces $3.700 \times 10^{10}$ disintegrations per second.

A 1-Ci point source emitting one 1-Mev photon per disintegration gives an exposure of 0.54 R $hr^{-1}$ at 1 m.

A 1-Ci $^{60}Co$ source gives 1.29 R $hr^{-1}$ at 1 m.

Photon flux at 1 m from a 1-Ci point source $= 1.059 \times 10^9$ $cm^{-2}$ $hr^{-1}$ (assuming one $\gamma$-ray photon per disintegration).

SI Units recommended by the international Commission on Radiation Units and measurements (I.C.R.U.) [*Brit. J. Radiology* **49,** 476(1976)] are the following:

— Absorbed dose: the Gray (Gy) = 100 rad = 1 J/kg
— Exposure: the coulomb per kilogram (no name given) = 1 C/kg
— Quantity activity: the Becquerel (Bq) = 1 $sec^1$ = $2.703 \times 10^{-11}$ Ci, with the old units to be abandoned over 10 years.

Basic units of biological dose from radiation are the rem and the sievert (Sv). 1 rem (radiation equivalent man) is the absorbed dose to the body of 1 rad weighted by a quality factor Q.F. that is dependent on the type of radiation involved. This is because the energy absorption from radiation is insufficient as a measure of biological damage.

$$1\,\text{rem} = \text{Q.F.} \times 1\,\text{rad}$$
$$1\,\text{Sv} = \text{Q.F.} \times 1\,\text{Gy}$$
$$\text{Q.F.} = 1 \text{ for X rays, } \gamma, \text{ or } \beta.$$

The values of Q.F. for $\alpha$, neutrons, and heavy particles are greater than 1.

## B. Useful General Constants

$$7\,\text{years} = 3.682 \times 10^6\,\text{min} = 2.209032 \times 10^8\,\text{sec}$$
$$1\,\text{year} = 5.25960 \times 10^5\,\text{min} = 3.155760 \times 10^7\,\text{sec}$$
$$1\,\text{day} = 1.440 \times 10^3\,\text{min} = 8.6400 \times 10^4\,\text{sec}$$
$$1000\,\text{Å} = 0.1\,\mu\text{m} = 100\,\text{nm}$$
$$1\,\text{mm} = 0.03937\,\text{in.}$$
$$0.001\,\text{in.} = 25.4\,\mu\,\text{m}$$
$$1\,\text{m}^3 = 10^6\,\text{cm}^3$$
$$1000\,\text{cm}^3 = 10^{-3}\,\text{m}^3$$
$$1\,\text{l} = 1000.028\,\text{cm}^{-3} = 0.219\,976\,\text{gal}$$
$$1\,\text{g cm}^{-3} = 1000\,\text{kg m}^{-3}$$
$$^*1\,\text{eV} = 1.602\ 176\ 462(63) \times 10^{-19}\,\text{J}$$
$$^*1\,\text{MeV} = 1.602\ 176\ 462(63) \times 10^{-13}\,\text{J}$$
$$1\,\text{J} = 10^7\,\text{erg}$$
$$1\,\text{cal} = 4.187\,\text{J}$$
$$1\,\text{eV /molecule} = 23.1\,\text{kcal/mol}$$

Permittivity of free space $\varepsilon_0 = 8.86 \times 10^{-14}\,\text{F cm}^{-1} = 8.86 \times 10^{-12}\,\text{F m}^{-1} = 55.4$ electronic charges $\text{V}^{-1}\,\mu\,\text{m}^{-1}$

Permeability of free space $\mu_0 = 1.26 \times 10^{-6}\,\text{H m}^{-1}$

$^*$Electronic charge $q = 1.602\,176\,462(63) \times 10^{-19}$ C

$^*1\,\text{C cm}^{-2} = 6.241\,509\,74 \times 10^{18}$ electrons $\text{cm}^{-2}$

$1\,\mu\text{A cm}^{-2} = 6.24 \times 10^{12}$ electrons $\text{cm}^{-2}\,\text{sec}^{-1}$

$^*$Velocity of light $= 2.997\,924\,58 \times 10^8\,\text{m sec}^{-1}$

$$1\,\text{N} = 10^5\,\text{dyn}$$
$$1\,\text{mm Hg} = 133.3224\ \text{N m}^{-2}$$

$^*$Boltzmann's constant $k = 1.380\,650\,3(24) \times 10^{-23}\,\text{J K}^{-1} = 8.617\,342(15) \times 10^{-5}\,\text{eV K}^{-1}$

$kT$ at room temperature $= 0.0259\,\text{eV}$

$^*$Planck's constant $h = 6.626\,068\,76(52) \times 10^{-34}\,\text{J sec}$

$^*$Avogadro's number $= 6.022\,141\,99(47) \times 10^{23}\,\text{mol}^{-1}$

$^*$Electron rest mass $m_c = 9.109\,381\,88(72) \times 10^{-31}\,\text{kg}$

$^*$Proton rest mass $m_\text{p} = 1.672\,621\,58(13) \times 10^{-27}\,\text{kg}$

$^*$From P. J. Mohr and B. N. Taylor (2000). The Fundamental Physical Constants, *Physics Today* **53**(No. 8, Part 2), BG6–BG13.

## C. Dose Rates From $\gamma$ Emitters[a]

| Nuclide | Half-life | Principal $\gamma$ energies (MeV) | Dose rate at 1 m from 1 Ci (rad/hr in tissue) |
|---|---|---|---|
| Antimony-124 | 60 day | 0.60; 0.72; 1.69; 2.09 | 0.94 |
| Arsenic-72 | 26 hr | 0.51;[b] 0.63; 0.835 | 0.97 |
| Arsenic-74 | 18 day | 0.51;[b] 0.596; 0.635 | 0.42 |
| Arsenic-76 | 26.5 hr | 0.56; 0.66; 1.21; 2.08 | 0.23 |
| Barium-140[c] | 12.8 day | 0.16; 0.33; 0.49; 0.54; 0.82; 0.92; 1.60; 2.54 | 1.19 |
| Bromine-82 | 35.4 hr | 0.55; 0.62; 0.70; 0.78; 0.83; 1.04; 1.32; 1.48 | 1.40 |
| Caesium-137 | 30 yr | 0.662 | 0.32 |
| Cobalt-58 | 71 day | 0.51;[b] 0.81; 1.62 | 0.53 |
| Cobalt-60 | 5.26 yr | 1.17; 1.33 | 1.27 |
| Gold-198 | 2.70 day | 0.412; 0.68; 1.09 | 0.22 |
| Iodine-131 | 8.04 day | 0.28; 0.36; 0.64; 0.72 | 0.21 |
| Iodine-132 | 2.3 hr | 0.52; 0.65; 0.67; 0.78; 0.95; 1.39 | 1.13 |
| Iridium-192 | 74 day | 0.296; 0.308; 0.316; 0.468; 0.605; 0.613 | 0.46 |
| Iron-59 | 45 day | 0.19; 1.10; 1.29 | 0.61 |
| Manganese-52 | 5.7 day | 0.51;[b] 0.74; 0.94; 1.43 | 1.79 |
| Manganese-54 | 314 day | 0.84 | 0.45 |
| Potassium-42 | 12.4 hr | 1.52 | 0.13 |
| Radium-226[d] | 1620 yr | 0.05–2.43 | 0.79 |
| Sodium-22 | 2.6 yr | 0.51;[b] 1.28 | 1.15 |
| Sodium-24 | 15.0 hr | 1.37; 2.75 | 1.77 |
| Tantalum-182 | 115 day | 0.068; 0.100; 0.222; 1.12; 1.19; 1.22; 1.23 | 0.64 |
| Thulium-170 | 127 day | 0.052; 0.084 | 0.002 |
| Zinc-65 | 245 day | 0.51;[b] 1.11 | 0.26 |

[a] Reprinted with permission from "The Radiochemical Manual." The Radiochemical Centre. Amersham, England, 1966.
[b] 0.51-MeV $\gamma$ rays from positron annihilation.
[c] Barium-140 in equilibrium with lanthanum-140.
[d] Radium-226 in equilibrium with daughter products; radiation filtered through 0.5 mm platinum; dose rate from l g.

## SEE ALSO THE FOLLOWING ARTICLES

ATOMIC PHYSICS • COSMIC RADIATION • DOSIMETRY • HEALTH PHYSICS • NEUTRINOS • NUCLEAR PHYSICS • NUCLEAR REACTOR MATERIALS AND FUELS • NUCLEAR SAFEGUARDS • RADIATION SHIELDING AND PROTECTION • SOLAR PHYSICS

## BIBLIOGRAPHY


Berger, M. J., and Hubbell, J. H. (1987). "XCOM: Photon Cross Sections on a Personal Computer," National Bureau of Standards (now National Institute of Standards and Technology) Report NBSIR 87-3597. Current version of this database available at: http://physics.nist.gov/PhysRefData/Xcom/Text/XCOM.html

Chilton, A. B., Shultis, J. K., and Faw, R. E. (1984). "Principles of Radiation Shielding," Prentice-Hall, Englewood Cliffs, NJ.

Christophorou, L. G. (1971). "Atomic and Molecular Radiation Physics," Wiley, New York.

Franklin, A. (2000). "The Road to the Neutrino," *Physics Today* **53**(2), 22–28.

Fuller, E. G., and Hayward, E. (1976). "Photonuclear Reactions," Dowden, Hutchinson & Ross, Stroudsburg, PA.

Greening, J. R. (1981). "Fundamentals of Radiation Dosimetry," Adam Hilger Ltd., Bristol, UK.

Harima, Y. (1993). "An historical review and current status of buildup factor calculations and applications," *Rad. Physics Chem*/5), 631–672.

Henke, B. L., Gullikson, E. M., and Davis, J. C. (1993). "X-ray interactions: Photoabsorption, scattering, transmission, and reflection at E = 50–30,000 eV, Z = 1–92," *Atomic Data Nucl. Data Tables* **54**(2), 181–342.

Hubbell, J. H., Bach, R. L., and Lamkin, J. C. (1960). "Radiation field from a rectangular source," *J. Res. Nat. Bureau Stand.* **64C**(2), 121–138.

Hubbell, J. H., Bach, R. L., and Herbold, R. J. (1961). "Radiation field from a circular disk source," *J. Res. Nat. Bureau Stand.* **65C**(4), 249–264.

Hubbell, J. H. (1963). "A power-series buildup factor formulation. Application to rectangular and off-axis disk source problems," *J. Res. Nat. Bureau Stand.* **67C**(4), 291–306.

Hubbell, J. H. (1969). "Photon Cross Sections, Attenuation Coefficients, and Energy Absorption Coefficients from 10 keV to 100 GeV," National Bureau of Standards Reference Data Series NSRDS-NBS 29.

Hubbell, J. H., Veigele, Wm. J., Briggs, E. A., Brown, R. T., Cromer, D. T., and Howerton, R. J. (1975). "Atomic form factors, incoherent scattering functions, and photon scattering cross sections," *J. Phys. Chem. Ref. Data* **4**(3), 471–538, Erratum in **6**(2), 615–617 (1977).

Hubbell, J. H., Gimm, H. A., and Øverbø, I. (1980). "Pair, triplet, and total atomic cross sections (and mass attenuation coefficients) for 1 MeV–100 GeV photons in elements Z = 1 to 100," *J. Phys. Chem. Ref. Data* **9**(4), 1023–1147.

Hubbell, J. H. (1982). "Photon mass attenuation and energy-absorption coefficients from 1 keV to 20 MeV," *Int. J. Appl. Rad. Isotopes* **33**(11), 1269–1290.

Hubbell, J. H. (ed.) (1993). "Radiation physics at 1993: A topical compendium," *Rad. Phys. Chem.* **41**(4/5), 579–789.

Hubbell, J. H., Trehan, P. N., Singh, N., Mehta, D., Garg, M. L., Garg, R. R., Singh, S., and Puri, S. (1994). "A review, bibliography, and tabulation of K, L, and higher atomic shell x-ray fluorescence yields," *J. Phys. Chem. Ref. Data* **23**(2), 339–364.

ICRU (1969). "Neutron Fluence, Neutron Spectra and Kerma," ICRU Report 13, ICRU Publications, Bethesda, MD.

ICRU (1984). "Stopping Powers for Electrons and Positrons," ICRU Report 37, ICRU Publications, Bethesda, MD.

ICRU (1993). "Stopping Powers for Protons and Alpha Particles," ICRU Report 49, ICRU Publications, Bethesda, MD.

ICRU (1998). "Fundamental Quantities and Units for Ionizing Radiation," ICRU Report 60, ICRU Publications, Bethesda, MD.

Jenkins, T. M., Nelson, W. R., and Rindi, A. (eds.) (1988). "Monte Carlo Transport of Electrons and Photons," Plenum, New York.

Johns, H. E., and Cunningham, J. R. (1983). "The Physics of Radiology," 4th ed., Charles C Thomas, Springfield, IL.

Knoll, G. F. (2000). "Radiation Detection and Measurement," 3rd ed., Wiley, New York.

Koch, H. W., and Motz, J. W. (1959). "Bremsstrahlung cross-section formulas and related data," *Rev. Mod. Phys.* **31**(4), 920–955.

Lamarsh, J. R. (1983). "Introduction to Nuclear Engineering," 2nd ed., Addison-Wesley, Reading, MA.

Morin, R. L. (ed.) (1988). "Monte Carlo Simulation in the Radiological Sciences," CRC Press, Boca Raton, FL.

Motz, J. W., Olsen, H. A., and Koch, H. W. (1969). "Pair production by photons," *Rev. Mod. Phys.* **41**(4), 581–639.

Schwarzschild, B. "The tau neutrino has finally been seen," *Physics Today* **53**(10), 17–19.

Seltzer, S. M. (1993). "Calculation of photon mass energy-transfer and mass energy-absorption coefficients,"*Rad. Res.* **136**(2), 147–170.

Winter, K. (ed.) (1991). "Neutrino Physics," Cambridge University Press, Cambridge, U. K. and New York.



# Superconductivity

**H. R. Khan**
*FEM and University of Tennessee at Knoxville*

I. Introduction
II. Superconducting Materials
III. Correlation: $T_c$ with the Electronic Structure of a Solid
IV. Flux Quantization
V. London Equation and Coherence Length
VI. Coherence Length and Energy Gap
VII. Thermodynamics of Superconductivity
VIII. Magnetic Superconductors
IX. Tunneling and the Josephson Effect
X. Theory of Superconductivity
XI. Applications of Superconductivity
XII. Recent Developments: High-Transition Temperature Superconductivity

## GLOSSARY

**Coherence length** Correlation distance of the superconducting electrons.
**Critical magnetic field** Above this value of an externally applied magnetic field, a superconductor becomes nonsuperconducting (normal).
**Energy gap** Gap in the low-energy excitations of a superconductor.
**Type I superconductor** When an external magnetic field is applied on this superconductor, the transition from a superconducting to a normal state is sharp.
**Type II superconductor** When an external magnetic field is applied, the transition from a superconducting to a normal state occurs after going through a broad "mixed-state" region.

**SUPERCONDUCTORS** are materials that lose all their electrical resistivity below a certain temperature and become diamagnetic. High values of an externally applied magnetic field are required to destroy the superconductivity. These electrical and magnetic properties of superconducting materials have found applications in lossless electrical transmission and generation of high-magnetic fields. Superconducting magnets are used where normal iron magnets are inadequate. These magnets are used as exciter magnets for homopolar generators or rotors in

large alternators, and much gain in efficiency and power density is obtained. Future fusion reactors will use superconducting magnets for confining deuterium and tritium plasma. Superelectron pairs in a superconductor can tunnel through a nonconducting thin layer. Based on this "Josephson effect," superconducting Josephson junctions are used as sensors, as high-energy electromagnetic radiation detectors, and in high-speed digital signal and data processing.

## I. INTRODUCTION

### A. Discovery of Liquid Helium Gas and Superconductivity

In 1908 Kammerlingh Onnes succeeded in liquifying helium gas, and this enabled him to measure the electrical resistivity of metals at lower temperatures down, to 4.2 K. The boiling temperature of liquid helium is 4.2 K. He measured the electrical resistivity of gold, platinum, and mercury and found that the electrical resistivity of mercury disappeared almost completely below 4.2 K. As shown in Fig. 1, the electrical resistivity of mercury is almost zero below 4.2 K. This state of a material in which the resistance is zero is called the *superconducting state*. The current flows without any attenuation in this state, and it has been estimated that the decay time of a current in a superconductor is about 100,000 years. The temperature below which a material loses its resistance is called the *superconducting transition* or *critical temperature* $T_c$.

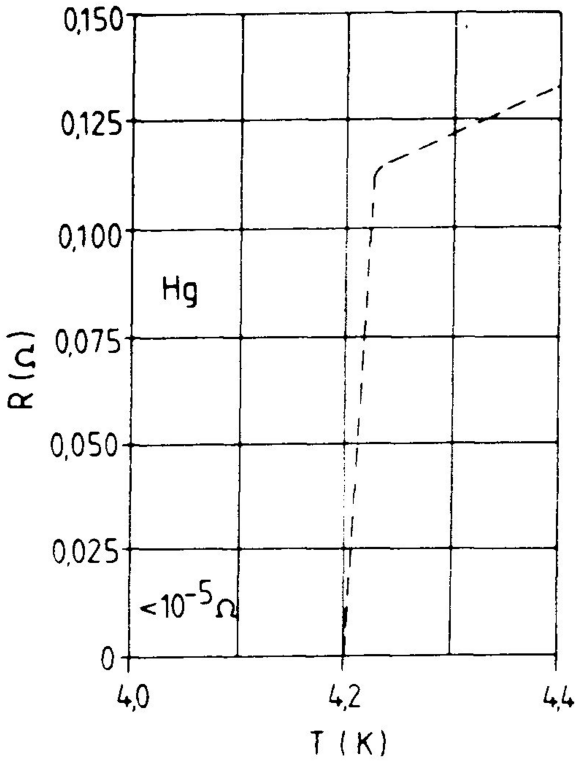



**FIGURE 1** Electrical resistance $R\,(\Omega)$ as a function of temperature for mercury metal (Hg).

### B. Effect of a Magnetic Field on Superconductivity and the Meissner Effect

Meissner discovered that a bulk superconducting material behaves like a perfect diamagnet with a zero magnetic induction in its interior. If a paramagnetic material is placed in a magnetic field, then the magnetic lines of force penetrate through the material. But when the same material is made superconducting by cooling to lower temperatures, then all the lines of force are expelled from the interior of this material. This is called the *Meissner effect*. Figure 2 shows a material in the normal and superconducting states in an externally applied magnetic field. When the strength of this externally applied magnetic field is increased slowly, a value is reached where the magnetic lines of force begin to penetrate the material and it becomes nonsuperconducting or normal. This particular value of the magnetic field above which the superconductivity is destroyed is called the *critical magnetic field* $H_c(T)$ and is also a function of the temperature of the material. A typical example of the variation of the $H_c(T)$ with temperature $T$ is shown in Fig. 3 for the metal mercury (Hg).

The variation of $H_c(T)$ with temperature can be expressed by the equation

$$H_c(T) = H_c(0)\left[1 - (T/T_c)^2\right],$$

which has a parabolic form. This expression can also be derived using thermodynamics.

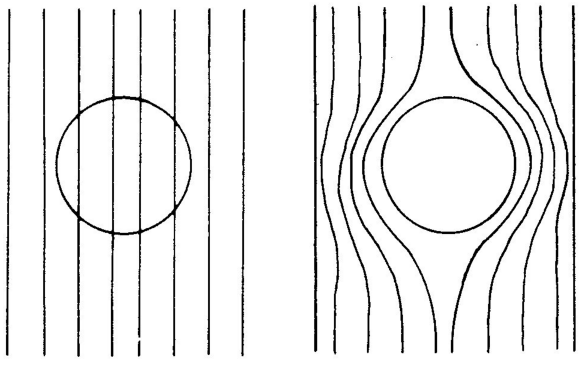


**FIGURE 2** Normal and superconducting states of a material in an external magnetic field.

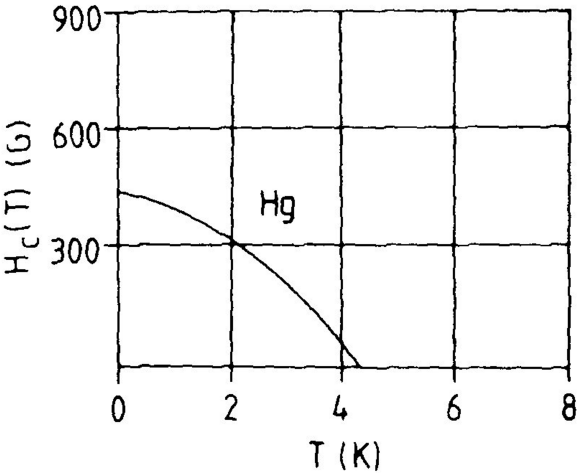



**FIGURE 3** Variation of the superconducting transition temperature of mercury in an extemally applied magnetic field.

### C. Type I and Type II Superconductors

Based on the Meissner effect, superconducting materials are classified as type I and type II. When the magnetic induction $4\pi M$ of a superconducting material in the form of a cylinder with its axis parallel to the apptied magnetic field is measured with an increasing magnetic field and if there is a sharp transition to the normal state above a certain value of the magnetic field $B_a$ as shown in Fig. 4, then this type of material is called a *type I superconductor*. This kind of behavior is shown in general by pure metals. On the contrary, the $4\pi M$ versus $B_a$ behavior of a *type II superconductor* is shown in Fig. 5. The magnetic flux

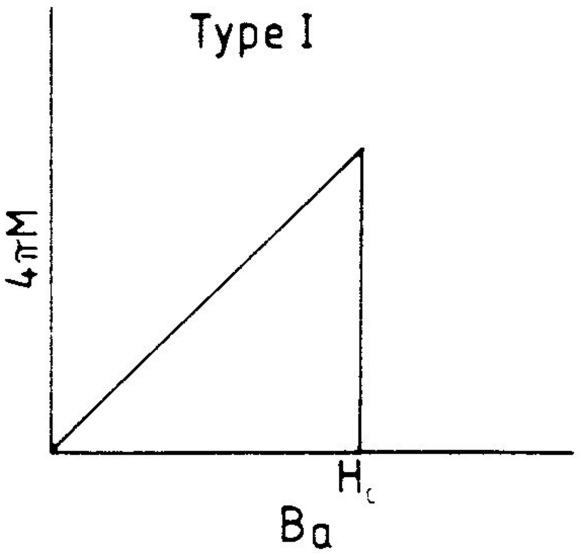



**FIGURE 4** Magnetization $4\pi M$ as a function of externally applied magnetic field $B_a$ of a type I superconductor.

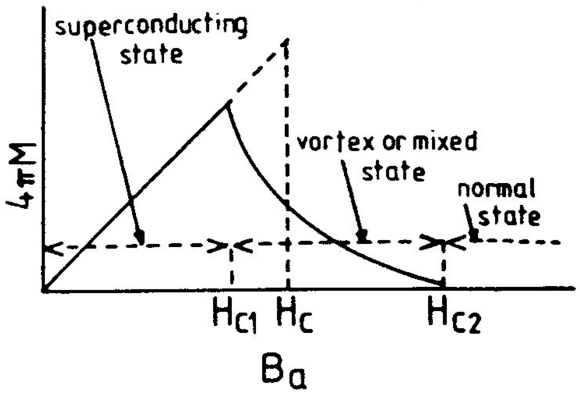



**FIGURE 5** Magnetization $4\pi M$ as a function of applied magnetic field $B_a$ of a type II superconductor.

penetrates the material slowly at a field value of $H_{c1}$ and continues up to $H_{c2}$, where the material is transformed to a normal state. The superconducting state between the field value $H_{c1}$ and the value $H_{c2}$ is called the *vortex* or *mixed state*. The $H_{c2}$ value can be 100 or more times greater than $H_c$. This type II superconducting behavior is shown in general by alloys and compounds that are called *dirty superconductors*.

Some of the superconducting alloys and compounds of special structures possess very high values of $H_{c2}$. For example, the $H_{c2}$ value of a compound of composition $Pb_1Mo_{5.1}S_6$ with a Cheveral phase structure is about 51 T. Very high magnetic fields can be generated by the solenoids of the wires made of superconducting materials of high $H_{c2}$. Commercial superconducting magnets capable of producing magnetic fields of more than 10 T are available and use wires of Nb–Ti and Nb–Sn alloys. The variation of $H_{c2}(T)$ with temperature of some high-$H_{c2}$ alloys is shown in Fig. 6. A type I superconductor can be transformed to a type II superconductor by alloying. A typical example is shown in Fig. 7. Here lead (Pb) is a type I superconductor, and when it is alloyed with indium (In), the alloys show type II behavior and the values of $H_{c1}$ and $H_{c2}$ are a function of the composition. The current flowing through the superconducting wire produces a magnetic field, and when the value of the current is increased slowly, a value is reached where the magnetic field becomes equal to the critical magnetic field. This value of current is called the *critical current*.

## II. SUPERCONDUCTING MATERIALS

### A. Elements, Compounds, and Alloys

The distribution of superconducting elements in the periodic system is shown in Fig. 8. Some elements do not

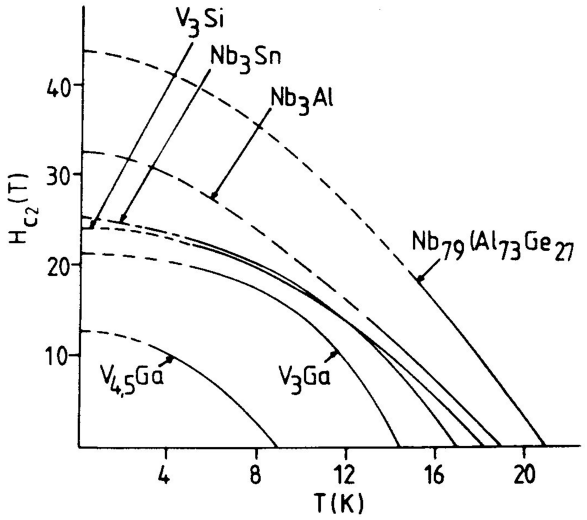



**FIGURE 6** Upper critical magnetic fields $H_{c2}$ as a function of temperature for some superconductors of $\beta$-W structure. [From Berlincourt, T. G., and Hake, R. H. (1963). *Phys. Rev.* **131,** 140; Fonner, S., McNiff, E. J., Jr., Matthias, B. T., Geballe, T. H., Willens, R. H., and Corenzwit, E. (1970). *Phys. Lett.* **3IA,** 349.]

become superconducting at all and others become so only under pressure. The superconducting transition temperatures along with their crystal structures and melting temperatures are listed in Table I. Niobium (Nb) metal has the highest superconducting transition temperature (9.2 K). The elements that become superconducting under pressure are listed in Table II.

The magnetic elements Mn, Fe, Co, and Ni do not become superconducting down to the lowest temperature available. Also, their presence in small amounts (of the order of parts per million) suppresses the superconducting transition temperature of other superconducting materials. The superconducting elements are classified into two groups. One group is the nontransition elements and consists of the elements Si, Ge, P, As, Sb, Bi, Se, and Tc, which are not superconducting under normal conditions but under pressure become superconducting. The other group consists of the transition elements and have unfilled $3d$, $4d$, and $5d$ shells. The crystal structure plays an important role in superconductivity. For example, as shown in Tables I and II, Bi is not superconducting, but different crystal modifications of it obtained by applying pressure exhibit superconductivity at temperatures ranging between 3.9 and 8.5 K.

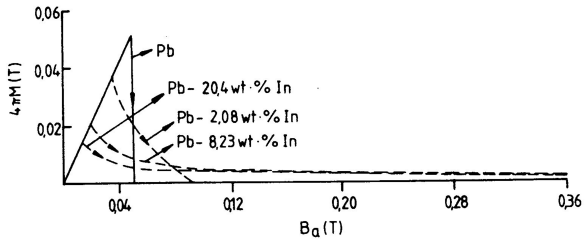



**FIGURE 7** Magnetization $4\pi M$ as a function of applied magnetic field $B_a$ and change of type I (Pb) to type II (Pb–In) alloy superconductors.

Multicomponent alloys and compounds of different crystal structures exhibit superconductivity. High-transition temperature superconductivity occurs with cubic structure, and the most favorable is the one with the $\beta$-W structure. Compounds and alloys with superconducting transition temperatures above 20 K form this structure. The $\beta$-W structure is cubic and is shown in Fig. 9. Each face of the cubic lattice is occupied by two atoms that form orthogonal linear atomic chains. The highest superconducting transition temperature is 23 K, and it is exhibited by a compound of composition $Nb_3Ge$ with the $\beta$-W structure. Here the Nb atoms form the linear atomic chains, and the Ge atoms occupy the center and corner sites of the cubic lattice.

The reason the materials of this particular structure show such high superconducting transition temperatures is explored by Labbé and Friedel. Their theoretical calculations based on tight binding approximation suggest that materials of this structure possess an unusually high electron density of states at the Fermi surface; this is also experimentally confirmed. In addition, the d-band of these materials is narrower and taller compared with that of the transition metals. These are the factors that cause the enhancement of the superconducting transition temperature. Some of the high superconducting transition temperature materials are listed in Table III.

There are other kinds of superconducting materials, including low-carrier density superconductors (semimetal or semiconductor), intercalated compounds, amorphous superconductors, and organic superconductors, and they are described separately as follows.

## B. Low-Carrier Density Superconductors

A class of materials that have carrier densities in the range of $10^{18}$ to $10^{21}$ are called *semimetals* because their carrier densities are between those of metallic conductors and semiconductors. Many of these materials are superconducting. For example, Fig. 10 shows that $La_3Se_4$, GeTc, SnTc, and $SrTi0_3$ are superconducting, and the superconducting transition temperature increases with increasing carrier density except in the case of $SrTi0_3$. For $SrTi0_3$, the superconducting transition temperature begins to decrease above a carrier density of $10^{20}$. This decrease is explained by the occurrence of the magnetic effect. All the above-mentioned materials investigated were in the form of single crystals.

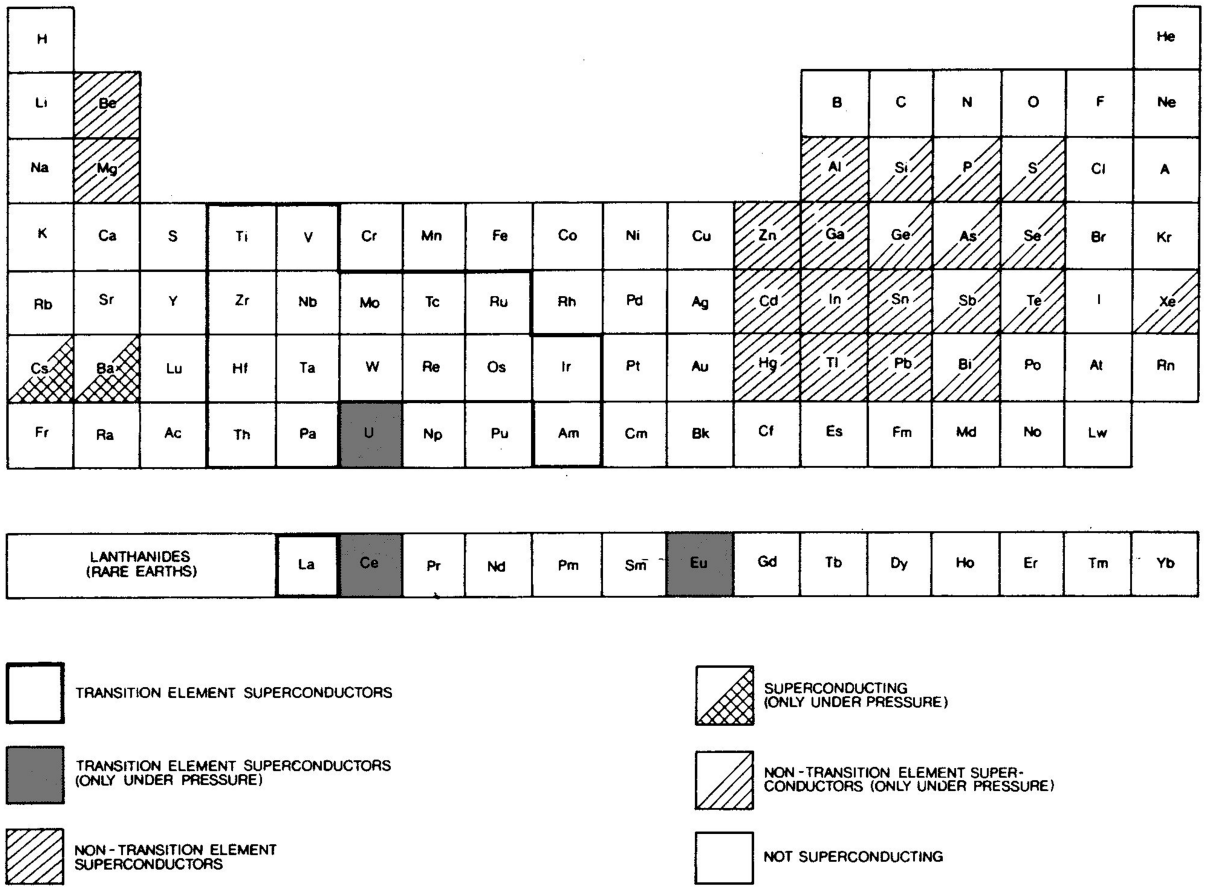



**FIGURE 8** Distribution of the superconducting elements in the periodic table. [From Khan, H. R. (1984). *Gold Bull.* **17**(3), 94.]

### C. Intercalated Compounds

A typical example of this class of materials is $TaS_2(C_5H_5N)_{1/2}$. This compound is formed when $TaS_2$ is intercalated with pyridine ($C_5H_5N$), and metallic layers about 6 Å thick are separated by pyridine layers of the same thickness. This intercalcated compound becomes superconducting at 3.5 K. A large number of transition metal chalcogenides exist that crystallize in the layered structures. These types of compounds show anisotropic superconducting properties parallel and perpendicular to the layer surface. The critical magnetic field is about 30 times higher in the direction parallel to the layer surface compared with the perpendicular direction, as shown in Fig. 11.

### D. Amorphous Superconductors

Unlike crystalline materials, amorphous or noncrystalline materials consist of atoms that do not form regular arrays and are randomly distributed. These amorphous materials can be obtained in the form of thin films by evaporation deposition on cold substrates. Amorphous materials in bulk can also be obtained by rapidly cooling an alloy melt. The amorphous materials obtained in this way are called *metallic glasses*. Materials of this class also exhibit superconductivity but are completely different from their crystalline counterparts. The superconducting transition temperature $T_c$ and the electron per atom ratio (e/a) of some amorphous nontransition metals and alloys are listed below. One sees that the $T_c$ values in the amorphous state are higher than those in the crystalline state.

| Alloy | $T_c$ (K) | e/a |
|---|---|---|
| Be | 9.95 | 2.0 |
| $Be_{90}A_{10}$ | 7.2 | 2.1 |
| Ga | 8.4 | 3.0 |
| $Pb_{90}Cu_{10}$ | 6.5 | 3.7 |
| Bi | 5 | 5 |

Amorphous films of transition metals and alloys have also been obtained, and the $T_c$ values are lower than those of the

**TABLE I Superconducting Transition Temperature $T_c$, Melting Temperature, and Crystal Structure of Elements**

| Element | $T_c$ (K) | Crystal structure[a] | Melting temperature (°C) |
|---|---|---|---|
| Al | 1.19 | f.c.c. | 660 |
| Be | 0.026 | Hex. | 1283 |
| Cd | *0.55* | Hex. | 321 |
| Ga | 1.09 | Orth. | 29.8 |
| | (6.5, 7.5) | | |
| Hg | *4.15* | Rhom. | −38.9 |
| | (3.95) | | |
| In | 3.40 | Tetr. | 156 |
| Ir | 0.14 | f.c.c. | 2450 |
| La | 4.8 | Hex. | 900 |
| | *(5.9)* | | |
| Mo | 0.92 | b.c.c. | 2620 |
| Nb | 9.2 | b.c.c. | 2500 |
| Os | 0.65 | Hex. | 2700 |
| Pa | 1.3 | — | — |
| Pb | 7.2 | f.c.c. | 327 |
| Re | 1.7 | Hex. | 3180 |
| Ru | *0.5* | Hex. | 2500 |
| Sn | 3.72 | Tetr. | 231.9 |
| | (5.3) | Tetr. | |
| Ta | 4.39 | b.c.c. | 3000 |
| Tc | 7.8 | Hex. | |
| Th | 1.37 | f.c.c. | 1695 |
| Ti | 0.39 | Hex. | 1670 |
| TI | 2.39 | Hex. | 303 |
| U($\alpha$) | 0.2 | Orth. | 1132 |
| V | 5.3 | b.c.c. | 1730 |
| W | 0.012 | b.c.c. | 3380 |
| Zn | 0.9 | Hex. | 419 |
| Zr | *0.55* | Hex. | 1855 |

[a] f.c.c., face-centered cubic; hex., hexagonal; orth., orthorhombic; rhom., rhombohedral; tetr., Tetrahedral; b.c.c., body-centered cubic.

same alloys in crystalline form. Metallic glass superconductors are classified into two main groups. One group consists of metal–metal compositions and the other of metal–metalloid compositions. The superconducting transition temperatures of some of the metallic glass superconductors are listed in Table IV. These metallic glass superconductors show some desirable properties. For example, they are ductile and possess a high strength, whereas their crystalline counterparts are brittle. The metallic glass superconductors also possess very high values of the critical magnetic field.

A practical superconductor capable of producing a magnetic field of about 10 T should have a critical current density of about $10^6$ A/cm$^2$. In general, the amorphous superconductors have low critical current densities. The current density can be increased by introducing some kind of inhomogeneities into the amorphous matrix. Some binary and ternary pseudoamorphous alloys of vanadium, hafnium, and zirconium metals possess reasonably high superconducting transition temperatures and very high values of critical magnetic fields and critical current densities. At the same time they also have good mechanical properties such as ductility and high tensile strength. These kinds of materials are promising future superconducting materials for generating magnetic fields above 10 T.

**TABLE II Superconducting Transition Temperature $T_c$ of Elements under Pressure**

| Element | $T_c$ (K) | Pressure (kbar) |
|---|---|---|
| As | 0.5 | 120 |
| Ba | 5.1 | 140 |
| | (1.8) | 55 |
| Bi II | 3.9 | 26 |
| Bi III | 7.2 | 27 |
| Bi V | 8.5 | 78 |
| Ce | 1.7 | 50 |
| Cs | 1.5 | 1000 |
| Ge | 5.4 | 110 |
| Lu | 0.1–0.7 | 130 |
| P | 4.6–6.1 | 100 |
| Sb | 3.6 | 85 |
| Se | 6.9 | 130 |
| Si | 6.7 | 120 |
| Te | 4.5 | 43 |
| Y | 1.5–2.7 | 120–160 |

## E. Organic Superconductors

In 1964 Little proposed that an organic polymer can also become a superconductor. His theory of superconductivity is based on a mechanism entirely different from that of the Bardeen *et al.* (1957) theory of metals and alloys. Little suggested that the electrons on the spine of the polymer chain are attracted to each other by an indirect process

**TABLE III Superconducting Transition Temperature $T_c$ of Some $\beta$-W Structure Compounds**

| Compound | $T_c$ (K) |
|---|---|
| $V_3Ga$ | 14.2–14.6 |
| $V_3Si$ | 17.1 |
| $Nb_3Au$ | 11.0–11.5 |
| $Nb_3Sn$ | 18.0 |
| $Nb_3Al_{0.8}Ge_{0.2}$ | 20.7 |

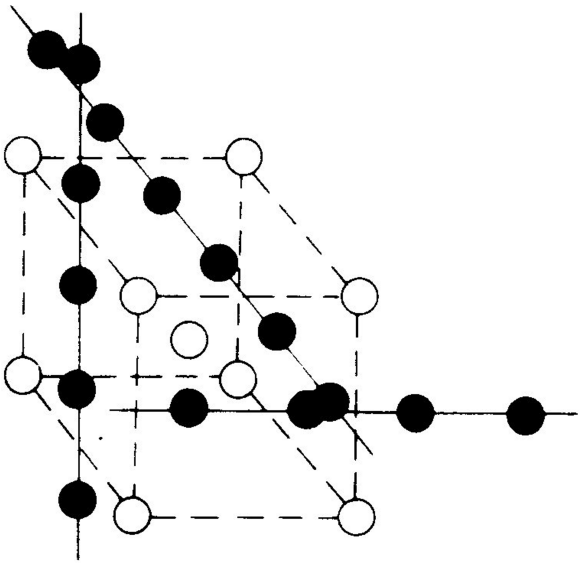


**FIGURE 9** The $\beta$-W lattice structure.

involving the fixed polar groups on the side branches of the polymer. He predicted a superconducting transition temperature of about 100 K using the molecular polarization mechanism. Experimentally, there are indications of a polymer becoming a superconductor. For example, a polymer of formula $(SN)_x$ shows superconductivity at 0.3 K and another organic compound called tetramethyltetrasulfofluoride (TMTSF) at 1 K under a pressure of 12 kbar.

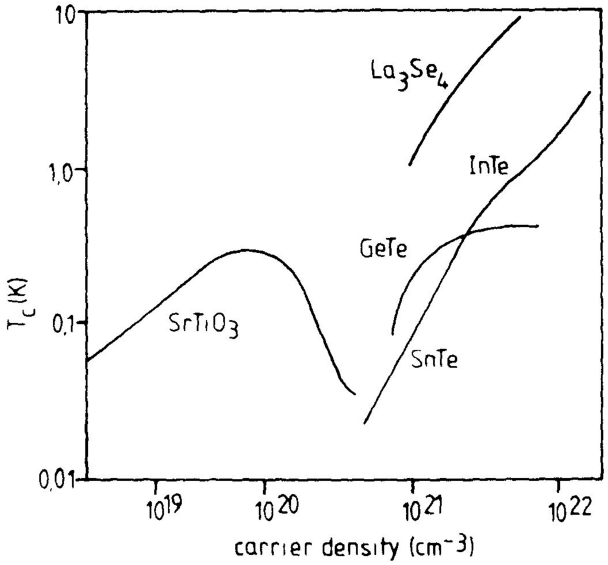



**FIGURE 10** Superconducting transition temperature $T_c$ as a function of carrier density. [From Hulm, J. K., Ashkin, M., Deis, D. W., and Jones, C. K. (1970). *Prog. Low Temp. Phys.* **VI,** 205.]

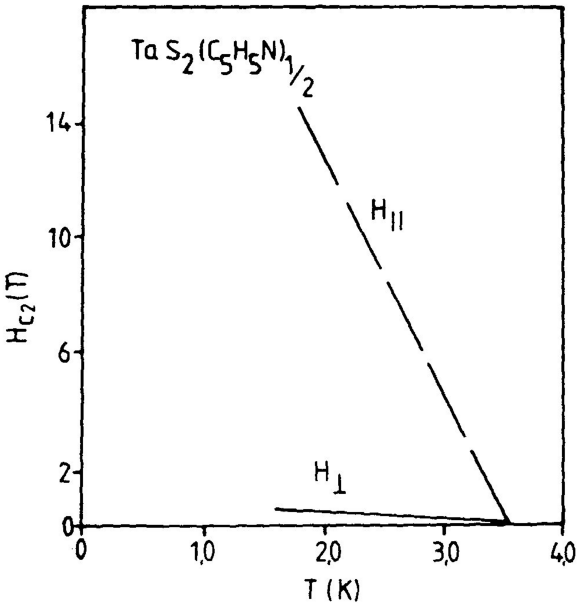



**FIGURE 11** Upper critical magnetic field $H_{c2}$ as a function of temperature parallel and perpendicular to the layered surface in $TaS_2(C_5H_5N)_{1/2}$. [From Gamble, F. R., *et al.* (1971). *Science* **174,** 493.]

## III. CORRELATION: $T_c$ WITH THE ELECTRONIC STRUCTURE OF A SOLID

Matthias proposed empirically that the superconducting transition temperature $T_c$ and the electron per atom ratio e/a of a solid are related. This Matthias empirical rule suggests that the maximum values of $T_c$ for transition metals occur at e/a values of 5 and 7, as shown in Fig. 12. In the case of solid solutions of transition metals, a slight shift of the first maximum to an e/a value of 4.5 occurs, as shown in Fig. 13. Amorphous materials consisting of transition metals show a different behavior. Amorphous materials based on the transition metals of an unfilled 4*d* shell show only one maximum at an e/a ratio of 6.4, whereas materials

**TABLE IV Superconducting Transition Temperature $T_c$ of Some Metallic Glasses**

| Metallic glass | $T_c$ (K) |
|---|---|
| $La_{80}Au_{20}$ | 3.5 |
| $La_{80}Ga_{20}$ | 3.8 |
| $Zr_{75}Rh_{25}$ | 4.55 |
| $Zr_{70}Pd_{30}$ | 2.4 |
| $Nb_{60}Rh_{40}$ | 4.8 |
| $(Mo_{0.8}Ru_{0.2})_{80}P_{20}$ | 7.31 |
| $(Mo_{0.6}Ru_{0.4})_{82}B_{18}$ | 6.05 |
| $(Mo_{0.8}Ru_{0.2})_{80}P_{10}B_{10}$ | 8.71 |

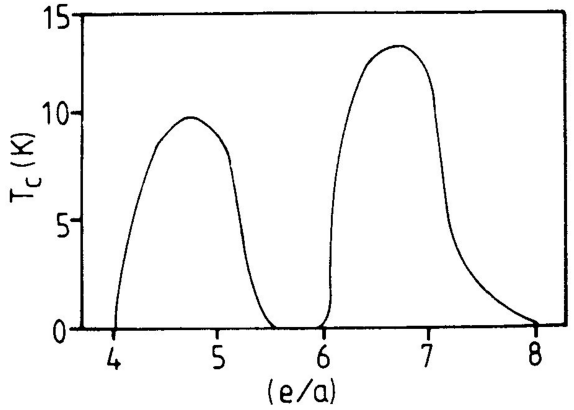



**FIGURE 12** Superconducting transition temperature $T_c$ as a function of electron per atom ratio e/a for the transition elements.

based on transition metals with an unfilled 5*d* shell have a maximum at an e/a of 7, as shown in Fig. 14. At these peak values of the superconducting transition temperatures, the values of the electron density of states are also maximum.

## IV. FLUX QUANTIZATION

In 1950 F. London suggested that the magnetic flux trapped by a superconducting ring is quantized and the flux quantum is given by $\Phi_0 = c\hbar 12e = 2 \times 10^{-7}$ G/cm$^2$, where $c$ is the velocity of light, $\hbar$ is Planck's constant, and $e$ is the electronic charge. The flux trapped in a superconductor is quantized and is equal to $n\Phi_0$. In the case of type I superconductors where the Meissner effect is perfect, the value of $n$ is zero. The flux quantization is observed only in the case of multiply connected geometries such as a superconducting ring. When the external magnetic field is removed, the magnetic flux trapped is equal to $n\Phi_0$. The flux quantization is expected to exist even in singly connected geometries in the case of type II superconductors because a mixed state exists in which the superconducting regions surround the lines of force and form a multiply connected system of filaments.

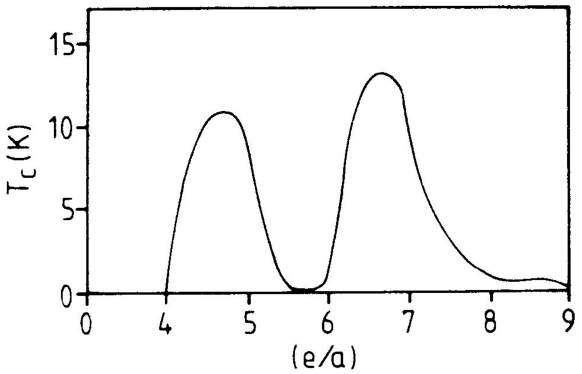



**FIGURE 13** Superconducting transition temperature $T_c$ as a function of electron per atom ratio e/a for solid solutions of the transition elements.

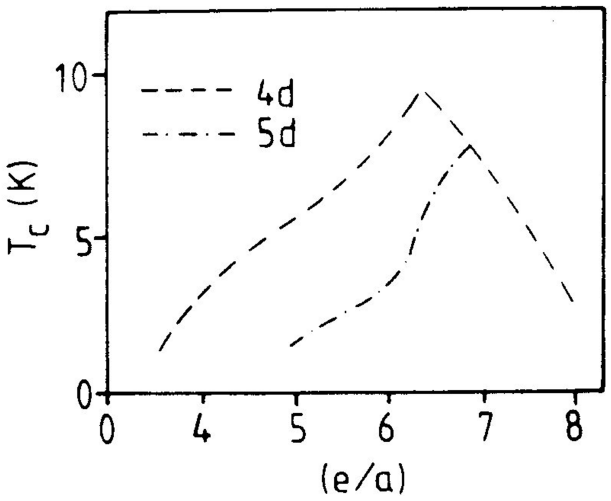



**FIGURE 14** Superconducting transition temperature $T_c$ as a function of the electron per atom ratio e/a for the amorphous 4*d* and 5*d* transition metals.

## V. LONDON EQUATION AND COHERENCE LENGTH

The magnetic field $\boldsymbol{H}$ and the supercurrent $\boldsymbol{J}_s$ in a superconductor are related by the equation

$$\nabla \times H = \frac{4\pi}{c} J_s,$$

where $c$ is the velocity of light. The free energy $F$ of a system is given as

$$F = F_s + E_{kin} + E_{mag},$$

where $|F_s|$ is the free energy of the electrons in the superconducting state. The kinetic energy $E_{kin}$ is

$$E_{kin} = \frac{1}{2} \int_{vol} m\mathbb{v}^2 n_s d\mathbf{r},$$

where $\mathbb{v}$ is the drift velocity of a parabolic band, $n_s$ is the number of superconducting electrons per unit volume, and $m$ is the effective mass of the electrons. The magnetic energy $E_{mag}$ in a magnetic field $\boldsymbol{H}$ is

$$E_{mag} = \int \frac{H^2}{8\pi} d\mathbf{r}.$$

The free energy $F$ can be written

$$F = F_s + \frac{1}{8\pi}\int \left[H^2 + \lambda_L^2 |\nabla \times \mathbb{H}|^2\right] d\tau,$$

where $\lambda_L$ is a constant,

$$\lambda_L = \left|\frac{mc^2}{4\pi n_s e^2}\right|^{1/2},$$

and is called the London penetration depth. London obtained an equation by minimizing the free energy with respect to the field distribution:

$$\mathbb{H} + \lambda_L^2 \nabla \times \nabla \times \mathbb{H} = 0.$$

When the current flows in the $y$ direction in a superconductor, then the magnetic field in the $z$ direction is

$$H_z = H(0)e^{-z/\lambda_L},$$

which shows that the magnetic field falls off exponentially inside a superconductor. The London penetration depth $\lambda(0)L$ at $T = 0K$ in terms of the Fermi velocity $\upsilon_F$ and the electron density of states $N(0)$ is

$$\lambda(0)L = \left[\frac{3C^2}{8\pi N(0)\upsilon_F^2 e^2}\right]^{1/2}.$$

## VI. COHERENCE LENGTH AND ENERGY GAP

Coherence length is a measure of the correlation distance of the superconducting electrons and is denoted $\xi_0$. The coherence length in terms of the Fermi velocity $\upsilon_F$, Boltzmann constant $K_B$, and superconducting transition temperature $T_c$ is

$$\xi_0 = \frac{\hbar \upsilon_F}{K_B T_c},$$

where $\hbar$ is Planck's constant.

One of the important features of superconductivity is the existence of a gap in the low-energy excitations, which is denoted $\varepsilon$. In most superconductors, an external energy $E$ must be supplied to create an electron-hole pair close to the Fermi surface.

This energy $E$ is

$$E \geq 2\varepsilon.$$

The coherence length $\xi_0$ and the energy gap are related by the equation

$$\xi = \frac{\hbar \upsilon_F}{\pi \varepsilon}.$$

Bardeen, Cooper, and Schrieffer (BCS), in 1957, related this energy gap to the formation of Cooper pairs. The formation of an energy gap in a superconductor is depicted in Fig. 15 for free electrons. According to the BCS theory,

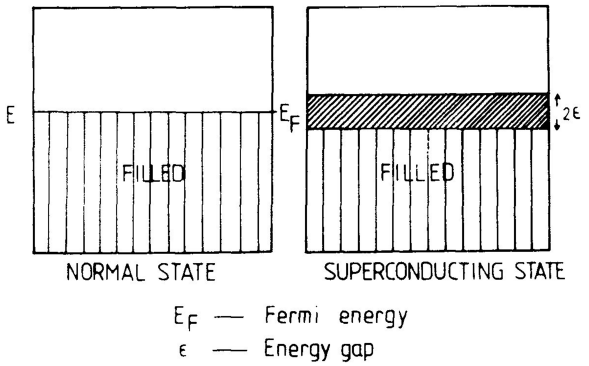



**FIGURE 15** Formation of an energy gap in the superconducting state.

electrons that have energies close to the Fermi energy form Cooper pairs easily. The paired states have a lower energy than the unpaired electrons that form them. The electron density of states $n(E)$-versus-energy $E$ curve of a normal metal as shown in Fig. 16 changes to the curve in Fig. 17 when the normal metal becomes a superconductor. From the BCS theory, a relationship among the energy gap at 0 K $\varepsilon(0)$, the Boltzmann constant $K_B$, and the superconducting transition temperature $T_c$ is

$$\varepsilon(0) = 1.76 K_B T_c.$$

The variation of the energy gap with temperature is shown in Fig. 18. The BCS theory is discussed later.

The energy gap of a superconductor can be measured experimentally as follows. The absorption coefficients of longitudinal ultrasonic waves in the normal state $\alpha_n$ and superconducting state $\alpha_s$ are related by

$$\frac{\alpha_s}{\alpha_n} = \frac{2}{1 + \exp(\varepsilon / K_B T)}.$$

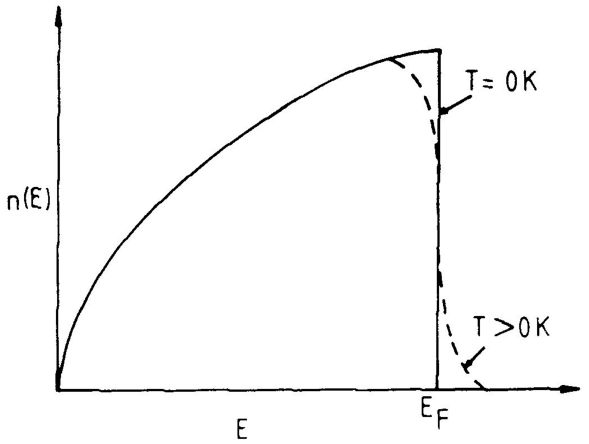



**FIGURE 16** Electron density of states $n(E)$ as a function of energy $E$ for a normal metal.

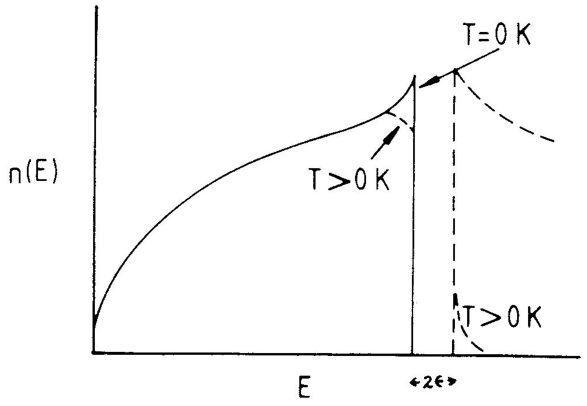



**FIGURE 17** Electron density of states $n(E)$ as a function of energy $E$ for an ideal superconductor.

The experimental determination of $\alpha_n$ and $\alpha_s$ at a particular temperature $T$ enables one to determine the value of $\varepsilon$.

Absorption of the electromagnetic waves in the far-infrared ($\lambda \approx 1$ mm) region occurs for photons of energy $h\nu = 2\varepsilon$. Determination of the frequency of absorption directly gives an energy gap as shown in Fig. 19. The specific heat is proportional to $\exp(-\varepsilon/K_B T)$, and the value of $\varepsilon$ can also be obtained from the specific heat measurements.

Tunneling experiments also give the value of $\varepsilon$ directly. An experimental arrangement for the determination of $\varepsilon$ is shown in Fig. 20. A superconductor with an energy gap $\varepsilon$ is depicted as A and is separated from a normal conductor C through a thin insulating layer B. The shaded areas represent the occupied states. The Fermi level is at the center of the energy gap in the case of the superconductor. When a potential difference is applied across the insulating layer, electrons tunnel through the barrier B from C to A. This potential difference that causes the onset of the tunneling current is a direct measure of the energy gap $\varepsilon$. When both of the materials across the insulating layer are superconductors, the energy gaps of these two superconductors can be measured simultaneously from the potential difference-versus-current curve. A typical arrangement for these kinds of measurements is shown in Figs. 21 and 22.

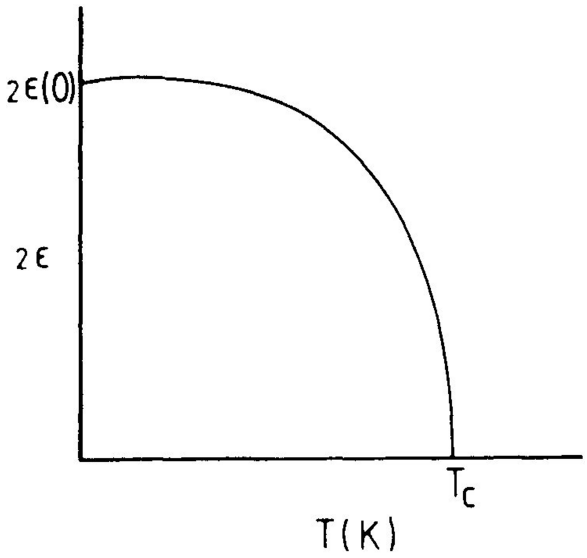



**FIGURE 18** Energy gap $2\varepsilon$ as a function of temperature $T$ for a superconductor.

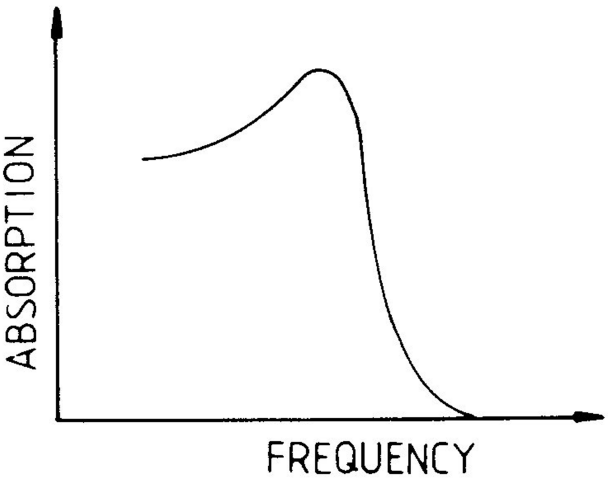



**FIGURE 19** Absorption of electromagnetic waves as a function of frequency in a superconductor.

## VII. THERMODYNAMICS OF SUPERCONDUCTIVITY

From the basic thermodynamic considerations, we derive relations among the critical magnetic field of a

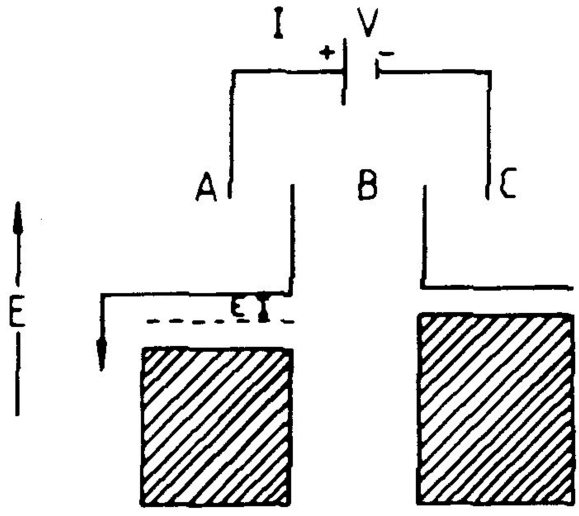



**FIGURE 20** Tunneling of electrons through a thin insulating layer, B, between two superconductors, A and C.

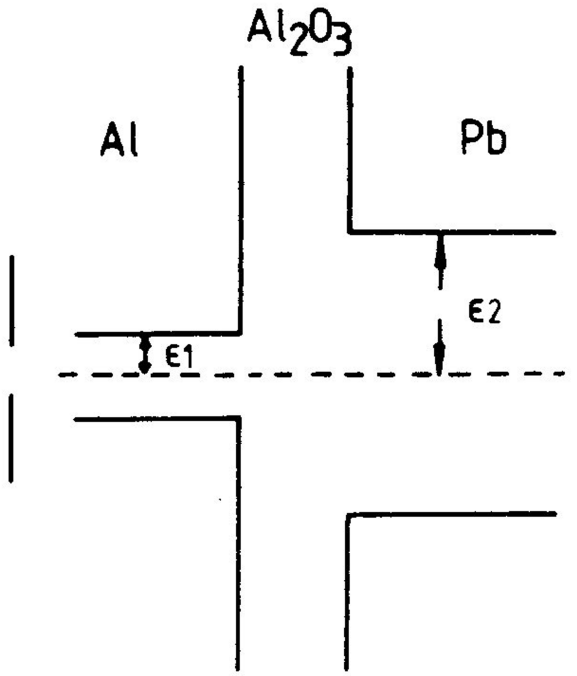



**FIGURE 21** Tunneling of electrons through a thin insulating layer of $Al_2O_3$ between the two superconductors lead (Pb) and aluminum (Al).

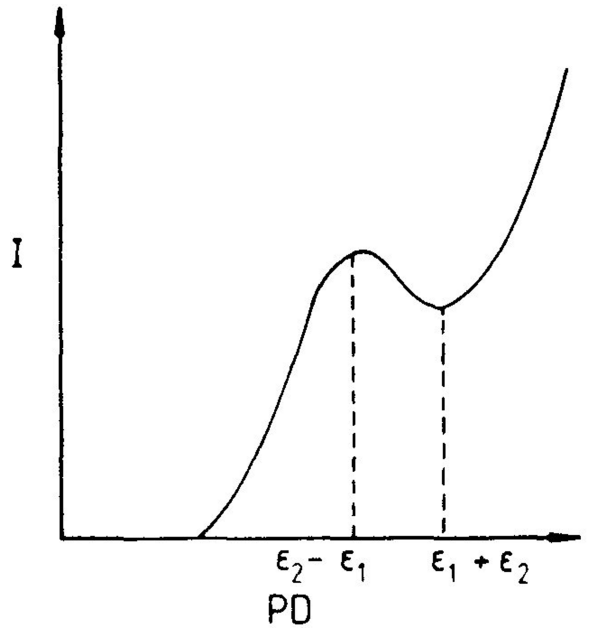



**FIGURE 22** Current $I$-versus-potential difference $PD$ plot of a tunnel junction consisting of two superconductors with energy gaps $\varepsilon_1$ and $\varepsilon_2$.

superconductor and the specific heats in the normal and superconducting states as well as the critical magnetic field and superconducting transition temperature $T_c$. Let us consider a material in the normal state with a negligible magnetization. Its Gibbs energy function $G_n$ in the normal state is

$$G_n = U - TS + PV,$$

where $U$ is the internal energy, $T$ the temperature, $S$ the entropy, $P$ the pressure, and $V$ the volume. For a superconductor in the presence of an external magnetic field, the magnetization is not negligible and the magnetic induction $\boldsymbol{B}$ is

$$B = H + 4\pi I,$$

where $H$ is the applied magnetic field and $I$ is the intensity of magnetization. In the case of a sharp superconducting transition for a long thin rod parallel to the field $B = 0$,

$$I = -H/4\pi.$$

The Gibbs function $G_s$ in the superconducting state per unit volume is

$$\begin{aligned} G_s(H) &= U - TS + PV - \int_0^H I\,dH \\ &= G_s + \int_0^H H\,dH/4\pi \\ &= G_s + H^2/8\pi. \end{aligned}$$

Assuming a negligible volume change at the transition,

$$G_n = G_s(H_c),$$

where $H_c$ is the critical magnetic field, and

$$G_n - G_s = H_c^2/8\pi.$$

Because

$$G = U + PV - TS \tag{1}$$

and

$$dG = dU + PdV + VdP - TdS - SdT,$$

using the first law of thermodynamics,

$$dQ = dU + PdV = TdS,$$

one obtains

$$dG = VdP - SdT,$$

which gives

$$S = -(\partial G/\partial T)P. \tag{2}$$

Combining Eqs. (1) and (2),

$$S_n - S_s = -H_c/4\pi \cdot \partial H_c/\partial T. \tag{3}$$

The difference of the normal-state and superconducting-state entropies is expressed in terms of the critical field $H_c$ and its slope $\partial H_c/\partial T$. The specific heat per unit volume is

$$C = dQ/dT = TdS/dT,$$

so that

$$C_n - C_s = T\partial(S_n - S_s)/\partial T,$$

and Eq. (3) reduces to

$$C_s - C_n = TH_c/4\pi \cdot \partial 2H/\partial T^2 + T/4\pi \cdot (\partial H_c/\partial T)^2.$$

At the transition $T = T_c$ and $H_c = 0$,

$$C_s - C_n = T_c/4\pi \cdot (\partial H_c/\partial T)^2. \tag{4}$$

where $C_s$ and $C_n$ are the specific heats in the superconducting and normal states. The specific heat $C_s$ follows the relationship

$$C_s = BT^3.$$

where $B$ is a constant, whereas $C_n$ is given as

$$C_n = AT^3 + \gamma T.$$

Because

$$dQ = dU + PdV = TdS,$$

$$S = \int dQ/T = \int C\,dT/T.$$

Therefore,

$$S_s = \int BT^2\,dT = (BT^3/3)$$

and

$$S_n = \int (AT^2 + \gamma)\,dT = (AT^3/3) + \gamma T,$$

and the difference is

$$S_n - S_s = \frac{1}{3}(A - B)T^3 + \gamma T.$$

At the transition temperature $T_c$ (in zero field)

$$S_n = S_s,$$

thus

$$\frac{1}{3}(B - A)T_c^2 = \gamma$$

and

$$S_n - S_s = \gamma\left(T - T^3/T_c^2\right). \tag{5}$$

From Eq. (4) it follows that

$$-H_c/4\pi \cdot \partial H_c/\partial T = \gamma\left(T - T^3/T_c^2\right)$$

or

$$\partial/\partial T/\left(H_c^2\right) = 8\pi\gamma\left(T^3/T_c^2 - T\right).$$

Since $H_c = H_o$ at $T = 0$ K and integrating,

$$H_c^2 = 8\pi\gamma\left(T^4/4T_c^2 - T^2/2\right) + H_0^2 \tag{6}$$

when $T = T_c$ and $H_c^2 = 0$, therefore

$$8\pi\gamma\left(T_c^2/2 - T_c^2/4\right) = H_0^2$$

or

$$\gamma = H_0^2/2\pi T_c^2. \tag{7}$$

Combining Eqs. (6) and (7),

$$H_c = H_0\left(1 - (T/T_c)^2\right).$$

This equation relates the critical magnetic field $\boldsymbol{H}_c$ of a superconductor with the critical temperature $T_c$ and has a parabolic form. This conforms to the experimental observation shown in Fig. 3 for a type I superconductor for which the relationship between $\boldsymbol{H}_c$ and $T_c$ was

$$H_c \cong H_0\left(1 - (T/T_c)^2\right).$$

## VIII. MAGNETIC SUPERCONDUCTORS

Ferromagnetism and superconductivity have been considered to be mutually exclusive phenomena. lt was assumed that the large internal magnetic field present in a ferromagnetic material would not allow it to become a superconductor. This is true, and so far none of the magnetic elements (for example, chromium, manganese, iron, cobalt, and nickel) have exhibited superconductivity. A search was made to find a material that exhibits superconductivity and ferromagnetism at different temperatures. Among the rare-earth elements, lanthanum is superconducting at 6 K. The other rare-earth elements are either paramagnetic or ferromagnetic, with magnetic moments that are due to $4f$ electrons. Matthias and co-workers dissolved gadolinium metal in lanthanum and measured the superconducting transition temperatures as a function of dissolved gadolinium. Figure 23 shows a plot of the superconducting transition temperature as a function of gadolinium dissolved

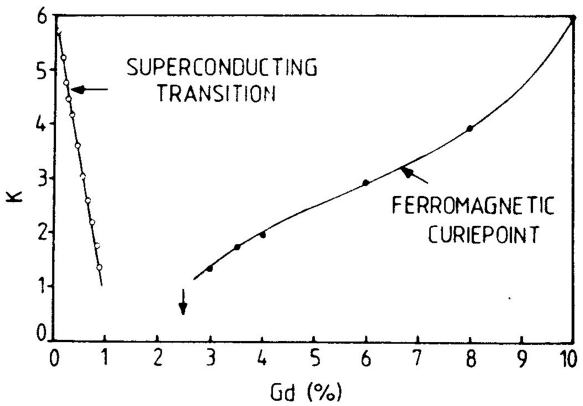



**FIGURE 23** Superconducting transition temperature and ferromagnetic Curie point as a function of gadolinium (Gd) in La–Gd alloys. [From Matthias, B. T., and Suhl, H. (1960). *Phys. Rev. Lett.* **4,** 51.]

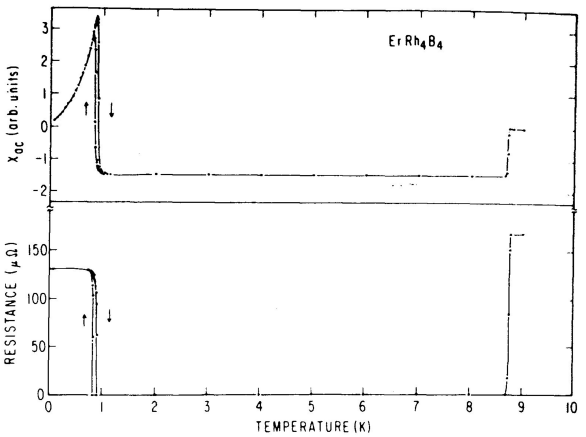



**FIGURE 24** Transitions to superconducting and ferromagnetic states in $ErRh_4B_4-$. [From Fertig, W. A., Johnston, D. C., Maple, M. B., and Matthias, B. T. (1977). *Phys. Rev. Lett.* **38,** 987.]

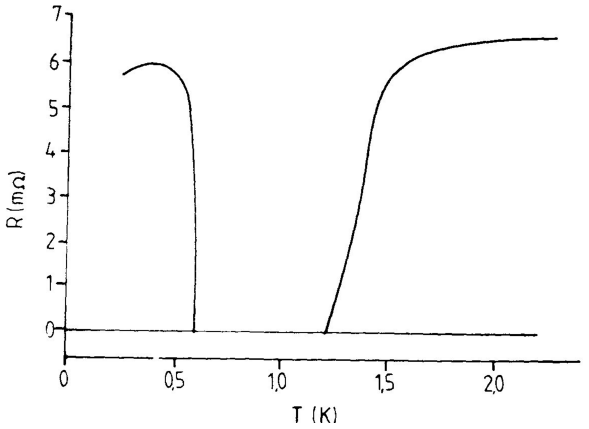



**FIGURE 25** Transitions to the superconducting and ferromagnetic states in $HoMo_6S_8-$. [From Ishikawa, M., and Fischer, O. (1977). *Solid State Commun.* **23,** 37.]

in lanthanum. The depression of $T_c$ is a linear function of gadolinium dissolved to approximately 1% gadolinium. More than 2.5% gadolinium in lanthanum makes it a ferromagnetic material.

These data suggest that an exchange interaction over conduction electrons leading to ferromagnetism is easy to bring about in an element that is itself a superconductor. This points to a possibility of a magnetic superconductor in which the phenomena of superconductivity and ferromagnetism overlap. A material of composition $ErRh_4B_4$ has been discovered that becomes superconducting at 8.7 K and shows ferromagnetic ordering at 0.93 K. The measurements of resistance and magnetic susceptibility as a function of temperature for $ErRh_4B_4$ are shown in Fig. 24. The anomalies at the temperatures of 0.93 and 8.7 K in the resistance and niagnetic susceptibility curve correspond to the ferromagnetic ordering and superconducting transition temperatures.

Another Chevrel phase compound of composition $HoMo_6S_8$ exhibits superconducting and ferromagnetic transitions at 2.15 and 0.6 K, as shown in Fig. 25. The discovery of the coexistence of ferromagnetism and superconductivity in these ternary rare-earth molybdenum chalcogenides and rare-earth rhodium borides has opened a new field of investigation on the interactions responsible for ferromagnetism and superconductivity.

## IX. TUNNELING AND THE JOSEPHSON EFFECT

In 1962 Josephson predicted theoretically that if two superconductors were separated by a thin (~10-Å) insulating film, then the superconducting electron pairs would tunnel through this junction. The tunneling current would flow without any voltage across the junction between the two superconductors. When the dc current is exceeded, a dc voltage would develop across the junction. This voltage is $2eV1\eta$, where $e$ is the charge on the electron, $\eta$ Planck's constant, and $V$ the frequency of the photon radiated by the electron pair while tunneling across the junction. The maximum zero-voltage current $\boldsymbol{J}$ across the junction is

$$J = J_0 \sin(\delta_0 + 2e/c\eta) \int A\, ds,$$

where $\delta_0$ is a constant. This shows that the current is a periodic function of the flux passing through the junction at right angles to the current and that the period is equal to the quantum of flux $\eta cl2e$. The Josephson prediction was experimentally proved by Anderson and Rowell, who showed that a zero-voltage current would flow through a thin insulating layer between the two superconductors. The maximum value of this current oscillates with the external magnetic field, as shown in Fig. 26.

## X. THEORY OF SUPERCONDUCTIVITY

In 1950 Ginzburg and Landau proposed a model for superconductors in which an order process in a superconductor is described in terms of an order parameter $\psi$, where $\psi$ represents the fraction of conduction electrons in the superconducting momentum state. This model contained expressions for the momentum and kinetic energy of superelectrons and described the magnetic behavior of superconductors very well, but a basic interaction mechanism was still lacking.

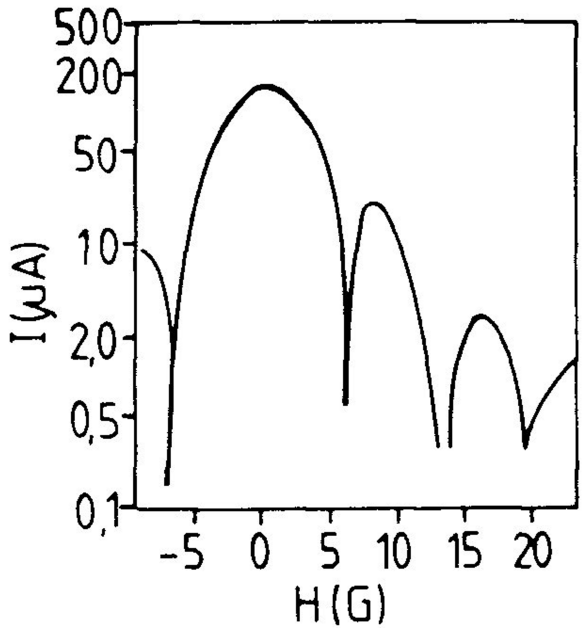



**FIGURE 26** Tunneling current as a function of applied magnetic field $H$ on a tunnel junction consisting of two superconductors. [From Rowell, J. M. (1963). *Phys. Rev. Lett.* **11,** 200.]

In 1957 Bardeen *et al*. proposed a theory of superconductivity in which they expressed the superconducting transition temperature in terms of an interaction between the electrons and the lattice vibrations of a solid. The quanta of lattice vibrations in a solid are called *phonons*. According to this theory, when the temperature of a solid is lowered, an interaction between the electrons and the phonons causes an attractive force between the conduction electron pairs called *Cooper pairs*. These Cooper pairs are paired states with equal and opposite momentum at zero supercurrent. When a current is applied to a superconductor, all the electron pairs have the same momentum directed parallel to the electric field. Due to this coherent motion, the pairs do not collide with the lattice and there is no electrical resistance. The expression for the superconducting transition temperature $T_c$ is

$$K_B T_c = 1.14\eta\omega_c \exp(-1/(N(0)V)). \tag{8}$$

This equation is valid for $N(0) \ll 1$. Here $N(O)$ is the electron density of states, $V$ the net attractive potential between the electrons, and $\omega_c$ the principal phonon frequency. The temperature $T_c$ is extremely sensitive to small changes in $V$. This theory successfully explains most of the physical property changes associated with the superconducting transition. lt is rather difficult to calculate the superconducting transition temperature itself using this theory. lt should be mentioned that in all critical phenomena, the critical temperatures are most difficult to calculate. For example, it is not easy to calculate the freezing or boiling point of water.

lt has been observed experimentally that the superconducting transition temperature of an element varies with the isotope mass. For example, for the isotopes of mercury, $T_c$ varies between 4.185 and 4.146 K, whereas the average atomic mass $M$ varies between 199.5 and 203.4. In Eq. (8) $\omega_c$ is proportional to $1/\sqrt{M}$, where $M$ is the atomic mass and $V$ is independent of $M$ in the BCS equation.

Therefore $T_c$ should be proportional to $1/\sqrt{M}$, and this dependence has been observed in the case of several elements such as tin, mercury, and indium. The term $N(0)V$ occurring in the BCS theory can be further expressed in terms of two parameters: the electron–phonon interaction parameter, $\lambda$; and $\mu^*$, which describes the normalized coulomb repulsion of electrons. This modification of the BCS theory was suggested by McMillan for the strong coupling superconductors $\lambda \gg \mu^*$, where the original BCS theory is not valid. The modified expression for $T_c$ is given by the expression

$$T_c = \Theta_D/1.45 \exp\left[-\frac{1.04(1+\lambda)}{\lambda - \mu^*(1+0.62\lambda)}\right], \tag{9}$$

where $\Theta_D$ is the Debye temperature. The electron–phonon interaction parameter $\lambda$ is

$$\lambda = \eta / M\omega_c^2,$$

where $\eta$ is a constant for a given structure class. Maximization of $T_c$ in Eq. (9) with respect to $\omega_c$ gives

$$T_c(\max) = (\eta/2M)^{1/2} \exp(-3/2). \tag{10}$$

Substituting suitable parameters into Eq. (10), a maximum $T_c$ value of 35 K is calculated. lt must be mentioned that, at present, a maximum $T_c$ value of 23 K exists for $Nb_3Ge$.

## XI. APPLICATIONS OF SUPERCONDUCTIVITY

Since its discovery, superconductivity has found many applications in technology. Because the electrical resistance in a superconductor is almost zero, large and homogeneous fields can be generated simply by winding the coils of the wires made from the high critical transition temperature and critical magnetic field superconducting materials. In the last decade much effort has gone into the development of these superconducting materials. Magnetic fields below and above 10 T can be produced using superconducting wires made from Nb–Ti and Nb–Sn alloys. These superconducting magnets have found a broad range of

applications where normal iron magnets are inadequate. These superconducting magnets can be used as exciter magnets for homopolar generators or rotors in large alternators, and a large gain in efficiency and power density is obtained.

Fusion reactors will employ superconducting magnets to confine plasma in which deuterium and tritium will be fused to produce energy. Soon the six D-shaped superconducting coils of dimensions $2.5 \times 3.5$ m manufactured in the United States and in European countries will be tested to produce a magnetic field of 8 T, which will be used to confine deuterium–tritium plasma to produce fusion energy. Superconducting magnets have found use in particle beam accelerators for high-energy particle physics research. Another application of superconducting magnets is in nuclear magnetic resonance tomography, which requires a homogeneous magnetic field; superconducting magnets are ideal for this purpose.

Another application of superconductors is as magnetic sensors. As mentioned earlier, a Josephson junction is an extremely nonlinear detector which, when connected to a loop of a superconducting wire, forms a superconducting quantum interference device (SQUID). These SQUIDs are extremely sensitive to small changes in magnetic fields. Based on the Josephson junction, high-frequency electromagnetic radiation detectors for frequencies in the range of microwaves have been developed. Josephson junction technology also finds applications in digital signal and data processing due to the high-speed and low-power dissipation compared to semiconductor technology. The Josephson junction can replace semiconductor technology where high speed, ultrahigh performance, reliability, lower power, and compactness are required. Other applications of superconductors include lossless transport of electrical energy and generation of magnetic fields for levitation and propulsion for high-speed ground transportation.

## XII. RECENT DEVELOPMENTS: HIGH-TRANSITION TEMPERATURE SUPERCONDUCTIVITY

Until April 1986, the maximum superconducting transition temperature measured in $Nb_3Ge$ was $\sim$23 K. This limited superconducting transition temperature allowed large- and small-scale applications of superconductors only with the use of liquid helium. Decades of experimental and theoretical research work showed that the phenomenon of superconductivity could be explained by the attraction of electrons caused by electron–phonon interaction (BCS theory). lt was suggested that, based on this mechanism, a superconducting transition temperature of $\sim$35 K could be achieved. These conclusions were based on research on about 24,000 superconducting inorganic phases. In 1986, J. G. Bednorz and K. A. Müller published a paper in *Zeitschrift für Physik* on the possibility of a superconducting transition temperature as high as 30 K in a mixture of lanthanum and barium–copper oxide $(La_{2-x}Ba_x\text{–}CuO_x)(x \sim 0.15)$ of tetragonal $K_2NiF_4$ structure. This discovery broke all previous records and received world attention, and the two authors received the 1987 Nobel Prize.

In a short time, superconducting oxides in the ranges 30–40, 90–100, and above 100 K were discovered. At present, the highest achievable superconducting transition temperature under normal conditions is about 133 K. The superconducting oxides of $\sim$90 K superconducting transition temperature are rare-earth barium–copper oxides of orthorhombic structure. The oxygen content in these oxides plays a major role in the superconductivity. When the oxygen content is reduced, the oxides transform to a tetragonal structure and become semiconducting. Superconducting transition temperatures above 100 K are observed in thalium-, bismuth-, strontium-, calcium-, and copper-based oxides.

All these materials are ceramics and brittle, not ductile like metals or alloys, and the electronic properties are highly anisotropic. The critical current density is high in one direction and low in the other, perpendicular, direction. The epitaxial thin films of some of these oxides show critical current densities of $10^6$ A/cm$^2$ at liquid nitrogen temperature. The critical current density of polycrystalline materials in the polycrystalline state is very low and not suitable for technical applications. The coherence length in these ceramic superconductors is quite small and is comparable to the lattice constants. These materials show rather strong electron–electron interactions, for example, as reported by Steiner *et al*. (1988). Therefore there is increasing evidence that the electron pairing in the superconducting state is of a pure electronic nature as suggested by Anderson (1987), and not caused by electron–phonon interaction.

The mechanical properties of these ceramic superconductors as well as their superconducting properties may be improved by the addition of silver metal as reported by Khan *et al*. At present, a worldwide effort is ongoing to improve the mechanical properties and to increase the critical current densities of these materials for large-scale applications. Once the mechanical properties of ceramic superconductors are improved and the critical current density is increased to a practical value, it is expected that these superconducting materials will revolutionize various technologies by working at liquid nitrogen, rather than liquid helium, temperatures.

## SEE ALSO THE FOLLOWING ARTICLES

CRYOGENIC PROCESS ENGINEERING • CRYOGENICS • FERROMAGNETISM • RARE EARTH ELEMENTS AND MATERIALS • SUPERCONDUCTING CABLES • SUPERCONDUCTING DEVICES • SUPERCONDUCTIVITY MECHANISMS • SUPERCONDUCTORS, HIGH TEMPERATURE • THERMOELECTRICITY

## BIBLIOGRAPHY


Anderson, P. W. (1987). *Science* **235,** 1196.

Bardeen, J., Cooper, L. N., and Schrieffer, J. R. (1957). *Phys. Rev.* **108,** 1175.

Barone, A., and Paterno, G. (1982). "Physics and Applications of the Josephson Effect," Wiley, New York.

Bednorz, J. G., and Müller, K. A. (1986). *Z. Phys*. **B64,** 189.

Buckel, W. (1972). "Supraleitung," Physik Verlag GmbH, Weinheim, Germany.

Khan, H. R. (1984). *Gold Bull*. **17**(3), 94.

Khan, H. R. (1998). *J. Superconduct* **11,** 1.

Khan, H. R., and Loebich, O., (1995). *Physica C.* **254,** 15.

Khan, H. R., and Raub, C. J. (1985). *Annu. Rev. Mater. Sci*. **15,** 21.

Kittel, C. (ed.) (1976). "Introduction to Solid State Physics," 5th ed. Wiley, New York.

Newhouse, V. L. (1964). "Applied Superconductivity," Wiley, New York.

Putlin, S. N., and Antipov, E. V. (1993). *Nature* **362,** 226.

Roberts, B. W. (1976). *J. Phys. Chem. Data* **5**(3), 581–821.

Saint-James, D., Sarma, G., and Thomas, E. J. (1969). "Type II Superconductivity," Pergamon, Oxford.

Steiner, P., *et al*. (1988). *Z. Phys*. **B69,** 449.



# Superconductivity Mechanisms

**Jozef Spałek**
*Jagiellonian University and Purdue University*

I. Introduction
II. The Bardeen–Cooper–Schrieffer (BCS) Theory: A Brief Summary
III. Normal and Magnetic States of Correlated Electrons
IV. Novel Mechanisms of Electron Pairing
V. Conclusions

## GLOSSARY

**Almost-localized Fermi liquid** A metallic system which, under a relatively small change of an external parameter such as temperature, pressure, or composition, undergoes a transition to the Mott insulating state. In such a metal electrons have a large effective mass. At low temperatures the system may order antiferromagnetically or undergo a transition to the superconducting state. Both nonstoichiometric oxides (such as $V_2O_{3-y}$) and heavy-fermion systems (e.g., $UPt_3$) are regarded as almost-localized Fermi liquids.

**Bardeen–Cooper–Schrieffer (BCS) theory** Theory describing properties of superconductors in terms of the concept of pairing of electrons with opposite spins and momenta. The pairing of electrons is mediated by a dynamic positive-ion lattice deformation, which produces resultant attractive interaction overcoming their mutual coulomb repulsion. At a critical temperature the electron system undergoes a phase transition to a condensed state of pairs which is characterized by a zero dc electrical resistance and a strong diamagnetism (Meissner–Ochsenfeld effect). The condensed state is destroyed by the application of an applied magnetic field (the critical fields $H_c$ and $H_{c2}$ for superconductors of the first and second kinds, respectively.

**Correlated electrons** Electrons with their kinetic (or band) energy comparable to or lower than the magnitude $U$ of electron–electron repulsion. This situation is described by the condition $U \gtrsim W$, where $W$ is the width of a starting (bare) energy band. Strictly speaking, we distinguish between the limits of almost–localized Fermi liquids, for which $U \lesssim W$, and the limits of strongly correlated electrons (Tomonaga–Luttinger or spin liquids), for which $U \gg W$. The term "correlated electrons" means that the motion of a single electron is correlated with that of others in the system (for example, its effective mass depends on the two-particle correlation function).

**Exchange interaction** Part of the coulomb interaction between electrons which depends on the resultant spin state of their partially filled $d$ or $f$ shells. If the spin–singlet configuration is favored in the ground state, then the interaction is called antiferromagnetic. The

exchange interaction provides a mechanism of magnetic ordering in Mott insulators; it may also correlate electrons into singlet or triplet pairs in the metallic state, particularly when the pair-exchange coupling $J$ of an electron pair is comparable to the kinetic energy of each of its constituents, as is the case for strongly correlated electrons. The superexchange (or kinetic exchange) is induced by a strong electron correlation.

**Fermi liquid** Term describing the state of interacting electrons in a metal. Equilibrium properties of such systems are modeled by a gas of electrons with renormalized characteristics such as the effective mass (they are called *quasiparticles*). The properties at low temperatures are determined mainly by electrons near the Fermi surface. The electron–electron interactions lead to specific contributions to the transport properties of such a system producing, e.g., sound-wave and plasmon excitations.

**High-temperature superconductors** Oxide materials of the type $La_{2-x}Sr_xCuO_4$ or $YBa_2Cu_3O_{7-x}$, which have a layer structure, with the principal role of electrons confined to the $CuO_2$ planes. The term "high-temperature superconductors" (HTS) was coined to distinguish these and other oxide superconductors with a critical temperature $T_c \gtrsim 20$ K from "classical" superconductors, which comprise metals and intermetallic compounds such as $Nb_3Ti$ with $T_c \lesssim 23$ K. At present, this class of materials (HTS) is characterized by the quasi-two-dimensonal structure of the normal metallic state above $T_c$ and strong deviations from either the normal Fermi-liquid or the BCS superconducting type of behavior in the corresponding temperature regimes $T > T_c$ and $T < T_c$, respectively.

**Hubbard subband** Term describing each of the two parts of an energy band in a solid which splits when the electron–electron repulsion energy is comparable to (or larger than) their kinetic (band) energy. The Hubbard splitting of the original band induced by the interaction explains in a natural way the existence of the Mott insulating state in the case of an odd number of electrons per atom (that is, when the atomic shells would normally form an only half-filled band; cf., e.g., CoO).

**Mott insulator** An insulator containing atoms with partially filled *3d* or *4f* shells. These systems order magnetically (usually antiferromagnetically) when the temperature is lowered. Thus, they differ from ordinary (Bloch–Wilson) or band insulators, which are weakly diamagnetic, and are characterized by filled atomic shells, separated from empty states by a gap. In the antiferromagnetic phase of the Mott insulators each electron with its (frozen) spin oriented up is surrounded by electrons with their spin down, and vice versa. The parent stochiometric materials for high-temperature superconductors (e.g., $La_2CuO_4$ and $YBa_2Cu_3O_6$) are antiferromagnetic Mott insulators with Néel temperatures ($T_N = 250$ and 415 K, respectively).

**Real-space pairing** Source of attraction or superconducting correlations that is not induced by lattice deformation (phonons). Such pairing may be provided by the density fluctuations within the interacting electron subsystem (e.g., by spin fluctuations or other excitations). By real-space pairing we mean the pairing of electron spins in correlated metals caused by exchange interactions (e.g., kinetic exchange) among electrons in coordinate space. The essence of the real-space pairing, not resolved as yet, is contained in the question, Can a strong short-range part of the coulomb repulsion (of range $a_0$) lead to an attraction (an effective binding) at intermediate distances ($2 \div 10a_0$), where strong singlet–spin correlations prevail?

**Strongly correlated electrons** Electrons describing the metallic state of high-temperature superconductors, some heavy-fermion systems (**non-Fermi liquids**), and, particularly, systems of low dimensonality, $d = 1$ and 2. In these systems, the concept of a Fermi liquid is inapplicable, and for $d = 1$, at least, the charge and spin degrees of freedom lead to separate quasiparticle representations—***holons*** and ***spinons***, respectively. The quantum liquid describing strongly correlated electrons composes a new quantum macrostate.

## I. INTRODUCTION

Superconductivity remains among the most spectacular manifestations of a macroscopic quantum state of electrons in a metal or plasma. Experimentally, one observes below a characteristic temperature $T_c$ a transition to a phase with nonmeasurable dc resistance (or with a persistent current), a perfect diamagnetism of bulk samples in a weak magnetic field, and quantum tunneling between superconductors separated by an insulating layer of mesoscopic ($\sim$1-nm) thickness. In the theoretical domain, one studies the quantum–mechanical (nonclassical) mechanisms of pairing of the microscopic particles (fermions) at a macroscopic scale. Here, we summarize briefly our present understanding of the Bardeen–Cooper–Schrieffer (BCS) theory of "classical" superconductors (see Section II) and we review the current theoretical approaches to new superconductors: the heavy-fermion materials and the high-$T_c$ magnetic oxides. The latter subject is discussed in Section IV, after we summarize normal-state properties of correlated electrons in Section III.

A brief characterization of the recent studies of superconductivity is in order. From the time of the first discovery (1911) of superconductivity in mercury (at temperature $T_c \simeq 4.2$ K) by Kammerlingh Onnes until 1986, studies were limited to low temperatures, $T < 25$ K. During the next 5 years, six classes of new superconducting compounds with critical temperatures $T_c = 30$ K (for $Ba_{1-x}K_xBiO_3$), 40 K (for $La_{2-x}Sr_xCuO_4$), 90 K (for $YBa_2Cu_3O_{7-\delta}$), 110 K (for $Bi_2Sr_2CaCu_2O_8$), and 125 K (for $Tl_2Ca_2Ba_2Cu_3O_{10-y}$), and 135 K ($HgBa_2Ca_2Cu_3O_8$) were discovered and/or thoroughly studied in a number of laboratories. In recent years the idea has also been applied to new systems such as Fermi condensated dilute gases and quark–gluon plasma in high-energy physics. Apart from the discovery of spin–triplet pairing in liquid $^3$He, evidence for it in $Sr_2RuO_4$ also opens new possibilities for pairing studies.

The starting point in both classical and new superconducting materials is the electronic structure that determines the metallic properties in the normal phase (that is, that above $T_c$). In this respect, the classical superconductors are well described by band theory and, in some cases, starting from the concept of the Fermi–liquid concept. In contrast, the new materials are characterized as those whose electrons are close to localization, that is, those close to the metal–insulator transition of the Mott–Hubbard type. The latter transition may be induced by a relatively small change in compound composition (cf. the behavior of $La_{2-x}Sr_xO_4$ or $YBa_2Cu_3O_{7-x}$ as a function of $x$). It is quite interesting to note that oxides such as $YBa_2Cu_3O_{7-x}$ may be synthesized in either insulating ($x \gtrsim 0.65$) or metallic states. Additionally, antiferromagnetic ordering of the *3d* electrons is observed close to the insulator–metal transition; the magnetic insulating state transforms into a superconducting state when $0 \leq x \lesssim 0.65$. Therefore, an account of our understanding of the almost-localized metallic state in a normal or magnetic (that is, nonsuperconducting) phase is highly desirable and summarized in Section III. The antiferromagnetic insulating, normal metallic, and superconducting states must all be treated on the same footing for a proper characterization of high-$T_c$ oxides. In this manner, the studies of those systems must incorporate the description of different quantum phase transitions. One can say that the theory of strongly correlated electrons and of the superconductivity in those systems poses one of the most challenging problems for physics of the 21st century.

Details of the electronic structure in high-$T_c$ oxides are also important for two additional reasons. First, as discussed later, in these superconductors the coherence length is quite small, that is, comparable to the lattice constant. Hence, the details of the wave function on the atomic scale become crucial. Second, a whole class of models (discussed in Section IV) relies on the electron pairing induced by short-range electron–electron interactions. These interactions are strong and also present in the normal phase. This is the reason one must develop a coherent theoretical picture of the correlated metallic state that undergoes a transformation either to the Mott insulating or to the superconducting state. Such a theory does yet not exist.

In this chapter, the properties of correlated electrons in normal, insulating, magnetic, and superconducting phases are reviewed and related to the parametrized models, starting from either Hubbard or Anderson-lattice Hamiltonians. These are the models that describe the properties of correlated metallic systems in terms of a few parameters, such as the band width $W$ of starting (uncorrelated, bare) electrons, the magnitude $U$ of short-range (intraatomic) coulomb interactions, etc. Such models provide an overall understanding of both the nature of correlated metallic and insulating ground states and the underlying thermodynamic properties of these systems. However, the guidance of detailed band structure calculations is often needed in choosing appropriate values for the microscopic parameters, as well as to understand the specific features of the compounds.

## II. THE BARDEEN–COOPER–SCHRIEFFER (BCS) THEORY: A BRIEF SUMMARY

The BCS theory [1–10] relies on three features of metallic solids: (1) the electron–lattice interaction; (2) the formation of an electron-pair bound state (the so-called Cooper pair state) due to the coupling of the electrons to the lattice; and (3) the instability of the normal metallic state with respect to the formation of a macroscopic condensed state of all pairs ($\mathbf{k}\uparrow, -\mathbf{k}\downarrow$) with antiparallel spins in momentum ($\mathbf{k}$) space. The condensed state exhibits the principal properties of superconductors, such as a perfect diamagnetism, zero dc resistance, etc. We first discuss these three features briefly and then summarize some consequences of the BCS theory. The BCS theory not only deals with one of the possible (phonon-mediated) mechanisms for superconductivity, but also provides proper language for the description of such a condensed state in general terms, independent of the particular pairing mechanism. One should also remark at the beginning that such a condensed state of pairs *cannot* be regarded as a Bose condensed state if the size of the bound-state wave function $\xi$ (*the coherence length*) is much larger than the interparticle distance $a = (V/N)^{\frac{1}{3}}$; this happens for the "classic" superconductors.

### A. From Electron–Phonon Coupling to the Effective Attractive Interaction between Electrons: Virtual Exchange of Phonons

The electron–lattice interaction can be described by introducing phonons as quasiparticles representing vibrational modes of the lattice. In this picture, an electron moving in a solid and scattering on the lattice vibration absorbs or emits a phonon with energy $\hbar\omega_{\mathbf{q}}$ and quasi-momentum $\hbar\mathbf{q}$. If during such processes the energy of the incoming electron (with energy $\epsilon_{\mathbf{k}}$ and momentum $\hbar\mathbf{k}$) and the scattered electron (with energy $\epsilon_{\mathbf{k}'}$) is conserved, then a real scattering process has taken place. Such events lead to the nonzero resistivity of metals at temperature $T > 0$. For these processes

$$\epsilon_{\mathbf{k}'} - \epsilon_{\mathbf{k}} = \pm\hbar\omega_{\mathbf{q}}, \tag{1}$$

where $-$ corresponds to the emission and $+$ to the absorption of the phonon. However, in the quantum-mechanical description of scattering processes, there also exist virtual processes that do not conserve energy. Such events involve the emission and subsequent reabsorption of a phonon in a time interval $\Delta t$ such that the uncertainty principle $\Delta E \cdot \Delta t \geq \hbar$ is not violated. The uncertainty of particle energies $\Delta E$ is related to the magnitude of the electron–phonon interaction. In effect, this leads to the following effective electron–electron interaction energy involving a pair $(\mathbf{k}, \mathbf{k}')$ of electrons:

$$V_{\mathbf{k}\mathbf{k}'\mathbf{q}} = |W_{\mathbf{q}}|^2 \frac{\hbar\omega_{\mathbf{q}}}{(\epsilon_{\mathbf{k}'} - \epsilon_{\mathbf{k}})^2 - (\hbar\omega_{\mathbf{q}})^2}, \tag{2}$$

where $(\mathbf{k}' - \mathbf{k}) = \mathbf{q}$, and $W_{\mathbf{q}}$ is the electron–phonon matrix element characterizing the process of single emission or absorption of the phonon by the electron subsystem. In many electron systems, one represents Eq. (2) by an effective electron–electron interaction, which can be written

$$H' = \sum_{\mathbf{k}\mathbf{k}'\mathbf{q}} V_{\mathbf{k}\mathbf{k}'\mathbf{q}} c^{+}_{\mathbf{k}+\mathbf{q}\sigma} c^{+}_{\mathbf{k}'-\mathbf{q}\sigma'} c_{\mathbf{k}'\sigma'} c_{\mathbf{k}\sigma}. \tag{3}$$

This is a phonon—mediated contribution to the interaction between electrons. More precisely, in this expression $c_{\mathbf{k}\sigma}$ symbolizes a destruction or annihilation of an electron in the initial single—particle state $|\mathbf{k}\sigma\rangle$, whereas $c^{+}_{\mathbf{k}'\sigma'}$ is the creation of an electron in the state $|\mathbf{k}'\sigma'\rangle$ after the scattering process has taken place. The processes represented in Eq. (3) of destruction of the electron pair in the states $|\mathbf{k}\sigma\rangle$ and $|\mathbf{k}'\sigma'\rangle$ and their subsequent reestablishment in the final states $|\mathbf{k}+\mathbf{q}\sigma\rangle$ and $|\mathbf{k}'-\mathbf{q}\sigma'\rangle$ are customarily represented by a diagram of the type in Fig. 1b. It symbolizes the phonon exchange between the two electrons moving through crystal. The virtual processes are composed of two parts: one describing phonon emission and the subsequent reabsorption process and one describing the reverse process.

One should note that if in Eq. (2) $|\epsilon_{\mathbf{k}'} - \epsilon_{\mathbf{k}}| < \hbar\omega_{\mathbf{q}}$, then $V_{\mathbf{k}\mathbf{k}'\mathbf{q}} < 0$, that is, the interaction is attractive. This happens, for example, on the Fermi surface, where $\epsilon_{\mathbf{k}} = \epsilon_{\mathbf{k}'} = \mu$. The sign of the interaction changes rapidly once we depart from the Fermi surface, since the electronic energies present are much higher than that of phonons. Hence, if only the magnitude of attraction overcomes the magnitude of the coulomb repulsion between the electrons in a given medium, this leads to a net attraction between the electrons. Such a net attractive interaction results in a stable superconducting state, as we shall see next.

### B. Instability of the Electron Gas State in the Case of Attractive Interaction between Electrons: Cooper Pairs

Following Fröhlich's discovery [11] that the electron–electron attraction can be mediated by phonons (cf. the previous discussion), the next step was taken by

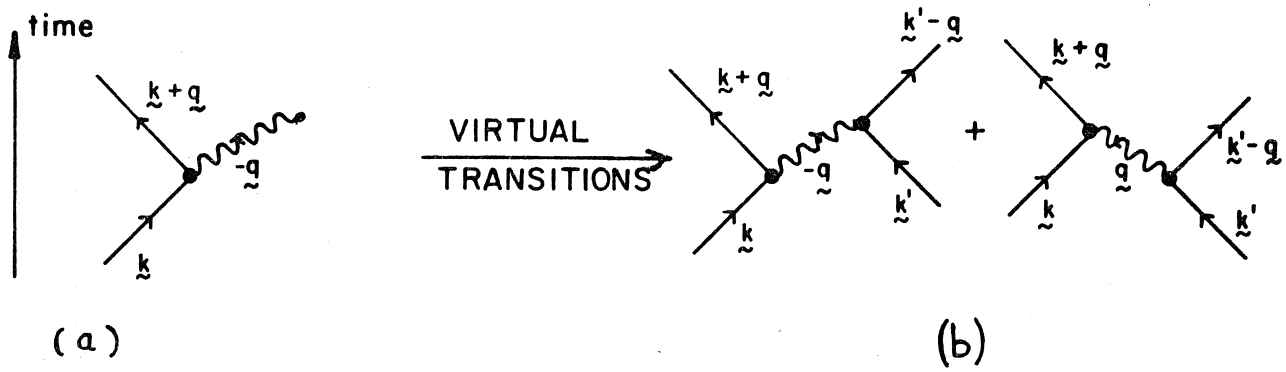



**FIGURE 1** (a) Scattering diagram of electrons with wave vectors $\mathbf{k} \rightarrow \mathbf{k}+\mathbf{q}$, accompanied by emission of the phonon of wave vector $-\mathbf{q}$. (b) Virtual emission and subsequent reabsorption of the phonon by electrons. The two processes drawn combine into the contribution [Eq. (2)] leading to the effective electron–electron attraction.

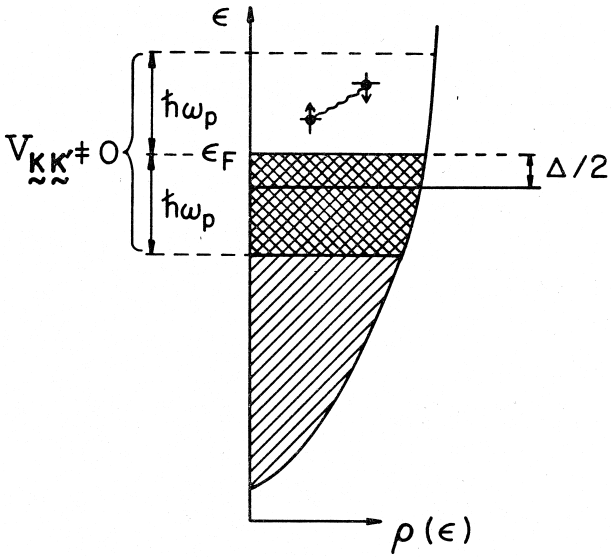



**FIGURE 2** Schematic representation of the conduction band filled with electrons up to the Fermi level $\epsilon_F$. The density of states is $\rho(\epsilon)$ (per one spin direction). The two electrons added to the system attract each other with the potential $V_{\mathbf{k}\mathbf{k}'} = -V$ if placed within the energy interval $\hbar\omega_p$ counting from $\epsilon_F$. The attraction leads to a binding energy $\Delta$ below $\epsilon_F$ for the pair configuration $(\mathbf{k}\uparrow, -\mathbf{k}\downarrow)$.

Cooper [12], who asked what happens when two electrons are added to an electron gas at $T = 0$. Because of the Pauli exclusion, they must occupy the states above the Fermi level, as shown in Fig. 2. Cooper showed that if the attractive potential in Eq. (2) is approximated by a negative nonzero constant $(-V)$ in the energy interval $2\hbar\omega_p$ around the Fermi level $\epsilon_F$ (cf. Fig. 2), then such a potential introduces a binding between these two electrons with a binding energy

$$\Delta = -2\hbar\omega_p \left[\exp\left(\frac{2}{\rho V}\right) - 1\right]^{-1} \simeq -2\hbar\omega_p \exp\left(-\frac{2}{\rho V}\right), \tag{4}$$

relative to the energy $2\epsilon_F$ of those two particles placed at the Fermi energy. In this expression, $\hbar\omega_p$ represents the average phonon energy (related to the Debye temperature $\theta_D$ through $\hbar\omega_p = k_B\theta_D$), and $\rho$ is the density of free—particle states at the Fermi energy $\epsilon_F$ for the metal under consideration.

A few important features of the bound state represented by Eq. (4) should be mentioned. First, the binding energy $\Delta$ is largest for the state of the pair at rest, that is, with the total pair momentum $\mathbf{k}_1 + \mathbf{k}_2 = 0$. Thus, $\Delta$ represents the binding energy of the pair $(\mathbf{k}, -\mathbf{k})$. Second, the spin of the pair is compensated, that is, a singlet state is assumed. Finally, the bound state has a lower energy than a pair of free particles placed at the Fermi level. Hence, the electron gas state is unstable with respect to such pair formation. A system of such pairs may condense into a superfluid state. However, the situation is not so simple since the size of the pair is of the order

$$\xi = \langle r^2 \rangle^{\frac{1}{2}} \approx \frac{2\hbar V_F}{\Delta} \approx \frac{\hbar V_F}{k_B T_C} \sim 10^{+4}\ \text{Å}, \tag{5}$$

where $V_F$ is the Fermi velocity for electrons. The quantity $\xi$ thus exceeds by far the average classical distance between the electrons, which is comparable to the interatomic distance $a \sim \frac{1}{2}2$ Å. In other words, the wave functions of the different pairs overlap very strongly, forming a condensed and coherent state of pairs in the superconducting phase. The properties of this condensed phase are discussed next. The new length scale $\xi$ appearing in the system when electrons are bound into Cooper pairs is called the coherence length.

### C. Properties of the Superconducting State: The Pairing Theory

The BCS theory [1] provides a method of calculating the ground state, thermodynamic, and electromagnetic properties of a superconductor treated as a condensed state of electron pairs with opposite momenta and spins. The starting microscopic Hamiltonian is

$$H = \sum_{\mathbf{k}\sigma} \epsilon_{\mathbf{k}} n_{\mathbf{k}\sigma} + \sum_{\mathbf{k}\mathbf{k}'} V_{\mathbf{k}\mathbf{k}'} c^+_{\mathbf{k}\uparrow} c^+_{\mathbf{k}\downarrow} c_{-\mathbf{k}'\downarrow} c_{\mathbf{k}'\uparrow}. \tag{6}$$

The first term describes the single-particle (band) energy, $\epsilon_{\mathbf{k}}$ being the energy per particle and $n_{\mathbf{k}\sigma} = c^+_{\mathbf{k}\sigma} c_{\mathbf{k}\sigma}$ the number of particles in the state $|\mathbf{k}\sigma\rangle$. The second term describes the pairing part [Eq. (3)] for the system of pairs that scatters from the state $(\mathbf{k}', -\mathbf{k}')$ into the state $(\mathbf{k}, -\mathbf{k})$. This term describes the dominant contribution of all processes contained in Eq. (3) (cf. Ref. 10).

To obtain eigenenergies of the Hamiltonian [Eq. (6)], one can use either the variational method due to Schrieffer [2], the transformation method developed by Bogoliubov and Valatin [13], or the two-component method due to Nambu [13]. To obtain quasiparticle states in the superconducting phase, one has to combine an electron in the state $|\mathbf{k}\uparrow\rangle$ with one in the time-reversed state $|-\mathbf{k}\downarrow\rangle$. More precisely, one defines new quasiparticle operators $\lambda^+_{\mathbf{k}0}$ and $\lambda^+_{\mathbf{k}1}$, which are expressed by the operators $c^+$ and $c$ in the following manner:

$$\lambda^+_{\mathbf{k}0} = u_{\mathbf{k}} c^+_{\mathbf{k}\uparrow} - v_{\mathbf{k}} c_{-\mathbf{k}\downarrow},$$

and

$$\lambda^+_{\mathbf{k}1} = v_{\mathbf{k}} c^+_{\mathbf{k}\uparrow} + u_{\mathbf{k}} c_{-\mathbf{k}\downarrow}.$$

The coefficients of the transformation fulfill the condition $u_{\mathbf{k}}^2 + v_{\mathbf{k}}^2 = 1$. One should note that the transformation *does not* conserve the particle number, so one has to add the term $(-\mu N)$ to the Hamiltonian (b), where $\mu \equiv \epsilon_{\mathrm{F}}$ is the chemical potential in the superconducting state.

The single-particle excitations in the superconducting phase are specified by

$$E_{\mathbf{k}} = \left[(\epsilon_{\mathbf{k}} - \mu)^2 + |\Delta_{\mathbf{k}}|^2\right]^{\frac{1}{2}}, \tag{7}$$

where $\mu$ is the chemical potential of the system and $|\Delta_{\mathbf{k}}|$ is the so-called superconducting gap determined from the self—consistent equation

$$\Delta_{\mathbf{k}} = -\sum_{\mathbf{k}} V_{\mathbf{k}\mathbf{k}'} \frac{\Delta_{\mathbf{k}}}{2E_{\mathbf{k}'}} \tanh\left(\frac{\beta E_{\mathbf{k}'}}{2}\right), \tag{8}$$

with $\beta \equiv (k_{\mathrm{B}}T)^{-1}$. One should note that if $V_{\mathbf{k}\mathbf{k}'}$ is approximated by a negative constant, then $\Delta_{\mathbf{k}} = \Delta$; Eq. (8) then yields as a solution either $\Delta \equiv 0$ or $\Delta \neq 0$, obeying the equation

$$1 = \frac{V}{N} \sum_{\mathbf{k}}{}' \frac{1}{2E_{\mathbf{k}}} \tanh\left(\frac{\beta E_{\mathbf{k}}}{2}\right), \tag{9}$$

where now

$$E_{\mathbf{k}} = \left[(\epsilon_{\mathbf{k}} - \mu)^2 + \Delta^2\right]^{\frac{1}{2}}. \tag{10}$$

The primed summation in Eq. (9) is restricted to the regime of $\mathbf{k}$ states where $V \neq 0$. Equations (9) and (10) constitute the simplest BCS solution for an isotropic ($\mathbf{k}$-independent) gap. One sees that $E_{\mathbf{k}}$ is always nonvanishing and reaches a minimum $E_{\mathbf{k}} = \Delta$ for electrons placed on the Fermi level, where $\epsilon_{\mathbf{k}} = \mu$. Thus, the meaning of the gap becomes obvious: it is the gap for the single—electron excitations from the superconducting (condensed) phase to a free—particle state. The presence of a gap $\Delta > k_{\mathrm{B}}T_{\mathrm{c}}$ in the spectrum of single—particle excitations suppresses the scattering of electrons with acoustic phonons. The thermally excited electrons across the gap do not yield nonzero resistivity because their contribution is short-circuited by the presence of the pair condensate that carries a current with no resistance. The same holds true even for the superconducting systems for which the gap vanishes along some lines or at some points in $\mathbf{k}$ space.

One should emphasize that all thermodynamic properties are associated with thermal excitations; the energies that are specified by Eq. (7) contain $|\Delta_{\mathbf{k}}|$ or $\Delta$ as a parameter to be determined self-consistently from Eq. (8) or (9), respectively. Next, we provide a brief summary of the results that may be obtained within the BCS theory.

## D. Summary of the Properties: The Homogeneous State

The solution of Eq. (9) provides the following properties.

1. At $T = 0$, Eq. (1) reads

$$1 = (V/2N) \sum_{\mathbf{k}}{}' E_{\mathbf{k}}^{-1}. \tag{11}$$

The value of $\Delta \equiv \Delta(T = 0)$ for $\rho V \ll 1$ is given by

$$\Delta_{\mathrm{o}} = \frac{\hbar\omega_p}{\sinh(1/\rho V)} \approx 2\hbar\omega_p \exp\left(-\frac{1}{\rho V}\right), \tag{12}$$

where $\hbar\omega_p \approx k_{\mathrm{B}}\theta_{\mathrm{D}}$[1] and $\rho$ is the density of States at the Fermi energy. One notes a striking similarity between Eq. (12) and Eq. (4), particularly for $\rho V \ll 1$ (this condition represents the so-called weak—coupling limit); The absence of factor 2 in (12) provides an enchancement of the gap in the condensed state due to the presence of other electrons.

2. We can choose the origin of energy at $\mu$. Then Eq. (9) can be transformed into an integral form:

$$1 = V \int_{\mathrm{o}}^{\hbar\omega_{\mathbf{p}}} \frac{\rho(\epsilon)\, d\epsilon}{(\epsilon^2 + \Delta^2)^{\frac{1}{2}}} \tanh\left(\frac{\beta}{2}\sqrt{\epsilon^2 + \Delta^2}\right). \tag{13}$$

Since $\hbar\omega_{\mathbf{p}} \ll \mu$, we may take $\rho(\epsilon) \approx \rho(\epsilon_{\mathrm{F}}) \equiv \rho$ within the range of integration. This allows for an analytic evaluation of the critical temperature for which $\Delta = 0$:

$$T_{\mathrm{c}} = 1.13\theta_{\mathrm{D}} \exp\left(-\frac{1}{\rho V}\right). \tag{14}$$

In all these calculations, it is implicitly assumed that $\rho V \ll 1$. Because of the presence of the exponential factor in Eq. (14), the critical temperature $T_{\mathrm{c}}$ is much lower than the Debye temperature characterizing the average energy of acoustic phonons. This is the principal theoretical reason that $T_{\mathrm{c}}$ is so low in the superconductors discovered in the period 1911–1986. The exponential dependence of $T_{\mathrm{c}}$ on the electronic parameter $\rho V$ also explains why the parameters pertaining to the electronic structure, which are of the order of 1 eV or more, respond to phase transitions on an energy scale that is three orders of magnitude smaller ($k_{\mathrm{B}}T_{\mathrm{c}} \sim 1$ meV). Effects with such a nonanalytic dependence of transition temperature on the coupling constant cannot be obtained in any order of perturbation theory starting with the normal state as an initial state. A similar type of effect is obtained in the studies of the Kondo effect (cf. Section IV).

3. Combining Eqs. (14) and (1.12) one obtains the universal ratio

$$\frac{2\Delta_{\mathrm{o}}}{k_{\mathrm{B}}T_{\mathrm{c}}} = 3.53, \tag{15}$$

[1] In actual practice, one assumes that $\hbar\omega_{\mathbf{p}} \approx 0.75 k_{\mathrm{B}}\theta_{\mathrm{D}}$ (cf. Meservey and Schwartz in Ref. 9).

which is frequently used as a test for the applicability of the BCS model. However, this value can also be obtained in the strong-coupling limit [15] for a particular strength of electron–phonon coupling.

4. By regarding energies $E_{\mathbf{k}}$ as representing electron excitations across the gap, one can write the expression for the entropy of a superconductor in the standard form:

$$S = -2k_{\mathrm{B}} \sum_{\mathbf{k}} [f_{\mathbf{k}} \ell \mathrm{n}\, f_{\mathbf{k}} + (1 - f_{\mathbf{k}})\, \ell \mathrm{n}(1 - f_{\mathbf{k}})], \quad (16)$$

where $f_{\mathbf{k}} \equiv f(E_{\mathbf{k}}) = [1 + \exp(\beta E_{\mathbf{k}})]^{-1}$ is the Fermi–Dirac distribution function $[1 + \exp(\beta E_{\mathbf{k}})]^{-1}$. Hence, the free energy of the superconducting state is

$$F_{\mathrm{S}} = 2 \sum_{\mathbf{k}} E_{\mathbf{k}} f_{\mathbf{k}} - TS. \quad (17)$$

One should note that the thermodynamic properties are determined fully only if the chemical potential $\mu = \mu(T)$ and the temperature dependence of the superconducting gap $\Delta_{\mathbf{k}} = \Delta_{\mathbf{k}}(T)$ are explicitly determined, since only then is the spectrum of single—particle excitations (characterized by the energies $\{E_{\mathbf{k}}\}$) uniquely determined. The quantity $\Delta(T)$ is determined from Eq. (13). The chemical potential is determined from the conservation of the number $N_{\mathrm{e}}$ of particles, that is, from the condition $\sum_{\mathbf{k}} f_{\mathbf{k}} = N_{\mathrm{e}}$. The temperature dependence of the gap in the isotropic case is shown schematically in Fig. 3.

5. By calculating this difference of the free energies $F_{\mathrm{S}} - F_{\mathrm{N}}$ in superconducting ($F_{\mathrm{S}}$) and normal ($F_{\mathrm{N}}$) phases and equating the difference with the magnetic free energy $H_{\mathrm{c}}^2 V/8\pi$ ($V$ is the volume of the system), one can obtain an approximate relation of the form

$$\frac{H_{\mathrm{c}}(T)}{H_{\mathrm{c}}(0)} \approx 1 - \left(\frac{T}{T_{\mathrm{c}}}\right)^2. \quad (18)$$

For the applied field $H > H_{\mathrm{c}}$, superconductivity is destroyed because in the thermodynamic critical field $H_{\mathrm{c}}$ the spin–singlet bound state is destroyed by the thermal fluctuations. The pair binding energy is then effectively overcome by the magnetic energy, so that the pairs break up into single particles. Strictly speaking, this type of behavior characterizes the so-called superconductors of the first kind.

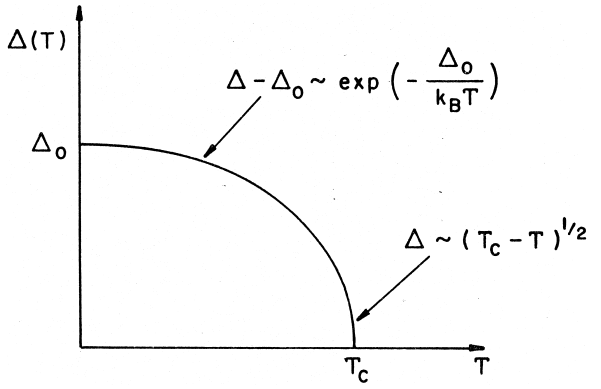



**FIGURE 3** Schematic representation of the temperature dependence of the superconducting gap for the isotropic change. $T_{\mathrm{c}}$ is the critical temperature for the transition, and $\Delta_0 \equiv \Delta\,(T = 0)$.

6. By calculating the specific heat from the standard thermodynamic analysis $[C_{\mathrm{S}} = -T(\partial^2 F_{\mathrm{S}}/\partial T^2)_{\mathrm{V}}]$, one obtains at $T = T_{\mathrm{c}}$ a discontinuity of the form

$$\frac{C_{\mathrm{S}} - C_{\mathrm{N}}}{C_{\mathrm{N}}} = 1.43, \quad (19)$$

where $C_{\mathrm{N}}$ is the specific heat at $T_{\mathrm{c}}$ for the material in its normal state. At low temperatures, the specific heat decreases exponentially:

$$C_{\mathrm{S}} \sim \exp\left(-\frac{\Delta_0}{k_{\mathrm{B}} T}\right), \quad (20)$$

for the special case of an isotropic gap. However, if the gap is anisotropic $[\Delta = \Delta_{\mathbf{k}}(T)]$ and has lines of zeros (along which $\Delta_{\mathbf{k}} = 0$), then the low—temperature dependence of $C_{\mathrm{S}}$ does not follow Eq. (20) but rather a power law $T^n$, with $n$ depending on the details of the gap anisotropy.

The specific heat grows with $T$ because the number of thermally broken pairs increases with rising temperature; eventually, at $T = T_{\mathrm{c}}$ ($k_{\mathrm{B}} T_{\mathrm{c}} \sim \Delta_{\mathrm{o}}$), all bound pairs dissociate thermally, at which point $C_{\mathrm{s}}$ reaches a maximum. If the temperature is raised further (above $T_{\mathrm{c}}$), the specific heat drops rapidly to its normal-state value since no pairs are left to absorb the energy. This type of behavior is observed in superconductors with an isotropic gap (cf., e.g., Hg and Sn). One should note that this interpretation of the thermal properties is based on the single—particle excitation spectrum [Eq. (10)]; we have disregarded any fluctuation phenomena near $T_{\mathrm{c}}$, as well as collective excitations of the condensed system. It can be shown that the large coherence length $\xi \sim 10^3/10^4$ Å encountered in classic superconductors [8] is related to the absence of critical behavior near $T_{\mathrm{c}}$. This is not the case in high-$T_{\mathrm{c}}$ superconductors (discussed in Section IV); hence, the new materials open up the possibility of studies of critical phenomena in superconducting systems.

7. The spin part of the static magnetic susceptibility vanishes as $T \to 0$. This is a direct consequence of the binding of electrons in the condensed state into singlet pairs. Therefore, the Meissner effect (the magnetic flux expulsion from the bulk of the sample) at $T = 0$ is present because the orbital part of the susceptibility is diamagnetic (roughly, it represents an electron-pair analogue of the Landau diamagnetism of single electrons in a normal electron gas). The expulsion of the magnetic flux from the bulk is measured in terms of the so-called London penetration depth $\lambda = \lambda(T)$, which characterizes the decay of

the magnetic induction inside the sample. It decays according to

$$B(z) = H_a \exp(-z/\lambda),$$

where the $z$ direction is perpendicular to the sample surface and the applied magnetic field $H_a$ is parallel to it. The temperature dependence of the penetration depth is given by

$$\frac{\lambda(T)}{\lambda(0)} = \left[\frac{\Delta(T)\tanh(\Delta/2k_BT)}{\Delta_o}\right], \approx \left[1-\left(\frac{T}{T_c}\right)^4\right]^{-\frac{1}{2}}. \quad (21)$$

This result has been derived under the assumption that the coherence length[2] $\xi_o \simeq \hbar V_F/\Delta_o$ is much larger than $\lambda$. One should note that for a bulk sample of dimension $d \gg \lambda$ the induction $B \equiv 0$ almost everywhere. This condition determines the magnetic susceptibility $\chi$ of a superconductor regarded as an ideal diamagnet; in cgs units, $\chi \equiv M/H = -1/(4\pi)$.

8. The relative ratio of the two characteristic lengths $\kappa \equiv \lambda/\xi$ determines the type of superconductivity behavior in a magnetic field. From the dependence $\xi \sim \Delta^{-1}$, we infer that as $T \to T_c$, $\xi \sim (T_c - T)^{-\frac{1}{2}}$. The same type of dependence for $\lambda(T)$ can be inferred from Eq. (21) when $T \to T_c$. Within the phenomenological theory of Ginzburg and Landau (which can be derived from the BCS theory as shown by Gorkov [8]), one can show that if $\kappa \lesssim 1/\sqrt{2}$, then the material is a superconductor of the first kind; if $\kappa \gtrsim 1/\sqrt{2}$, then the material is of the second kind. The value of $\kappa$ is directly related to the penetration depth $\lambda(T)$. The thermodynamic critical magnetic field Eq. (18) has the form

$$H_c(T) = \Phi_0 \frac{\sqrt{2}}{\kappa\lambda^2(T)} \quad (22a)$$

or, equivalently,

$$H_c(T) = \frac{\Phi_o}{2\pi\sqrt{2}\xi(T)\lambda(T)}, \quad (22b)$$

where $\Phi_0 = \hbar c/2e$ is the magnetic—flux quantum. This value of the field terminates superconductivity of the first kind. For superconductors of the second kind, the corresponding field is given by

$$H_{c2} = \kappa\sqrt{2}H_c = \frac{\Phi_0}{2\pi\xi(T)^2} i. \quad (22c)$$

[2]The coherence length in a superconductor can be estimated by using the uncertainty relation $\Delta_p \cdot \xi_o = \hbar$ where $\Delta_p$ is a change of the electron momentum (at $\epsilon = \epsilon_F$) due to the attractive interaction, which can be estimated from the corresponding change of the particle kinetic energy $\Delta E = v_F \Delta_p$. Taking $\Delta E \simeq \Delta_0$, we obtain the desired estimate of $\xi_o$.

For fields $H_{c1} < H_a < H_{c2}$ [with $H_{c1} \equiv H_c(0)\,\ell n\,\kappa/(\sqrt{2}\kappa)$], the superconducting phase is inhomogeneous, composed of the lattice of vortices, each of the form of a tube containing one flux quantum, penetrating the sample. All of the newly discovered high-$T_c$ superconductors are of the second kind, with very small values of $H_{c1}$ and very large values of $H_{c2}$. This means that the value of the coherence length $\xi$ is very small in those systems.

9. The sound absorption coefficient $\alpha_s$ in the superconducting phase is related to that in the normal phase $\alpha_N$ by

$$\frac{\alpha_S}{\alpha_N} = \frac{2}{1+\exp(\Delta/k_BT)}.$$

This is a very simple result; hence, experimental results for $(\alpha_S/\alpha_N)$ are used to determine the temperature dependence of the gap $\Delta$.

A complete discussion of superconducting states within the BCS theory is provided in Refs. 1–10.

## E. Strong—Coupling Effects: The Eliashberg Approach

The BCS theory provides a complete though approximate theory of both thermal and dynamic properties of superconductors in the weak-coupling limit $\rho V \ll 1$. The electron–electron interactions deriving from the electron–lattice interaction are treated in the lowest order and the electron–electron correlations are decoupled in the mean field-type approximation. Generalizations of the BCS treatment concentrate on two main problems—(1) inclusion of the repulsive coulomb interaction between the electrons [14] and (2) extension of the BCS theory to the situation with arbitrarily large electron–phonon coupling [15]—by generalizing the treatment of normal metals, with electron–lattice interactions incorporated in a systematic fashion [16]. Both of these factors have been included in the Eliashberg approach to superconductivity [15].

The coulomb repulsive interaction reduces the effective attractive interaction between the electrons, so that, instead of Eq. (14), one obtains in the BCS approximation

$$T_c = 1.14\theta_D \exp\left(-\frac{1}{\lambda-\mu^*}\right), \quad (23)$$

where $\lambda = \rho V$ is the effective electron–phonon coupling and $\mu^*$ is the so-called coulomb pseudopotential [14] multiplied by $\rho$.

The Eliashberg correction to the BCS theory must be evaluated numerically. The numerical solution of the Eliashberg equation representing higher-order corrections to the BCS theory may be represented by [17]

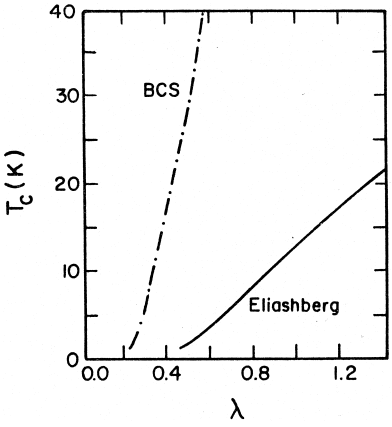



**FIGURE 4** Numerical solution of $T_c$ versus the electron–phonon coupling constant $\lambda$ for the coulomb pseudopotential $\mu^* = 0.1$. The other parameters are taken as for the superconducting element niobium. Note that the Eliashberg theory gives a much slower increase in $T_c$ than does the BCS theory.

$$T_c = \frac{\theta_D}{1.45} \exp\left[-\frac{1.04(1+\lambda)}{\lambda - \mu^*(1+0.62\lambda)}\right]. \qquad (24)$$

Figure 4 illustrates the difference in the values of $T_c$ obtained by the BCS vs the Eliashberg theory [18]. We see that the repulsive coulomb interaction and the higher-order electron–phonon effects combine to reduce the superconducting transition temperature drastically. This and other results [19] have led to the conclusion that the value of $T_c$ determined within the phonon–mediated mechanism has an upper limit of the order of 30 K.

One should mention a very important feature of the phonon-mediated electron pairing. Namely, the transition temperature is proportional to the Debye temperature $\Theta_D$. Hence, $T_c$ given by expression (23) depends on the mass $M$ of the atoms composing the lattice. In the simplest situation we expect that $T_c \sim M^{-\frac{1}{2}}$. A dependence of $T_c$ on the mass $M$ was demonstrated experimentally [20] by studying the isotope influence on $T_c$. These observations provided a crucial argument in favor of the lattice involvement in the formation of superconducting state. If the Coulomb repulsion between electrons is taken into account, then the relation is $T_c \sim M^{-\alpha}$ with [17]

$$\alpha = \frac{1}{2}\left\{1 - \frac{(1+\lambda)(1+0.62\lambda)}{[\lambda - \mu^*(1+0.62\lambda)]^2}\right\}.$$

In the strong coupling limit ($\lambda \geq 1$) the exponent $\alpha$ is largely reduced from its initial value $\frac{1}{2}$. Therefore, if the value of $\alpha$ is small, one may interpret this fact as the evidence for either strong electron–phonon coupling or that a new nonphonon mechanism is needed to explain the superconductivity.

### F. Where Do We Go from Here?

The BCS theory is a microscopic theory providing a description of thermodynamic and electrodynamic properties as a function of two parameters: $T/T_c$ and $H_a/H_c$, where $T_c$ contains the effective attraction strength $|V|$. Such a simple approach is not possible for high-temperature superconductors, as one can see from already existing books and review articles [cf. Refs. 21a–k]. In the next two sections we summarize briefly the principal features of strongly correlated systems. This discussion provides us with new phenomena and some new terms describing them. This overview by no means contains a full discussion of papers published during the last 12 years. Rather, we sketch different paths of approaching the problems encountered in dealing with strongly correlated fermions.

## III. NORMAL AND MAGNETIC STATES OF CORRELATED ELECTRONS

### A. Narrow—Band Systems

The modern theory of metals derives from the concept of a free electron gas, which obeys the Pauli exclusion principle. The principal influences of the lattice periodic potential on the individual electron states are to renormalize their mass and to change the topology of the Fermi surface. Landau [22] was the first to recognize the applicability of the electron–gas concept to the realistic situation where the repulsive coulomb interaction between particles is not small compared to the kinetic energy of electrons near the Fermi surface. He incorporated the interaction between electrons into a further (many–body) renormalization of the effective mass and investigated the physical properties, such as specific heat, magnetic susceptibility, sound propagation, and thermal and electric conductivities in the terms of quasiparticle contributions.

An important next development was contributed by Mott [23], who pointed out that if the coulomb interaction between the electrons is sufficiently strong (that is, comparable to the band energy of the quasiparticles), then electrons in a solid would have to localize on the atoms, e.g., with one valence electron per atom. This qualitative change of the nature of single–electron states from those for a gas to those for atoms is called the metal–insulator or the Mott transition. An empty (unoccupied) state in the Mott insulator (that is, that without electrons available) will act as a mobile hole. In these circumstances, the transport of charge takes place via the correlated hopping of electrons through such hole states. In the Mott insulator limit, those hole states play a crucial role in establishing the superconductivity of oxides, as discussed in Section IV.

The paramagnetic or magnetically ordered states of electrons comprising the Mott insulator distinguish this class of materials from ordinary band (Bloch–Wilson) insulators or intrinsic semiconductors; the latter are characterized at $T = 0$ by a filled valence band and an empty conduction band, separated by a gap. The electrons in the filled valence band are spin-paired into $|\mathbf{k}\uparrow, \mathbf{k}\downarrow\rangle$ singlets; hence Bloch–Wilson insulators are diamagnetic.

The basic question now arises whether one can treat Mott insulators and metals within a single microscopic description of electron states by generalizing the band theory of electron states so as to describe Mott insulators within the same microscopic model. The first step in this direction was proposed by Hubbard [25], who showed by the use of a relatively simple model that as the interaction strength (characterized by the magnitude $U$ of the intraatomic coulomb repulsion) increases and becomes comparable to the band energy per particle (characterized by the bare bandwidth $W$), the original band of single-particle states splits into two halves. Thus, the Mott insulator may be modeled by a lattice of hydrogeniclike atoms with one electron per atom, placed in the lowest *1s* state. The distinction between the normal metallic and the Mott insulating states is shown schematically in Figs. 5a and b, where the metal (a) is depicted as an assembly of electrons represented by the set of plane waves characterized by the wave vector $\mathbf{k}$ and spin quantum number $s = \sigma/2$, where $\sigma \equiv \pm 1$.

The transformation to the Mott localized state may take place only if the number of electrons in the metallic phase is equal to the number of parent atoms, that is, when the starting band of free electron-like states is half-filled. The collection of such unpaired spin moments will lead to the paramagnetic Curie–Weiss behavior at high temperatures. As the temperature is lowered, the system undergoes a magnetic phase transition; in the case of the Mott insulators, the experimentally observed transition is almost always to antiferromagnetism, as shown in Fig. 5b, where each electron with its spin moment up is surrounded by electrons on nearest-neighboring sites with spins in the opposite direction (down). Such a spin configuration reflects a two–sublattice (Néel) antiferromagnetic state. The actual magnetic structure of $Cu^{2+}$ ions in $La_2Cu\,O_4$, taken from Ref. 24, is shown in Fig. 6. The expectation value of the spin is reduced by 40% from the value $s^z = \frac{1}{2}$.

If the number of electrons in the band is smaller than the number of available atomic sites, then electron localization cannot be complete because empty atomic sites are available for hopping electrons. However, for the half-filled band case, as the ratio $U/W$ increases, half of the total number of single-particle states in the starting band is gradually pushed above the Fermi level $\epsilon_F$. An increase in the ratio $U/W$ may be achieved by lengthening the

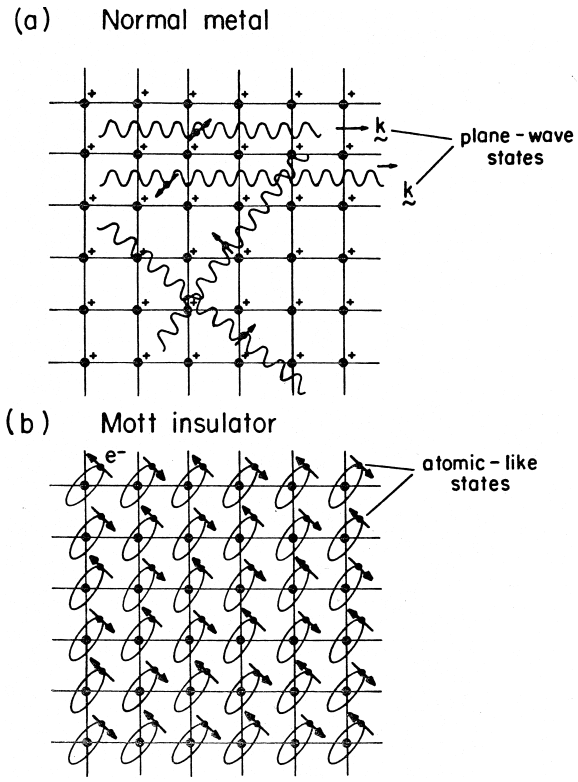



**FIGURE 5** (a) Schematic representation of a normal metal as a lattice of ions and the plane waves, with wave vector $\mathbf{k}$ representing free electron states. (b) Model of the Mott insulator as a lattice of atoms with electrons localized on them. Note that the ground–state configuration is usually antiferromagnetic (with the spins antiparallel to each other).

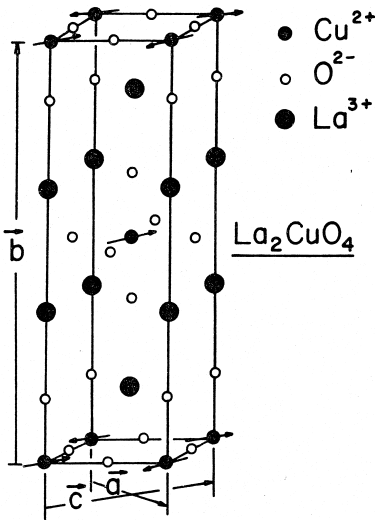



**FIGURE 6** The magnetic structure of $La_2CuO_4$. The neighboring $Cu^{2+}$ ions in the planes have their spins (each representing the $3d^9$ configuration) antiparallel to each other. The antiferromagnetic structure is three–dimensional. (From Endoh *et al.* [24].)

interatomic distance, thus reducing $W$, which is directly related to the wave–function overlap for the two states located on the nearest-neighboring sites. The splitting of the original band into two Hubbard subbands eliminates the paired (spin–singlet) occupations of the same energy state $\epsilon$. Effectively, this pattern reflects the situation of electrons being separated from each other as far as possible; however, the correspondence between the Hubbard split–band situation, shown schematically in Fig. 7, and the electron disposition in the spin lattice in real space (cf. Fig. 5b) is by no means obvious and requires a more detailed treatment that relates these two descriptions of the Mott insulator. This problem is dealt with in the following section.

### 1. The Hubbard Model

In discussing narrow band systems, one usually starts from the model Hamiltonian due to Hubbard [25], which appears to be complicated but can really be interpreted in simple terms, namely,

$$H = \sum_{\mathbf{k}\sigma} \epsilon_{\mathbf{k}\sigma} n_{\mathbf{k}\sigma} + U \sum_{i} n_{i\uparrow} n_{i\downarrow}, \tag{25}$$

where $\epsilon_{\mathbf{k}}$ is the single-particle (band) energy per electron with the wave vector $\mathbf{k}$, $U$ is the magnitude of intraatomic coulomb repulsion between the two electrons located on the same atomic site $i$, $n_{\mathbf{k}\sigma}$ is the number of electrons in the single–particle state $|\mathbf{k}\sigma\rangle$, and $n_{i\sigma}$ is the corresponding quantity for the atomic state $|i\sigma\rangle$. This simple Hamiltonian describes the localization versus delocalization aspect of electron states since the first term provides the gain in energy ($\epsilon_{\mathbf{k}} < 0$) for electrons in the band state $|\mathbf{k}\sigma\rangle$, whereas the second accounts for an energy loss ($U > 0$) connected with the motion of electrons throughout the system that is hindered by encounters with other electrons on the same atomic site. The competitive aspects of the two terms are expressed explicitly if the first term in Eq. (24) is transformed by the so–called Fourier transformation to the site $\{|i\sigma\rangle\}$ representation. Then Eq. (24) may be rewritten

$$H = \sum_{ij\sigma} t_{ij} a^{+}_{i\sigma} a_{j\sigma} + U \sum_{i} n_{i\uparrow} n_{i\downarrow}, \tag{26}$$

where

$$t_{ij} = \frac{1}{N} \sum_{\mathbf{k}} \epsilon_{\mathbf{k}} \exp[i\mathbf{k} \cdot (\mathbf{R}_j - \mathbf{R}_i)], \tag{27}$$

is the Fourier transform of the band energy $\epsilon_{\mathbf{k}}$ and $a^{+}_{i\sigma}$ ($a_i\sigma$) is the creation (annihilation) of electrons in the atomic (Wannier) state centered on the site $\mathbf{R}_j$. The first term in Eq. (25) represents the motion of an electron through the system by a series of hops $j \to i$, which are described in terms of destruction of the particle at site $j$ and its subsequent recreation on the neighboring site $i$. The width of the corresponding band in this representation is given by

$$W = 2 \sum_{j(i)} |t_{ij}| \approx 2z|t|, \tag{28}$$

where $z$ is the number of nearest neighbors (n.n.), and $t$ is the value of $t_{ij}$ for the n.n. pair $\langle ij \rangle$. Thus, the Hamiltonian [Eq. (25)] is parameterized through the bandwidth $W$ and the magnitude $U$. In actual calculations, it is the ratio $U/W$ that determines the localized versus collective behavior of the electrons in the solid.

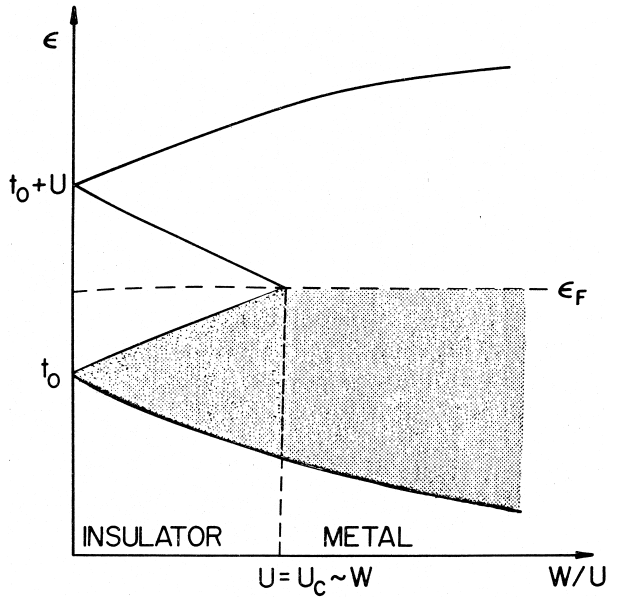



**FIGURE 7** The Hubbard splitting of the states in a single, half–filled band for the strength of the intraatomic coulomb interaction $U > U_c$. The state with a filled lower Hubbard subband for $U > U_c$ is identified with that of the Mott insulator. [From Ref. 25.]

### 2. Hubbard Subbands and Hole States

The normal–metal case is represented in Eq. (24) by the limit $W/U \gg 1$; the first (band) term then dominates. On the other hand, the complementary limit $W/U \ll 1$ corresponds to the limit of well-separated atoms, since the excitation energy of creating double occupancy on a given atom (with the energy penalty $\epsilon \sim U$) far exceeds the band energy of individual particles. The transition from the metallic to the atomic type of behavior takes place when $W \sim U$; this is also the crossover point where the single band in Fig. 7 splits in two. The actual dependence of the density of states for interacting particles is shown in Fig. 8 (taken from Ref. 26). These curves were drawn for the Lorentzian shape of the density of states (DOS), that is, for a starting band with a characteristic width $\Delta$:

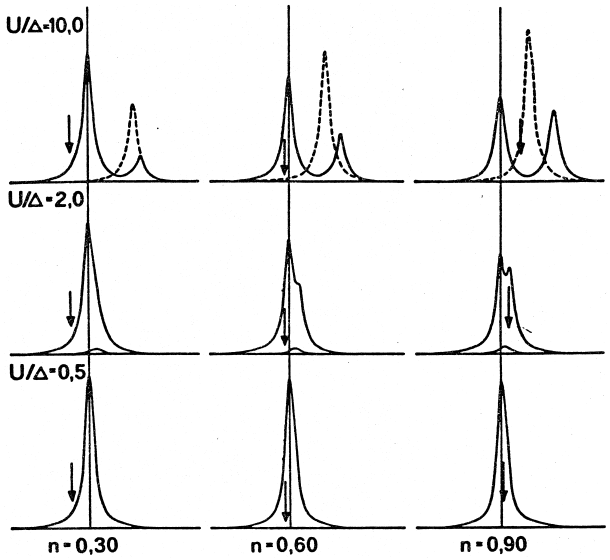



**FIGURE 8** The Hubbard splitting of the states for different band fillings, $n = 0.3$, 0.6, and 0.9, and for different $U/W$ ratios, 0.5, 2, and 10, respectively. The $x$ axis is the particle energy value; the $y$ axis is the density of states value. The arrow indicates the position of the Fermi energy, whereas the dashed line represents the inverse lifetime of the quasiparticle state in the pseudogap. [From Ref. 26.]

$$p^{o}(\epsilon) = \frac{\Delta}{\pi} \frac{1}{(\epsilon - t_0)^2 + \Delta^2}, \tag{28a}$$

where $t_0$ determines the position of the center of the band (usually chosen as $t_0 = 0$). Detailed calculations [28] show that with a growing magnitude of interaction $(U/\Delta)$, the DOS [Eq. (28a)] splits into two parts described by the density of states:

$$\rho(\epsilon) = \frac{\Delta}{\pi}\left[\frac{1-(n/2)}{(\epsilon - t_0)^2 + \Delta^2} + \frac{n/2}{(\epsilon - t_0 - U)^2 + \Delta^2}\right]. \tag{28b}$$

The first term describes the original DOS [Eq. (28a)], with the weighting factor $(1-(n/2))$, whereas the second represents the upper subband (on the energy scale), with the weighting factor $(n/2)$ and shifted by an amount $U$. These two terms and the corresponding two parts of the DOS in Fig. 8 describe the Hubbard subbands. The dashed line in Fig. 8 represents the inverse lifetime of single–electron states placed in the pseudogap, while the arrows point to the position of the Fermi energy in each case. For $n = 1$, the Fermi level falls in a pseudogap, where the lifetime of those quasiparticle states is very short. This is reminiscent of the behavior encountered in an ordinary semiconductor, where the states in the band gap are those with a complex wave vector $\mathbf{k}$. The lifetime may qualitatively simulate the atomic disorder-producing spread (Lorentian-shape) form of the bare density states.

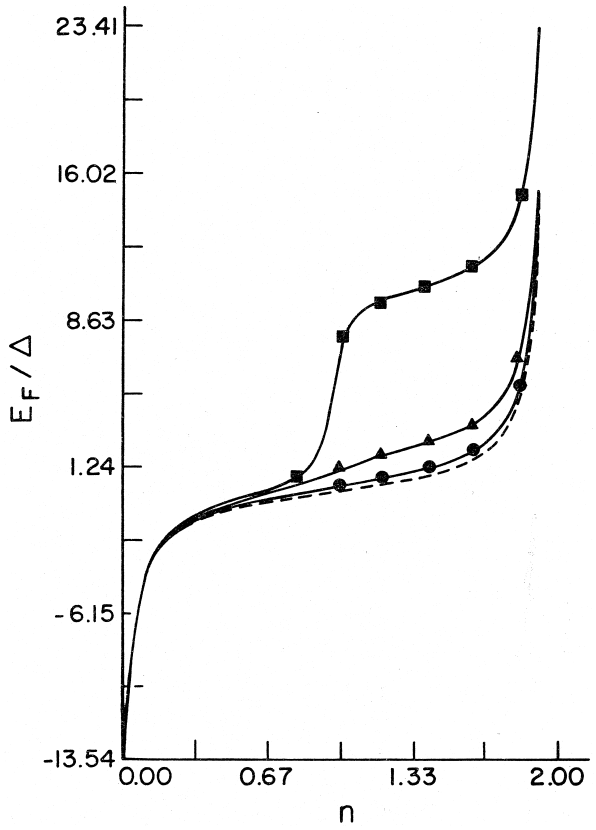



**FIGURE 9** The position of the Fermi level $\epsilon_F$ as a function of the band filling $n$ for different values of interaction (from the bottom to the top curve), $U/\Delta = 0$,0.5, 2, and 10. For $U/\Delta = 10$, the Fermi level jumps between the subbands when $n \approx 1$. [From Ref. 26.]

To display the of Mott insulator as a two-band system in which the Hubbard subbands assume a role similar to that of the valence and conduction bands in an ordinary semiconductor, we have plotted in Fig. 9 the position of the Fermi level as a function of the numbers of electrons $n$ per atom in the system. As $n$ moves past unity, a jump in $\epsilon_F$ occurs for $U/\Delta \gg 1$. This is exactly what happens in the ordinary semiconductor when the electrons are added to the conduction band. This feature shows once more that the states near the upper edge of the lower Hubbard subband (that is, the states near $\epsilon_F$ for $n$ close to but less than unity) can be regarded as hole states. We will see that those states are the ones with a high effective mass.

It should be emphasized that the Hubbard subband structure is characteristic of magnetic insulators and cannot be obtained with a standard band theoretical approach to the electron states in solids. The $N$ states in the lower Hubbard subband are almost singly occupied; this is directly related to the picture of unpaired spins in Fig. 5b and is one of the reasons for calling the electron states for such interacting systems correlated electronic states.

The other reason (discussed in detail later) arises because the proper description of electronic states near the localization threshold (the Mott transition) requires that one incorporates two-particle correlations into the quasiparticle states. The Hubbard split-band picture is only the first step in the proper description of the electron states. Those additional correlations will lead to a very heavy mass of quasiparticles near the Mott transition; the heavy mass indicates a strong reduction of the bare bandwidth $W$ as the localization threshold is approached from the metallic side.

## 3. Localized versus Itinerant Electrons: Metal–Insulator Transitions

The Hubbard split–band picture of unpaired electronic states in a narrow band, shown in Figs. 7 and 8, provides a rationale for the existence of a paramagnetic insulating ground state of the interacting electron system. The corresponding experimentally observed metal–insulator transitions (MITs) at a finite temperature are very spectacular, as demonstrated in Fig. 10, where the resistivity (on a logarithmic scale) is plotted as a function of the inverse temperature for a canonical system $(V_{1-x}Cr_x)_2O_3$ (the data are from Ref. 27). The number of transitions (one, two, or three) depends on the Cr content. Note the presence of an intervening metallic state between the antiferromagnetic insulating (AFI) and the paramagnetic insulating (PM) states, as well as the reentrant metallic behavior at high temperatures for $0.005 \lesssim x \lesssim 0.0178$. To rationalize these data, we discuss the physical implications of a model of interacting narrow-band electrons for $U \sim W$ starting from the Hamiltonian [Eq. (25)]. We summarize here the main features of the detailed discussion presented in Refs. 28–30, which provide the main features of the ground-state and thermodynamic properties.

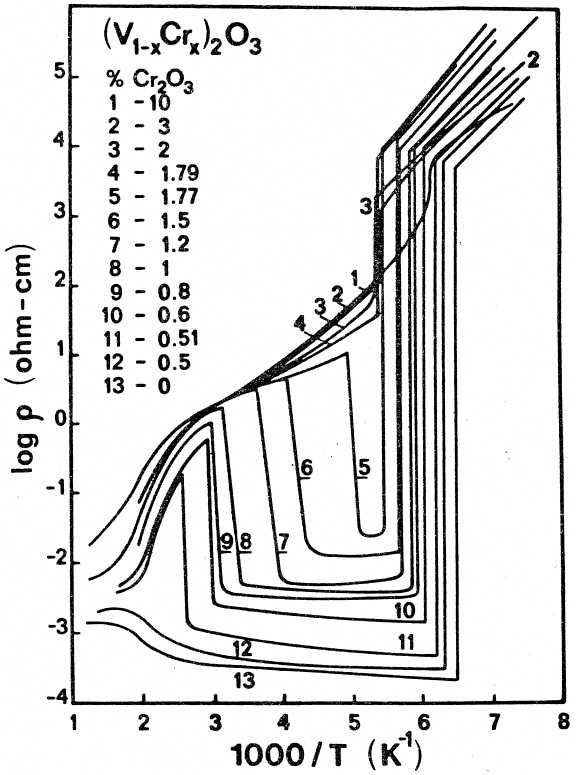



**FIGURE 10** Experimental measurements [27] pertaining to the variation of resistivity $\rho$ on a logarithmic scale with inverse temperature $1000/T$ for the $(V_{1-x}Cr_x)_2O_3$ system. The atomic content of $Cr_2O_3$ in $V_2O_3$ for each curve is specified.

In the absence of interactions ($U = 0$), the band energy per particle is $\bar{\epsilon} = -(W/2)n(1-n)/2)$, where $0 \le n \le 2$ is the degree of band filling; for $n = 1$, this reduces to $\bar{\epsilon} = -W/4$. When the interactions are present, the band narrows; this is because of a restriction on the electron motion caused by their repulsion, as described earlier. One way of handling this restriction is to adjoin to the bare bandwidth a multiplying factor $\Phi$. This leads to a renormalized DOS for quasiparticles, as illustrated in Fig. 11. The factor $\Phi$ is a function of the particle–particle correlation function $\eta \equiv \langle n_{i\uparrow} n_{i\downarrow} \rangle$, the expectation value for the double occupancy of a representative lattice site. The quantity $\eta$ is calculated for $T = 0$ self-consistently by minimizing the total energy $E_G$ (per site), composed of the band energy $E_B = \Phi\epsilon$ and the coulomb repulsion energy $U\eta$, where the parameter is specified by $\Phi = 8\eta(1-2\eta)$ [28, 29]. These two energies represent the expectation values of the two terms in Eq. (25) for the case

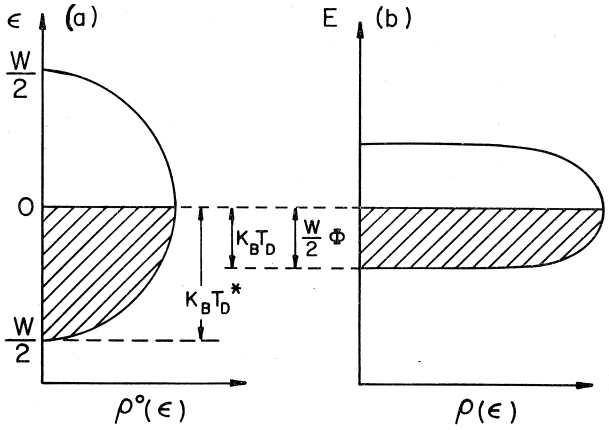



**FIGURE 11** Schematic representation of the bare ($\rho_0$) and quasiparticle ($\rho$) densities of states. The band narrowing factor $\Phi$ for interacting electrons (b) is specified. The degeneracy temperature $T_D$ for the interacting electrons and that corresponding to noninteracting electrons ($T_D^*$) are also indicated. The situation drawn corresponds to the half–filled case ($n = 1$), for which the Fermi energy can be chosen as $\epsilon_F = 0$.

of a half-filled band. The optimal values of the quantities are given by

$$\eta_0 = \frac{1}{4}(1 - (U/U_c)), \tag{29a}$$

$$\Phi_0 = 1 - (U/U_c)^2, \tag{29b}$$

and

$$E_G = \left(1 - (U/U_c)^2\right)\bar{\epsilon}, \tag{29c}$$

with $U_c = 8|\bar{\epsilon}| = 2W$. Thus, as $U$ increases, $\eta_0$ decreases from $\frac{1}{4}$ to 0. At the critical value $U = U_c$, $E_B = 0$ and there are no double occupancies for the same lattice site; this signals the crossover by the system from the itinerant (band) to the localized (atomiclike) state. The point $U = U_c$ corresponds to a true phase transition at $T = 0$; the last statement can be proved by calculating the static magnetic susceptibility, which is [29]

$$\chi = \chi_0 \Big/ \left\{\Phi_0\left[1 - U\rho\frac{(1+I)/2}{(1+I)^2}\right]\right\}, \tag{30}$$

where $I \equiv U/U_c$, $\rho$ is the density of bare band states at $\epsilon = \epsilon_F$, and $\chi_0$ is the magnetic susceptibility of band electrons with energy $\epsilon_{\mathbf{k}}$ at $U = 0$. As $\Phi \to 0$ (that is, $U \to U_c$), the susceptibility diverges. The localized electrons are represented in this picture by noninteracting magnetic moments for which the susceptibility is given by the Curie law $\chi = C/T \to \infty$ as $T \to 0$. Thus, the MIT is a true phase transition; $\eta_0$ may be regarded as an order parameter, and the point $U = U_c$ as a critical point. We concentrate now on a more detailed description of the metallic phase, which permits a generalization of the previous results to the case $T > 0$. First, as has been said, the increase in magnitude of interaction $U$ reduces the band energy according to $E_B = -W\Phi_0/4$. Eventually, $E_B$ becomes comparable to the interaction part $U\eta$; they exactly compensate each other at $U = U_c$. The resultant electronic configuration (localized versus itinerant) is then determined at $T > 0$ by the very low entropy and the exchange interaction contributions. The entropy of the metallic phase in the low–temperature regime may be estimated by using the linear specific heat expression for electrons in a band narrowed by correlations, namely, $C_v \equiv \gamma T = (\gamma_0 \mid \Phi_0)T$, where $\gamma_0 = 2\pi^2 k_B^2 \rho/3$ is the linear specific heat coefficient (per one atom) for uncorrelated electrons (that is, $U = 0$). Hence, the entropy $S = \gamma T = C_v$. Combining this relation with the resultant energy at $T = 0$, given by Eq. (29c), one can write an explicit expression for the free energy of the metallic phase [30]:

$$\frac{F}{N} = \left(1 - \frac{U}{U_c}\right)^2 \bar{\epsilon} - \frac{1}{2}\frac{\gamma_0}{\Phi_0}T^2. \tag{31}$$

This is the free energy per one atomic site. On the other hand, if the exchange interaction between the localized moments is neglected, then each site in the paramagnetic state is randomly occupied by an electron with its spin either up or down. The free energy $F_I$ for such an insulating system of $N$ moments is provided by the entropy term for randomly oriented spins, that is,

$$\frac{F_I}{N} = -k_B T \ell n\, 2 \tag{32}$$

Now, a system in thermodynamic equilibrium assumes the lowest $F$ state. The condition for the transition from the metallic to the local-moment phase is specified by $F = F_I$. The phase transition determined by this condition can be seen explicitly when we note that the free energy varies with $T$ either parabolically [Eq. (31)] or linearly [Eq. (32)], depending on whether the system is a paramagnetic metallic (PM) or a paramagnetic simulating (PI) phase. As illustrated in Fig. 12, several of those curves

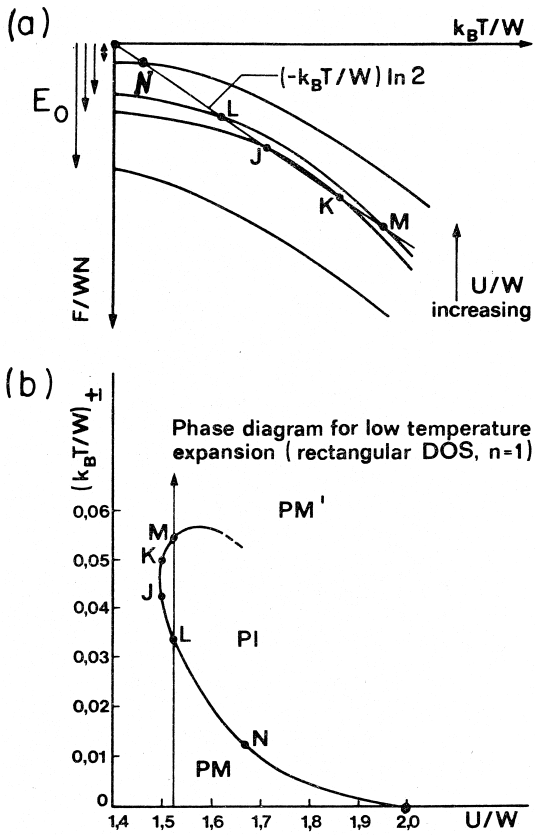



**FIGURE 12** (a) Plots of the free energies for the paramagnetic Mott insulator (the straight line starting from the origin) and the correlated metal (the parabolas). The parabolic curves' points of crossing at L and J correspond to a discontinuous metal–insulator transition, while those crossing at K and M correspond to the reverse. (b) Schematic representation of the phase diagram between paramagnetic metallic (PM and PM′ and paramagnetic insulating (PI) phases. The points of crossing from a are also shown. The vertical arrow represents a sequence of the transitions shown in Fig. 10 for $0.005 \lesssim x \lesssim 0.018$ and in the paramagnetic phase.

intersect at one or two points depending on the value $U/W$. These intersection points determine the stability limits of the PM and PI phases. The lowest curve for the PM phase lies below the straight line for the PI state; there is no transition, that is, the metallic Fermi liquid state with the effective mass enhancement $m^*/m_0 \lesssim 2.5$ is stable at all temperatures. As $U/W$ increases, the parabolas fall higher on the free energy ($F/WN$) scale and the possibilities for transitions open up. The higher two curves illustrate the case in which the intersections with the straight lines occur at J and K and at L and M, respectively; at low and high temperatures the parabola lies below the straight line for $F_I/WN$, so that the metallic phase is stable in those $T$ regions. At intermediate temperatures, the PI phase is stable. The loci of the intersections move farther apart on the $k_BT/W$ scale as $U/W$ is increased, as shown in Fig. 12b, where the phase boundaries are drawn; this part of the figure represents the temperature of the transitions (the intersection points in Fig. 12a) versus the relative magnitude of interaction $U/W$. We see that the PM phase is stable at low temperatures; thus, reentrant metallic behavior is encountered at high $T$. The explicit form of (the curve in Fig. 12b) is obtained from the coexistence condition $F = F_I$, which leads to the following expression for the transition temperatures [29]:

$$\frac{k_BT_+}{W} = \frac{3}{2\pi^2}\left[1-\left(\frac{U}{U_c}\right)^2\right]\left\{\ell n\, 2 \pm \left[(\ell n\, 2)^2 + \frac{\pi^2}{3}\frac{1-(U/U_c)}{1+(U/U_c)}\right]^{\frac{1}{2}}\right\}. \quad (33)$$

The root $T_-$ represents the low-temperature part, that is, that for $k_BT/W \leq 0.049$. The $T_+$ part is the one above the point where both curves meet; this takes place at the lower critical value of $U = U_{\ell c}$ such that

$$\frac{U_{\ell c}}{U_c} = 1 - \frac{3\sqrt{2}}{2\pi}\frac{1}{(\rho|\bar{\epsilon}|)^{\frac{1}{2}}} \approx 0.75. \quad (34)$$

Below the value of $U = U_{\ell c}$, the correlated Fermi liquid is stable at all temperatures. Ultimately, for $1.58 \leq U/W \leq 2.0$, only one intersection (at low $T$) of the curves remains. This means that in this regime of $U/W$ the reentrant metallic behavior is achieved gradually as the temperature increases. The above-described transitions are observed when changing the magnitude of interaction ($U/W$ ratio). In the case of high-temperature superconductors we observe the transition from a Mott insulator to a superconductor as a function of doping (carrier concentration). This case is discussed next.

*Note added in August 2000.* In recent years, the Mott localization in the limit of infinite dimension has been discussed extensively (e.g., Gebhardt, Ref. 30). A central peak is located between the Hubbard subbands, which carries the main part of the quasiparticle weight. There are two problems with the application of this solution to concrete systems. First, the upper critical dimensonality is not known for Mott systems. Second, the disappearance of the central peak at the localization threshold is being debated.

## 4. Strongly Correlated Electrons: Kinetic Exchange Interaction and Magnetic Phases in Three-Dimensonal Space

In the limit $W \ll U$, the ground state of the interacting electron system will be metallic only if the number of electrons $N_e$ in the system differs from the number $N$ of atomic sites. Simply, only then can charge transport take place via the hole states in the lower Hubbard subband (for $N_e < N$), that is, when the transport of electrons can be represented via hopping from site to site, avoiding the doubly occupied configurations on the same site. This restriction on the motion of individual electrons is described above in terms of the band narrowing factor $\Phi$, which, in the normal phase, is now of the order [28, 29] $\Phi = (1-n)/((1-n)/2)$. This shows that the effective quasiparticle bandwidth $W^* \equiv W\Phi$ is nonzero only if the number of holes $\delta \equiv 1-n > 0$.

For $W \ll U$, there is one class of dynamic processes that is important in determining the magnetic interactions between strongly correlated itinerant electrons, namely, the virtual hopping processes, with the formation of a doubly occupied site configuration in the intermediate state. Such processes are depicted in Fig. 13, where one electron hops onto the site occupied by an electron with opposite spin and then hops back to the original site. During such processes, the electrons can exchange positions (and the yields to the spin reversal of the pair with respect to the original configuration) or the same electron can hop back and forth. The corresponding effective Hamiltonian, including the virtual-hopping processes in first nontrivial order, has the form

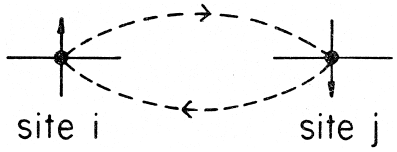



**FIGURE 13** Virtual hopping processes between singly and doubly occupied atomic sites that lead to an antiferromagnetic exchange interaction between the neighboring sites. This interaction is responsible for the antiferromagnetism in most of the Mott insulators.

$$H = \sum_{\mathbf{k}\sigma} \Phi_\sigma \epsilon_{\mathbf{k}} n_{\mathbf{k}\sigma} + \sum_{ij} \left(2t_{ij}^2/U\right)\left(\mathbf{S}_i \cdot \mathbf{S}_j - \frac{1}{4} n_i n_j\right), \tag{35}$$

where in general the band narrowing factor $\Phi_\sigma = (1-n)/(1-n_\sigma)$, $n_\sigma = \langle n_{i\sigma}\rangle$ is the average number of particles per site with the spin quantum number $\sigma$, and $n_i \equiv n_{i\uparrow} + n_{i\downarrow}$ is the operator of the number of particles on given site $i$. Note that in the paramagnetic state $n_\sigma = n_{-\sigma} = n/2$, and $\Phi_\sigma$ reduces to $\Phi = (1-n)/((1-n)/2)$, the value for the normal state.

One should note that the effective Hamiltonian [Eq. (35)] represents approximately the original Hubbard Hamiltonian for $W \ll U$ (for more precise treatment, see Ref. 31 and Section IV.A). When $n \to 1$, $\phi \to 0$, and Eq. (35) reduces to the Heisenberg Hamiltonian with antiferromagnetic interaction, which is the reason why most Mott insulators order antiferromagnetically. In the limit of a half–filled band, we also find that the effective bandwidth $W^* \equiv W\Phi = 0$, thus proving that the electrons in that case are localized on atoms. The nature of the wave function for these quasi–atomic states has not yet been satisfactorily analyzed, though some evidence given later shows that they should be treated as soliton states.

For $n < 1$, the normalized band (the first term) and the exchange parts in Eq. (35) do not commute with each other. This means that for the narrow-band system of electrons represented by the spin dynamics influences the nature of itinerant quasiparticle states of energies $\Phi\epsilon_{\mathbf{k}}$. What is even more striking is that, as $n \to 1$, the two terms in Eq. (35) may contribute equally to the total energy. The critical concentration of electrons $n_c$ for which these two terms are comparable is

$$n_c \simeq 1 - \frac{1}{2z}\frac{W}{U} \sim 0.02 \div 0.05.$$

In Fig. 14, we have plotted schematically the commonly accepted phase diagram for three dimensional systems describing the possible magnetic phases on the plane $n - (U/W)$. Close to the case of one electron per atom, the antiferromagnetic (AF) phase is stable for any arbitrary strength of interaction. At intermediate filling, the ferromagnetic (F) phase may be stable. On the low-interaction side ($W/U > 1$), the ferromagnetic phase terminates at points where the Stoner criterion is met, that is, when $\rho^0(\epsilon_F)U = 1$, where $\rho^0(\epsilon_F)$ is the value of the bare density of states (per spin) at the Fermi level $\epsilon_F$.

Peculiar features appear in the corner where $n \approx 1$, and $W/U \ll 1$, that is, where the number of holes is small, so that the exchange interaction contribution to the total system energy is either larger than or comparable to the band energy part $\Phi\bar{\epsilon}$. In such a situation, a mixed ferromagnetic–antiferromagnetic phase is possible [32]. When the number of holes is very small, each hole may form a magnetic polaron with a ferromagnetic cloud accompanying it: the hole is self-trapped within the cloud of ferromagnetic polarization it created. We consider those objects next.

## B. Magnetic Polarons

### 1. The Classical Approach

It has been proved by Nagaoka [32] that in the limit $W/U \to 0$ the ground state of the Mott insulator with one hole involves ferromagnetic ordering of spins. This is because in this limit the antiferromagnetic exchange term

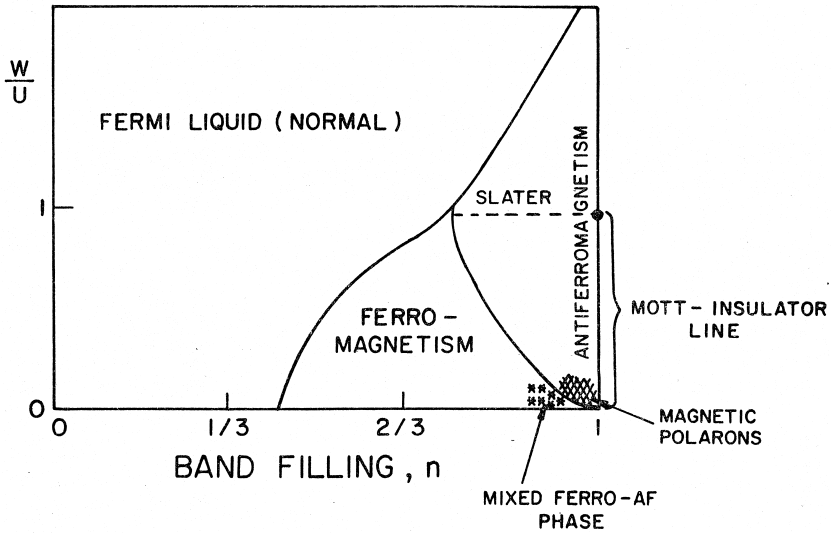



**FIGURE 14** Commonly accepted magnetic phase diagram for strongly correlated electrons on the $n$–($W/U$) plane.

in Eq. (35) vanishes and the band energy is lowest when $\Phi_{\sigma=\uparrow} = 1$ and $\Phi_{\sigma=\downarrow} = 1 - n$. We can thus choose an equilibrium state with $n_\uparrow = n$, $n_\downarrow = 0$, that is, a state with all spins pointing up.

Mott, and Hertier and Lederer [23], has pointed out that if $W/U$ is small but finite, a hole may create locally a ferromagnetic polarization of the spins in a sphere of radius $R$, surrounded by a reservoir of antiferromagnetically ordered spins. The situation is shown schematically in Fig. 15. The energy of such a hole accompanied by a cloud of saturated polarization can be estimated roughly as

$$E(R) = -\frac{W}{2} + \pi^2 |t| \left(\frac{a}{R}\right)^2 + \frac{4\pi}{3}\left(\frac{R}{a}\right)^3 \frac{zt^2}{U}, \quad (36)$$

where $a$ is the lattice constant and $t$ is the hopping integral $t_{ij}$ between the $z$ nearest neighbors. In this expression, the first term is the band energy of a free hole in a completely ferromagnetic medium, the second represents the kinetic energy loss due to the hole confinement, and the third involves the antiferromagnetic exchange energy penalty paid by polarizing the spins ferromagnetically within a volume $4\pi R^3/3$. Minimizing this equation with respect to $R$, we obtain the optimal number of spins contained in the cloud,

$$N = \frac{4\pi}{3}\left(\frac{\pi U}{W}\right)^{3/5}, \quad (37a)$$

and the polaron energy,

$$E_0 = -\frac{W}{2}\left[1 - \frac{5\pi^2}{3z}\left(\frac{W}{U}\right)^{2/5}\right]. \quad (37b)$$

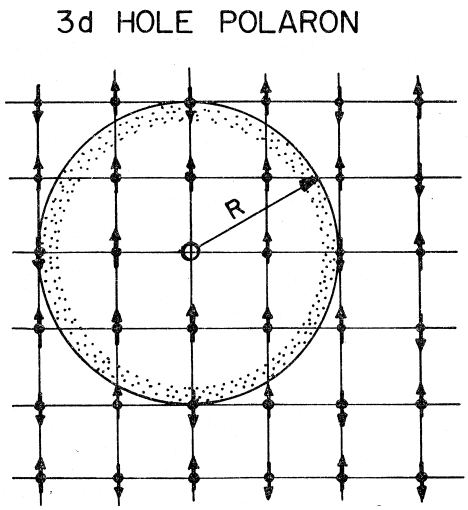



**FIGURE 15** Representation of the magnetic polaron state, that is, one hole in the antiferromagnetic Mott insulator. This hole produces ferromagnetic polarization around itself and may become self-trapped.

Equation (36) holds for a three-dimensional system; for a planar system, the factor $(4/3)\pi(R/a)^3$ in the last term should be replaced by the area $\pi(R/a)^2$. One then obtains the corresponding optimal values,

$$N = \pi\left(\frac{2\pi U}{W}\right)^{\frac{1}{2}} \quad (38a)$$

and

$$E_0 = -\frac{W}{2}\left[1 - \frac{2\pi^2}{z}\left(\frac{W}{2\pi U}\right)^{\frac{1}{2}}\right]. \quad (38b)$$

These size estimates will be needed later when discussing the hole states at the threshold for the transition from antiferromagnetism to superconductivity in high-$T_c$ oxide materials. One should note that $U/W$ must be appreciably larger than unity to satisfy the requirement $N \gg 1$. In other words, the condition $R \gg a$ must be met, so that the spin subsystem (and the hole dynamics) may be treated in the continuous–medium approximation, the condition under which Eq. (36) can be derived.

### 2. The Quantum Approach: Two Dimensions

The motion of a single hole in the Mott insulator is much more subtle than the formation of the polaron discussed above. Namely, if we consider $n$ holes in the lower Hubbard subband, then the probability of electron hopping around is $\approx n(1-n)$, so effectively, the bandwidth of such itinerant states is $W_{\text{eff}} = zt(1-n)$. For small $n$, we have $W \leq J$, where $J$ is the magnitude of the kinetic exchange. In the limit of a single hole the dynamics is determined by the magnitude of exchange interactions $J$, since $W_{\text{eff}} \to 0$. In effect, we have a hole moving slowly in the background of antiferromagnetically ordered spins. This picture seems to be a good representation of the hole motion in highly insulating magnetic oxides such as NiO and CoO. Instead, in high-temperature superconductors individual polaronic states must overlap appreciably for ($n_c \sim 0.95$) when the magnetic insulator $\to$ metal transition takes place. Therefore, some sort of homogeneous state must be formed in the metallic phase. This is particularly so since high-temperature superconductors evolve from a *charge-transfer insulator*, for which the gap for $2p \to 3d^{n+1}(O^{2-} \to Cu^{1+})$ transitions is smaller than the Hubbard gap $\Delta \simeq U - W$. In effect, the hole states are hybridized *3d–2p* states (in proportions 2:1), not pure *3d* states due to copper ions. As a result, few alternative pictures of the fermionic liquid of strongly correlated electrons in the normal (metallic) phase arise, starting from the phenomenological pictures of marginal Fermi liquid (Varma *et al.* [33a]) and nearly antiferromagnetic Fermi liquid (NAFL) [33b, 33c] to a mean-field picture of

strongly correlated electrons coupled to a gauge field (cf. Lee and Nagaosa [33d]). A separate class of models composes cluster calculations including realistic structures of $CuO_2$ planes (cf. Ref. 33e). Another class of models forms bosonic models with preformed pairs of bound bipolarons [33f]. The latter class of models requires that the bipolaron radius is $R < (a_0/x)^{\frac{1}{2}}$, where $x$ is the hole concentration, and $a_0$ is the Cu–Cu interatomic distance. This, in turn, requires a rather strong attractive interaction, which most probably can be furnished only by the combined effect of magnetic and rather strong electron–phonon interactions. Finally, there exists a substantial number of papers on numerical diagonalization for small clusters [33g]. A separate class of models are those involving the stripe structure [33h].

The lack of microscopic theory of normal properties transforms into the arbitrariness in selecting the pairing potential, as we shall see in the next section. The linear resistivity in the full temperature range at optimal doping [34a], the spin-gap existence [34b] in underdoped systems, and the anomalous (non-Drude) form of the optical conductivity all speak in favor of the non-Fermi-liquid (absence of quasiparticles) behavior of correlated electrons [34c] in two spatial dimensions. The role of disorder has not been explained properly either.

## 3. The Spin Liquid

The difference between an electron liquid of strongly correlated electrons (represented, for example, by the holes in the lowest Hubbard subband) and a Fermi liquid can be shown clearly in the limit of relatively high temperatures $W^* \ll k_B T \ll U$, where the quasiparticle band states with energies $(\Phi\epsilon_{\mathbf{k}})$ are populated equally, independent of their energy. Namely, if $N_e$ electrons are placed into $N$ available states of almost the same energy, then the number of configurations for a phase with excluded double occupancies of each state is [34]

$$2^{N_e}\frac{N!}{N_e!(N-N_e)!}. \tag{39a}$$

The first factor is the number of spin configurations for the singly occupied sites, while the second specifies the configurational entropy—the number of ways to distribute $N_e$ spinless particles among $N$ states. This leads to molar entropy in the form

$$S_L = R[n\,\ell n\,2 - n\,\ell n\,n - (1-n)\,\ell n(1-n)], \tag{39b}$$

where $n = N_e/N$ is the degree of subband filling and $R$ is the gas constant. The above reduces to $S_L = R\,\ell n\,2$ for $n = 1$, that is, to the entropy of the $N$ spins $(\frac{1}{2})$ on the lattice. In contrast, in a Fermi liquid that obeys the Fermi–Dirac distribution, double occupancies are not excluded,

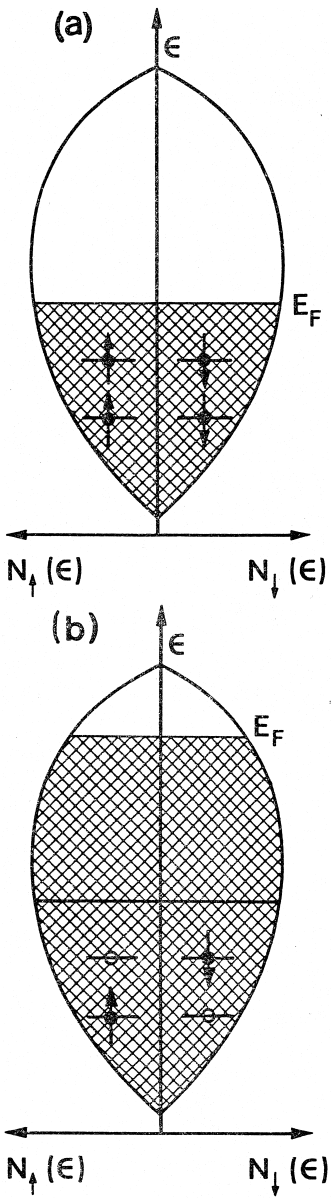



**FIGURE 16** Schematic representations of the difference in the **k**–space occupation for ordinary fermions (a) and strongly correlated electrons (b) The spin subbands with $\sigma = \uparrow$ and $\downarrow$ are drawn. Note that the holes drawn in b do not appear; they are shown only to indicate the single occupancy of each single–particle state. The position of the Fermi level is different for the same number of electrons in the two situations. [From Ref. 34.]

as illustrated in Fig. 16a. The corresponding number of configurations is then

$$\left(\frac{N}{N_e/2}\right)^2 = \frac{N!}{(N_e/2)!(N-N_e/2)!}, \tag{40a}$$

with the corresponding molar entropy,

$$S_F = R[2\,\ell n\,2 - n\,\ell n\,n - (2-n)\,\ell n(2-n)]. \tag{40b}$$

Hence, for $n = 1$, $S_F = 2S_L = 2R\,\ell n\,2$. One should emphasize that only the value for $S_L$ reproduces correctly the entropy of $N$ localized paramagnetic spins (the electronic part of the entropy for magnetically disordered states of the Mott insulator). Hence, in accord with intuitive reasoning, the Fermi–Dirac distribution, which allows for double state occupancy, cannot be applied to a strongly correlated electron liquid, which we call a spin liquid. The state of such a liquid reduces to that of the spin system on the lattice if $N = N_e$ (for the Fermi-liquid case, the ground state is then a metal with a half–filled band).

One should now ask how these results may be generalized to handle the regime of low temperatures and of an arbitrary number of holes. One observes that in Fig. 8 the band states for $U \ll W$ are split for any arbitrary degree of band filling [cf. also Eq. (28b)]. Therefore, in enumerating the distribution of particles in the lower Hubbard subband, one must exclude double occupancies of the same energy ($\epsilon$) state. Since the quasiparticle energy is labeled by the wave vector $\mathbf{k}$, one can equivalently exclude the double occupancies of given state $|\mathbf{k}\rangle$. Under this assumption, the statistical distribution is given by [34]

$$\bar{n}_{\mathbf{k}\sigma} = (1 - \bar{n}_{\mathbf{k}-\sigma})\frac{1}{1 + \exp[\beta(E_{\mathbf{k}\sigma} - \mu)]}, \tag{41a}$$

where $\beta = (k_B T)^{-1}$, $\bar{n}_{\mathbf{k}\sigma}$ is the average occupancy of the state $|\mathbf{k}\sigma\rangle$, and $\mu$ is the chemical potential that is determined from the conservation of the total number of particles

$$N_e = \sum_{\mathbf{k}\sigma} \bar{n}_{\mathbf{k}\sigma}. \tag{41b}$$

The corresponding molar entropy is now given by

$$S_L = -\frac{R}{N}\sum_{\mathbf{k}}[(1 - \bar{n}_{\mathbf{k}})\,\ell n(1 - \bar{n}_{\mathbf{k}}) + \bar{n}_{\mathbf{k}\uparrow}\ell n\,\bar{n}_{\mathbf{k}\uparrow} + \bar{n}_{\mathbf{k}\downarrow}\,\ell n\,\bar{n}_{\mathbf{k}\downarrow}], \tag{41c}$$

with $\bar{n}_{\mathbf{k}} = \bar{n}_{\mathbf{k}\uparrow} + \bar{n}_{\mathbf{k}\downarrow}$.

One should note that the distribution function [Eq. (41c)] differs from the ordinary Fermi–Dirac formula by the factor $(1 - \bar{n}_{\mathbf{k}-\sigma})$, which expresses the conditional probability that there should exist no second particle with the spin quantum number $\mathbf{k}(-\sigma)$ if the state $\mathbf{k}\sigma$ is to be occupied by an electron, as shown in Fig. 16b. If $E_{\mathbf{k}\sigma} \equiv E_{\mathbf{k}}$ (that is, when the particle energy does not depend on its spin direction), Eq. (41a) reduces to

$$\bar{n}_{\mathbf{k}} = \frac{1}{1 + \left(\frac{1}{2}\right)\exp[\beta(E_{\mathbf{k}} - \mu)]}. \tag{41d}$$

This is the same type of formula that applies to the occupation number of simple donors, if the index $\mathbf{k}$ is dropped and $\epsilon$ represents the position of the donor level with respect to the bottom edge of the conduction band.

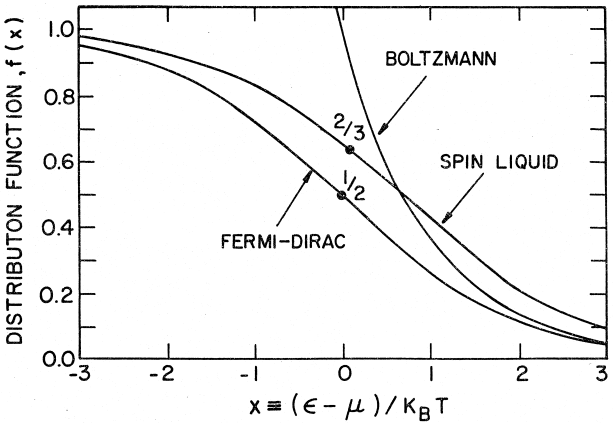



**FIGURE 17** Comparison of the Fermi–Dirac and Boltzmann distributions for $\bar{n}_{\mathbf{k}\sigma}$ with that for strongly correlated electrons (the spin—liquid phase); the total occupancy $n_{\mathbf{k}} = n_{\mathbf{k}\uparrow} + n_{\mathbf{k}\downarrow}$ is taken in the latter case.

At $T = 0$, each state is singly occupied. This is the principal feature by which the present formula differs from the Fermi–Dirac distribution at $T = 0$, as illustrated in Fig. 17. The distribution [Eq. (41a)] leads to a doubling of the volume enclosed by the Fermi surface in the spin-liquid state compared to the Fermi-liquid state. At low temperatures, application of the distributions [Eq. (41a) or (41c)] yields Fermi liquid-like properties: a linear $T$ dependence of the specific heat (of large magnitude if $n \to 1$) of the entropy. At high temperatures, the new distribution leads to entropy of the form of Eq. (39b) and local–moment behavior in the form of the Curie–Weiss law for susceptibility. Hence, the properties of the spin liquid governed by the distribution [Eq. (41a) or (41d)] interpolate between those of a metal and those of local moments. Such behavior is observed in many correlated systems, for example, in heavy fermions.

One should note that the entropy expression [Eq. (41c)] can be rewritten for the paramagnetic state in the following form:

$$S_L = -nR\,\ell n\,2 - k_B\sum_{\mathbf{k}}[n_{\mathbf{k}}\,\ell n\,n_{\mathbf{k}} + (1 - n_{\mathbf{k}})\,\ell n(1 - n_{\mathbf{k}})]. \tag{42}$$

The first part represents the entropy of spin moments; the second, the entropy of spinless fermions. An alternative decomposition has been put forward [35] in which the dynamics of correlated electrons is decomposed into that of neutral fermions called spinons and the charged bosons called holons. Within this picture, the onset of superconductivity is considered as a combined effect of Bose condensation of the holons with the simultaneous formation of a coherent paired state by the fermion counterpart [36–38]. This problem is discussed in more detail in Section IV.A.

The above treatment of the spin liquid deals only with its statistical properties in the $U \to \infty$ limit. The problem

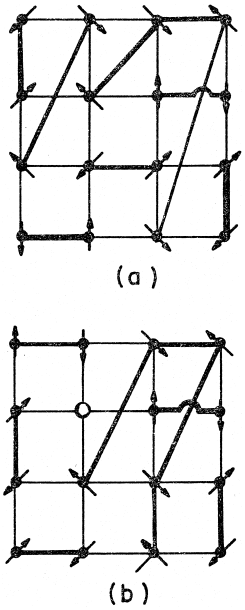



**FIGURE 18** Schematic representation of singlet–spin pairing forming the RVB state. All paired configurations should be taken to calculate the actual ground state. (a) The RVB state for the Mott insulator; (b) that with one hole. The latter case will contain an unpaired spin, as indicated.

now arises as to what happens when the spin part of the form of the second term in Eq. (35) is explicitly included. The problem of the resultant quantum ground state of holes in a Mott insulator is a matter of intensive debate [36–38]. The state called the resonating valence-bond (RVB) state has been involked [36] specifically to deal with this problem; this state is shown schematically in Fig. 18 for the case without holes (a) and with one hole (b). The connecting lines represent bonds, across which the two electron from spin–singlet pairs. The resonating nature of bonds is connected with the idea that the RVB ground state is a coherent superposition of all such paired configurations. The dynamic nature of this spin dimerization is connected with the terms $(S_i^+ S_j^- + S_i^- S_j^+)$ in the exchange part of the Hamiltonian [Eq. (35)]. There is the possibility that the RVB state [which, for obvious reasons, differs from the ordinary (Néel) antiferromagnet] is a ground state for the planar $CuO_2$ planes in high-$T_c$ oxides, such as $La_{2-x}Sr_xCuO_4$, where the long–range magnetic order is destroyed for $x \approx 0.02 \div 0.03$. We return to this problem in Section IV when discussing the boundary line between antiferromagnetism and superconductivity for high-$T_c$ oxides.

## C. Hybridized Systems

Most of the strongly correlated systems are encountered in oxides and in several classes of organic and inorganic compounds. In oxides the *3d* orbitals of cations such as $Cu^{2+}$ and $Ni^{2+}$ hybridize with the *2p* orbitals of oxygen, particularly if the atomic *3d* states are energetically close to the *2p* states. The properties of correlated and hybridized states can be properly discused in terms of the Anderson lattice model Hamiltonian, which is of the form

$$H = \epsilon_f \sum_{i\sigma} N_{i\sigma} + \sum_{\mathbf{k}\sigma} \epsilon_{\mathbf{k}} n_{\mathbf{k}\sigma} + U \sum_i N_{i\uparrow} N_{i\downarrow} + \frac{1}{\sqrt{N}} \sum_{\mathbf{k}\sigma} \left( V_{\mathbf{k}} e^{i\mathbf{k}\cdot\mathbf{R}_i} a_{i\sigma}^+ c_{\mathbf{k}\sigma} + \text{H.C.} \right). \quad (43)$$

In this Hamiltonian, the first term describes the energy of atomic electrons positioned at $\epsilon_f$, the second represents the energy of band electrons, the third represents the intraatomic coulomb repulsion between two electrons of opposite spins, and the last describes the mixing of atomic with band electrons due to the energetic coincidence (degeneracy) of those two sets of states (H.C. refers to the Hermitian conjugate part of the hybridization part). In heavy fermions, the atomic states are $4f$ states, whereas they are *3d* states of $Cu^{2+}$ ions in high-$T_c$ systems; the band states are *5d–6s* and *2p* states, respectively. Note that $N_{i\sigma} = a_{i\sigma}^+ a_{i\sigma}$ and $n_{\mathbf{k}\sigma} = c_{\mathbf{k}\sigma}^+ c_{\mathbf{k}\sigma}$ are the number of particles on given atomic $(i)$ or $\mathbf{k}$ states, respectively. In this Hamiltonian, the following parameters appear: the atomic–level position $\epsilon_f$, the width $W$ of starting band states with energies $\{\epsilon_{\mathbf{k}}\}$, the magnitude $U$ of the coulomb repulsion for two electrons located in the same atomic site, and the degree of hybridization (mixing), $V_{\mathbf{k}}$, characterized by its magnitude $V$.

Two completely different situations should be distinguished from the outset: (1) $U > W > |\epsilon_f| \gg |V|$, and (2) $U > W > |V| \gtrsim |\epsilon_f|$. Case 1 applies when the starting (bare) atomic level is placed deeply below the Fermi level and the atomic states admix weakly to the band states. In case 2, the hybridization is large and is responsible for strong mixing of the two starting sets of states. The band structure corresponding to the hybridized band states in the absence of electron–electron interactions (that is, $U = 0$) is depicted in Fig. 19. We observe a small gap in the hybridized band structure; it occurs around the bare atomic level position $\epsilon_f$ and separates two hybridized bands. Those two bands, which have the energies

$$E_{\mathbf{k}\pm} = \frac{\epsilon_{\mathbf{k}} + \epsilon_f}{2} \pm \left[ \left( \frac{\epsilon_{\mathbf{k}} - \epsilon_f}{2} \right) + |V_{\mathbf{k}}|^2 \right],$$

correspond to the bonding and antibonding types of states in molecular systems. The structure of the hybridized bands is demonstrated explicitly in Fig. 20 of the DOS for each band. One sees that strongly peaked structures occur in the regions near the gap. If the Fermi level falls within these peaks, a strong enhancement of the effective mass

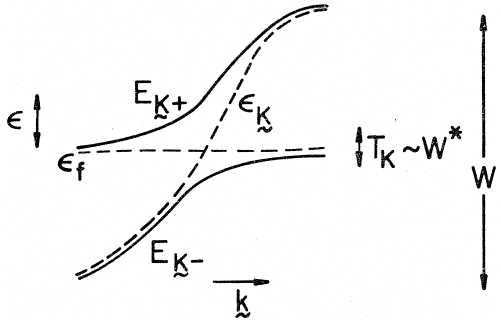



**FIGURE 19** Schematic representation of the hybridized bands with energies $E_{\mathbf{k}\pm}$, which are formed by mixing the band states (with energy $\epsilon_{\mathbf{k}}$) and atomic states (located at $\epsilon = \epsilon_f$). The original band has width $W$, much wider than the peaked structure, of width $W^*$.

should takes place solely because of these peculiarities of the band structure. In some situations only a pseudogap caused by the hybridization is formed, as shown in Fig. 21. This is so if the hybridization matrix element $V$ depends on the wave vector $\mathbf{k}$ and if, along some directions in reciprocal space, $V_{\mathbf{k}} = 0$.

The inclusion of the interaction term in Eq. (43) renders the treatment of the Anderson lattice Hamiltonian much more complicated; up to now this problem has not been solved rigorously. A large variety of approximate treatments has been proposed and reviewed recently [39–42],

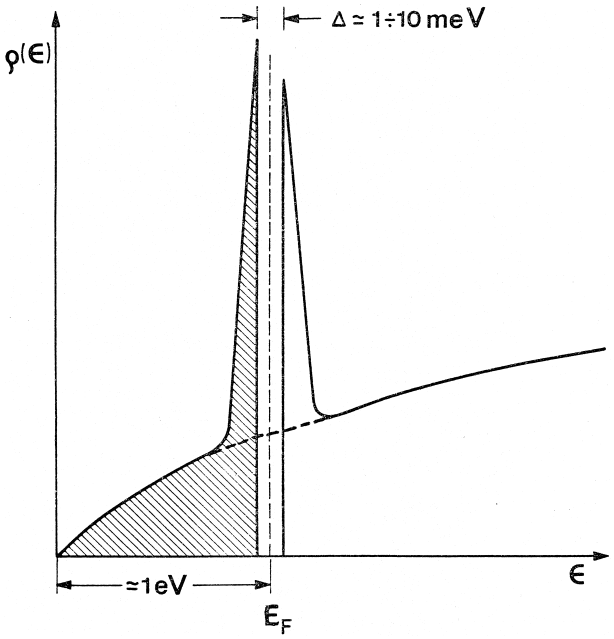



**FIGURE 20** Density $\rho(\epsilon)$ of hybridized states versus particle energy $\epsilon$. Note that the hybridization gap $\Delta_h$ may be very small compared to the total width of the band states. The position of the Fermi level $\epsilon_F$ corresponds to the filled lower band.

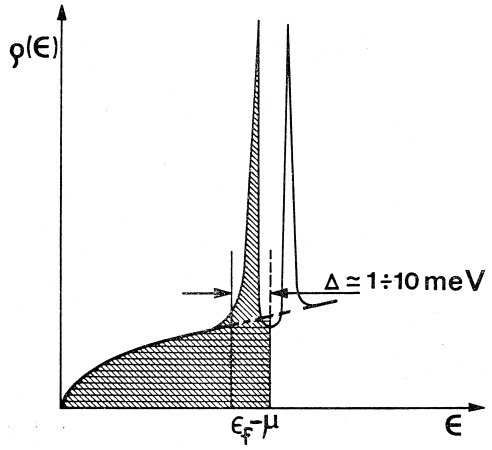



**FIGURE 21** Same as Fig. 20 but with the pseudogap among the hybridized bands.

in all of which the principal task was to provide a satisfactory description of heavy-fermion materials [43]. In effect, the limiting case of almost localized strongly correlated electrons was studied, which, among others, provides a quasiparticle electronic structure similar to that shown in Fig. 20, with a very strong enhancement of the DOS near the Fermi surface. This yields to very heavy quasiparticles, which, in some systems, may undergo transitions either to antiferromagnetism or to superconducting states. In this respect, heavy-fermion materials are analogous to high-$T_c$ systems, though with much lower transition temperatures.

## D. The Electronic States of Superconducting Oxides

The high-$T_c$ superconducting oxides, such as $La_{2-x}Sr_xCuO_4$ (the so-called 214 compounds) and $YBa_2Cu_3O_{7-\delta}$ (the so-called 123 compounds), have one common structural unit: the quasi–two–dimensional structure that is approximated by $CuO_2$ planes, one of which is shown schematically in Fig. 22. We discuss mainly the role of these planes since it is widely accepted that the electronic properties of these subsystems are the main factor determining the observed superconductivity, antiferromagnetism, and localization effects in those materials. In stoichiometric $La_2CuO_4$ or $YBa_2Cu_3O_7$, the formal valence of Cu is 2+, that is, it corresponds to a one-hole ($3d^9$) electron configuration. In a strictly cubic structure, with $Cu^{2+}$ surrounded by $O^{2-}$ ions in an octahedral arrangement, the highest band is doubly degenerate and of $e_g$ symmetry, that is, composed of $d_{x^2-y^2}$ and $d_{3z^2-r^2}$ orbitals. However, in high-$T_c$ materials, the octahedra are largely elongated in the direction perpendicular to the $CuO_2$ planes, so that the bands are further split; it is commonly assumed that the

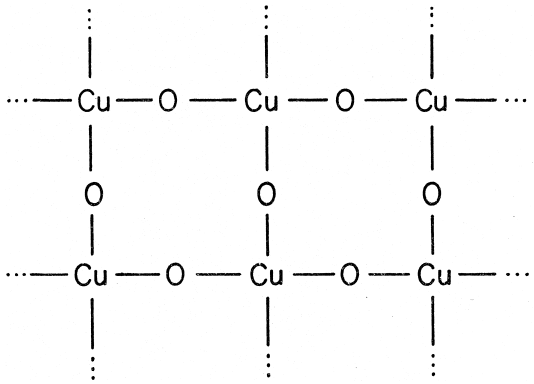



**FIGURE 22** Schematic representation of the $CuO_2$ planes in superconducting oxides in the tetragonal phase. The Cu–Cu distance is ≈1.9 Å for $La_2CuO_4$.

antibonding orbital $d_{x^2-y^2}$ is higher in energy and hence half–filled. These $d$ states hybridize with the oxygen $2p_x$ and $2p_y$ orbitals of $\sigma$ type, as shown schematically in Fig. 23; both the bonding and the antibonding configurations are shown; the latter corresponds to the signs of the two $p$ orbitals shown in parentheses.

A simple description of the electronic states for the planar $CuO_2$ system is obtained by introducing a single band representing Cu $d$ electrons in the tight-binding approximation. For the square configuration of the Cu atoms (which reflects the tetragonal structure of $La_2CuO_4$), such a dispersion of band energies has the form

$$\epsilon_{\mathbf{k}} = 2t(\cos k_x a + \cos k_y a), \tag{44}$$

where $t$ is the so–called hopping or Bloch integral $\langle i|V|j\rangle$ between the nearest neighboring ions $i$ and $j$, and $a$ is the Cu–Cu distance. For $La_2CuO_4$ and $YBa_2Cu_3O_{6.5}$, this band is half–filled, with the Fermi surface for bare (noninteracting) electrons determined from the condition $\epsilon_{\mathbf{k}} = \mu = 0$. As shown in Fig. 24, this leads to a square in reciprocal space connecting the points $(\pi/a)(\pm 1, 0)$ with the points $(\pi/a)(0, \pm)$. The oxygen electrons in the $2p$ states are regarded as playing only a passive role of a transmitter of the individual $d$ electrons from one $d_{x^2-y}$ state to its neighbor (note that the $O^{2-}$ valence state has completely filled $p$ shells). If the number of electrons in that band is decreased (for example, by substituting Sr for La in 214 compounds), then the Fermi surface shrinks and gradually transforms into a circle, as shown in Fig. 24 [44]. Within such a model, $La_2CuO_4$ should be metallic. However, at $T < T_N \simeq 240$ K, this compound orders antiferromagnetically [24], and the ground state is then insulating. The fact that this system remains insulating above the Néel temperature $T_N$ means that the stoichiometric $La_2CuO_4$ and $YBa_2Cu_3O_{6.5}$ are Mott insulators, not a Slater split–band antiferromagnet; for the latter, the split–band structure for $T < T_N$ should coalesce into one band as $T \to T_N$. The presence of the paramagnetic insulating state for both $La_2CuO_4$ and $YBa_2CuO_{6+\delta}$ supports the view that those oxides should be regarded as

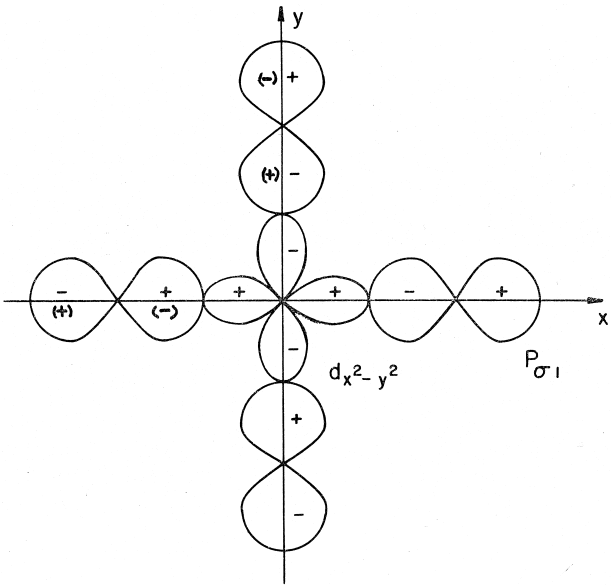



**FIGURE 23** The configuration of the $3d_{x^2-y^2}$ and $p_\sigma$ orbitals for bonding configurations. The reverse signs for the two $p$—orbitals (that is, those in parentheses) represent the hybridized configuration for the antibonding state.

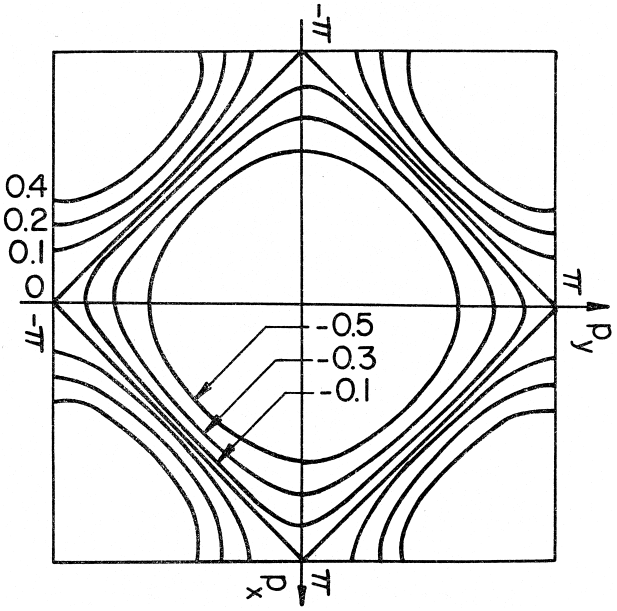



**FIGURE 24** The shape of the two—dimensional Fermi surface for band energy of the form of Eq. (44). The values specified represent $\mu/t$ as a parameter. The square shape corresponds to $\mu = 0$ or, equivalently, to $n = 1$. [From Ref. 44.]

narrow–band systems characterized by strong electron–electron interactions ($U > U_c$), as originally proposed by Anderson [45]. An antiferromagnetic ground state is then expected since the kinetic exchange interaction between the strongly correlated electrons takes place [31].

A principal problem appears when holes occur in the Mott insulator, that is, when we consider a real situation $La_{2-x}Sr_xCuO_4$ or $YBa_2Cu_3O_{6.5+x}$. We have already seen (at the end of Section III.A) that for small $x$ the kinetic energy of the holes and the exchange energy of electrons may become comparable or the latter may become even larger than the former. In such situations, the motion of the holes will be influenced by the setting-in of almost-instantaneous spin–spin correlations. This means that such metallic states (if formed) cannot be regarded as the Fermi liquid with slowly evolving spin fluctuations; instead, the resonance between various spin configurations must be built into the electron wave function characterizing its itinerant state. The decomposition of the resonating spin configurations into spin pair–singlet configurations constitutes an important characteristic of the RVB theory of the normal state [36, 45]. Some experimental evidence for the quantum spin–liquid state above the Néel temperature has been provided by neutron quasi–elastic scattering [46]; these results were subsequently interpreted [47].

The interpretation of the metallic state in terms of a single, narrow band requires the presence of both $3d^9$ ($Cu^{2+}$) as well as $3d^8$ ($Cu^{3+}$) states. Most of the X–ray spectroscopical studies [48] conclude that the satellite peak corresponding to a $3d^8$ configuration is actually absent. Therefore, to explain both the insulating properties of $La_2Sr_xCuO_4$ and the metallic properties of $La_{2-x}Sr_xCuO_4$ for $0.04 \div 0.05$, one introduces hybridized $2p$–$3d$ states for the holes introduced by the doping. The proper model of such states is then the Anderson lattice type of model [Eq. (43)]. Band–structure calculations by Mattheis [49] for $La_{2-x}Sr_xCuO_4$, shown in Fig. 25, justify a reasonable description within a simple two-dimensional tight-binding model with only the Cu $3d_{x^2-y^2}$ and $p_\sigma$ orbitals on oxygens taken into account. Namely, the structures denoted A and B in Fig. 25 correspond, respectively, to antibonding and bonding hybridized bands, with respective band energies

$$E_{\mathbf{k}\pm} = \frac{\epsilon_p + \epsilon_d}{2} \pm \left[ \left( \frac{\epsilon_p - \epsilon_d}{2} \right)^2 + 4V^2 \left( \sin^2 \frac{k_x a}{2} + \sin^2 \frac{k_y a}{2} \right) \right]^{\frac{1}{2}}, \quad (45)$$

where $\epsilon_p$ and $\epsilon_d$ are atomic level positions for the $3d$ and $2p$ states, respectively. Detailed calculations [49] lead to a nonzero bandwidth of the $2p$ band because of the $p$–$p$ overlap; then $\epsilon_p \to \epsilon_p + \epsilon_{\mathbf{k}}$, with

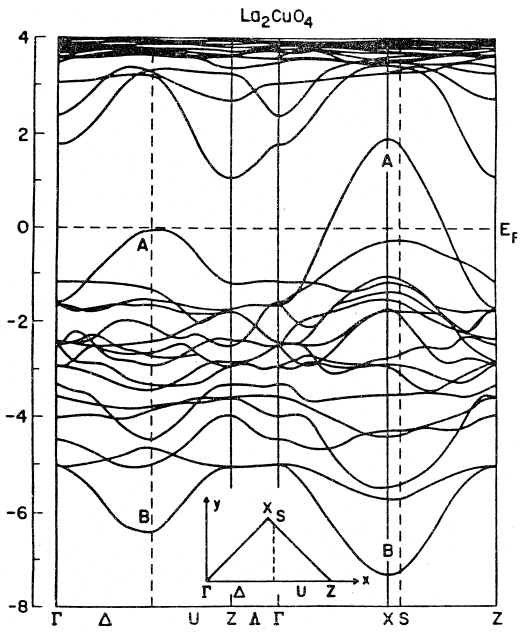



**FIGURE 25** Energy bands for $La_2Cu_4$ calculated within a local—density approximation for the assumed crystal structure are body—centered tetragonal. A portion of the $x$–$y$ plane in the extended Brillouin zone scheme is shown in the inset. Portions B and A correspond to the bonding and antibonding parts of the hybridized band discussed in the text. [From Ref. 49.]

$$\epsilon_{\mathbf{k}} = 2t_p[\cos(k_x a/\sqrt{2}) + \cos(k_y a/\sqrt{2})].$$

The band structure calculations should be regarded as providing input parameters for the parametrized models which include electron correlations more accurately. On the basis of various estimates [50] of those parameters, one can assume that they fall in the range

$$|\epsilon_p - \epsilon_d| \simeq 3.6\ \text{eV}, \quad |V| \approx 1.3 \div 1.5\ \text{eV}, \quad |t| \simeq 0.5\ \text{eV},$$
$$|t_p| \simeq 0.6\ \text{eV}, \quad \text{and} \quad U \simeq 8\text{–}10\ \text{eV}.$$

From these estimates of the parameters, one sees that $|V| \sim |\epsilon_p - \epsilon_d|$. Hence, one may not be able to use the perturbation expansion in $V/(\epsilon_p - \epsilon_d)$ of the Anderson lattice Hamiltonian. Such a perturbation expansion was used [51] when transforming the hybridized model represented by the $p$–$d$ Hamiltonian into an effective narrow–band model, which is represented by the effective Hamiltonian for the electrons in the $CuO_2$ plane,

$$H = t \sum_{i,j=\text{n.n},\sigma} \tilde{a}^+_{i\sigma} \tilde{a}_{j\sigma} + t' \sum_{i,j=\text{n.n.n},\sigma} \tilde{a}^+_{i\sigma} \tilde{a}_{j\sigma} + J \sum_{i,j=\text{n.n}} \left( \mathbf{S}_i \cdot \mathbf{S}_j - \frac{1}{4} \tilde{n}_i \tilde{n}_j \right),$$

where $t$ and $t'$ are the hopping integrals between the nearest (n.n.) and the next-nearest neighbors, $J$ is the value of exchange integral for the kinetic superexchange, and

the tilded operators mean that they are projected onto subspace of singly occupied lattices sites. The three parameters take the values $t \simeq -0.5$ eV, $t' \simeq 0.05$–$0.1$ eV, and $J \simeq 0.13$ eV.

On the basis of the facts that the present-day band calculations do not provide paramagnetic insulating states for stochiometric materials, such as $La_2CuO_4$, and that the antiferromagnetic ground state is difficult to achieve within the local–density approximation [49], we conclude that an approach based on the parametrized models discussed in the preceding two sections should be treated in detail. The microscopic parameters obtained from the band–structure calculations should be treated as input parameters in those models. A review of properties obtained within the parametrized models and relevant to high-$T_c$ systems is given in Section IV.

## IV. NOVEL MECHANISMS OF ELECTRON PAIRING

The binding of two fermions into either a bosonic or a more complicate bound state is a prerequisite for the condensation of microscopic particles into a coherent (superfluid) macroscopic state. This condensation may take the form of Bose–Einstein condensation if the interaction energy between the pairs is much lower than the binding energy of a single pair (also, the pairs must be well separated spatially). Such a Bose condensed state of charged particles may exhibit the principal properties of the superconducting state such as the Meissner–Ochsenfeld effect [52]. In the BCS theory (discussed in Section II), pair condensation occurs under a completely different condition, namely, when the states of different pairs overlap strongly so that the motion of one widely separated pair takes place in the mean field of almost all other pairs.

The pairing of particles in the BCS theory is described in momentum (reciprocal) space, where it is assumed that the quasiparticle states with a well-defined Fermi surface are formed first; the pairing involves electrons from the opposite points on the Fermi surface ($\mathbf{k}$, $-\mathbf{k}$) and generates either a simple spin–singlet state (as in the classic superconductors) or a higher angular-momentum state, e.g., $L = S = 1$ (as in superfluid $^3$He [53]). Because of the small coherence length ($\xi \sim 10$ Å), the new superconductors offer an opportunity for exploring the possibility of pairing in real (coordinate) space. Moreover, since the carrier concentration determined from the Hall-effect measurements [54] for high-$T_c$ oxides is at least one order of magnitude lower than that for ordinary metals, it is tempting to describe the onset of the superconducting state as a Bose condensation of preexisting pairs. In fact, the situation is not that simple. For example, in $La_{2-x}Sr_xCu_4$, with $x = 0.04$, the average distance between holes in the normal phase is $\approx 5a$, a magnitude comparable to $\xi$. These circumstances, combined with antiferromagnetism and localization effects, render the new superconducting materials unique in the sense that their description requires a unification of theoretical approaches to phenomena previously regarded as disparate.

The accumulated evidence for rather strong electron–electron interaction in high-$T_c$ oxides [36,48] and in heavy–fermion systems [39–43] makes it unlikely that electron pairing in these materials is caused by extremely strong electron–phonon interaction. Furthermore, the electron–phonon interaction does not allow for a connection (or, strictly speaking, competition) between the observed superconductivity and antiferromagnetism [24]. This is one of the reasons for an intensive search for a purely electronic mechanism of pairing. We now discuss some of the mechanisms that have been proposed. The main emphasis so far has been placed on an exchange–mediated pairing for strongly correlated electrons [45], since for such systems, the pairing, antiferromagnetism, and MITs to the Mott localized phase are derived from a single theoretical scheme. The latter two phases have been discussed in Section III; here, we concentrate on the spin–singlet pairing among strongly correlated electrons. Later we discuss charge transfer- and phonon-mediated pairings. Finally, we classify the types of correlated states and metallic states in solids. This classification provides a concise way of characterizing specific properties of these systems by which the almost-localized systems differ from ordinary metals.

### A. Exchange Interactions and the Real-Space Pairing

#### 1. Narrow–Band Systems

In Section III.A we provided an approximate Hamiltonian [Eq. (34)], which includes the antiferromagnetic exchange interactions between the correlated electrons in the limit $U/W \gg 1$. The precise form of this Hamiltonian to second order in $W/U$ is [31, 55]

$$H = \sum_{ij\sigma}{}' t_{ij} b_{i\sigma}^{+} b_{j\sigma} + \sum_{ij}{}' \frac{2t_{ij}^2}{U}\left(\mathbf{S}_i \cdot \mathbf{S}_j - \frac{1}{4}\nu_i \nu_j\right) + (\text{three–site terms}), \tag{46}$$

where the primed summation means that $i \neq j$. In this Hamiltonian, doubly occupied site $i$ configurations $|i\uparrow\downarrow\rangle$ are excluded. This exclusion is reflected by the presence of creation ($b_{i\sigma}^{+}$) and annihilation ($b_{i\sigma}$) operators for electrons in the state $|i\sigma\rangle$, which are defined as

$$b_{i\sigma}^{+} \equiv a_{i\sigma}^{+}(1-n_{i-\sigma}) \quad \text{and} \quad b_{i\sigma} \equiv a_{i\sigma}(1-n_{i-\sigma}), \quad (47)$$

so that

$$v_{i\sigma} = b_{i\sigma}^{+} b_{i\sigma} \qquad \text{and} \qquad v_i = \sum_{\sigma} v_{i\sigma}. \qquad (48)$$

The spin operator is defined as

$$\mathbf{S}_i \equiv \left(S_i^{+}, S_i^{-}, S_i^{z}\right) \equiv \left[a_{i\uparrow}^{+} a_{i\downarrow}, a_{i\downarrow}^{+} a_{i\uparrow}, (n_{i\uparrow} - n_{i\downarrow})/2\right].$$

Note that the same representation of the operator $\mathbf{S}_i$ can be written in terms of projected operators $b_{i\sigma}^{+}$ and $b_{i\sigma}$. The factor $(1-n_{i-\sigma})$ in Eq. (47) imposes explicitly the restriction that the creation or the annihilation of electrons in the state $|i\sigma\rangle$ can take place only if there is no second electron already on the same site. Thus, $v_i = \sum_{\sigma} n_{i\sigma}(1-n_{i-\sigma})$ enumerates only the singly occupied sites ($v_i = 0$ or 1). In other words, the $N$ states corresponding to the doubly occupied site configurations have been projected out. Thus, Eq. (46) describes the dynamics of strongly correlated electrons for $N_e \leq N$ of electrons. Also, in performing the summations in Eq. (46), one usually considers only the pairs $\langle ij\rangle$ of nearest neighbors; in this approximation the parameters $J_{ij} = J$ and $t_{ij} = t$ can be chosen as constants.

The first term in Eq. (46) describes the single–particle hopping of electrons from the singly occupied to the empty atomic sites; the second describes the exchange interaction induced by virtual hopping between site $i$ and site $j$, while the three-site part describes the motion of electron with spin $\sigma$ from the singly occupied site located at $i$ to the next-nearest neighboring empty site $k$ via the occupied configuration (with electron of opposite spin) located at site $j$. The various contributions to Eq. (46) are represented graphically in Fig. 26.

If one introduces a new pair of creation and annihilation operators in coordinate space by

$$\tilde{b}_{ij}^{+} = \frac{1}{\sqrt{2}}(b_{i\uparrow}^{+} b_{j\downarrow}^{+} - b_{i\downarrow}^{+} b_{j\uparrow}^{+}) \qquad (49a)$$

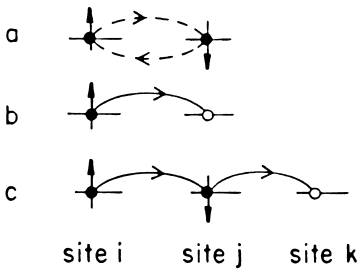



**FIGURE 26** Various hopping processes in narrow-band systems in a partial band-filling case: (a) virtual hopping processes leading to a kinetic exchange interaction; (b) single-particle hopping representing the band energy of correlated electrons; (c) contribution to the pair hopping—this process gives the pairing contribution in Eq. (50) with $k \neq i$.

and

$$\tilde{b}_{ij} = \frac{1}{\sqrt{2}}(b_{i\downarrow} b_{j\uparrow} - b_{i\uparrow} b_{j\uparrow}), \qquad (49b)$$

then the Hamiltonian [Eq. (46)] with inclusion of the three–site part can be written in the following very suggestive closed form [55]:

$$H = \sum_{ij\sigma}{}' t_{ij} b_{i\sigma}^{+} b_{j\sigma} - \sum_{ij} (2t_{ij}t_{jk}/U)\tilde{b}_{ij}^{+}\tilde{b}_{kj}. \qquad (50)$$

The first term represents, as before, the dynamics of single electrons moving between the empty sites regarded as holes; the second term combines the last two terms in Eq. (46) and expresses the dynamics of the singlet pairs [cf. Eqs. (49a) and (49b)]. The division in Eq. (50) into single–particle and pair parts is in analogy to the BCS Hamiltonian; however, here, the operators are expressed in coordinate space. The term with $i = k$ in the pairing part enumerates the spin–singlet pairs of neighboring spins; the terms with $i \neq k$ represent pair hopping of such singlet pair bonds. Thus, in the language of operators [Eqs. (49)], one adds the bond dynamics to that of single electrons. Moreover, the forms of Eqs. (46) and (50) are completely equivalent; hence, the pairing effect and the antiferromagnetism should be directly linked within this formalism (they are two different expressions of the same part of $H$).

It is difficult to diagonalize the Hamiltonian [Eq. (50)] to obtain the eigenvalues of the system. Part of the problem arises from the fact that the single–particle operators $b_{i\sigma}$ and $b_{j\sigma}^{+}$ do not obey the fermion anticommutation relation and that the pair operators $\tilde{b}_{ij}^{+}$ and $\tilde{b}_{ij}$ do not obey boson commutation relations. Additionally, the two terms in (3.5) do not commute, so that the itinerant characteristics of the electrons and the pair-binding effects combine and produce a paired metallic phase, particularly if the two terms are of comparable magnitude. We have seen in Section III that if the number of holes $\delta \equiv 1 - n < \delta_c \sim 0.02$, then the pairing (or exchange) part dominates and antiferromagnetism sets in. Detailed calculations [32] lead to the boundary line between the antiferromagnetic and the ferromagnetic phase, as shown in Fig. 27. The energy of the completely saturated ferromagnetic phase (CF) indicated does not depend on the value of exchange integral $J_{ij} \equiv 2t_{ij}^2/U$.

We now discuss the superconducting phase for which the pairing part in Eq. (50) plays a crucial role. To make the problem tractable at this point, one replaces the operators [Eqs. (49)] by fermion operators [45, 56], that is,

$$b_{i\sigma}^{+} \rightarrow a_{i\sigma}^{+}, \qquad b_{i\sigma} \rightarrow a_{i\sigma}, \qquad (51a)$$

and introduces the replacement

$$\tilde{b}_{ij}^{+} \rightarrow b_{ij}^{+} = \frac{1}{\sqrt{2}}\left(a_{i\downarrow}^{+} a_{j\uparrow}^{+} - a_{i\downarrow}^{+} a_{j\uparrow}^{+}\right) \qquad (51b)$$

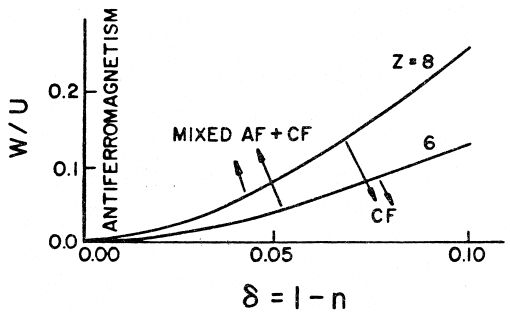



**FIGURE 27** Phase boundary between the mixed ferromagnetic (CF)–antiferromagnetic (AF) phase and the pure ferromagnetic phase for simple cubic ($z=6$) and b.c.c. cubic ($z=8$) structures. A similar type of phase boundary can be obtained for other structures. [From Ref. 63.]

and

$$\tilde{b}_{ij} \rightarrow b_{ij} = \frac{1}{\sqrt{2}}(a_{i\downarrow}a_{j\uparrow} - a_{i\uparrow}a_{j\downarrow}). \qquad (51c)$$

Simultaneously, one renormalizes the parameters $t_{ij}$ and $J_{ij}$ in such a manner that they contain the restrictions on particle dynamics due to the projection of doubly occupied site configurations in the expression for the ground-state energy. Within the Gutzwiller–Ansatz approximation [29], Eqs. (49) reduce the starting Hamiltonian to the form

$$H = \delta \sum_{ij\sigma} t_{ij} a^{+}_{i\sigma} a_{j\sigma} \sum_{ijk} (2t_{ij}t_{jk}/U) b^{+}_{ij} b_{kj}, \qquad (52)$$

where $\delta = 1 - n$. This Hamiltonian has been solved within the mean-field approximation equivalent to the BCS approximation [56, 57] and with neglect of the pairing terms with $k \neq i$. This leads to the following self-consistent equations for $\Delta_k \neq 0$:

$$\frac{J}{N} \sum_{\mathbf{k}} \frac{\gamma^2_{\mathbf{k}}}{E_{\mathbf{k}}} \tanh\left(\frac{\beta E_{\mathbf{k}}}{2}\right) = 1, \qquad (53)$$

with $J = 2t^2/U$, $E_{\mathbf{k}} = [(\epsilon_{\mathbf{k}} - \mu)^2 + |\Delta_{\mathbf{k}}|^2]^{\frac{1}{2}}$, and $\gamma_{\mathbf{k}} = \cos(k_x a) + \cos(k_y a)$ for a planar configuration of the lattice. This equation must be supplemented with the equation for the chemical potential in the superconducting phase of the form

$$\frac{1}{N} \sum_{\mathbf{k}} \left(1 - \frac{\epsilon_{\mathbf{k}}}{E_{\mathbf{k}}}\right) \tanh\left(\frac{\beta E_{\mathbf{k}}}{2}\right) = n. \qquad (54)$$

In solving Eq. (53), solutions of the following type have been considered:

1. extended s–wave [56, 58],

$$\Delta^{(s)}_{\mathbf{k}} = \Delta[\cos(k_x a) + \cos(k_y a)]; \qquad (55a)$$

2. d–wave [57, 58],

$$\Delta^{(d)}_{\mathbf{k}} = \Delta[\cos(k_x a) - \cos(k_y a)]; \qquad (55b)$$

3. mixed s and d phases [59],

$$\Delta^{(sd)}_{\mathbf{k}} = s\Delta^{(s)}_{\mathbf{k}} + d\Delta^{(d)}_{\mathbf{k}}. \qquad (55c)$$

The mixed phase was found to be the most stable close to the half-filled band case. For the half-filled band case, the ground-state energies for s– and d–wave states are the same.

The type of solution obtained within the mean-field approximation (cf. Section II for details) is illustrated in Fig. 28, where the temperature dependence of the specific heat is shown for a different number of holes $\delta$ and for $|t|/U = 0.1$ and with the inclusion of the nearest-neighbor repulsive coulomb interaction $V$. A discontinuity of $C(T)$ at $T = T_c$ takes place for each $\delta$. For comparison, the dotted lines represent the specific heat for the normal phase.

There is a major problem with the standard mean-field solutions discussed in Refs. 55–59, namely, it yields a nonzero (in fact, maximal or almost-maximal) value of the superconducting transition temperature $T_c$ for the half-filled band case, which corresponds to the Mott insulating state. This is a spurious result; it appears because by performing the transformation Eqs. (51), the double-site occupancies reappear again for $n < 1$. To remove some of the unphysical features of the mean-field solution, a new formalism has been proposed [58–60] in which auxiliary (slave) bosons are introduced. In this formalism, some

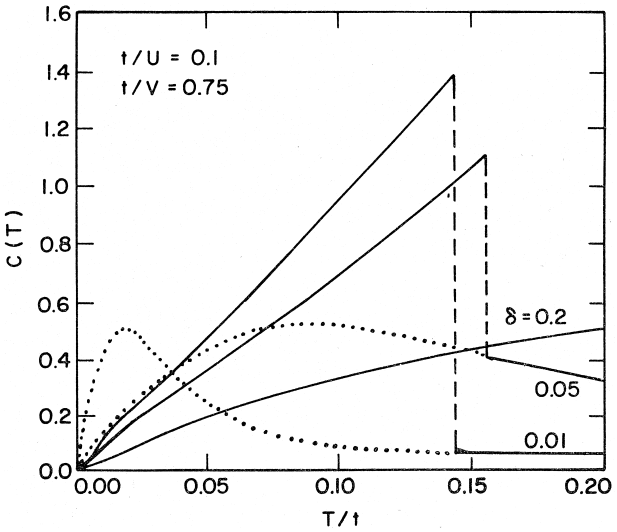



**FIGURE 28** Temperature dependence of the specific heat $C(T)$ within the mean-field approach to the exchange-mediated pairing in a narrow band. The dotted line represents $C(T)$ for the normal phase, while the discontinuity occurs at the transition to the superconducting phase. [From Ref. 58.]

of the properties of the projected operators [Eqs. (47) and (49)] are already preserved in the mean field-type approximation involving boson and fermion fields on the same footing. The transition temperature $T_c$ now vanishes, as it should, in the limit $n = 1$. According to Ref. 61, the slave bosons represent holes in the Mott insulators and are regarded as charged, while the fermions are neutral. These entities are called holons and spinons, respectively.

The holon–spinon language is introduced formally by noting that the projected operators [Eq. (47)] are represented as

$$b_{i\sigma}^{+} \equiv b_i f_{i\sigma}^{+} \qquad \text{and} \qquad b_{i\sigma} = b_i^{+} f_{i\sigma}, \tag{56}$$

where $b_i$ and $b_i^{+}$ are annihilation and creation boson operators located at the atomic site $i$, while $f_{i\sigma}^{+} \equiv a_{i\sigma}^{+}$ and $f_{i\sigma} \equiv a_{i\sigma}$ are the commonly used fermion operators. Substituting Eq. (56) into Eq. (46), one obtains

$$\begin{aligned} H = {} & \sum_{ij\sigma}{}' t_{ij} b_i b_j^{+} f_{i\sigma}^{+} f_{j\sigma} - \sum_{ij}{}' \left(2t_{ij}^2/U\right) b_{ij}^{+} b_{ij} \\ & + \sum_i \lambda_i \left( b_i^{+} b_i + \sum_\sigma f_{i\sigma}^{+} f_{i\sigma} - 1 \right) \\ & - \mu \sum_{i\sigma} n_{i\sigma} + (\text{three–site terms}). \end{aligned} \tag{57}$$

The first two terms represent, respectively, the coupled holon–spinon hopping and the binding of spinons into singlet pairs. The third term expresses the fact that the number of holons and spinons is equal to unity on each site; the Lagrange multiplier $\lambda_i$ thus explicitly provides formally the removal of double occupancies. The fourth term represents the conservation of the number of electrons. Now, in a further approximation, one decouples fermions from bosons and then solves the two parts self-consistently. The mean-field treatment discussed earlier corresponds to the approximation in which $\lambda_i$ is taken as the same at each site ($\lambda_i \to \lambda$) and in which one introduces the replacement $\langle b_i b_i^{+} \rangle = \langle b_i \rangle \langle b_i^{+} \rangle = |\langle b \rangle|^2 = 1 - n$. The superconducting solution is described in terms of two correlation functions: $\Delta_B \equiv \langle b_i^{+} b_j^{+} \rangle \approx \langle b^{+} \rangle^2$, characterizing the Bose condensation of holons, and $\Delta_F \equiv \langle f_{i\uparrow} f_{j\downarrow} \rangle$, characterizing the gap in the spectrum of fermion excitations (the site indices $i$ and $j$ denote a pair $\langle ij \rangle$ of nearest neighbors). The nonzero $\Delta_B$ occurs only below a temperature $T_B$, which we call the Bose condensation temperature, whereas the nonzero $\Delta_F$ appears only below $T = T_{RVB}$, characterizing the mean-field solution within the RVB theory [56]. The superconducting phase is characterized by nonzero values of both $\Delta_B$ and $\Delta_F$ simultaneously. This is because in the mean—field approximation $\langle b_{i\uparrow} b_{i\downarrow} \rangle = \Delta_B \Delta_F$. Hence, the lower of the two temperatures ($T_B$ and $T_{RVB}$) determines the superconducting transition temperature. In Fig. 29, taken from Ref. 63, we have plotted these two temperatures as a function $\delta = 1 - n$. One should note that to have $T_B \neq 0$, a small nonzero overlap $t_z = 0.1t$ was taken in the direction perpendicular to the square planar configuration of the atoms. We see that $T_c \to 0$ as $\delta \to 0$, as should be the case.

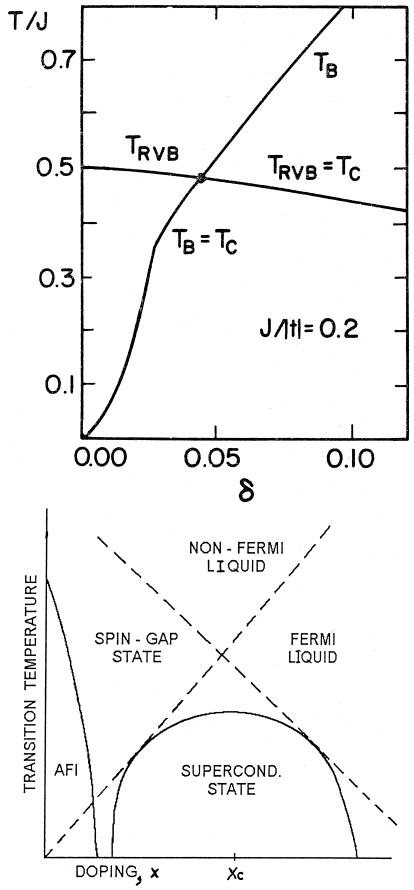



**FIGURE 29** (a) Critical temperature $T_c$ (the thick line) versus $\delta = 1 - n$. The temperatures $T_{RVB}$ and $T_B$ are those characterizing the onset of coherency for spinons and the Bose condensation of holons. Note that $T_c$ is determined by the lower of the two temperatures. [From Ref. 62.] (b) Schematic theoretical phase diagram obtained within the gauge theory.

*Note added in August 2000.* In the last 10 years the slave-boson approach evolved into the gauge-theory approach to doped Mott insulators [33d]. This approach leads to the phase diagram shown schematically in Fig. 29b. The details of this phase diagram go beyond the scope of this article. The other factors are the detailed role of the van Hove singularity in the two-dimensonal idensity of states for a square lattice [67a, b] in increasing even the BCS value of the critical temperature, as well as the determined d-wave symmetry reflecting the strong on-site Coulomb repulsion, which produces a node in the spatial dependence of the gap $\Delta(\mathbf{r})$ [67c, d]. Finally,

the role of the interlayer Josephson tunneling has been stressed [67e, f], although there is some discussion [67g] about the magnitude of the condensation energy due to the formation of (a) truly three-dimensonal paired state from a two-dimensonal normal metal. This means that a simple type of Lawrence–Doniach–Ginzburg–Landau approach [67h] and other models based on the interplanar Josephson tunneling [67e, f] may not be sufficient.

### 2. Hybridized Systems

The electron states near the Fermi surface in high-$T_c$ oxides such as $La_2CuO_4$ involve hybridization of electrons of atomiclike $3d_{x^2-y^2}$ states of copper with $2p_\sigma$ of oxygen (cf. Fig. 23 and Section III.D). These electron states can be described by the Anderson lattice Hamiltonian of the type of Eq. (43), with a width of the bare p–band $W \approx 4$ ev, the position of the $3d^9$ level at $\epsilon_f \equiv \epsilon_d - \epsilon_p \sim 1$ eV, $U \leq 10$ eV, and hybridization magnitude $|V| \simeq 1.5$ eV [68]. The hybridization is intersite in nature, that is, it involves the $2p$ and $3d$ orbitals located on different sites. Therefore, the effective hybridization energy is $Vz \simeq 6$ eV, where $z = 4$ is the number of nearest-neighboring O atoms in the plane for a given Cu atom. We see that $Vz > \epsilon_f$; hence, the $3d$ and $2p$ states mix strongly, that is, the $d$ electrons can be promoted to $2p$–hole states, and vice versa. Additionally, $2p$ electrons can be promoted to form the $3d^{10}$ configurations of the excited states. If $Vz \gtrsim \epsilon_f$, but $|V|z \ll \epsilon_f + U$, the above two promotion–mixing events are low– and high–energy processes, respectively. The situation is shown schematically in Fig. 30, where the parameter $U$ is assumed to be by far larger than $|\epsilon_f|$, $W$, or $|V|z$. We consider this limiting situation first [68].

The high–energy processes take place only as virtual events, that is, with electron hopping from the $p$ state to the highly excited $3d$ state and back. Such virtual $p$–$d$–$p$ processes are shown schematically in Fig. 31, where site $m$ labels the $2p_\sigma$ state of the oxygen anion $O^{2-}$ centered at $\mathbf{R_m}$ and site $i$ labels $3d_{x^2-y^2}$ due to the $Cu^{2+}$ ion centered at $\mathbf{R_i}$. Then the effective Hamiltonian can be rewritten in the real–space language and for large $U$ reads [68]

$$H = \sum_{\mathbf{k}\sigma} \epsilon_{\mathbf{k}} n_{\mathbf{k}\sigma} + \epsilon_f \sum_{i\sigma} b^+_{i\sigma} b_{i\sigma} + \sum_{im\sigma} V_{im} \left( b^+_{i\sigma} c_{m\sigma} + V^*_{im} c^+_{m\sigma} b_{i\sigma} \right) - \sum_{imn} \frac{2V^*_{mi} V_{im}}{U + \epsilon_f} B^+_{im} B_{in}. \quad (58)$$

The first term describes the band energy of itinerant $(2p_\sigma)$ electrons, while the third represents the residual mixing pairing since, as in the case of narrow-band electrons, the operators $(b^+_{i\sigma})$ and $(b_{i\sigma})$ are projected operators [Eq. (47)] for the starting $3d$ states. The last term represents the so-

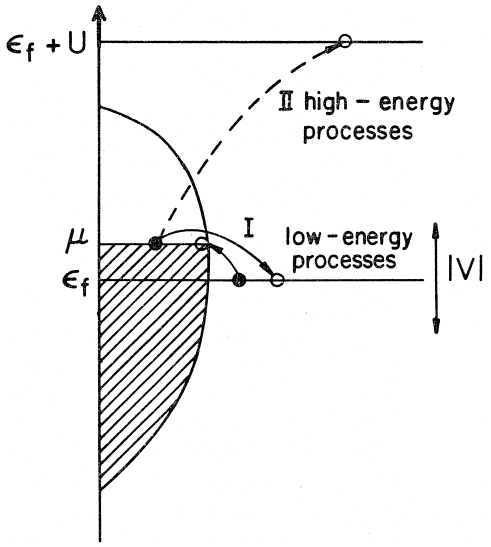



**FIGURE 30** Division of the charge-transfer ($p$–$d$) processes into low- and high-energy parts. The processes labeled II give rise to Kondo and superexchange interactions when treated perturbationally to second and fourth order, respectively.

called hybrid interorbital pairing with the pairing operators

$$B^+_{im} = \frac{1}{\sqrt{2}} \left( b^+_{i\uparrow} c^+_{m\downarrow} - b^+_{i\downarrow} c^+_{i\uparrow} \right) \quad (59a)$$

and

$$B_{im} = \frac{1}{\sqrt{2}} (b_{i\uparrow} c_{m\downarrow} - b_{i\downarrow} c_{i\uparrow}). \quad (59b)$$

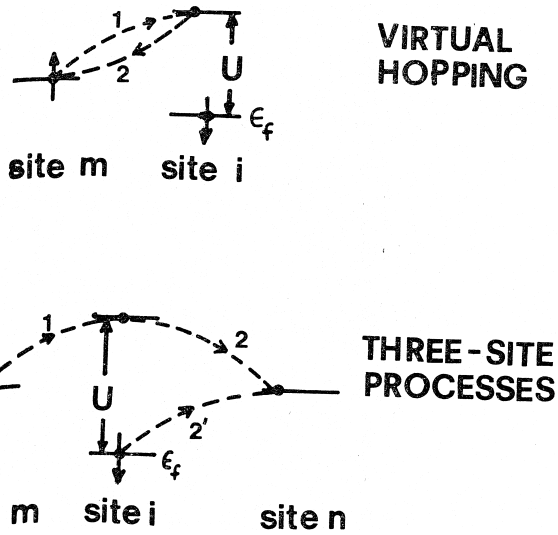



**FIGURE 31** Schematic representation of the hopping processes induced by high-energy mixing processes. The hoppings labeled 2 and 2′ are alternative processes.

The meaning of the effective Hamiltonian [Eq. (58)] is as follows. The first three terms provide eigenvalues representing the hybridized quasiparticle states with the structure discussed in Section III.C. The last term provides a singlet pairing for those hybridized states. It expresses (for $m = n$) the Kondo interaction between the $p$ and the $3d$ electrons of the form

$$\sum_{im} \frac{2|V_{im}|^2}{U + \epsilon_f} \left( \mathbf{S}_i \cdot \mathbf{s}_m - \frac{1}{4} \nu_i n_m \right).$$

It is antiferromagnetic in nature, with the exchange integral

$$J_{im} \equiv \frac{2|V_{im}|^2}{U + \epsilon_f} \sim 0.5 \text{ eV},$$

hence, the pairing results in a spin–singlet state. It must be underlined that Eq. (58) represents hybridized correlated states in the so-called fluctuating-valence regime in which $U \gg |V_{im}| \gtrsim \epsilon_f$. This is the reason why we cannot completely transform out the hybridization. Also, the occupancy $n_f$ of the atomic level is a noninteger because the strong hybridization induces a redistribution of the particles among starting atomic and band states.

When both $U$ and $|\epsilon_f|$ are much larger than $|V_{im}|$, one can transform out the hybridization completely and obtain, instead of Eq. (58), the following effective Hamiltonian:

$$H = \sum_{\mathbf{k}\sigma} \epsilon_{\mathbf{k}} n_{\mathbf{k}\sigma} + \epsilon_f \sum_{i\sigma} b_{i\sigma}^+ b_{i\sigma} + \sum_{ijm\sigma} \frac{V_{mi}^* V_{mi}}{\epsilon_f} \times b_{i\sigma}^+ b_{j\sigma} (1 - n_{m\sigma}) \sum_{im\sigma} \frac{2V_{mi}^* V_{in} U}{\epsilon_f(\epsilon_f + U)} b_{im}^+ b_{in}. \quad (60)$$

We now have a two–band system: the $3d$ electrons acquire a bandwidth $W^* \sim (V^2/\epsilon_f)(1 - \langle n_{m\sigma} \rangle)$. The spin–singlet pairing is again of the interband type. The part with $m = n$ in the last term is equivalent to the Kondo interaction derived a long time ago for magnetic impurities [69]. Here, the lattice version of this Hamiltonian provides both pairing and itinerancy to the bare atomic electrons.

Note that the hybrid pairing introduced in this section expresses both the Kondo interaction (the two–site part) and pair hopping. It is therefore suitable for a discussion of the superconductivity of Kondo lattice effects in heavy-fermion systems. The pairing part supplements the current discussions of the Anderson lattice Hamiltonian in the $U \to \infty$ limit [40–42]. One may state that the Kondo interaction-mediated pairing introduced above represents the strong–coupled version of spin fluctuation-mediated pairing for almost-localized systems introduced previously [70].

An approach using the slave-boson language for hybridized systems has also been formulated [71] and contains a principal feature of the effective Hamiltonian [Eq. (58)]; the solution in the mean-field approximation has been also discussed. Figure 32 illustrates the dependence of the superconducting transition temperature $T_c$ versus the hole concentration $x_h$; this is compared with experimental data [72]. Dependence of $T_c$ over the full concentration range of holes is shown in Fig. 33. The superconductivity appears for $La_{2-x}Sr_xCuO_4$ only for $0.04 \lesssim x_h < 0.34$. The full phase diagram comprising localization and antiferromagnetism (LM phase) and superconductivity (SC) is provided in Fig. 33.

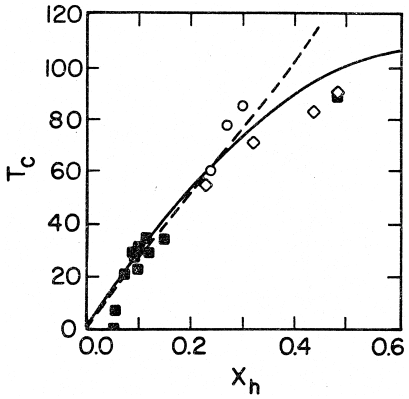



**FIGURE 32** Superconducting transition temperature $T_c$ versus hole concentration $x_h$. Squares, experimental data for $La_{2-x}Sr_xCuO_4$; circles and diamonds, data for $YBa_2Cu_3O_{7-y}$. [From Ref. 71.]

### 3. An Overview

Two alternative models and mechanisms of exchange–mediated pairing have been discussed so far: the narrow–band model, with $d$–$d$ kinetic exchange-mediated pairing, and the hybridized model, with $d$–$p$ Kondo interaction-mediated pairing. The hybridized model should be regarded as a basis of narrow–band behavior in real oxides and in heavy–fermion systems since the direct $d$–$d$ (or $f$–$f$) overlap of the neighboring atomic wave functions is extremely small. Next, we give a brief overview of the

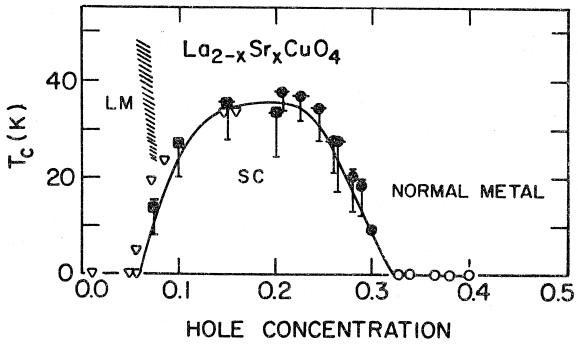



**FIGURE 33** Superconducting transition temperature $T_c$ versus hole concentration for $La_{2-x}Sr_xCuO_4$ over the full range. LM, the regime of local moments (insulating phase). [From Ref. 72.]

narrow–band properties of the correlated electrons starting from the hybridized (Anderson lattice) model.

First, we discuss the quasiparticle states in the $U \to \infty$ limit. The simplest approximation is to reintroduce ordinary fermion operators $a_{i\sigma}^{+}$ and $a_{i\sigma}$ in Eq. (58) and readjust the hybridization accordingly [73]. In effect, one obtains the hybridized bands of the form of Eq. (45), that is,

$$E_{\mathbf{k}\pm} = \frac{\epsilon_{\mathrm{f}} + \epsilon_{\mathbf{k}}}{2} \pm \left[ \left( \frac{\epsilon_{\mathrm{f}} - \epsilon_{\mathbf{k}}}{2} \right)^2 + 4|\tilde{V}_{\mathbf{k}}|^2 \right]^{\frac{1}{2}}, \qquad (61)$$

where $\tilde{V}_{\mathbf{k}} \equiv q^{\frac{1}{2}} V_{\mathbf{k}}$, and $q \equiv (1 - n_{\mathrm{f}})/(1 - n_{\mathrm{f}}/2)$ for $0 \leq n_{\mathrm{f}} \leq 1$, while $V_{\mathbf{k}}$ is the space Fourier transform of $V_{im}$. For the case of the $CuO_2$ layers [74],

$$|\tilde{V}_{\mathbf{k}}|^2 = q V^2 \left[ \sin^2\left( \frac{k_x a}{2} \right) + \sin^2\left( \frac{k_y a}{2} \right) \right]. \qquad (62)$$

If the Fermi level falls into the lower hybridization band and $n_{\mathrm{f}} = 1 - \delta$, with $\delta \ll 1$, then it can be shown that the quasiparticles describing the hybridized states are of mainly quasi-atomic character. In other words, the effective Hamiltonian [Eq. (58)] is approximately of the narrow-band form [Eq. (52)]. The pairing takes place between heavy quasiparticles. This limiting situation describes qualitatively the situation in heavy fermions with Kondo interaction mediating the pairing. In contrast, if the Fermi level falls close to the top of the upper hybridization band (as is the case for high-$T_{\mathrm{c}}$ superconductors, since the $p$ band is almost full and the *3d* level is almost half-filled), then the pairing is due mainly to the band electrons (*2p* holes in the case of high-$T_{\mathrm{c}}$ oxides). These results are obtained by constructing explicItly the eigenstates corresponding to the eigenvalues [Eq. (61)] and taking the limits corresponding to heavy fermions ($n_{\mathrm{f}} \to 1$) and high-$T_{\mathrm{c}}$ systems ($n = n_d + n_p \approx 3$, which also corresponds to the situation of one hole in the system).

*a. Mott–Hubbard insulators, charge–transfer insulators, and mixed–valent systems.* The next problem concerns the Mott localization in systems with hybridized $d$–$p$ states. The systems such as NiO, CoO, and MnO regarded as classic Mott insulators are, strictly speaking, hybridized *3d–2p* systems. However, these cases are, to a good approximation, ionic systems in the sense that the electronic configuration in, for example, NiO, is $Ni^{2+}O^{2-}$. Then, the valence *2p* band is completely full and plays only a passive role in effective $d$–$d$ charge transfer processes [75], since a $2p \to 3d$ transfer is followed by $3d \to 2p$ transfer from the neighboring *3d* shell of $Ni^{2+}$. In effect, the antiferromagnetic exchange interaction in Eq. (46) expresses formally the superexchange interaction that has been known for a long time [75, 76]. In this approach, the kinetic exchange interaction between $d$ electrons (induced by virtual $d$–$d$ transitions; cf. Section III.A) is expressed as a fourth-order effect in the hybridization $V$ since the virtual $d$–$d$ transition involves a sequence of $d$–$p$ and $p$–$d$ transitions in the fourth order.

The possible macroscopic states of hybridized systems are illustrated in Fig. 34 as a schematic classification of possible states of hybridized systems modeled by the periodic Anderson Hamiltonian [Eq. (43)]. The parameter $W/U$ characterizes the degree of correlation of quasi–atomic electrons that may acquire a nonzero bandwidth due mainly to hybridization; the parameter $V/\epsilon_{\mathrm{f}}$ characterizes the degree of mixing of the states involved. If the $d$ (or $f$) atomic level lies deeply below the top of the valence band ($V/\epsilon_{\mathrm{f}} \ll 1$), then we have either Mott–Hubbard (M–H) or charge–transfer (C–T) insulators; for the former the band gap $\Delta$ is due to $d^n \to d^{n+1}$ excitations (that is, $\Delta \sim U - W$), whereas for the latter it is due to $d^n p^2 \to d^{n+1} p^1$ charge–transfer transitions. The atomic *3d* (or *4f*) electrons are unpaired in both the C–T and the M–H states. If $V/\epsilon_{\mathrm{f}} \gtrsim 1$, and $W/U \ll 1$ then we enter mixed valent (M–V) and (close to the border with M–H) heavy–fermion regimes. On the other hand, if $W/U \gtrsim 1$, then irrespective of the value of $V/\epsilon_{\mathrm{f}}$, we encounter the correlated–metal regime that we call an almost–localized Fermi liquid (AL–FL). Both heavy–fermion and high-$T_{\mathrm{c}}$ systems are close to the line separating M–H and M–V regimes. Such a classification scheme for transition-metal oxides has been proposed in Ref. [77].

The classification shown schematically in Fig. 34 provides only a distinction between insulating and metallic states. A complete magnetic phase diagram for the high-$T_{\mathrm{c}}$ system $La_{2-x}Sr_xCuO_4$ is shown schematically in Fig. 35 (taken from Ref. 79a). Stoichiometric or doped $La_2CuO_4$,

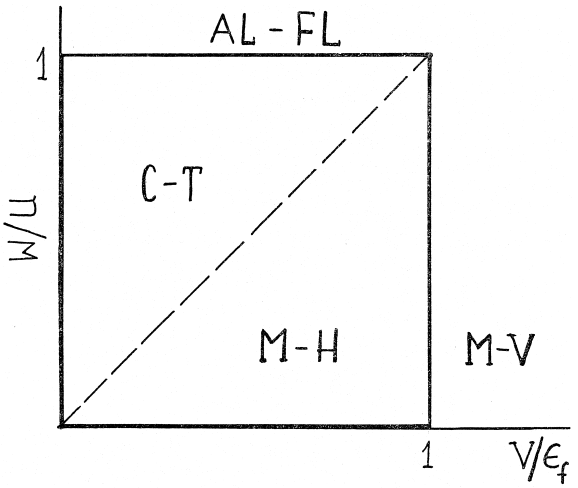



**FIGURE 34** Schematic representation of the regimes of stability of the charge-transfer (C–T) and Mott–Hubbard (M–H) insulating states, as well as of the mixed-valent (M–V) and almost-localized Fermi-liquid (AL–FL) metallic states.

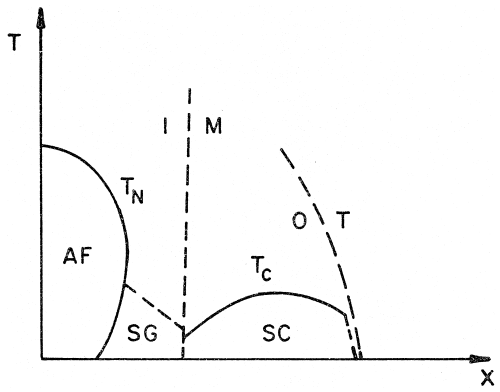



**FIGURE 35** Schematic phase diagram on the plane $T$–$x$ for $La_{2-x}Sr_xCuO_4$. Antiferromagnetic (AF), spin-glass (SG), superconducting (SC), insulating (I), and metallic phases are drawn, as well as the boundary between the orthorhombic (O) and the tetragonal (T) crystallographic phases. [From Ref. 80.]

with $x \lesssim 0.02$, exhibits antiferromagnetism (AF). In the regime of $0.02 \lesssim x \lesssim 0.04$, the inhomogeneous (SG) magnetic insulating phase sets in, while for $x \gtrsim 0.04$, a transition from insulating (I) to metallic (M) takes place and the system is superconducting until a transition from an orthorhombic (O) to a tetragonal (T) crystallographic structure occurs. A similar phase diagram was established for $YBa_2Cu_3O_{6+x}$ [77b]. Those phase diagrams combine all the features we have discussed separately so far. The main features of this phase diagram are explained next.

*b. Magnetic interactions hybrid, polarons, and pairing.* To address the phase diagram shown in Fig. 35 within the hybridized $p$–$d$ model, we note first that antiferromagnetism is stable only close to the half-filling of the $d$–band (cf. Fig. 27 and the discussion in Section III.A). In the case of the hybridized model, one has to calculate explicitly the contributions to the $d$–$p$ and $d$–$d$ interactions. Within the perturbation expansion for the Anderson lattice model but with only the high–energy mixing processes (cf. Fig. 30) treated in this manner [68, 70], we obtain the magnetic part of the effective Hamiltonian to fourth order as

$$\mathcal{H}_m \simeq J_{pd} \sum_{im} \left( \mathbf{S}_i \cdot \mathbf{s}_m - \frac{1}{4} n_i n_m \right) + J_{dd} \sum_{\langle ij \rangle} \left( \mathbf{S}_i \cdot \mathbf{S}_j - \frac{1}{4} N_i N_j \right) + J_{pp} \sum_{\langle mm' \rangle} \left( \mathbf{s}_m \cdot \mathbf{s}'_m - \frac{1}{4} n_m n'_m \right), \tag{63}$$

where the first term represents the $p$–$d$ Kondo–type interaction, with the exchange integral

$$J_{pd} \approx \frac{2|V|^2}{U + \epsilon_f} \left[ 1 - \frac{|V|^2}{U + \epsilon_f} (n_d + n_p + 1) \right]. \tag{64}$$

The second term expresses the $d$–$d$ (kinetic exchange) interaction, with $J_{dd} = |V|^4/(U + \epsilon_f)^3$, and the last term represents the interaction between $p$ holes, with $J_{pp} = |V|^4_{n_d}/(U + \epsilon_f)^3 \approx J_{dd}$. The antiferromagnetic $p$–$d$ and $d$–$d$ interactions are not compatible; in the hole language, the $p$ hole polarizes its surroundings ferromagnetically, as shown in Fig. 16 (note that the hole may be located in any $O^{1-}$ ion, so its position with the volume of radius $R$ is not fixed). A simple estimate [79] of the canting angle $\theta$ between the neighboring $3d$ spins $\mathbf{S}_i$ and $\mathbf{S}_j$ caused by the hole polarization gives

$$\cos \frac{\theta}{2} \approx \frac{J_{pd} - 2J_{dd}}{2J_{dd}}. \tag{65}$$

Taking $J_{pd} \approx 0.5$ eV and $J_{dd} \simeq 50$ K, we obtain the average canting angle $\theta$ through the relation $\cos(\theta/2) \approx 25 x_p$. The energy $E_c$ of the system with a single hole canting the surrounding spins is

$$E_c = -\frac{1}{2} \frac{(J_{pd} - 2J_{dd})^2}{J_{dd}} z - J_{dd} z. \tag{66}$$

This energy is lower than the energy $(-J_{dd}z)$ of the antiferromagnetic (Néel) state of antialigned $d$ spins due to $Cu^{2+}$ copper ions. Next, we estimate the radius $R$ of the hole polaron with aligned spins, as depicted in Fig. 36. Applying the same type of reasoning as in Section III.A, we obtain the expression for the energy $E_p$ of a single polaron:

$$E_p = \frac{E_0}{(R/a)^2} - \frac{1}{2} \frac{(J_{pd} - 2J_{dd})^2}{J_{dd}} z \cdot \bar{x}_p^2, \tag{67}$$

where now $\bar{x} = (a/R)^2$ is the probability of finding a $p$ hole on a given oxygen atomic site within the radius $R$. Minimization with respect to $R$ for the two–dimensional case leads to

$$\frac{R}{a} = \left[ \frac{(J_{pd} - 2J_{dd})^2 z}{J_{dd} E_0} \right]^{\frac{1}{2}} \approx J_{pd} \sqrt{\frac{z}{J_{dd} e_0}} \approx 4. \tag{68}$$

A MIT takes place when the neighboring polarons overlap, that is, when $R x_{pc}^{-1/2} = 1$; this yields the critical hole concentration $x_c \approx 0.07$. One can also estimate this critical concentration by equating the band energy of holes, which is $-(W/2) x_p (1 - x_p)$, with the magnetic energy gain per hole due to aligning the neighboring $d$ spins $(-J_{pd}^2/2J_d d) z x_p^2$. This leads again to $x_c \approx 0.068$, in rough agreement with the observed value $x_c \simeq 0.04 \div 0.05$. For $x > x_c$, the ground state of the system is metallic, and the pairing described in Sections IV.A.1 and IV.A.2 can take place. Within the exchange–mediated mechanism, all interactions in Eq. (63) are antiferromagnetic. Hence, in general, one has $p$–$d$ pairing characterized by the operators of Eqs. (59), $d$–$d$ pairingcharacterized by the operators

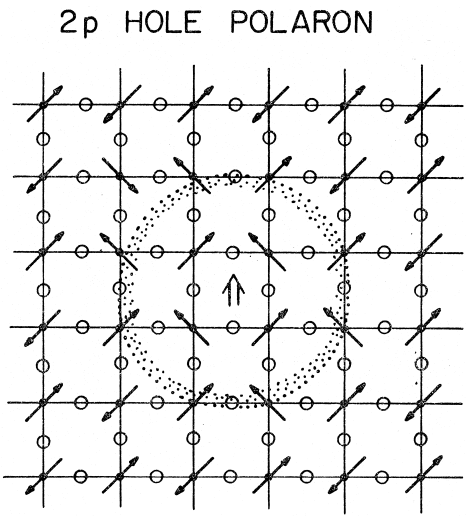



**FIGURE 36** Schematic representation of a *2p*-hole polaron in a planar $CuO_2$ structure. $Cu^{2+}$ ions are indicated by arrows, while $O^{2-}$ ions are indicated by open circles. The hole creates a canted spin configuration with resultant ferromagnetic polarization and autolocalizes in it. This is the reason the high-$T_c$ oxides remain insulating when the concentration of the hole does not exceed $x_c \sim 0.04 \div 0.05$.

of Eqs. (49), and $p$–$p$ pairing [80] characterized by the operators

$$p^+_{mm'} \equiv \frac{1}{\sqrt{2}}\left(c^+_{m\uparrow}c^+_{m'\downarrow} - c^+_{m\downarrow}c^+_{m'\uparrow}\right) \tag{69a}$$

and

$$p_{mm'} \equiv \frac{1}{\sqrt{2}}(c_{m\downarrow}c_{m'\uparrow} - c_{m\uparrow}c_{m'\downarrow}). \tag{69b}$$

All three types of pairing may contribute to the superconducting ground state. However, the $d$–$p$ interaction is much stronger, hence, the $d$–$p$ hybrid type of pairing is in the limit $U > W > |V| \gtrsim \epsilon_f$, the dominant one. As stated, this type of pairing may appear effectively as a $d$–$d$ or $p$–$p$ type of pairing in the hybridized basis, depending on whether the Fermi level lies close to the top of the lower or upper hybridized bands, respectively. For the sake of completeness, we write down the full effective Hamiltonian with all the pairings specified, namely,

$$\begin{aligned}\mathcal{H} = &\sum_{\mathbf{k}\sigma}\epsilon_{\mathbf{k}}n_{\mathbf{k}_\sigma} + \epsilon_f\sum_{i\sigma}b^+_{i\sigma}b_{i\sigma} + \sum_{im\sigma}\left(V_{im}b^+_{i\sigma}c_{m\sigma}\right.\\ &\left. + V^*_{im}c^+_{m\sigma}b_{i\sigma}\right) + J_{pd}\sum_{\langle im\rangle}\tilde{B}^+_{im}\tilde{B}_{in}\\ &+ J_{dd}\sum_{\langle ijk\rangle}\tilde{b}^+_{ij}\tilde{b}_{kj} + J_{pp}\sum_{\langle mm'\rangle}p^+_{mm'}p_{mm'}.\end{aligned} \tag{70}$$

In deriving this result, one does not assume that $|V| \ll \epsilon_f$; therefore Eq. (70) isapplicable to the situation with fluctuating valence. Next, by introducing a slave-boson representation [Eq. (56)], we obtain the most general Hamiltonian for treatment of pairing in correlated systems [79]. We should be able to witness a decisive progress in the near-future concerning the relative role of hybrid $p$–$d$, $d$–$d$, and $p$–$d$ pairings in high-$T_c$ systems using the slave-boson or gauge-field approaches to Eq. (70). Also, Eq. (70) should serve as a basis for the discussion of antiferromagnetism and superconductivity in heavy–fermion systems; in that situation, the role of itinerant *2p* states is played by hybridized *5d–6s* conduction bands, while the role of *3d* electrons is played by the *4f* electrons due to Ce or by *5f* electrons in uranium compounds.

*c. Coexistence of antiferromagnetism and superconductivity.* In the previous analysis, above we have treated antiferromagnetism (AF) and superconductivity (SC) separately. Detailed calculations [43, 62, 81], within the mean–field theory discussed, point to the possibility of the coexistence of AF and SC phases. It is possible to visualize this coexistence by considering a narrow-band model with the two-dimensional (almost-square) Fermi surface as shown in Fig. 37. Namely, the band energy of electrons located on the Fermi surface has the property $\epsilon_{\mathbf{k}+\mathbf{Q}} = -\epsilon_{\mathbf{k}}$, where $\mathbf{Q} \equiv (\pi/a, \pi/a) = 2\mathbf{k}_F$. This is the so-called nesting condition; any system with this property is unstable with respect to the formation of the spin density-wave (SDW) state with the wave vector $\mathbf{Q}$. One should note (cf. Fig. 37) that $\mathbf{Q}$ connects two single-particle states on the opposite sides of the Fermi surface since $-\mathbf{k}_F + \mathbf{Q} = \mathbf{k}_F$. Furthermore, both SDW and SC states couple electrons with the opposite spins. This is why two sublattice AF and SC states are compatible only for $n \approx 1$, i.e., for the half-filled band. There is no

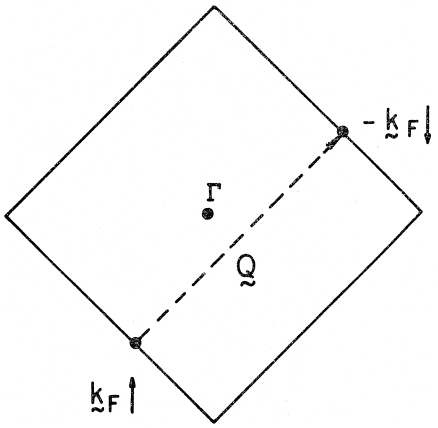



**FIGURE 37** Two-dimensional Fermi surface for a half-filled band. The opposite points of the surface are related by the wave vector $\mathbf{Q} = (\pi/a)(1, 1)$.

clear experimental evidence that these two phases coexist in a high-$T_c$ system, though there is some evidence from muon–spin rotation that it is so [82]. A clear detection of such coexistence would demonstrate directly the importance of exchange interactions in a superconducting phase. Namely, within exchange-mediated superconductivity, one can show [55] that close to the half-filled narrow-band case, $T_N/T_c \sim 6 \div 8$ (this is a mean-field-approximation estimate). The analysis of AF–SC coexistence conditions within the Anderson lattice Hamiltonian has not yet been performed satisfactorily, even though those two phases coexist in heavy-fermion compounds such as $UPt_3$ and $URu_2Si_2$.

## B. Phonons and Bipolarons

After the discoveries of superconductivity in the 40 and 90 K ranges [83], the obvious question was posed whether the phonon–mediated mechanism of pairing, so successful in the past, can explain the superconductivity with such a high value of $T_c$. It was realized from the outset that one should include specific properties of these compounds, such as the quasi-planar ($CuO_2$) structure with a logarithmic (Van Hove) singularity in the density of states $\rho(\epsilon)$ at the middle point of the two–dimensional band [83–85], the polar nature of the CuO bonds rendering applicable the tight–binding representation of the electronic states [49, 84], and strong electron lattice coupling [85–87], leading to the local formation of small bipolarons (that is, two–electron pairs) [89] that may undergo Bose condensation when the metallic state is reached [in a more refined version a mixed-fermion model is used (cf. Ranniger, Ref. 90)].

There is no clear evidence for the phonon–mediated mechanism of pairing in classic high-$T_c$ superconductors since the isotope effect in both La systems [90] and Y systems [91] is quite small. However, the recently discovered superconductors $Ba_{1-x}K_xBiO_3$ [92] exhibit a large isotope effect [93] and superconductivity with 20 K $\leq T_c \lesssim$ 30 K in the concentration range $0.25 \lesssim x \lesssim 0.4$. Also, the proximity of superconductivity and the charge density—wave (CDW) state is observed [94].

The last property, as well as the observed diamagnetism in the insulating phase $x \lesssim 0.25$, is very suggestive [95] that small trapped polarons are formed before the electron subsystem condenses into a superconducting phase. Condensation takes place when the percolation threshold for the insulator–metal transition is reached[3] (at $x \sim 0.2$).

[3]The actual percolation threshold for the onset of the metallic phase is $x_c/2 \sim 0.12$ since the bipolarons reside on every alternate Bi lattice site. Also, the holes introduced by K doping must be present in a Bi–O hybridized band for $x > x_c$ to render the bipolarons mobile for $x > x_c$.

Three specific features of $Ba_{1-x}K_xBiO_3$compounds should be noted. First, the diamagnetic nature of the parent compound $BaBiO_3$ distinguishes the systems from the parent compounds $La_2CuO_4$ and $YBa_2Cu_3O_7$, which are both antiferromagnetic. Second, the $Ba_{1-x}K_xBiO_3$ systems are copper-free and have a truly three–dimensional cubic structure in the SC phase [92]. Third, their main superconducting properties are in accordance with the prediction of the standard BCS theory [96].

The theory of the $Ba_{1-x}K_xBiO_3$ compound must incorporate three additional obvious facts. First, the pairing process $2Bi^{4+} \rightarrow Bi^{3+} + Bi^{5+}$ is possible when the electron–lattice coupling leads to an attraction overcoming the $e$–$e$ repulsion in the $Bi^{3+}$ state relative to the $Bi^{5+}$ state [89]. It involves a relaxation of the $O^{2-}$ octahedra, that is, an optical, almost dispersionless, breathing mode. This can provide a local (on-site) attractive interaction between *6s* electrons of the type $\lambda n_{i\uparrow} n_{n\downarrow}$, which leads to a scalar (**k**–independent) pairing potential $V_{\mathbf{k}\mathbf{k}'} = \lambda$, which, in turn, provides a justification for the observed properties reflecting an isotropic shape of the gap ($\Delta_{\mathbf{k}} \equiv \Delta$), as in the standard BCS theory (cf. Section II).

Second, from the fact that the parent compound $BaBiO_3$ is an insulator, we conclude that either the magnitude $V$ of the coulomb repulsion between the electrons on nearest-neighboring Bi atoms exceeds the width $W^*$ of the bipolaron band [96] or the small bipolarons are self-trapped in the potential created by interaction with nearest-neighboring oxygens. The onset of the metallic phase at concentrations near the percolation threshold $x_c \sim 0.1$ for n.n. interaction means that both effects may be important. In either case, the CDW state will set in, so the entropy of the bipolaron lattice vanishes at $T = 0$ (at least, for $x = 0$). The CDW phase plays the same role here as does AF ordering in $La_2CuO_4$ and $YBa_2Cu_3O_6$. The properties of $Ba_{1-x}K_xBiO_3$ are instead similar to those of the $Ba_{1-x}Pb_xBiO_3$ compounds discovered over a decade earlier [98].

Third, the fact that the onset of the superconductivity coincides with the transition from the CDW insulator to an SC metal speaks in favor of preexisting electron pairs already present in the insulating phase. However, the bipolaron concentration is large, and hence, the interpretation of the superconducting transition as Bose condensation of bipolarons may be inapplicable even when the coherence length is small. The overall theoretical situation is nonetheless much clearer for $Ba_{1-x}K_xBiO_3$ compounds than for either the $La_{2-x}Sr_xCuO_4$ or the $YBa_2Cu_3O_{7-\delta}$ series since the accumulated (so far) experimental evidence indicates that (optical?) phonon–mediated pairing takes place [99].

The $Ba_{1-x}K_xBiO_3$ compounds seem to be natural candidates for a bipolaronic mechanism of electron pairing

[100]. This is because the diamagnetic (and charge–ordered) parent system $Ba_2Bi^{3+}Bi^{5+}O_3$ can be regarded as an ordered lattice of locally bound two–electron pairs (bipolarons) located on alternate $Bi^{3+}$ sites; these pair states are stabilized by a strong relaxation of the surrounding oxygen anions. Effectively, the two electrons are attracted to each other. The effect of potassium doping is to make these pairs mobile by diminishing the number of bipolarons per Bi site from the value (1/2) [101]. In essence, the lattice distortion is responsible for the bipolaron formation in the same manner as in the case of the copper pairs; the difference is due to the circumstance that the bipolarons are locally bound complexes in a direct space that undergo a Bose condensation from an incoherent state of preexisting and moving pairs. The temperature of such condensation is $T_c \sim x^{\frac{2}{3}}$, where $x$ is the dopant (K) concentration [102]. A key feature of the bipolaron theory of superconductivity is that the Bose-condensed state develops from the CDW insulating state, not from the SDW (antiferromagnetic) state; the latter situation takes place for the cuprates. Further studies are necessary to calculate the physical properties of a bipolaron superconductor and, in particular, the differences from an ordinary (phonon–mediated) superconductor.

### C. Charge Excitations

In 1964 Little [103] introduced the idea that virtual electron–hole (exciton) excitations may lead to a pairing with a high value of $T_c$. This idea has been reformulated recently in the context of high-$T_c$ superconductivity by considering the role of charge transfer ($P \rightarrow d$ and $d \rightarrow p$) fluctuations [104], as well as of intraatomic (Cu $d \rightarrow d$) excitations [105]. The charge–transfer fluctuations involve both $Cu^{2+}$–$O^-$ and $Cu^{3+}$–$O^{2-}$ low–energy configurations and $Cu^+O^-$ states. The former two configurations are particularly important if the energy difference $|\epsilon_p - \epsilon_d|$ is comparable to the magnitude $|V|$ of the $2p$–$3d$ hybridization. This is the limit we have considered within the hybridized model in Section IV.A, probably extended to include the $3d$–$2p$ coulomb interaction directly. The method of approach is therefore similar to that in Section IV.A in the limit of strongly correlated electrons. In the limit of weakly interacting electrons (that is, for $U \ll W$), the perturbation expansion in the powers of $U$ provides an effective pairing potential in an explicit form. The processes leading then to the pairing are virtual exciations involving charge and antiferromagnetic spin fluctuations [106]. At the moment, it is difficult to see clearly the difference between exchange-mediated and charge transfer-mediated types of pairing for strongly correlated hybridized systems.

## V. CONCLUSIONS

In this article we have concentrated mainly on reviewing the properties of correlated electrons in normal, antiferromagnetic, and superconducting phases, in copper-containing systems in which the last two are phases caused by antiferromagnetic exchange interactions. Two theoretical models have been discussed in detail: the Hubbard model of correlated narrow-band ($3d$) electrons and the Anderson lattice model of correlated and hybridized electrons, involving $2p$ and $3d$ states in the case of high-$T_c$ oxides. The latter model is regarded as more general and applicable to both high-$T_c$ and heavy–fermion systems; in some limiting situations discussed previously, hybridized bands exhibit a narrow–band behavior.

The principal novel feature of the metallic phase involving either $3d$ (in high-$T_c$ oxides) or $4f$ (in heavy–fermion systems) electrons is that for the half–filled band configuration the itinerant electron states transform into a set of localized states constituting the Mott insulator. The difference between the Fermi liquid (FL) and the liquid of correlated electrons [the statistical spin liquid (SL)] is illustrated in Fig. 38, where the high–temperature value of the entropy has been plotted for these two phases as a function of the number $n$ of electrons per atom [the statistical distribution Eq. (41d) was used to calculate the entropy $S(n)$ for the latter phase]. Only the spin–liquid case correctly reproduces the entropy of localized moments when the Mott insulator limit is reached for $n \rightarrow 1$. This limiting value of the entropy per mole, $S = R\,ln\,2$ for $n \rightarrow 1$, represents one of the necessary conditions to be fulfilled by any theory claiming to describe properly the situation near the Mott insulator limit. Additionally, those systems are characterized by pseudo-particles with a very heavy

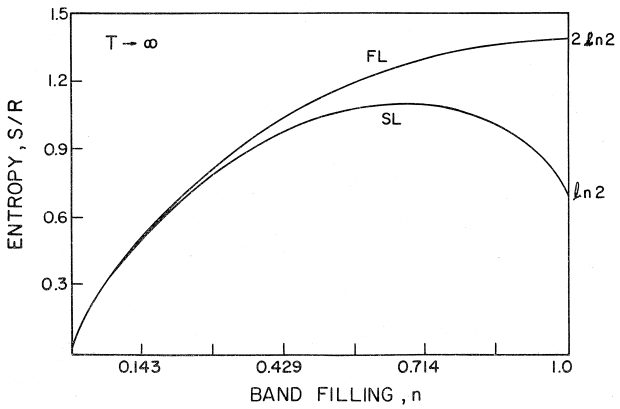



**FIGURE 38** The high-temperature limiting value for the entropy (in units of the gas constant $R$) as a function of $n$ for the Fermi liquid (FL) and the spin liquid (SL). Note the difference in the values of a factor of two in the limit $n \rightarrow 1$.

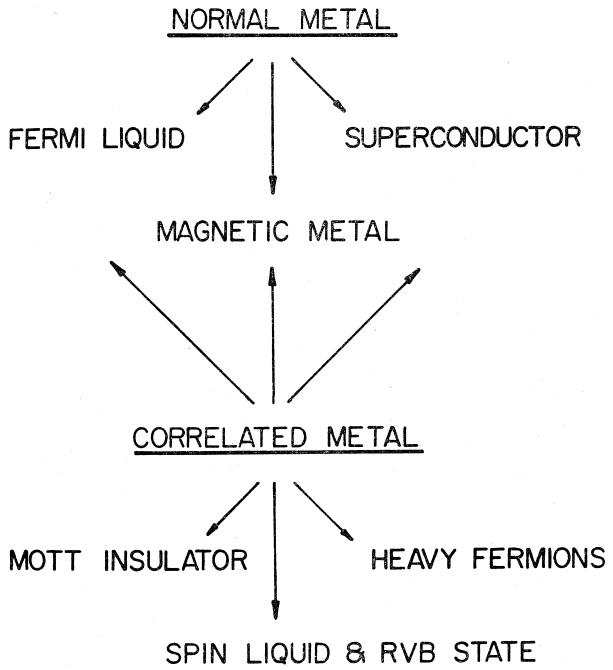



**FIGURE 39** Schematic representation of the difference between a normal metal and a correlated metal. Only the latter state may lead to Mott localization, as well as to heavy-fermion and spin-liquid metallic phases.

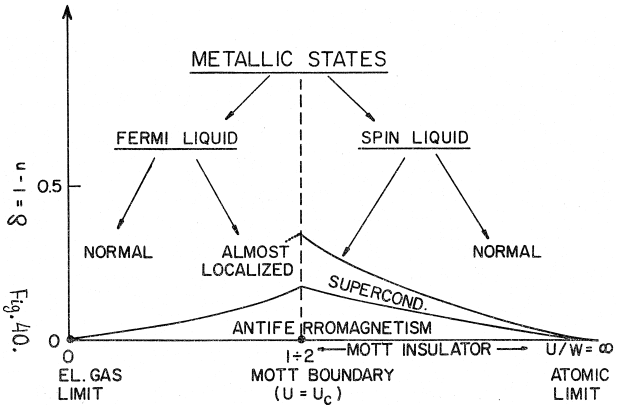



**FIGURE 40** Qualitative distinction between the Fermi-liquid and the spin-liquid states. The Mott boundary $U = U_c$ roughly separates the two limiting phases.

effective mass $m^* \sim \delta^{-1}$ or $W^* \sim \delta$. For $\delta \ll 1$, the band energy becomes comparable to the kinetic exchange characterized by $J = W^2/(Uz)$. Itinerant systems for which $J \gtrsim W^*$ are called quantum spin–liquid systems. The Mott insulator, the spin–liquid, and the heavy–fermion states are the primary phases of correlated electrons different from the normal–metal state. This difference is sketched out in Fig. 39, where the arrows point both to common features for normal and correlated metals and to those specific to the correlated systems.

The correlated systems that interested us here may also be called almost-localized systems. As discussed in Section II, there are two classes of such systems, separated roughly by the Mott–Hubbard boundary $U = U_c \sim W$: those for which the coulomb interaction $U < U_c$ are regarded as Fermi liquids have been treated extensively in Refs. 29 and 30, while those systems for which $U > U_c$ are the spin liquids. This qualitative division is sketched in Fig. 40, where the various thermodynamic phases have been specified for each class (cf. also Fig. 14 for all magnetic phases). The complementary regimes are those with $U/W \ll 1$ and $U/W \gg 1$. Most of the metallic systems can be located between these two limiting situations. It remains to be proven more precisely that the Mott–Hubbard boundary separating, for $n = 1$, Fermi liquid from the Mott insulator extends to the part of the diagram with $n \neq 1$, where Fermi liquid transforms with increasing interaction into non-Fermi liquid. This is a fundamental problem, related, in the case of strongly interacting systems, to the question of the validity of the Luttinger theorem[4] and to the problem of the existence of local magnetic moments in the itinerant–electron picture, that is, to the problem of the validity of the Bloch theorem for a correlated metal. Also, the question of applicability of the Fermi–liquid concept in the limit $U/W \gg 1$ is connected with that concerning the properly defined existence of fermion quasiparticles,[5] interacting only weakly among themselves. One should emphasize that the discussion of the standard mean-field treatment of superconductivity presented in Section IV reduces the whole problem to the single-particle approach with a self-consistent field $\sim \Delta_{\mathbf{k}}$. It is not yet completely clear what types of collective excitations (antiferromagnetic spin fluctuations? stripes?) are needed to make the theory complete. The introduction of holons as bosons and spinons as fermions [35] seems to be just one possibility; more natural seems to be a treatment of holons as spinless fermions and of spinons as boson operators that reflect magnonlike properties of local moments.

Early studies of high-$T_c$ oxides revealed that some of their characteristics are close to those provided by the BCS theory. Namely, the value of $2\Delta_0/k_bT_c \simeq 4 \div 6$ is

[4]The Luttinger theorem states that, as long as the metallic state is stable, the volume encircled by the Fermi surface remains independent of the strength of the electron–electron interaction. This theorem is not valid when the Mott transition takes place, as the Fermi surface then disappears. The volume also doubles when metal is described by a statistical spin liquid discussed in Section II (cf. Figs. 16a and b).

[5]The holons and spinons cannot be regarded as quasiparticles, since the Green function describing them has branching cuts rather than poles.

indicated [107], the temperature dependence of the London penetration depth is close to $[1-(T/T_c)^4]^{-\frac{1}{2}}$ over a wide temperature range [108], and the electron pairing is in the spin-singlet state [109]. Additionally, the shape of the Fermi surface for $YBa_2Cu_3O_7$, as determined by the positron annihilation technique [110], agrees with the predictions of the band-structure calculations for an even ($n \approx 4$) number of electrons. These results do not necessarily eliminate the principal features obtained from the theory of strong electron correlations. We think that before discarding the theory based on electron correlations, we must show clearly that the stoichiometric $La_2CuO_4$ or $YBa_2Cu_3O_6$ compounds are not insulating in the paramagnetic phase; actually, they seem to be paramagnetic insulators with well–defined magnetic moments (that is, with the Curie–Weiss law for the magnetic susceptibility obeyed), which supports strongly the view that they are Mott insulators. In this respect, the situation in heavy–fermion systems is rather clear since the recent theoretical results [39–42] based on the theory of strongly correlated and hybridized states provide a reasonable rationalization of most of the properties of their normal state. The mechanism of pairing in superconducting heavy–fermion systems has not yet been determined fully; but in view of the circumstances that some of the superconductors (for example, $UPt_3$) are antiferromagnetic and exhibit pronounced spin fluctuations in the normal state, the spin–fluctuation mechanism in the version outlined in Section IV.A.2 is a strong candidate [111]. In the coming years one should be able to see a clarification of these problem.

Let us end with a methodological remark concerning the analogy of the studies of magnetism and superconductivity. In 1928, Heisenberg introduced the exchange interaction $J_{ij}\mathbf{S}_i \cdot \mathbf{S}_j$ between the magnetic atoms with spins $\{\mathbf{S}_i\}$. The ferromagnetic state was understood in terms of a molecular field $\mathbf{H}_i \sim \langle \mathbf{S}_i \rangle$ which was related to the direct exchange integral $J_{ij}$. Later, various other exchange interactions have been introduced, such as superexchange, double exchange, RKKY interaction, the Bloembergen–Rowland interaction, Hund's rule exchange, and kinetic exchange, to explain magnetism in specific systems, such as oxides, rare-earth metals, and transition metals. However, all these new theories provided a description in terms of a single–order parameter—the magnetization $\langle \mathbf{S}_i \rangle$; the particular feature of the electron states in each case (localized states, itinerant states, or a mixture of the two states) is contained only in the way of defining this order parameter or the exchange integral. By analogy, the BCS theory provided a concept of a superconducting order parameter ($\Delta_{\mathbf{k}}$), which is universal for all theories of singlet superconductivity. New mechanisms of pairing should provide a novel interpretation to the coupling constant $V_{\mathbf{k}\mathbf{k}'}$ as well as supplying some details concerning the specific features of the system under consideration: the gap anisotropy, the role of hybridization, etc. It remains to be seen if some qualitative differences arise if superconductivity should occur as a result of Bose condensation of the preexisting pairs. This question is particularly important in the case when the coherence length is small, as in high-$T_c$ systems.

In the coming years, one should see detailed calculations within the exchange mechanism and comparisons with the experiment concerning the complete phase diagram, as well as the thermodynamic and electromagnetic properties of the new superconductors $La_{2-x}Sr_xCuO_4$ and $YBa_2Cu_3O_{7-\delta}$. It would not be surprising if the final answer for these systems came from a detailed analysis of the model outlined in Section IV.A. The systems $Ba_{1-x}K_xBiO_3$ will probably be described satisfactorily within the standard phonon–mediated mechanism. On the other hand, it is too early to say anything definite about Bi and T$\ell$ compounds with $T_c > 100$ K, though the suggested influence of the electronic structure near $\epsilon_F$ by the $CuO_2$ planes seems to indicate a nontrivial role of the exchange interactions also in those systems when coupled with interlayer pair tunneling. One of the missing links between the properties of the last two classes of compounds and those of $La_{2-x}Sr_xCuO_4$ is the conspicuous lack of evidence for antiferromagnetism in the $Bi_2Sr_2CaCu_2O_8$ and the $Tl_2Ca_2Ba_2Cu_3O_{10-y}$ compounds.

*Note added in August 2000.* This article was originally written almost 12 years ago. During those years a tremendous number of papers has been published, but the questions concerning either the pairing mechanism or the non-Fermi liquid behavior have not been clearly resolved, either for high-$T_C$ or for heavy-fermion superconductors. Nonetheless, a number of experimental results have been obtained in a clear form for high-$T_C$ systems. Let us mention two additional results. First, the role of the hopping between the next neighbors is important for obtaining the open Fermi surface (cf. Fig. 41). This Fermi surface is obtained from the photoemission and encompasses all electrons [112], i.e., not only the hole states in the doped Mott insulator. Thus, the principal question is how to reconcile the strong-correlation nature of the electrons in the $CuO_2$ plane, as reviewed above with the Luttinger theorem, which seems to be obeyed in optimally doped and overdoped systems, as concluded from the photoemission data. Does this mean that the photoemission experiment samples states physically different from those involved in thermally induced transport properties? The Fermi-liquid features seem clearly to break down in underdoped systems [113], where a pseudogap related to the superconducting gap is also observed [114].

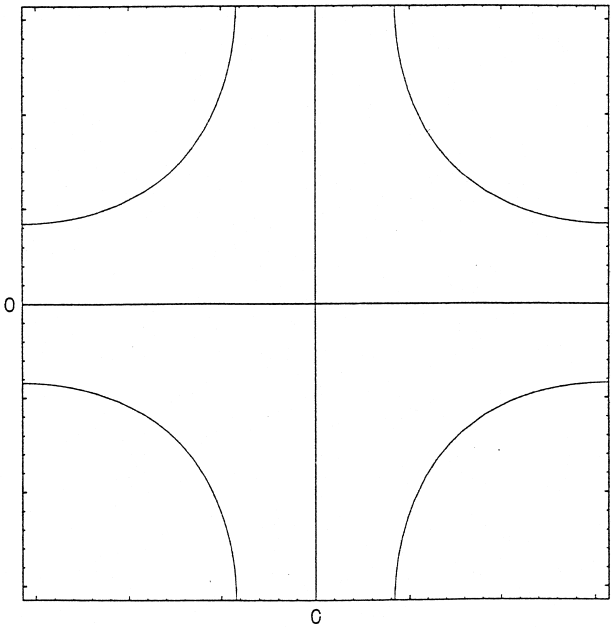



**FIGURE 41** Schematical representation of a two-dimensonal Fermi surface for a nonzero amplitude of second-nearest-neighbor hopping.

## ACKNOWLEDGMENTS

I would like to thank Leszek Spalek for technical help. This work was supported by KBN (Poland) Project No. 2PO3B O92 18.

## SEE ALSO THE FOLLOWING ARTICLES

ELECTRONS IN SOLIDS • SUPERCONDUCTING CABLES • SUPERCONDUCTING DEVICES • SUPERCONDUCTIVITY • SUPERCONDUCTORS, HIGH TEMPERATURE

## BIBLIOGRAPHY

1. Bardeen, J., Cooper, L. N., and Schrieffer, J. R. (1957). *Phys. Rev*. **106,** 162. Bardeen, J., Cooper, L. N., and Schrieffer, J. R. (1957). *Phys. Rev*. **108,** 1175.
2. Schrieffer, J. R. (1964). "Theory of Superconductivity," W. A. Benjamin, Reading, PA.
3. De Gennes, P. G. (1966). "Superconductivity of Metals and Alloys," W. A. Benjamin, Reading, PA.
4. Tinkham, M. (1996). "Introduction to Superconductivity," 2nd ed., McGraw–Hill, New York.
5. Rickayzen, G. (1965). "Theory of Superconductivity," Wiley, New York.
6. Blatt, J. M. (1964). "Theory of Superconductivity," Academic Press, New York.
7. Barone, A., and Paterno, G. (1982). "Physics and Applications of the Josephson Effect," John Wiley, New York.
8. Landau, L. D., Lifshitz, E. M. (1950). "Statistical Physics," 2nd ed., Part 2, Chap. 5, Pergamon, Oxford, Abrikosov, A. A., Gorkov, L. P., and Dzyaloshinski, I. E. (1963). "Methods of Quantum Field Theory in Statistical Physics," Chap. 7, Dover, New York.
9. Parks, R. D. (ed.) (1969). "Superconductivity" (2-vols.), Dekker, New York.
10. Kuper, C. G. (1968). "An Introduction to the Theory of Superconductivity," Clarendon Press, Oxford, New York. Rose—Innes, A. C., and Rhoderick, E. H. (1969). "Introduction to Superconductivity," Pergamon Press, Oxford.
11. Fröhlich, H. (1952). *Proc. Roy. Soc. A* **215,** 291.
12. Cooper, L. N. (1956). *Phys. Rev*. **104,** 1189.
13. Bogoliuboy, N. N. (1958). *Nuovo Cimento* **7,** 794. Valatin, J. G. (1958). *Nuovo Cimento* **7,** 843. Nambu, Y. (1960). *Phys. Rev*. **117,** 648.
14. Morel, P., and Anderson, P. W. (1962). *Phys. Rev*. **125,** 1263. See also Scalapino, D. J., in Ref. 9, Chapt. 10.
15. Eliashberg, G. M. (1966). *Zh. Eksp. Teor. Fiz*. **38,** 966 [*Sov. Phys. JETP* **11,** 696 (1960)]. For review see Allen, P. B., and Mitrovic, B. (1982). *In* "Solid State Physics" (H. Ehrenreich, F. Seitz, and D. Turnbull, eds.), pp. 2–92, Academic Press, New York.
16. Migdal, A. B. (1958). *Zh, Eksp. Teor. Fiz.* **34,** 1438 [*Sov. Phys. JETP* **7,** 996(1958)].
17. McMillan, W. L. (1968). *Phys. Rev.* **167,** 331.
18. Khan, F. S., and Allen, P. B. (1980). *Solid State Commun*. **36,** 481.
19. Cohen, M. L., and Anderson, P. W. (1972). *AIP Conf. Proc. No. 4* (D. H. Douglass, ed.), p. 17, (AIP, New York).
20. Maxwell, E. (1950). *Phys. Rev.* **78,** 477. Reynolds, C. A., Serin, B., Wright, W. H., and Nesbitt, L. B. (1950). *Phys. Rev.* **78,** 487.
21. Some of the monographs and reviews which appeared during the last decade are as follows. (a) Ginsberg, D. H. (ed.) (1989–1995). "Physical Properties of High Temperature Superconductors," Vols. 1–5, World Scientific, Singapore. (b) Battlogg, B., *et al*., (1996). "Proceedings of the 10th Anniversary Workshop on Physics, Materials and Applications, World Scientific, Singapore. (c) Tsuneto, T. (1998). "Superconductivity and Superfluidity," Cambridge University Press, Cambridge. (d) Cyrot, M., and Pavuna, D. (1992). "Introduction to Superconductivity and High-$T_C$ Materials," World Scientific, Singapore. (e) Carbotte, J. P. (1990). *Rev. Mod. Phys.* **62,** p. 1027ff. (f) Waldram, J. R. (1996). "Superconductivity of Metals and Cuprates," IOP, Bristol, Philadelphia. (g) Narlikar, A. (ed.) (1990–). "Studies in High Temperature Superconductors," Vols. 1–16, Nova, Science, New York, Budapest. (h) Anderson, P. W. (1997). "The Theory of Superconductivity in High-$T_C$ Cuprates," Princeton University Press, Princeton, NJ. (i) Bisarsh, A. (ed.) (1999). "Superconductivity: An Annotated Bibilography with Abstracts," Nova, Science, New York. ( j) Poole, C. P. (1999). "Handbook of Superconductivity," Academic Press, San Diego, CA. (k) Plakida, N. M. (1995). "High Temperature Superconductivity," Springer Verlag, Berlin.
22. For review see, e.g., Baym, G., and Pethick, C. (1978). *In* "The Physics of Liquid and Solid Helium" (K. H. Bennemann and J. B. Ketterson, eds.), Chap. 3, Wiley, New York; Pines, D., and Nozieres, P. (1996). "The Theory of Quantum Liquids," W. A. Benjamin, New York.
23. Mott, N. F. (1974). "The Metal–Insulator Transitions," Taylor and Francis, London. Héritier, M., and Lederer, P. (1977). *J. Phys. (Paris)* **38,** L209.
24. Vaknin, D., *et al.* (1987). *Phys. Rev. Lett.* **58,** 2802. Endoh, Y., *et al.* (1987). *Phys. Rev. B* **37,** 7443.
25. Hubbard, J. (1964). *Proc. R. Soc. London A* **281,** 401.
26. Acquarone, M., Ray, D. K., and Spałek, J. (1982). *J. Phys. C* **15,** 959.
27. Kuwamoto, H., Honig, J. M., and Appel, J. (1980). *Phys. Rev. B* **22,** 2626.

28. Spałek, J., Oleś, A. M., and Honig, J. M. (1980). *Phys. Rev. B* **28,** 6802.
29. Brinkman, W. F., and Rice, T. M. (1970). *Phys. Rev. B* **2,** 4302.
30. Spałek, J., Datta, A., and Honig, J. M. (1987). *Phys, Rev. Lett*. **59,** 728. Spałek, J., Kokowski, M., and Honig, J. M. (1989). *Phys. Rev. B* **39,** 4175. For $d = \infty$ solution see, e.g., Gebhard, F. (1997). "The Mott Metal–Insulator Transition," Springer Verlag, Berlin.
31. For the Mott insulators, the kinetic exchange interaction was introduced by in Anderson, P. W. (1959). *Phys. Rev.* **115,** 2. This result was extended to the case of a strongly correlated metal by Chao, K. A., Spałek, J., and Oleś, A. M. (1977). *J. Phys. C* **10,** L271. Cf. also Spałek, J., and Oleś, A. M. (1976). Jagiellonian University preprint SSPJU–6/76, Oct. Zaanen, J., and Sawatzky, G. A. (1990). *J. Solid State Chem.* **88,** 8ff.
32. Vischer, P. B. (1974). *Phys. Rev.* **10,** 943. Spałek, J., Oleś, A. M., and Chao, K. A. (1981). *Phys. Stat. Sol. (b)* **108,** 329. Nagaoka, Y. (1966). *Phys. Rev.* **147,** 392.
33. (a) Varma, C. M., *et al.* (1989). *Phys. Rev. Lett*. **63,** 1996; (1989). *Int. J. Mod. Phys. B* **3,** 2083–2118; Littlewood, P. B., and Varma, C. M. (1991). *J. Appl. Phys.* **69,** 4979. (b) For review see Pines, D. (1998). *In* "The Gap Symmetry and Fluctuations in High-$T_C$ Superconductors" (J. Bok *et al.*, eds.), Plenum Press, New York, NATO ASI Series, Ser. B, Vol. 371, and references therein. (c) Moriya, T., and Ueda, K. (2000). *Adv. Phys.* **49,** 555–606, and references therein. (d) Lee, P. A., and Nagaosa, N. (1992). *Phys. Rev. B* **46,** 5621; for review see Nagaosa, N. (1999). "Quantum Field Theory in Strongly Correlated Electronic Systems," Chap. 5, Springer Verlag, Berlin; Lee, P. A. (1996). *J. Low Temp. Phys.* **105,** 581. (e) van den Brink, J., *et al.* (1995). *Phys. Rev. Lett.* **75,** 4658; (1996). *Phys. Rev. Lett.* **76,** 2826. (f ) Bała, J., *et al.* (1995). *Phys. Rev. B* **52,** 4597. (g) Dagotto, E. (1994). *Rev. Mod. Phys.* **66,** 763, and references therein; Dagotto, E., and Rice, T. M. (1996). *Science* **271,** 618; Dagotto, E. (1998). *J. Phys. Chem. Solids* **59,** 1699. (h) Zaanen, J. (1998). *J. Phys. Chem. Solids* **59,** 1769; Castellani, C., *et al.*, *ibid*., p. 1694; Kivelson, S. A., *ibid*. p. 1705, and references therein.
34. (a) Takagi, H., *et al.* (1992). *Phys. Rev. Lett.* **69,** 2975. (b) Rossat-Mignot, J., *et al.* (1991). *Physica C* **185–189,** 86; Bourges, P., *et al.* (1997). *Phys. Rev. B* **56**(11) 439. (c) Anderson, P. W. (1997). "The Theory of Superconductivity in the High-$T_C$ Cuprates," Princeton University Press, Princeton, NJ; Byczuk, K., Spałek, J., and Wójcik, W. (1998). *Acta Phys. Polonica B* **29,** 3871; Varma, C. M. (1997). *Phys. Rev. B* **55**(14), 554.
35. Spałek, J., and Wójcik, W. (1988). *Phys. Rev. B* **37,** 1532.
36. Kivelson, S. A., Rokhsar, D. S., and Sethna, J. P. (1987). *Phys. Rev. B* **35,** 8865. Anderson, P. W., and Zou, Z. (1988). *Phys. Rev. Lett.* **60,** 132.
37. Anderson, P. W. (1988). Cargese 1988, Lecture Notes. Also (1987). *In* "Proceedings of the International School 'Enrico Fermi'—1987: Frontiers and Borderlines in Many—Particle Physics," North-Holland, Amsterdam.
38. For a recent review, see Fukuyama, H., Hasegawa, Y., and Suzumura, Y. (1988). *Physica C* **153–155**.
39. Anderson, P. W. (1988). *Physica C* **153–155**.
40. Lee, P. A., Rice, T. M., Serene, J. W., Sham, L. J., and Wilkins, J. W. (1986). *Comments Condens. Matter. Phys.* **12,** 99.
41. Rice, T. M. (1987). "*In* Proceedings of the International School of Physics 'Enrico Fermi'—1987: Frontiers and Borderlines in Many—Particle Physics," North-Holland, Amsterdam.
42. Newns, D. M., and Read, N. (1987). *Adv. Phys.* **36,** 799.
43. Fulde, P., Keller, J., and Zwicknagl, G. (1988). *In* "Solid State Physics" (H. Ehrenreich and D. Turnbull, eds.), Vol. 41, Academic Press, San Diego, CA.
44. For a review of experimental properties of heavy fermions, see Stewart, G. R. (1984). *Rev. Mod. Phys.* **56,** 755. Fisk, Z., *et al.* (1986). *Nature (London)* **320,** 124. Steglich, F. (1955). *In* "Theory of Heavy Fermions and Valence Fluctuations" (T. Kasuya and T. Saso, eds.), p. 23ff, Springer-Verlag, Berlin, New York.
45. Wróbel, P., and Jacak, L. (1988). *Mod. Phys. Lett. B* **2,** 511.
46. Anderson, P. W. (1987). *Science* **235,** 1196.
47. Shirane, G., *et al.* (1988). *Phys. Rev. Lett.* **59,** 1613.
48. Chakravarty, S., *et al.* (1988). *Phys. Rev. Lett.* **60,** 1057.
49. Nucker, M., *et al.* (1987). *Z. Phys. B* **67,** 9. Fujimori, A., *et al.* (1987). *Phys. Rev. B* **35,** 8814. Steiner, P., *et al.* (1988). *Z. Phys. B* **69,** 449.
50. Mattheis, L. F. (1987). *Phys. Rev. Lett*. **58,** 1028. Yu, J., Freeman, A. J., and Xu, J.-H. (1987). *Phys. Rev. Lett.* **58,** 1035. Szpunar, B., and Smith, V. H., Jr. (1988). *Phys. Rev. B* **37,** 2338. For review see Hass, K. C. (1989). *In* "Solid State Physics" (H. Ehrenreich and D. Turnbull, eds.), Vol. 42, Academic Press, San Diego, CA.
51. Fulde, P. (1988). *Physica* **153–155,** 1769. Hybertsen, M. S., *et al.* (1994). *Phys. Rev. B* **41,** 11068.
52. Zhang, F. C., and Rice, T. M. (1988). *Phys. Rev. B* **37,** 3759.
53. Schafroth, M. R. (1955). *Phys. Rev.* **100,** 463.
54. For a review see Leggett, A. (1975). *Rev. Mod. Phys*.
55. Shafer, M. W., Penney, T., and Olson, B. L. (1987). *Phys. Rev. B* **36,** 4047.
56. Spałek, J, (1988). *Phys. Rev. B* **37,** 533. Acquarone, M. (1988). *Solid State Commun.* **66,** 937.
57. Baskaran, G., Zou, Z., and Anderson, P. W. (1987). *Solid State Commun.* **63,** 973.
58. Cyrot, M. (1987). *Solid State Commun*. **62,** 821.
59. Ruckenstein, A. E., Hirschfeld, P. J., and Appel, J. (1987). *Phys. Rev. B* **36,** 857.
60. Kotliar, G. (1988). *Phys. Rev. B* **37,** 3664.
61. Isawa, Y., Maekawa, S., and Ebisawa, H. (1987). *Physica* **148B,** 391.
62. Zou, Z., Anderson, P. W. (1988). *Phys. Rev. B* **37,** 627.
63. Inui, M., Doniach, S., Hirschfeld, P. J., and Ruckenstein, A. E. (1988). *Phys. Rev. B* **37,** 2320.
64. Suzumura, Y., Hasegawa, Y., and Fukuyama, H. (1988). *J. Phys. Soc. Jpn.* **57,** 2768.
65. Kotliar, G., and Liu, J. (1988). *Phys. Rev. B* **38,** 5142.
66. Nagaosa, N. (1966). *In* "Proceedings of the 10th Anniversary HTS Workshop on Physics, Materials, and Applications" (B. Batlogg *et al.*, eds.), pp. 505ff, World Scientific, Singapore, and references therein.
67. (a) Labbé, J., and Bok, J. (1987). *Europhys. Lett*. **3,** 1225; Newns, D. M., *et al.* (1992). *Comments Cond. Mat. Phys.* **15,** 273. (b) Bouvier, J., and Bok, J. (1998). *In* "The Gap Symmetry and Fluctuations in High-$T_C$ Superconductors" (J. Bok *et al.*, ed.), pp. 37–54, Plenum Press, New York. (c) Van Harlingen, D. J. (1995). *Rev. Mod. Phys*. **67,** 515. (d) Annett, J. F., Goldenfeld, N., and Legett, A. J. (1996). *In* "Physical Properties in High-Temperature Superconductors," Vol. 5 (D. M. Ginsberg, ed.), pp. 375–461, World Scientific, Singapore. (e) Chakravarty, S., *et al.* (1993). *Science* **261,** 337. (f ) Byczuk, K., and Spałek, J. (1996). *Phys. Rev. B* **53,** R518. (g) Leggett, A. J. (1998). *J. Phys. Chem. Solids* **59,** 1729; Tsvetkow, A. A., *et al.* (1998). *Nature* **395,** 360. (h) Lawrence, W. E., and Doniach, S. (1971). *In* "Proc. 12th Int. Conf. Low Temp. Phys." (E. Kanda, ed.), Keigaku, Tokyo; see also Tinkham, M., in Ref. 4.
68. Mila, F. (1988). *Phys. Rev. B* **38,** 11358, and references therein.
69. Spałek, J. (1988). *Phys. Rev. B* **38,** 208. Spałek, J. (1988). *J. Solid State Chem.* **76,** 224.
70. Schrieffer, J. R., and Wolff, P. A. (1966). *Phys. Rev.* **149,** 491.

71. Anderson, P. W. (1984). *Phys. Rev. B* **30,** 1549. Miyake, K., Schmitt–Rink, S., and Varma, C. M. (1986). *Phys. Rev. B* **34,** 6554. Scalapino, D. J., Loh, E., and Hirsch, J. E., (1986). *Phys. Rev. B* **34,** 8190. Norman, M. R. (1987). *Phys. Rev. Lett.* **59,** 232. Norman, M. R. (1988). *Phys. Rev. B* **37,** 4987.
72. Newns, D. M. (1987). *Phys. Rev. B* **36,** 5595. Newns, D. M. (1988). *Phys. Scripta T* **23,** 113.
73. Torrance, J. B., *et al.* (1988). *Phys. Rev. Lett.* **61,** 1127.
74. Rice, T. M., and Ueda, K. (1985). *Phys. Rev. Lett.* **55,** 995. Rice, T. M., and Ueda, K. (1986). *Phys. Rev. B* **34,** 6420.
75. Miyake, K., Matsuura, T., Sano, K., and Nagaoka, Y. (1988). *J. Phys. Soc. Jpn.* **57,** 722.
76. Anderson, P. W. (1959). *Phys. Rev.* **115,** 2. Also (1963). *In* "Solid State Physics" (F. Seitz and D. Turnbull, eds.), Vol. 14, pp. 99–213, Academic Press, New York. Cf. also Zaanen, J., and Sawatzky, G. A. (1990). *J. Solid State Chem.* **88**.
77. For a review see, e.g., Vonsovskii, S. V. (1974). "Magnetism" Wiley, New York.
78. Zaanen, J., Sawatzky, G. A., and Allen, J. W. (1985). *Phys. Rev. Lett.* **55,** 418; (1986). *J. Magn. Magn. Mat.* **54–57,** 607.
79. (a) Aharony, A., *et al.* (1988). *Phys. Rev. Lett.* **60,** 1330. (b) Tranquada, J. M., *et al.* (1988). *Phys. Rev. Lett.* **60,** 156.
80. Spałek, J., and Honig, J. M. (1990). *In* "Studies of High-Temperature Superconductors" (A. Narlikar, ed.), Vol. 4, Nova Science, New York.
81. The p–p pairing was discussed first by Emery, V. J. (1987). *Phys. Rev. Lett.* **58,** 2794, and also by Emery, V. J., and Reiter, G. (1988). *Phys. Rev. B* **38,** 4547.
82. The coexistence of SDW and SC states has also been discussed within BCS theory: Baltensperger, W., and Strassler, S. (1963). *Phys. Kondens. Mater.* **1,** 20; Nass, M. J., *et al.* (1981). *Phys. Rev. Lett.* **46,** 614. Overhauser, A. W., and Daemen, L. (1988). *Phys. Rev. Lett.* **61,** 1885. The corresponding problem for exchange–mediated superconductivity has been outlined by Parmenter, R. H. (1987). *Phys. Rev. Lett.* **59,** 923.
83. Weidinger, A., *et al.* (1989). *Phys. Rev. Lett.* **62,** 102. Brewer, J. H., *et al.* (1988). *Phys. Rev. Lett.* **60,** 1073.
84. Bednorz, J. G., and Muller, K. A. (1986). *Z. Phys. B* **64,** 189. Chu, C. W., *et al.* (1987). *Phys. Rev. Lett.* **58,** 405. Uchida, S., *et al.* (1987). *Jpn. J. Appl. Phys.* **26,** L1. Wu, M. K., *et al.* (1987). *Phys. Rev. Lett.* **58,** 908.
85. abbé, J., and Bok, J. (1987). *Europhys. Lett.* **3,** 1225.
86. Prelovšek, P., Rice, T. M., and Zhang, F. C. (1987). *J. Phys. C* **20,** L229.
87. Jorgensen, J. D., *et al.* (1988). *Phys. Rev. Lett.* **58,** 1024.
88. Weber, W. (1987). *Phys. Rev. Lett.* **58,** 1371. Barisic, S., Batistic, J., and Friedel, J. (1987). *Europhys. Lett.* **3,** 1231.
89. For review of phonon- and bipolaron-mediated superconductivity see, e.g., Oguri, A. (1988). *J. Phys. Soc. Jpn.* **57,** 2133; de Jongh, L. J. (1988). *In* "Proc. 1st Int. Symp. Superconduct.," Nagoya, Springer—Verlag, New York.
90. Alexandrov, A., and Ranninger, J. (1981). *Phys. Rev. B* **24,** 1164. Alexandrov, A., Ranninger, J., and Robaszkiewicz, S. (1986). *Phys. Rev. B* **33,** 4526. For review see Micnas, R., Ranninger, J., and Robaszkiewicz, S. (1990). *Rev. Mod. Phys*. **62,** 113; Ranninger, J. (1998). *J. Phys. Chem. Solids* **59,** 1759, and references therein.
91. Batlogg, B., *et al.* (1987). *Phys. Rev. Lett.* **59,** 912. Faltens, T. A., *et al.* (1987). *Phys. Rev. Lett.* **59,** 915.
92. Leary, K. J., *et al.* (1987). *Phys. Rev. Lett.* **59,** 1236.
93. Mattheis, L. F., Gyorgy, E. M., and Johnson, D. W., Jr. (1988). *Phys. Rev. B* **37,** 3745. Cava, R. J., *et al.* (1988). *Nature* **332,** 814. Hinks, D. G., *et al.* (1988). *Nature* **333,** 6176.
94. Hinks, D. G., *et al.* (1988). *Nature* **335,** 419.
95. Pei, S., *et al.*, preprint.
96. Rice, T. M. (1988). *Nature* **332,** 780.
97. Dąbrowski, B., personal communication.
98. Varma, C. M. (1988). *Phys. Rev. Lett.* **61,** 2713.
99. Sleight, A. W., Gillson, J. J., and Bierstedt, P. E. (1975). *Solid State Commun.* **17,** 27.
100. For critical estimates of isotope shifts of the $T_c$ value, see Allen, P. B. (1988). *Nature* **335,** 396.
101. Chakraverty, B. K. (1979). *J. Phys. Lett.* **40,** L99, Alexandroy, A. S., and Ranninger, J. (1981). *Phys. Rev. B* **23,** 1796. Alexandrov, A. S., Ranninger, J., and Robaszkiewicz, S. (1986). *Phys. Rev. B* **33,** 4526.
102. Rice, T. M. (1988). *Nature* **332,** 780.
103. Prelovšek, P., Rice, T. M., and Zhang, F. C. (1987). *J. Phys. C* **20,** L229.
104. Little, W. A. (1964). *Phys. Rev.* **134A,** 1416. Ginzburg, V. L. (1970). *Sov. Phys. Uspekhi* **13,** 335.
105. Varma, C. M., Schmitt-Rink, S., Abrahams, E. (1987). *In* "Proceedings of the Conference on Novel Mechanisms of Superconductivity" (S. A. Wolff and V. Z. Kresin, eds.), p. 355, Plenum Press, New York; (1987). *Solid State Commun.* **62,** 681.
106. Weber, W. (1988). *Z. Phys. B* **70,** 323.
107. Scalpino, D. J., Loh, E., Jr., and Hirsch, J. E. (1987). *Phys. Rev. B* **35,** 6694. Schrieffer, J. R., Wen, X. G., and Zhang, S. C. (1988). *Phys. Rev. Lett.* **60,** 944. White, S. R., and Scalpino, D. J. (1997), *Phys. Rev. B* **55,** 6504.
108. For a critical review see Little, W. A. (1988). *Science* **242,** 1390.
109. Fiory, A. T., Hebard, A. F., Mankiewich, P. M., and Howard, R. E. (1988). *Phys. Rev. Lett.* **61,** 1419.
110. Niemeyer, J., Dietrich, M. R., and Politis, C. (1987). *Z. Phys. B* **67,** 155.
111. Smedskjaer, L. C., Liu, J. Z., Benedek, R., Legnini, D. G., Lam, D. J., Stahulak, M. D., and Bansil, A. (1988). *Physica C* **156,** 269.
112. Cf. also Mathur, N. D., *et al.* (1998). *Nature* **394,** 39; Fisk, Z., and Pines, D. *ibid*., p. 22.
113. For review see Shen, Z.-X., and Dessau, D. S. (1995). *Phys. Rep.* **253,** 1ff; Ding, H., *et al.* (1995). *Phys. Rev. Lett.* **74,** 2784; (1996). *Phys. Rev. B* **54,** R9878; Campuzano, J.-C., *et al.* (1998). *In* "The Gap Symmetry and Fluctuations in High-$T_C$ Superconductors" (J. Bok *et al.*, eds.), Plenum Press, New York.
114. Fujimori, A., *et al.* (1998). *J. Phys. Chem. Solids* **59,** 1892, and references therein.
115. Ding, H., *et al.* (1998). *J. Phys. Chem. Solids* **59,** 1888, and refrences therein; Norman, M. R., *et al.*, *ibid*. p. 1902.



# Superconductors, High Temperature

**John B. Goodenough**
*University of Texas*

I. Introduction
II. Superconductivity
III. Metallic Oxides
IV. High-$T_c$ Superconductors

## GLOSSARY

**Bohr magneton $\mu_B$** Magnetic moment of an electron.
**Brillouin zone** Volume in **k** space containing two valence electrons per atomic valence orbital per atom of a primitive unit cell of crystal lattice.
**Correlation energy** Electrostatic electron-electron interactions not accounted for in Hartree-Fock one-electron band theory of an itinerant electron.
**Debye temperature $\Theta_D$** Characteristic temperature proportional to maximum vibrational frequency of atoms of a solid ($k\Theta_D = \hbar\omega_{max}$).
**Isotope** Same element with different nuclear masses.
**k-space** Momentum (or reciprocal-lattice) space in which electron momenta and energies can be plotted.
**Magnetic flux** Lines of magnetic-field strength defining field direction; their density defines field strength.
**Phonon** Quantum of lattice vibrational energy $\hbar\omega$.
**Quasi-particle** Electron of a one-electron energy band renormalized by electron-electron and/or electron-lattice interactions.
**Wave function $\psi$** Quantum-mechanical descriptor of an electron; $|\psi(\mathbf{r})|\,2$ is the probability of finding an electron at position **r**.

## I. INTRODUCTION

In 1908, the Dutch physicist Heike Kammerlingh Onnes succeeded in liquifying helium. This accomplishment made possible the exploration of the low-temperature properties of matter; and in 1911 he reported a phase transition in metallic mercury from a normal state to a superconductive state below a critical temperature $T_c$. What Kammerlingh Onnes observed was an abrupt change in the direct-current (dc) resistance of mercury at a $T_c = 4.15$ K; the normal state exhibited an electrical resistance $R_n$ with attendant joule heating $I^2R_n$ on passing a current $I$, whereas the superconductive state was a "perfect" conductor with no measurable resistance ($R_s = 0$). Moreover, in the absence of a magnetic field, $T_c$ is independent of the shape or the size of the

sample; superconductivity is an intrinsic property of the material.

Since 1911, many materials have been surveyed to determine whether they are superconductors and, if so, the value of their $T_c$. Although there is still little understanding to guide in the search for, or design of, high-$T_c$ superconductors, extensive investigations of the elements, of alloys, of compounds, and of polymers had, by 1985, resulted in several empirical guidelines.

1. Only metals are superconductors.
2. Superconductivity is associated with a dynamic electron–phonon coupling.
3. The highest values of $T_c$ are associated with partially filled $d$ bands, and $T_c$ varies sensitively with the electron–atom ratio for the partially filled band.
4. $T_c$ is suppressed where the conduction electrons exhibit magnetic order at low temperatures.
5. $T_c$ is suppressed where the electron–phonon coupling becomes static, inducing a phase transformation to a nonmetallic state.

From 1911 to 1986, the critical temperature $T_c$ remained below 25 K, increasing by less than 0.3 K per decade (see Fig. 1). Moreover, the existing theory—applicable to nearly all known superconductors—predicted a ceiling

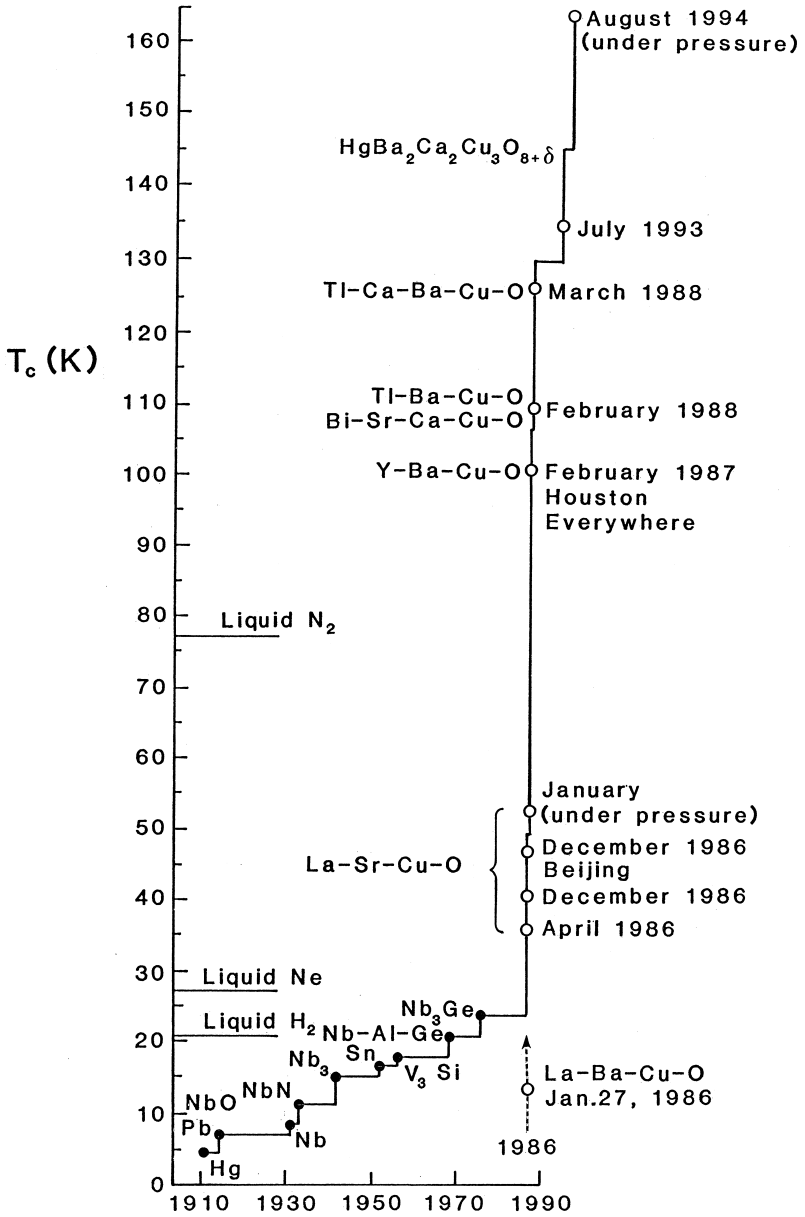



**FIGURE 1** Maximum known $T_c$ versus date of discovery.

for $T_c$ in the neighborhood of 30 K. Nevertheless alternate mechanisms for enhancing $T_c$ had been suggested, and a few experimentalists persisted in the hope of finding a material that exhibited such an enhancement.

Bednorz and Müller, of IBM Zürich, were two who persisted, and in 1986 they reported the existence of superconductivity above 30 K in a multiphase oxide containing Ba, La, and Cu. Although their discovery was initially overlooked within their own corporation, it was immediately pursued by Kitazawa and Uchida of the University of Tokyo, who identified the superconductor phase as $La_{2-x}Ba_xCuO_{4-y}$ having a well-known intergrowth structure. Announcement of this identification in December 1986 at conferences in Boston, Massachusetts, and Bangalore, India, triggered an excited effort to reproduce, extend, and "explain" this breakthrough. Within weeks, substitution of Sr for Ba had increased the $T_c$ to 40 K and attempts to substitute Y for La had resulted in a polyphase mixture containing a new superconductor with a $T_c \approx 90$ K. First announced in the *New York Times* by Chu of the University of Houston—but found independently at the same time by workers in Tokyo, Peking, and Bangalore—the latter discovery electrified the entire solid-state community. A $T_c$ higher than 77 K, the boiling point of nitrogen, introduced an entirely new technical dimension, and conventional theory clearly could not be stretched to include this new finding without some radical modification. A race to articulate this theoretical modification, to establish it, and to use it to find new high-$T_c$ superconductors had begun. Simultaneously, the problem of processing these new materials for technological exploitation began to be addressed in more than 1000 laboratories around the world. After 12 years of intensive effort by many groups, there is no consensus yet even on the character of the normal state out of which the superconductive pairs condense, and processing the brittle ceramic materials into flexible wires or tapes that can remain superconducting in high magnetic fields remains a technical challenge. This article can be only a personal commentary on this activity.

## II. SUPERCONDUCTIVITY

### A. Phenomenology

#### 1. Nomenclature

A superconductor is any material that undergoes a transition from the normal state to the superconductive state below a critical temperature $T_c$. It is superconducting when it is carrying a resistance-free ($R_s = 0$) current (i.e., a supercurrent) in the superconductive state.

#### 2. The Normal State

Superconductors are metallic in the normal state. Each conduction electron of a metal is said to be itinerant because it belongs equally to all like atoms at energetically equivalent lattice positions in a crystal; each may also belong, to a lesser extent, to other atoms in the crystal. Because their position in real space is poorly defined, itinerant electrons are characterized by their momentum vector $\mathbf{k}$, where the momentum $\mathbf{p}$ transforms to $\hbar\mathbf{k}$ ($\hbar = h/2\pi$, where $\hbar$ is Planck's constant) in the absence of a magnetic field. Where the like-atom interatomic interactions are much stronger than the intraatomic electron–electron interactions, each itinerant electron may be described as a single particle moving in the average electrostatic potential created by the atomic nuclei and all the other electrons; they therefore occupy one-electron states, each having an energy $\varepsilon_{\mathbf{k}}$ and, in the absence of a magnetic field, a twofold spin degeneracy. Moreover, the one-electron energies for an $N$-atom array are grouped into energy bands containing $2N/n$ states per atomic orbital, where $n \geq 1$ is an integer that depends on the translational symmetry of the crystal. The density of one-electron states $N(\varepsilon)$ per infinitesimal energy interval $d\varepsilon$ is a fundamental parameter; so also is the effective mass entering the relationship $\varepsilon_{\mathbf{k}} - E_0 = \hbar^2 k^2/2m^*$, where $E_0$ is a band-edge reference energy. The Pauli exclusion principle allows one electron per state, so at $T = 0$ K the electron states are successively occupied from the bottom of an energy band until all the electrons are accounted for.

What distinguishes a metal from a semiconductor such as silicon is that, in a metal, an occupied band of itinerant-electron states is only partially filled (Fig. 2). In this case there is an abrupt change in the electron population at a surface in $\mathbf{k}$ space; this surface is called the Fermi surface, and the energy of the Fermi surface is called the Fermi energy $E_F$. At finite temperatures, electrons are thermally

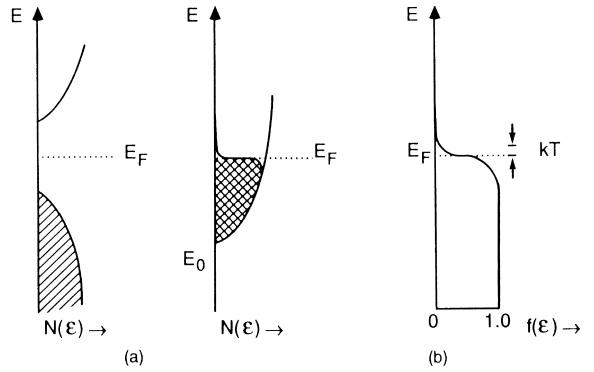



**FIGURE 2** Energy versus (a) the density $N(\varepsilon)$ of one-electron states for a semiconductor and a metal and (b) the Fermi–Dirac distribution function $f(E)$.

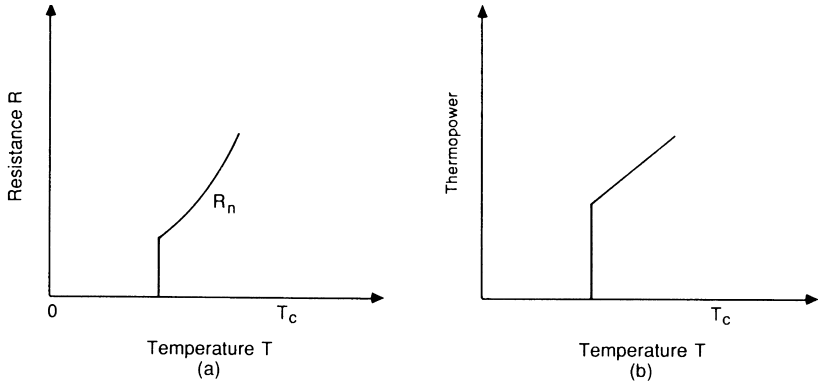



**FIGURE 3** Temperature dependence of (a) the resistance and (b) the thermopower of a superconductor.

excited from occupied to unoccupied states in the band: this process gives rise to a Fermi–Dirac statistical distribution that, for $k_B T \ll E_F$ ($k_B$ is Boltzmann's constant), leaves intact the concept of a Fermi surface (Fig. 2b).

As the ratio of interatomic/intraatomic interaction energies decreases toward unity, the band of one-electron energies narrows, increasing $N(E_F)$ for a given $E_F$; single electrons become transformed into quasiparticles that cannot be described by the average potential of all other electrons because correlations between the electrons on nearest like neighbors reflect the intraatomic electron–electron interactions and/or because electron–phonon interactions "dress" the electrons in local crystallographic distortions. Although these interactions reduce the discontinuity in the electron population at $E_F$, an identifiable Fermi surface remains as long as the specimen is metallic.

Partially occupied, narrow energy bands of quasiparticle states may lose their Fermi surface at $E_F$ by inducing a diffusionless phase transition at low temperatures that splits the band into bands of occupied and empty states separated by a finite energy gap $E_g$. Three types of transitions cause such a splitting: (1) an atomic clustering that changes the translational symmetry of the crystallographic structure, (2) a magnetic ordering that changes the translational symmetry and/or the degree of electron localization at atomic positions, and (3) the onset of superconductivity caused by a pairing of one-particle states having energies near $E_F$ into an ordered condensate of two-particle states. Since the first two processes compete with the onset of superconductivity, any realization of a high $T_c$ must involve a mechanism that suppresses the stabilization of atomic clustering and of magnetic ordering of the conducting electrons.

## 3. The Superconductive State

The critical temperature $T_c$ marks the boundary between two distinguishable thermodynamic states of the material, each with its own set of properties. The superconductive state is distinguished from the normal state by its electric, magnetic, thermodynamic, and tunneling properties.

*a. Electric.* The dc resistance $R$ of a superconductor wire drops abruptly at $T_c$, from $R_n > 0$ in the normal state to $R_s = 0$ in the superconductive state (Fig. 3a).

In the normal state, the potential difference $V$ between the ends of a wire of length $l$ and resistance $R_n$ is, by Ohm's law,

$$V = IR_n \tag{1}$$

if the wire carries a current $I$. By definition, a constant electric field $E = V/l$ then exists in the wire. In the superconductive state, on the other hand, $R_s = 0$ makes $V = E = 0$. There is no constant electric field in, or potential difference across, a superconducting wire. Consequently all the thermoelectric effects present in the normal state vanish abruptly at $T_c$. For example, in the normal state an applied temperature gradient $\Delta T$ gives rise to an electric field $E$ in the conductor; the thermoelectric power, defined as $E/\Delta T$, vanishes with $E$ below $T_c$ (Fig. 3b).

The resistance $R_s$ of the superconductive state is strictly zero only for direct currents of a constant value. If the current changes with time, as in an alternating-current (ac) application, then $R_s$ is not zero. Nevertheless at temperatures $T \ll T_c$, $R_s$ remains much less than the resistance $R_n$ of the normal state for frequencies $\nu < E_g/h$, where $E_g$ is the energy gap (see Section II.C) at the Fermi energy of the superconductive state. The ratio $R_s/R_n$ increases from a small value to nearly 1 in a finite frequency interval $\Delta\nu$ (Fig. 4). The width $\Delta\nu$ broadens and its midpoint shifts to a lower frequency as $T$ increases to $T_c$.

*b. Magnetic.* The magnetization $M$ of a substance is defined as its magnetic moment per unit volume. The magnetic susceptibility per unit volume is defined as

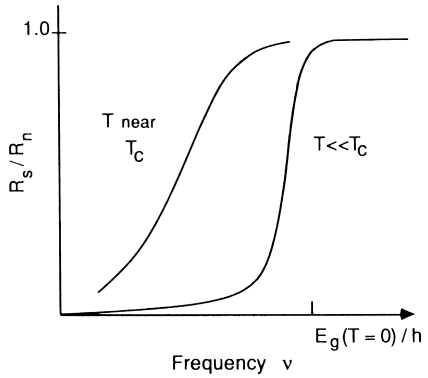



**FIGURE 4** Ratio of superconductive-to-normal resistance versus frequency for different values of $T < T_c$.

$$\chi = \mu_0 M / B_a, \tag{2}$$

where $B_a/\mu_0$ is the intensity of an applied magnetic field.

Substances with a negative magnetic susceptibility are called diamagnetic; those with a positive susceptibility are called paramagnetic. Diamagnetism reflects changes in electron motion that oppose the applied magnetic field; paramagnetism reflects an increase in the populations of electron spins (or of localized atomic moments) oriented parallel to the applied magnetic field. The inner, closed-shell atomic cores retain spin-paired electrons; they always give a small, temperature-independent diamagnetic contribution $\chi_{core} < 0$ to the susceptibility. However, in superconductors the dominant contribution is made by the conduction electrons.

In the normal state of a superconductor, an applied magnetic field $\mathbf{B}_a$ defines the orientations of the one-electron spin states and stabilizes the parallel spin states relative to the antiparallel spin states of the conduction band by an energy $2\mu_B B_a/\mu_0$, where $\mu_B$ is the magnetic moment imparted by a single-electron spin (the Bohr magneton) (Fig. 5). The resulting change in electron population of

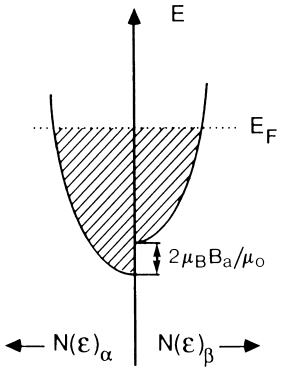



**FIGURE 5** Shifting of $\alpha$-spin and $\beta$-spin energies in an applied magnetic field strength $B_a/\mu_0$.

parallel ($\alpha$) versus antiparallel ($\beta$) spin states creates, to lowest order, a paramagnetic magnetization

$$M \approx \mu_B^2 N(E_F) \cdot B_a/\mu_0, \tag{3}$$

where $N(E_F)$ is the density of one-electron states at $E_F$. First derived by Pauli, this contribution to the total magnetization is called the Pauli spin magnetization. Equation (3) applies to broad energy bands with $E_F \gg k_B T$.

Changes in the motions of the conduction electrons introduce a diamagnetic contribution. Landau has shown that where Eq. (3) applies, this contribution to $M$ is minus one-third the Pauli spin magnetization, so the total conduction-electron contribution to the normal-state susceptibility is paramagnetic and temperature independent:

$$\chi_{cond} = \mu_0 M/B_a = (2/3)\mu_B^2 N(E_F) > 0. \tag{4}$$

It generally dominates the total temperature-independent susceptibility,

$$\chi = \chi_{cond} + \chi_{core}. \tag{5}$$

If the energy bands are narrow, it is necessary to introduce into $\chi_{cond}$ a temperature-dependent enhancement factor.

In the superconductive state the situation is quite different. Meissner and Ochsenfeld found that, if a superconductor is cooled in a magnetic field to below $T_c$, the magnetic flux within the superconductor in the normal state is pushed out of the superconductive state as illustrated in Fig. 6. This phenomenon is called the Meissner effect.

The extent to which the internal magnetic flux is expelled by the Meissner effect depends not only on the temperature and the magnitude of the applied magnetic field $\mathbf{B}_a$, but also on the sample shape and its orientation with respect to $\mathbf{B}_a$. A long, thin cylinder (or wire) oriented with its long axis parallel to $\mathbf{B}_a$ has a negligible demagnetizing field within it, and the internal magnetic field is

$$B = B_a + \mu_0 M, \tag{6}$$

where $M$ is the induced magnetization. Complete expulsion of $B$ would make $B = 0$ inside the superconductor, to give perfect diamagnetism, with

$$M = -B_a/\mu_0 \quad \text{and} \quad \chi = -1. \tag{7}$$

The magnetization curve for such a situation at $T < T_c$ is illustrated in Fig. 7a. It is found to apply quantitatively to pure specimens for applied fields less than a critical field strength $H_c(T)$:

$$B_a/\mu_0 \leq H_c. \tag{8}$$

Two types of superconductors can be distinguished. Type I, originally termed soft superconductors, exhibit an abrupt loss of the Meissner effect at $H_c$. Type II superconductors exhibit two critical field strengths (Fig. 7b): $H_{c1}$,

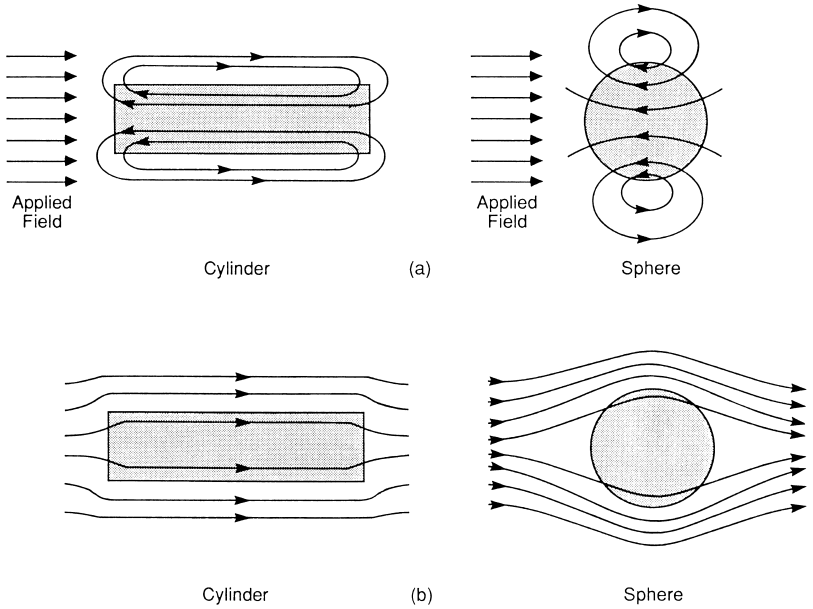



**FIGURE 6** (a) The magnetic field in a superconductive cylinder and sphere induced by a uniform applied field $B_a$. (b) The total induced plus applied field.

beyond which the Meissner effect is less than complete; and $H_{c2}$, beyond which it disappears completely.

In the absence of a magnetic field, the transition at $T_c$ is always sharp; the material is, essentially, wholly superconductive at temperatures $T < T_c$ and wholly normal at $T > T_c$. At temperatures $T < T_c$, the change from the superconductive to the normal state at the critical field strength $H_c(T)$ is not sharp in type II superconductors, and for most geometries it is not sharp even in type I superconductors. Penetration of magnetic flux occurs between $H_{c1}$ and $H_{c2}$ in Fig. 7b; in this range of applied field, parts of the specimen are in the normal state and parts are in the superconductive state. It is possible to distinguish type I from type II superconductors by the way in which the normal-state regions penetrate the superconductive state with increasing $B_a/\mu_0$.

The distinction between the two types of superconductors is illustrated in Fig. 8 for the case of a $\mathbf{B}_a$ applied perpendicular to a plane slab of a superconductor. If the superconductor is type I, the normal regions enter as relatively thick, parallel laminae; and if both normal and superconductive states coexist, the superconductor is said to be in an intermediate state. If the superconductor is type II, the normal regions enter as numerous, extremely thin tubular filaments separated by small distances ($\leq 10^{-5}$ mm), and for $H_{c1} < H < H_{c2}$ the superconductor is said to be in a

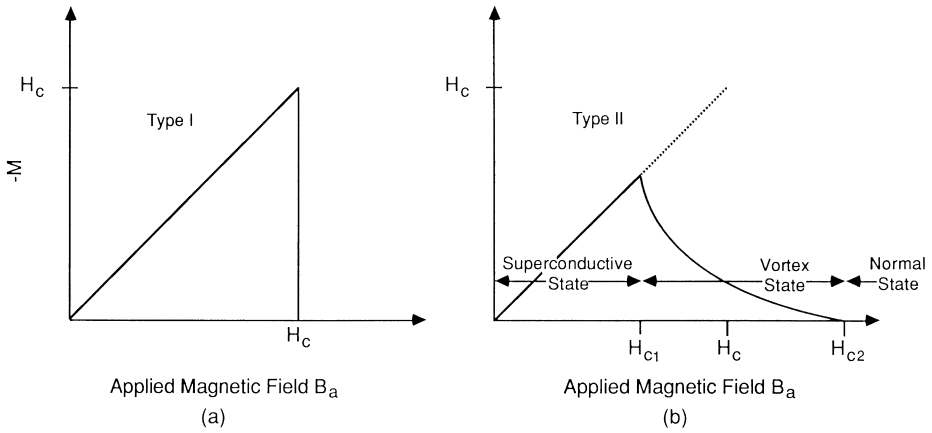



**FIGURE 7** Magnetization versus applied magnetic field $B_a$ for a bulk superconductive cylinder with its axis parallel to $B_a$ for (a) type I and (b) type II superconductivity.

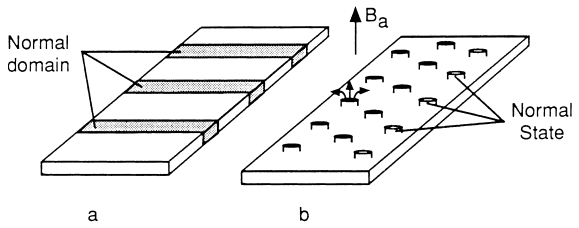



**FIGURE 8** (a) Intermediate state of a type I superconductor. (b) Vortex state of a type II superconductor.

vortex state (or in a mixed state). The field $H_{c2}$ in Fig. 7b depends on the mobility of the vortices. Large values of $H_{c2}$ are obtained by introducing crystalline imperfections that pin the vortices; flux pinning introduces hysteresis in Fig. 7b between the curve obtained by increasing and that obtained by decreasing $B_a$. A hard superconductor is a type II superconductor exhibiting a large magnetic hysteresis due to vortex pinning.

*c. Thermodynamic: Order parameter.* In the absence of an applied magnetic field, the transition from the superconductive to the normal state is second order: there is no discontinuity at $T_c$ in either entropy (no latent heat) or volume (no thermal hysteresis), but there is a sharp discontinuity $\Delta C$ in the heat capacity $C$ (Fig. 9).

A decrease in entropy on going from the normal to the superconductive state shows that the superconductive state is more ordered and can be described by an order parameter that varies smoothly with temperature from unity at 0 K to 0 at $T = T_c$. A natural choice for the order parameter in classical physics is $n_s/n_0$, the local density of superconductive electrons normalized to its value at 0 K. However, superconductivity is a quantum—not a classical—phenomenon, and a more profound choice is the corresponding quantum physics wave function

$$\psi(r) = |\psi(r)| \exp[i\phi(r)], \tag{9}$$

where $\phi(r)$ is a phase factor and $n_s \equiv |\psi|^2$.

*i. Persistent currents and flux Quantization.* The most basic implication of the existence of a phase factor in $\psi(r)$ is the quantization of magnetic flux in a superconducting ring. Consider first the macroscopic ring in Fig. 10. The application perpendicular to the ring of a uniform magnetic field of flux density $B_a$ that varies with time $t$ creates a voltage that induces a current $I(t)$ to circulate in the ring. According to Lenz's law

$$-A_r \frac{dB_a}{dt} = RI(t) + L\frac{dI(t)}{dt}, \tag{10}$$

where $A_r$ is the area enclosed by the ring, $R$ the resistance of the ring, and $L$ the inductance of the ring. If there is no applied magnetic field ($B_a = 0$), then the solution of Eq. (10) is

$$I(t) = I(0)\exp(-Rt/L), \tag{11}$$

which shows that any initial current circulating in the ring decays exponentially to zero in the normal state. However, in the superconductive state an $R = R_s = 0$ makes $I(t) = I(0)$, and the initial current $I(0)$ continues to circulate around the ring without any change in its magnitude. Such currents are known as persistent currents; and any current $I$ circulating around the ring produces a magnetic flux threading the ring equal to $LI$. In the presence of $B_a$, the total flux $\Phi$ threading the ring is $\Phi = A_r B_a + LI$, and Eq. (10) reduces to

$$d\Phi/dt = -IR. \tag{12}$$

In the superconductive state, $R = 0$ gives

$$\Phi = \text{constant} = A_r B_s, \tag{13}$$

where $B_s$ is the ring magnetic field.

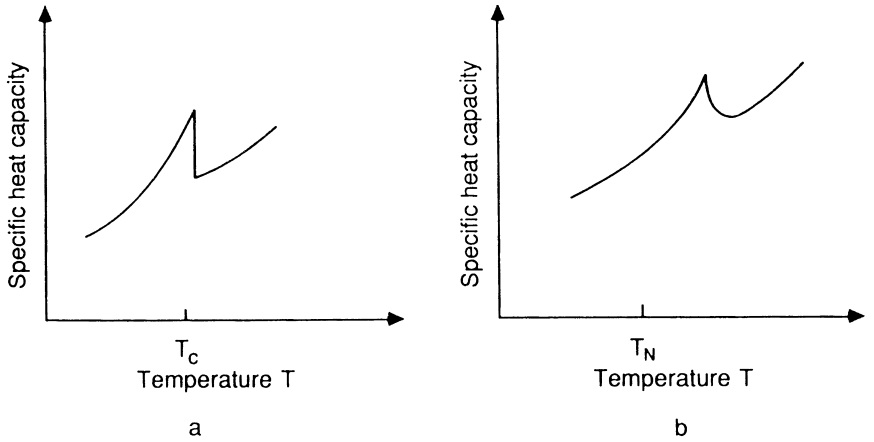



**FIGURE 9** Temperature dependence of the specific heat capacity for (a) a BCS superconductor and (b) a typical antiferromagnet.

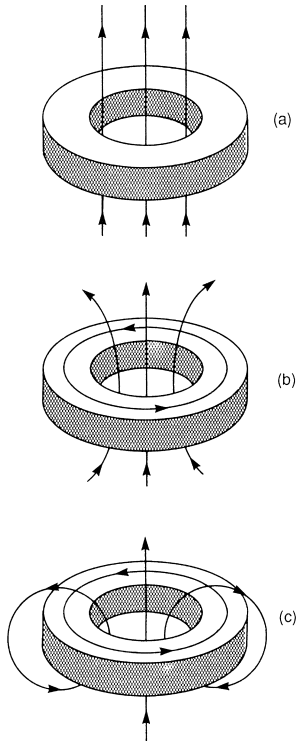



**FIGURE 10** Production of persistent current in a superconductive ring: (a) $B_a$ applied normal to a ring at $T > T_c$; (b) ring cooled to $T < T_c$; (c) $B_a$ removed, leaving persistent current.

A persistent current and its associated magnetic field are established in a macroscopic ring by introducing a current $I(t)$ into the ring before it is cooled to below $T_c$; the external circuit is switched off only after the ring is in its superconductive state. In a type II superconductor below $T_c$, the penetration of flux in a $B_a/\mu_0 > H_{cl}$ is accomplished by the movement of a normal-state filament into the superconductive state; the normal-state region contains flux, and its movement into the superconductor creates a microscopic persistent current within the surrounding superconductive state that traps the flux within the normal-state core of the vortex. The amount of flux within a microscopic vortex is quantized because of the phase factor $\phi$ in Eq. (9).

For a superconducting ring, the single-valuedness of $\psi$ requires that $\phi(r)$ return to itself modulo $2\pi$ on going once around the circuit; that is, if the orbit of a superconductive electron is quantized to a path length that is an integral number of electron wavelengths, then the electron neither gains nor loses energy. For a superconductive particle in an orbit of radius $r$, the condition for quantization, and hence the existence of a supercurrent, is

$$\mathbf{p} \cdot 2\pi r = N\hbar, \tag{14}$$

where $\hbar$ is Planck's constant, $N$ is an integer, and the canonical momentum in a local magnetic field $\mathbf{B} = \nabla \times \mathbf{A}$ is

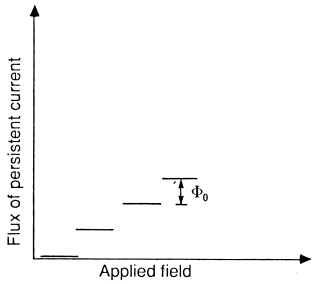



**FIGURE 11** Quantum steps in flux of persistent current versus applied field for a type II superconductor.

$$\mathbf{p} = \hbar(\mathbf{k}_1 + \mathbf{k}_2 \cdots \mathbf{k}_n) + ne\mathbf{A} \tag{15}$$

for a superconductive particle consisting of $n$ electrons. If the superconducting particle consists of a pair of electrons having opposite momentum vectors $\mathbf{k}$ and $-\mathbf{k}$, then $n = 2$ and $\mathbf{p} = 2e\mathbf{A}$. Moreover, the flux enclosed by a vortex is $\Phi = 2\pi r A$, so that Eq. (14) reduces to

$$\Phi = Nh/2e. \tag{16}$$

This result is of outstanding importance. It means that if the superconductive state consists of paired electrons, then in a closed superconducting circuit the flux is quantized in units of

$$\Phi_0 = h/2e = 2.07 \times 10^{-15}\ \text{Wb}. \tag{17}$$

The existence of flux quantization and the magnitude of $\Phi_0$ have been confirmed experimentally (Fig. 11); these experiments demonstrate not only the quantum character of the ordering, but also that ordering in the superconductive state consists of the formation of pairs of electrons having opposite momentum vectors $\mathbf{k}$ and $-\mathbf{k}$. Moreover, the perfect diamagnetism associated with this order indicates pairing of $s = \frac{1}{2}$ and $s = -\frac{1}{2}$ spins of a superconductive electron pair. The critical field strength $H_c$ is that required to decouple the spin pairing of a superconductive particle.

A localized atomic moment interacts with the conduction-band electrons via spin–spin "exchange" to produce a local magnetic field; if the local field strength exceeds $H_c$, superconductivity is suppressed. Ferromagnetic ordering of localized moments generally suppresses superconductivity, but antiferromagnetic ordering may not be incompatible with superconductivity.

*ii. Energy gap.* The electronic heat capacity in the superconductive state $C_{es}$, normalized to its value $\gamma T_c$ in the normal state at $T = T_c$, is commonly found to vary exponentially as $-1/T$ at temperatures $T \ll T_c$:

$$\begin{aligned} C_{es}/\gamma T_c &= a \exp[-bT_c/T] \\ &= a \exp[-\Delta_0/k_B T], \end{aligned} \tag{18}$$

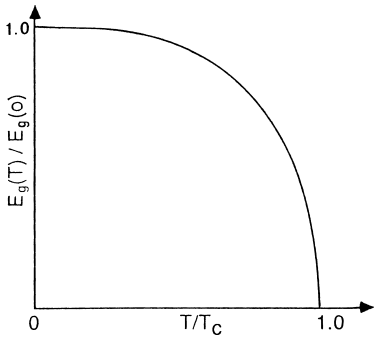



**FIGURE 12** Normalized energy gap $E_g(T)/E_g(0)$ versus normalized temperature $T/T_c$.

which is suggestive of an excitation of electrons across an energy gap $E_g = 2\Delta_0$. An energy gap $E_g$ has been measured independently by spectroscopic techniques; it compares favorably at the lowest $T$ with the calorimetric data (i.e., $b \approx \Delta_0/k_B T_c$). Moreover, $E_g = 2\Delta(T)$ is found to decrease with increasing temperature, from $2\Delta_0$ at 0 K to 0 at $T = T_c$ (Fig. 12) with a Brillouin-function dependence, which makes $\Delta(T)/\Delta_0$ a measure of the order parameter in a mean-field description of the transition. In addition, the presence of an energy gap at the Fermi energy $E_F$ shows that the superconductive electron pairs have been formed by condensing out single-electron states from the vicinity of $E_F$.

*iii. Isotope effect.* The $T_c$ for mercury, and most other elemental superconductors, varies smoothly with the average atomic mass $M$ as the isotope mix is varied:

$$M^{\alpha} T_c = \text{constant}. \tag{19}$$

This correlation of $T_c$ and $M$ is known as the isotope effect. This early observation shows that for conventional superconductors, electron–phonon interactions play an important role in the binding of superconductive pairs of electrons. In the simplest theory, only the electronic states within an energy $k_B\Theta_D$ of $E_F$, where $\Theta_D$ is the Debye temperature, can be coupled by electron–phonon interactions. This simplest theory limits the magnitude of the energy gap to a specific multiple of $k_B T_c$ and predicts, for an elemental superconductor, an $\alpha = \frac{1}{2}$ in Eq. (19) [see Eqs. (42) and (43)]. Although an $\alpha = \frac{1}{2}$ has been observed for mercury, there is nothing sacred about this value even for the elements; for example, an $\alpha = 0$ for Zr and Ru does not signal the absence of a phonon mechanism in these two superconductors.

An electron–lattice mechanism for binding a superconductive pair leads to an upper limit for $T_c$ of about 30–40 K; a higher $T_c$ requires either another type of superconductive pair, the bipolaron, which is stabilized in the limit of strong electron–lattice coupling, or an electronic enhancement of the electron–lattice mechanism. An electronic enhancement would replace $\Theta_D$ with $\Theta_e \approx \hbar\omega_e/k_B$, where $\hbar\omega_e \ll E_F$ is the energy of the electronic excitations that enhance the pairing potential energy.

*iv. Many-body condensate.* Significantly, the usual superconductor transition is much sharper than other second-order transitions. A second-order magnetic-ordering transition, for example, exhibits a substantial temperature range of short-range order above the critical temperature for long-range order. In this case, each atomic moment interacts strongly with only a few near neighbors, so thermodynamic fluctuations not treated by a mean-field theory play an important role. In conventional superconductors, only small vestiges of superconductivity remain above $T_c$, and any resistivity remaining in the superconductive state is infinitesimally small. This observation indicates that each electron pair in the superconductive state is strongly coupled to all the other pairs in a many-body condensate. To break the binding of a given electron to the condensate costs a minimum energy $\Delta_0$. This many-body aspect of the superconductive condensate makes it difficult to depict in real space the nature of the electron ordering that is occurring. An inability to picture the condensate in real space has hindered formulation of a chemical guide for the search for new high-$T_c$ materials.

*v. Temperature dependence of* $H_c$. In the presence of a magnetic field $B_a$, the transition at $T_c$ becomes first order. In a type I superconductor, the increase in free energy at $B_a/\mu_0 = H_c$ is, from Eqs. (7) and (8),

$$\Delta G = \int_0^{H_c} \mu_0 M dH = \tfrac{1}{2}\mu_0 H_c^2, \tag{20}$$

and the latent heat at the transition becomes

$$Q = T\,\Delta S = T\frac{d(\Delta G)}{dT} = \mu_0 T H_c \frac{dH_c}{dT}, \tag{21}$$

which vanishes at $T = T_c$ where $H_c = 0$. It is found empirically that at temperatures below $T_c$, the entropy difference is described by

$$\Delta S = \gamma T\left[1 - (T/T_c)^2\right]. \tag{22}$$

Equating Eqs. (21) and (22) and integrating with respect to the boundary conditions $H_c = 0$ at $T = T_c$ and $H_c = H_0$ at $T = 0$ K gives the relation

$$H_c = H_0\left[1 - (T/T_c)^2\right] \tag{23}$$

for the transition between the superconductive and the normal state in the presence of an applied field (Fig. 13). The extent of the intermediate-state region depends on the shape of the sample and its orientation with respect to $\mathbf{B}_a$.

*d. Tunneling.* If two metals are separated by an insulator, the insulator acts as a barrier to electron flow from

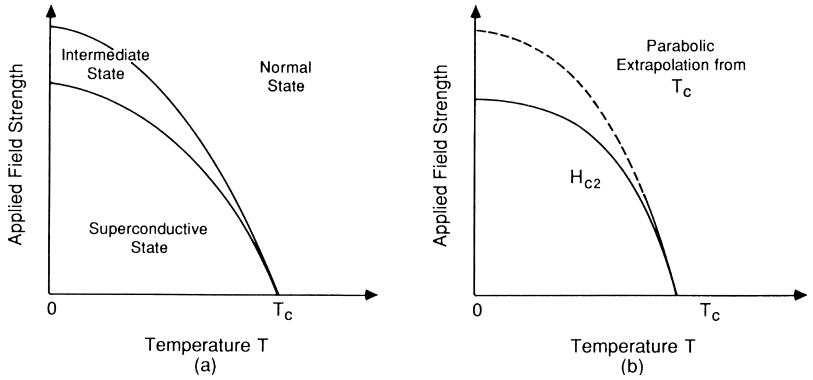



**FIGURE 13** Variation with temperature of (a) $H_c$ of type I and (b) $H_{c2}$ of type II superconductors.

one metal to the other. However, the conduction-electron wave functions extend beyond the metal surface, decaying exponentially in magnitude with the distance from the surface. If the insulating layer is thin enough (less than 10 to 20 Å), a significant amplitude extends through the insulating layer into the other metal. If an empty electronic state of equal energy is also available in the other metal, then there is a finite probability that an electron impinging on the barrier will pass through the insulating layer. This phenomenon is called tunneling.

If both metals are superconductors, two types of particles may tunnel: single quasiparticles and paired superconductive particles. Tunneling of single quasiparticles has been used to measure the energy gap in the superconductive state; tunneling of superconductive particles—called Josephson tunneling—exhibits unusual quantum effects that have been exploited in a variety of quantum devices.

In 1962, Josephson proposed that a tunnel junction between two superconductors—each in their superconductive state—should exhibit a zero-voltage supercurrent in the direction $x$ perpendicular to the junction,

$$I_x = I_{0x} \sin \gamma, \tag{24}$$

due to the tunneling of superconductive electron pairs. Both the phase differences $\phi_2 - \phi_1$ of the wave function on either side of the insulating layer and the canonical momentum of Eq. (15) in the presence of a magnetic flux enter into

$$\gamma = (\phi_2 - \phi_1) - \frac{2\pi}{\Phi_0} \int_1^2 A_x \, dx. \tag{25}$$

A maximum dc flows in the absence of any electric or magnetic field. This is the dc Josephson effect.

Josephson further predicted that if a voltage difference $V$ is applied across the junction, the parameter $\gamma$ becomes time dependent,

$$\gamma(t) = \gamma(0) - (4\pi eVt/h), \tag{26}$$

which means that the current oscillates with a frequency

$$\nu = 2eV/h. \tag{27}$$

This is the ac Josephson effect.

These predictions have been verified experimentally and shown to apply to any sufficiently thin "weak link" in a superconducting circuit. A weak link can be any planar defect at which $T_c$ is sharply reduced from its value in the bulk superconductor. Such weak links appear to limit the supercurrents in the new high-$T_c$ superconductors.

## B. Applications

The technical applications of superconductivity have exploited all its basic properties. However, an extensive commercial potential has been made possible only by the discovery of type II superconductors and Josephson tunneling.

### 1. High Magnetic Field, High Direct Current

The discovery of zero dc resistance, which makes possible macroscopic persistent currents, immediately raised the hope of building a solenoid magnet of superconductive wire capable of generating an intense magnetic field at manageable power levels. Although no energy is expended by a static magnetic field, the energy required to create and sustain an intense magnetic field with a normal conductor is prohibitive.

Attempts to exploit this concept encountered the intrinsic limitation imposed by $H_c$. A cylindrical wire of radius $r_w$ carrying a current $I$ has, at its surface, a magnetic field strength produced by the current

$$H_{surf} = I/2\pi r_w. \tag{28}$$

A supercurrent may increase until $H_{surf} = H_c$; for any current higher than the critical current,

$$I_c = 2\pi r_w H_c. \tag{29}$$

The surface of the wire is transformed to the normal state. In type I superconductors, critical field strengths $H_0 \leq 80$ kA/m (1 k0e) are not sufficient to replace an iron-core magnet. However, a type II hard superconductor is capable of remaining superconductive to high magnetic fields $H_{c2}$; and the generation of high magnetic fields with type II superconductors is now used in a wide range of applications.

### 2. Alternating-Current Devices

A low $R_s/R_n$ ratio requires ac operation at $T \ll T_c$ and $\nu \ll E_g/h$. Type I superconductors may retain an $R_s \approx 0$ up to 100 MHz. This property has enabled the realization of very high-frequency linear electron accelerators with magnifications up to $10^{10}$; they can operate continuously with only a fraction of the power requirements of conventional accelerators.

Larger energy gaps $E_g$ in type II superconductors permit low-loss ac transmission over superconductive strip lines to even higher frequencies. On the other hand, attempts to use superconductors in ac power devices remain restricted to specialty applications such as space vehicles where, with type II superconductors, high current densities in high fields permit significant reductions in weight and size.

### 3. Levitation

The Meissner effect is demonstrated in the classroom by levitation of a bar magnet over a superconductive bowl. The experiment begins with a bar magnet resting on the bottom of a shallow bowl of superconductor in its normal state. The bowl is then cooled to below $T_c$; expulsion of a magnetic field from the bowl creates an "image" magnet that exerts a repulsive force on the real magnet, causing it to rise until this force is balanced by the weight of the magnet.

A most spectacular application of this principle is the "levitated train," which requires high magnetic fields and, therefore, a hard, type II superconductor.

### 4. Bolometer

A bolometer detects electromagnetic radiation by an absorption of radiation that increases its temperature. The temperature increase $\Delta T$ is related to the energy $\Delta E$ absorbed per unit mass via the specific heat capacity $C_v$:

$$\Delta T = \Delta E / C_v. \tag{30}$$

At low temperatures, a low $C_v$ enhances $\Delta T$ for a given $\Delta E$. A type I superconductive detector is designed to operate in the intermediate state where a small $\Delta T$ gives rise to a large resistance change $\Delta R$ as illustrated in Fig. 14.

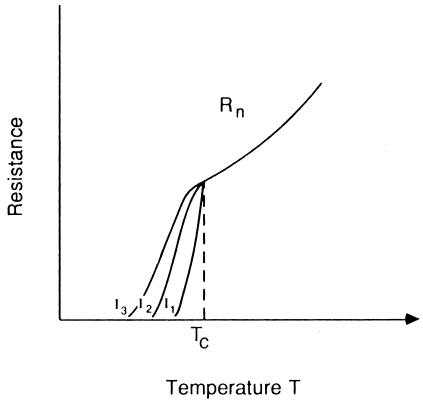



**FIGURE 14** Temperature variation of resistance for a wire carrying a current $I_1 < I_2 < I_3$.

The ability to amplify $\Delta R$ electronically makes the superconductive bolometer an extremely sensitive radiation detector; it is particularly important in the far-infrared region of the spectrum, where most other types of radiation detectors are inoperative.

### 5. Josephson Tunneling

Practical application of the dc Josephson effect has been realized in very sensitive galvanometers and magnetometers. The SQUID (superconducting quantum interference device) magnetometer, for example, is used for measuring small magnetic fields, with extensive use in geological surveying. A laboratory SQUID was the first practical demonstration of the high-$T_c$ superconductor oxides.

The ac Josephson effect has been used in precision determinations of the value of $h/e$.

Applications in the computer field promise higher-density, lower-power components; however, their realization in practice requires an exquisite control of materials processing that is particularly demanding for the new high-$T_c$ superconductor oxides.

## C. Theory

### 1. History

Once the basic phenomena of zero resistance and the Meissner effect had been established, the experimental strategies responsible for our understanding of superconductivity were guided by theory. The theory began with purely phenomenological equations; these equations introduced fundamental length parameters as well as the order parameter, and their application permitted Abrikosov to explain the distinction between type I and type II superconductivity.

The quantum theory of Bardeen, Cooper, and Schrieffer (BCS) introduced numerical values for three universal

ratios relating $T_c$ to $\Delta_0$, $H_0$, and the jump $\Delta C(T_c)$ in the specific heat at $T_c$. Refinement of the BCS theory by Eliashberg has expressed these ratios in terms of two parameters, and with this refinement the agreement between theory and experiment for the three universal ratios is truly remarkable over a wide range of conventional superconductors.

The phenomenon of high-$T_c$ superconductivity cannot be accounted for with the Eliashberg theory if the theory is restricted to binding of the superconductive electron pairs by a dynamic coupling to phonons. This breakdown of existing theory has split theorists into two camps: one camp would extend the BCS or Eliashberg theory by introducing into the pair-binding potential energy an electronic mechanism together with the phonon mechanism; the other camp would break from the "weak-coupling" theory to construct a "strong-coupling" theory in which electron pairs form as a disordered array of "bipolarons" at temperatures $T > T_c$, ordering into the superconductive state occurring only below $T_c$.

## 2. London Equation

In order to account for zero resistance and the Meissner effect in the superconductive state, the London brothers postulated that the local current density in the superconductive state is proportional to the vector potential $\mathbf{A}$, where $\mathbf{B} = \nabla \times \mathbf{A}$:

$$\mathbf{j}_s = \left(1/\mu_0 \lambda_L^2\right)\mathbf{A}. \tag{31}$$

Applying the Maxwell equation $\nabla \times \mathbf{B} = \mu_0 \mathbf{j}_0$, applicable under static conditions, to Eq. (31) gives, on taking the curl of both sides of Maxwell's equation,

$$\nabla^2 \mathbf{B} = \mathbf{B}/\lambda_L^2 \tag{32}$$

for a superconductive state. This equation accounts for the Meissner effect because it does not allow a solution uniform in space unless $B = 0$. Moreover, Maxwell's equation shows that $j = 0$ wherever $B = 0$.

On the other hand, Eq. (32) does allow a solution for $B$ that is nonuniform in space. If a field $B_a$ is applied parallel to an external surface, as illustrated in Fig. 15, then Eq. (32) gives the solution

$$B(x) = B(0)\exp(-x/\lambda_L), \tag{33}$$

where $x$ is the vertical distance into the superconductor from the surface and $B(0)$ is the value of $B_a$ at $x = 0$. Thus $\lambda_L$ measures the depth of penetration of the magnetic field; it is known as the London penetration depth. The current flowing in the superconductor responsible for expelling $B$ is confined to a thin surface layer.

A measure of the magnitude of $\lambda_L$ can be obtained from the canonical momentum $\mathbf{p} = m\mathbf{v} + e\mathbf{A}$. If the average superconductive-particle momentum is zero in the ground state, then the average velocity is $\langle \mathbf{v}_S \rangle = -e\mathbf{A}/m$. If the number density of electrons participating in the rigid ground state is $n_s$, then the local superconductive current density becomes

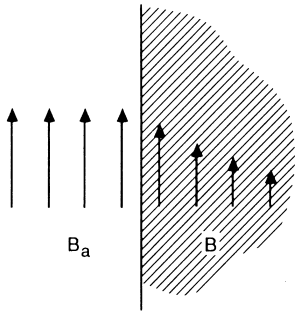



**FIGURE 15** Penetration of an applied magnetic field into a semi-infinite superconductor. The penetration depth $\lambda_L$ is the distance at which $B$ decays to $B_a/e$.

$$\mathbf{j}_s = n_s e\langle \mathbf{v}_s \rangle = -n_s e^2 \mathbf{A}/m, \tag{34}$$

and comparison of Eq. (34) with Eq. (31) gives

$$\lambda_L = \left(m^*/\mu_0 n_s e^2\right)^{1/2}. \tag{35}$$

However, careful measurements of $\lambda_L(T)$ near $T = 0$ K indicate that $\lambda_L(0)$ is larger than the prediction of Eq. (35), which suggests a reduced $n_s$ and hence a rigidity of the condensate only over a finite volume defined by a characteristic length $\xi_0$.

## 3. Coherence Length

The concept of a characteristic dimension $\xi_0$ was introduced by Pippard to formulate a nonlocal generalization of the London equation. He estimated this length from the Heisenberg uncertainty principle: only electrons having energies within $\sim k_B T_c$ of $E_F$ can play a major role in a phenomenon that sets in at $T_c$; these electrons have a momentum range $\Delta p \approx k_B T_c / v_F$, where $v_F = \hbar k_F / m^*$ is the velocity of an electron with Fermi energy $E_F$. From the uncertainty principle, $\Delta x \geq \hbar/\Delta p \approx \hbar v_F / k_B T_c$ defines a characteristic length

$$\xi_0 = a\hbar v_F / k_B T_c, \tag{36}$$

where $a$ is a numerical constant of order unity. The length $\xi_0$ plays a role analogous to the mean free path $l$ in the nonlocal electrodynamics of normal metals; and in the presence of scattering, the characteristic length is called the coherence length, where

$$(1/\xi) = (1/\xi_0) + (1/l). \tag{37}$$

In fact, Ginzburg and Landau were the first to introduce the idea of a characteristic length. In 1950, they

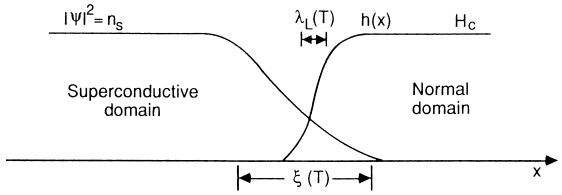



**FIGURE 16** Interface between superconductive and normal domains in the intermediate state; *h(x)* is the local magnetic-field strength.

introduced the order parameter $\psi(r)$ in Eq. (9) defined by $n_s = |\psi|^2$ to obtain an equation for the supercurrent density $\mathbf{j}_s$ in terms of $\psi(r)$. (This theory was later shown to be a limiting form of the microscopic BCS theory first presented in 1957.) With this formalism they were able to treat two features that were beyond the scope of the London formalism: (1) nonlinear effects in fields strong enough to change $n_s$ and (2) the spatial variation of $n_s$.

A major triumph of the formalism was its description of the intermediate state of a type I superconductor in which superconductive and normal domains coexist in an applied field strength $B_a/\mu_0 \approx H_c$ (Fig. 16). The interface between the two domains is characterized by two lengths: the penetration depth $\lambda_L(T)$, over which the local magnetic field is varying, and the coherence length

$$\xi_{GL}(T) = \hbar/|2m^*\alpha(T)|^{1/2}, \tag{38}$$

over which $\psi(r)$ can vary without an undue energy increase. In pure superconductors at $T \ll T_c$, the Ginzburg–Landau coherence length approaches the Pippard coherence length [i.e., $\xi_{GL}(T) \approx \xi_0$], but $\xi_{GL}(T)$ diverges as $(T_c - T)^{-1/2}$ near $T_c$ since $\alpha$ vanishes as $(T_c - T)$. Since $\lambda_L(T)$ also diverges as $(T_c - T)^{-1/2}$, the ratio $\lambda_L/\xi_{GL}$ is nearly independent of temperature. Therefore the Ginzburg–Landau parameter is

$$\kappa = \lambda_L/\xi_{GL}. \tag{39}$$

In type I superconductors, a $\kappa \ll 1$ results in a positive interface energy, which stabilizes a macroscopic domain pattern.

Abrikosov investigated what would happen if the Ginzburg–Landau parameter is greater than unity. He found that for $\kappa > 1/\sqrt{2}$, the energy of the interface between normal and superconductive domains becomes negative and the superconductor is type II. With a negative interface energy, field strengths $B_a/\mu_0 \geq H_{c1}$ perpendicular to a superconducting slab cause flux to penetrate within cylindrical, normal-state domains; persistent supercurrents surrounding the normal-state regions form vortices. The vortex concentration increases with $B_a/\mu_0 \geq H_{c1}$ until $H_{c2} = \sqrt{2}\kappa H_c$; above $H_{c2}$ the vortices are merged into a single normal-state phase.

## 4. BCS Theory

In 1956, Cooper showed that so long as there exists an attractive interaction between pairs of electrons, the Fermi sea of electrons is unstable against the formation of at least one bound pair formed from states with $\mathbf{k} > \mathbf{k}_F$. Moreover, he argued that the two-electron wave function for a superconductive pair is a singlet (paired spins), spherical state containing a weighted sum over $\mathbf{k} > \mathbf{k}_F$ of product wave functions with momentum $\mathbf{k}$, $-\mathbf{k}$ for each product and that the maximum contribution comes from states with $\mathbf{k} \approx \mathbf{k}_F$. The two-electron binding energy relative to $2E_F$ was shown to be ($\hbar\omega_c = h\nu_c$)

$$E_{bind} \approx 2\hbar\omega_c \exp[-2/VN(E_F)] \tag{40}$$

in the weak-coupling limit $VN(E_F \ll 1$. In this derivation, Cooper made the approximation that the coupling energy $V$ is a constant for all values of $\mathbf{k}$ out to a cutoff energy $\hbar\omega_c$ away from $E_F$. Since $E_{bind}$ is of the order of $k_B T_c$, an argument similar to that preceding Eq. (36) suggests that the size of the Cooper-pair state is approximately $\xi_0$, which is much larger than the interparticle distance. Thus the Cooper pairs are strongly overlapping, which is why they form a rigid condensate.

The interaction between electrons of a pair always contains an electrostatic repulsive energy $U_p$ between the two electrons; a high dielectric constant introduces an electronic screening that reduces $U_p$, but it is always repulsive. The problem is to identify an attractive mechanism.

In 1950 Frölich suggested that electron–lattice interactions were responsible for the attractive potential, and this idea was confirmed experimentally with the discovery of the isotope effect. The physical idea in the BCS treatment of this suggestion is that the first electron polarizes the crystal by attracting the positive atomic cores; the polarization in turn attracts a second electron provided that it arrives in the polarized region of the crystal before the lattice has had time to relax to its initial state. This time constraint limits the size of a Cooper pair to a characteristic length of order $\xi_0$. Moreover, the cutoff energy $\hbar\omega_c$ in Eq. (40) is, for this mechanism, the Debye energy $\hbar\omega_D = k_B\Theta_D$, which characterizes the cutoff of the phonon spectrum. If the attractive energy $V_C$ exceeds the electrostatic repulsive energy in magnitude, then the net BCS potential

$$V_{BCS} = V_C - U_p \tag{41}$$

is attractive.

The BCS theory involves a calculation of the ground state of the system in the presence of a net attractive potential $V_{BCS}$. Condensation of Cooper pairs changes the state of the Fermi sea (the collection of one-particle states), and at some point the binding energy for an additional pair has

gone to zero. The simplest form of the theory contains one adjustable parameter, $V_{BCS}$; all other parameters entering the theory are independently measurable. Its principal deductions are the following:

$$T_c \approx 1.13\Theta_D \exp[-1/V_{BCS}N(E_F)], \quad (42)$$

where $\Theta_D \approx M^{-1/2}$ is the Debye temperature, and

$$2\Delta_0/k_B T_c = 3.53, \quad (43)$$

$$\Delta C(T_c)/\gamma T_c = 1.43, \quad (44)$$

$$\gamma T_c^2 / H_0^2 = 0.168, \quad (45)$$

where $\Delta_0$ and $H_0$ are the gap parameter $E_g = 2\Delta$ and the critical field strength $H_c$ at $T = 0$ K, $\Delta C(T_c)$ is the discontinuity in the specific heat at $T = T_c$, and $\gamma$ is the Sommerfeld constant of the electron gas in the normal state (i.e., $\gamma T_c$ is the electronic specific heat of the normal state at $T = T_c$). It is found experimentally that $\alpha$ in Eq. (19) is not universally $\frac{1}{2}$ and that Eqs. (43) to (45) do not hold quantitatively in many high- and intermediate-$T_c$ materials, which indicates a need to extend the simplest BCS theory.

## 5. Beyond BCS

The limitation of the BCS theory is that it is a one-parameter theory in which $V_{BCS}$ is assumed to be constant in an energy region about $E_F$ of width $\pm\hbar\omega_D$; and there is no prescription available for calculating $V_{BCS}$ from microscopic theory. An important extension of BCS theory has been given by Eliashberg.

Whereas BCS theory simply postulates an attractive potential $V_{BCS}$, Eliashberg theory treats properly the microscopic electron–phonon interactions responsible for the pairing potential in conventional superconductors. The Eliashberg theory contains two parameters. One is the pseudopotential $\mu^*$ for the Coulomb electron–electron repulsions; it is adjusted to give the correct value of $T_c$ for a given electron–phonon interaction. The other is the electron–phonon spectral density $\alpha^2 F(\omega)$, where $F(\omega)$ is the number of phonon modes (lattice vibrations) having an energy between $\hbar\omega$ and $\hbar(\omega + d\omega)$; $\alpha^2 F(\omega)$ is a phonon frequency distribution weighted by the strength of the electron–phonon interaction for that mode. It is possible to measure $\alpha^2 F(\omega)$ accurately with superconductive-state–insulator–normal-state tunneling experiments.

The Eliashberg equations determine not only the tunneling characteristic of a tunnel diode, but also all of the thermodynamics of a particular superconductive material provided that $\alpha^2 F(\omega)$ and $\mu^*$ are known. In this theory, the BCS universal ratios [Eqs. (43)–(45)] have become transformed to

$$2\Delta_0/k_B T_c = 3.53[1 + 12.5m^2 ln(1/2m)], \quad (46)$$

$$\Delta C(T_c)/\gamma T_c = 1.43[1 + 53m^2 ln(1/3m)], \quad (47)$$

$$\gamma T_c^2 / H_0^2 = 0.168[1 - 12.2m^2 ln(1/3m)], \quad (48)$$

where $m = k_B T_c/\hbar\omega_{ln}$ contains a parameter $\omega_{ln}$ that represents a weighted measure of the significant phonon frequencies appearing in $\alpha^2 F(\omega)$. The agreement between theory and experiment for elemental and alloy superconductors, as given by Carbotte, is displayed in Fig. 17 for conventional weak to strong coupling regimes $m \leq 0.25$.

Within the Eliashberg theory,

$$2\Delta_0/k_B T_c < 9 \quad (49)$$

would reach its maximum value only in the unrealistic situation that the entire spectral weight occurs at an optimum vibrational energy $\hbar\omega_E = 0.75$ meV. The critical temperature

$$T_c = C(\mu^*)A_p/k \quad (50)$$

increases with the strength of the effective electron–phonon interaction

$$A_p = \int_0^\infty \alpha^2 F(\omega)\, d(\hbar\omega), \quad (51)$$

where $C(\mu^*)$ decreases smoothly with increasing $\mu^*$. $T_c$ increases with $A_p$ until a lattice instability freezes out a static distortion of the structure. Thus the theory of superconductivity itself does not put an upper limit on $T_c$; however, the conditions for a high $T_c$ appear to be the same as those for the stabilization of competitive mechanisms.

The electronic density of states $N(E_F)$ at the Fermi energy plays an important role, as in the BCS theory, since

$$A_p = N(E_F)\langle g^2 \rangle, \quad (52)$$

where $\langle g^2 \rangle$ involves a double Fermi-surface average of the square of the electron–atomic core interaction. In the absence of any physical intuition as to how to enhance $\langle g^2 \rangle$, efforts to increase $T_c$ have traditionally concentrated on increasing $N(E_F)$, but this strategy is frustrated by the appearance of spontaneous magnetism or atomic-clustering lattice instabilities as competing processes. Significantly the high-$T_c$ copper oxides have a relatively small $N(E_F)$, which implies that a large $\langle g^2 \rangle$ is enhancing the ratio $m$ into a very strong-coupling regime. This observation requires, in turn, either an electronic enchancement of the electron–phonon interaction or an entirely novel mechanism for the formation of electron pairs.

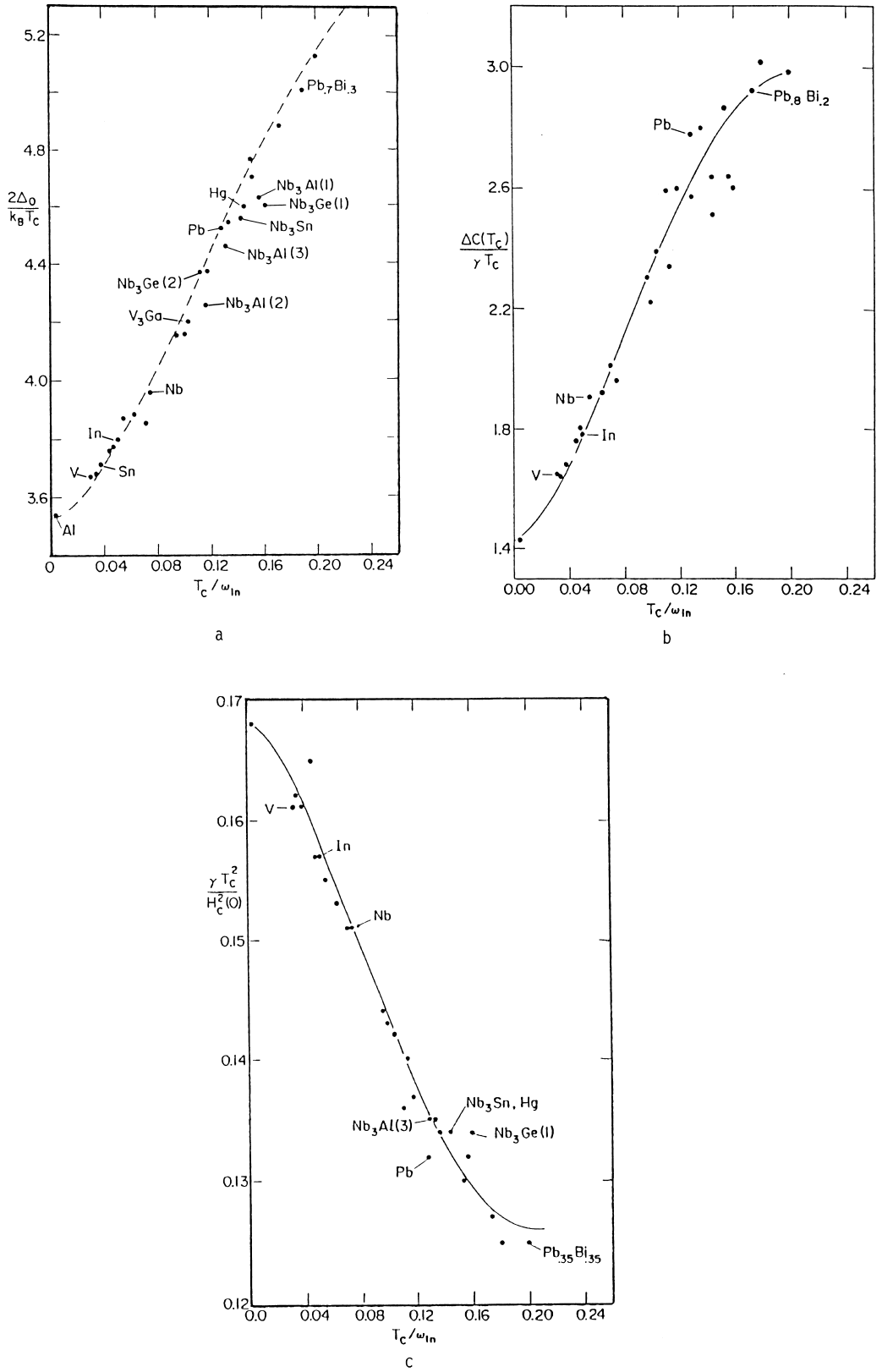



**FIGURE 17** The ratios (a) $2\Delta_0/k_B T_c$, (b) $\Delta C(T_c)/\gamma T_c$, and (c) $\gamma T_c^2/H_0^2$ versus $m = kBT_c/\hbar\omega_{In}$. [After Carbotte, J. P. (1987). *Sci. Prog. Oxf.* **71,** 327.]

## III. METALLIC OXIDES

The new high-$T_c$ superconductors are oxides. Since only metals are superconductors, this fact may appear anomalous, as most of the common oxides are either insulators or low-mobility semiconductors. However, many oxides are metallic, and a few of these have extensive commercial application. For example, the metallic cathode of the lead–acid battery is $PbO_2$; and $Bi_2Ru_2O_7$ has a temperature-independent resistivity in the vicinity of room temperature, which makes it an important resistor material in the electronics industry. Discussion of the high-$T_c$ superconductors rightfully begins, therefore, with a review of the conditions that must be satisfied if an oxide is to be metallic.

### A. Electron Energies in a Typical Insulator

Figure 18 shows schematically the construction of an energy diagram for the insulator MgO. The $O^{2-}$ ion is not stable in free space; a negative electron affinity places the $O^{-/2-}$ redox energy above the lowest energy $E_{vac}$ for a free electron in a vacuum. An energy $E_1$ is required to remove the outer $Mg = 3s^1$ electron from a free $Mg^+$ ion to a free $O^-$ ion to create free $Mg^{2+}$ and $O^{2-}$ ions. This cost in energy is more than compensated by the electrostatic energy $E_M$ gained by ordering the $Mg^{2+}$ and $O^{2-}$ ions into a crystal structure; the Madelung energy $E_M$ is calculated for a lattice of point charges. The crystalline electric field raises the $Mg^{2+/+}$ level and lowers the $O^{-/2-}$ level; crossing of these two energies ensures stabilization of the crystalline phase with a charge transfer from magnesium to oxygen.

In the real MgO crystal, transfer of an integral electronic charge does not occur; a quantum-mechanical covalent component in the Mg—O bond transfers a fraction of the $O^{2-}$-ion electronic charge back onto the $Mg^{2+}$ ion. However, the reduction in $E_M$ caused by this lowering of the effective ionic charges is compensated by the quantum-mechanical covalent-mixing repulsion between the two ionic energy levels. Therefore the point-charge model gives a good first approximation to the binding energy of the solid. The covalent component to the bonding introduces an O-2$p$ character into the Mg-3$s$ states and a Mg-3$s$ character into the O-2$p$ states, but without changing the number of electron states at each energy level. Even where this "mixing" is large, it is customary to identify the energy levels by their ionic component only (i.e., as $O^{2-}:2p^6$ and $Mg^{2+}:3s^0$ levels); the wave functions describing the mixed Mg-3$s$ and O-2$p$ states are referred to as crystal-field orbitals so as to distinguish them from the atomic orbitals of a point-charge model.

The crystal-field energies reflect the point-group symmetry of the near-neighbor Mg—O interactions. The final step is to introduce the like-atom interatomic interactions, which broaden the energy levels of the crystal-field orbitals into energy bands of one-electron states. Whereas the crystal-field orbitals are localized to discrete atomic sites, the one-electron states are itinerant with a well-defined momentum $\mathbf{p} = \hbar\mathbf{k}$ in the absence of a magnetic field $\mathbf{B}_a = \nabla \times \mathbf{A}$. Each band contains $2N/n$ one-electron states per atomic orbital (the factor 2 reflects the twofold spin degeneracy of an orbital) for an array of $N$ like atoms containing $n$ atoms per primitive unit cell. Thus the band states reflect the translational space–group symmetry of the crystal.

In MgO there is one magnesium and one oxygen atom per primitive unit cell, so that $O^{2-}:2p^6$ and $Mg^{2+}:3s^0$ levels are each broadened into single bands with a bandwidth much broader than the small spin-orbit splitting of the threefold degeneracy of the oxygen 2$p$ crystal-field orbitals. Therefore the highest occupied band of one-electron states is represented as an orbitally threefold-degenerate $O{:}2p^6$ band, which is full; the lowest unoccupied band is identified as the $Mg^{2+}:3s^0$ band.

In tight-binding theory, the width of a band of one-electron states is

$$W \cong 2zb, \tag{53}$$

where $z$ is the number of like nearest neighbors on energetically equivalent lattice sites and

$$b \equiv (\psi_i, H'\psi_j) \cong \varepsilon_{ij}(\psi_i, \psi_j) \tag{54}$$

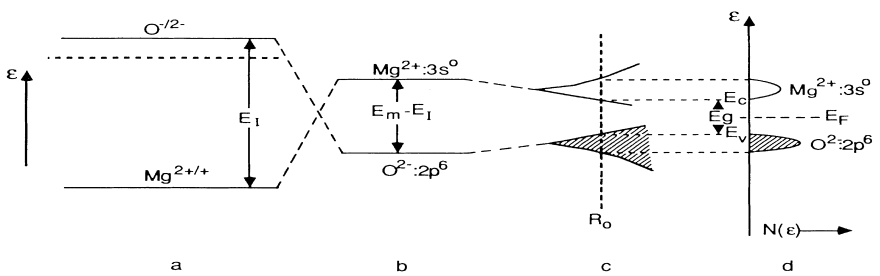



**FIGURE 18** Electron energies for MgO: (a) free ions; (b) point-charge model; (c) band model; (d) density of states.

is a measure of the strength of the interatomic interactions between nearest-neighbor like atoms at positions $R_i$ and $R_j$ in the lattice. The perturbation $H'$ of the potential at $R_j$ by the presence of a like atom at $R_i$ factors out as a one-electron energy $\varepsilon_{ij}$, which increases with the overlap integral $(\psi_i, \psi_j)$ for crystal-field wave functions $\psi_i$ and $\psi_j$ at $R_i$ and $R_j$. Therefore, the overlap integral becomes the guiding qualitative indicator of the strength of the interatomic-interaction parameter $b$. (Chemists call this parameter a resonance integral; physicists, an electron-energy transfer integral.)

Neglected in the band description of crystalline orbitals is the intraatomic electron–electron electrostatic energy $U$ associated with the creation of polar states (e.g., $O^-$ and $O^{3-}$ in the $O^{2-}$:$2p^6$ band); the tight-binding theory admixes polar and nonpolar states equally in a first-order perturbation theory. So long as the condition

$$W \gg U \tag{55}$$

is valid, the assumption of $U \approx 0$ is a useful approximation. Outer $s$ and $p$ electrons participating in near-neighbor chemical bonding satisfy Eq. (55).

The right-hand side in Fig. 18 indicates the density $N(\varepsilon)$ of one-electron states versus the energy $E$ for the equilibrium lattice constant. Since the $O^{2-}$:$2p^6$ band is filled and the Mg:3$s$ band is empty, the Fermi energy lies near the middle of a large energy gap $E_g = E_c - E_v$ between the two bands, which makes MgO an insulator. The highest occupied band $O^{2-}$:$2p^6$ is called the valence band; the lowest unoccupied band $Mg^{2+}$:$3s^0$ is called the conduction band.

Attempts to render MgO conducting by doping with aliovalent impurities, as in semiconductor technology, are frustrated by the energetic inaccessibility of both $E_c$ and $E_v$; it is energetically favorable for the crystal to incorporate a native defect that charge compensates for the dopant so as to retain $E_F$ near the middle of the energy gap $E_g$. The cost of introducing a native defect is less in an ionic crystal than in a covalent solid, which is why oxides with large band gaps $E_g$ tend to be good insulators. It follows that the first requirement for metallic conduction in an oxide is the introduction of energetically accessible electron energies.

## B. Problems with 5s and 6s Electrons

Heavy group B metals such as Sn and Pb have a relatively large separation of 5$s$ from 5$p$ or 6$s$ from 6$p$ states; it is therefore chemically straightforward to stabilize $Sn^{2+}$:$5s^2$ and $Pb^{2+}$:$6s^2$ configurations in oxides, which demonstrates that with these cations the outer 5$s$ or 6$s$ states have become energetically accessible. Therefore, electrons can be introduced into the 5$s$ band of $SnO_2$ and the 6$s$ band of $PbO_2$. In $SnO_{2-x}$ (Fig. 19a) the 5$s$ conduction band is over 3 eV above $E_v$; but an oxygen deficiency introduces oxygen vacancies that trap out from the conduction band Sn-5$s$ electrons in shallow, two-electron Sn-5$s$ donor states. In $PbO_2$ (Fig. 19b) the Pb-6$s$ conduction band appears to overlap the O-2$p$ valence band ($E_v > E_c$), thus eliminating $E_g$. On the other hand, introduction of additional conduction electrons, as is done by hydrogen insertion into $PbO_2$ on battery discharge, renders the system unstable with respect to a disproportionation reaction represented by

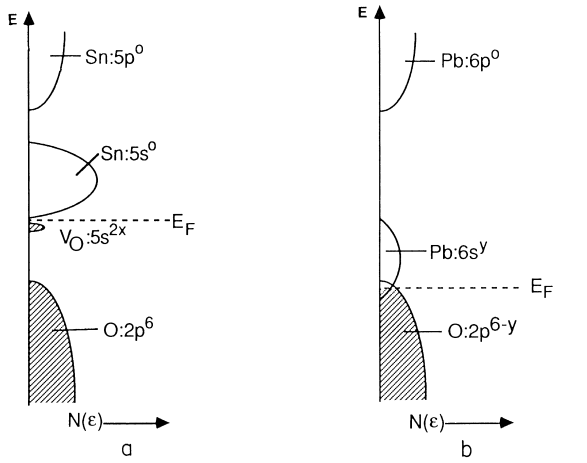



**FIGURE 19** Electron energies for (a) $SnO_{2-x}$ and (b) $PbO_2$. $V_o$ is an oxygen vacancy trapping two electrons from the Sn:5$s$ band.

$$2Pb^{3+} \rightarrow Pb^{2+} + Pb^{4+} \tag{56}$$

Lattice instabilities associated with trapping out the conduction electrons as pairs at specific $Pb^{2+}$ sites plague efforts to increase the conduction-electron density to any significant concentration. In $SnO_{2-x}$ they are already trapped as pairs at oxygen vacancies.

## C. Problems with Valence-Band Holes

Alternatively it is possible to gain access to the $O^{2-}$:$2p^6$ valence band with strongly electropositive cations such as the alkali-metal ions $A^+$ and the larger alkaline-earth ions $Sr^{2+}$ and $Ba^{2+}$. However, in this case holes introduced into the valence band become trapped as pairs in the homopolar O—O bonds of peroxide ions $(O_2)^{2-}$. Only where the covalent component of the M—O bond is strong and there is some overlap of the conduction and valence bands, as in $PbO_2$, are the valence-band holes not trapped out by O—O dimerization.

## D. Transition-Metal Oxides

Transition-metal cations may have $d^n$ or $f^n$ configurations with energies lying within $E_g$ that offer the possibility of

obtaining partially filled $d$ or $f$ bands. However, the wave functions of partially filled $d$ or $f$ shells have smaller radial extensions, which reduces the interatomic-interaction parameter $b$ of Eq. (54) and increases the intraatomic-interaction parameter $U$, so Eq. (55) may no longer be satisfied.

### 1. Problems with Outer 4*f* Electrons

The $4f$ electrons of a rare earth ion are tightly bound to their atomic nucleus and they are screened from near neighbors by closed $5s^2 5p^6$ shells, so the condition

$$W \ll U \tag{57}$$

prevails in the rare earth oxides. Under these conditions the $4f$ orbitals remain as crystal-field orbitals; they are not transformed into itinerant-electron states. Therefore $4f^n$ configurations remain localized; they impart localized atomic moments to the ions that are essentially identical to the localized atomic moments they impart to the free ions or atoms. In this situation, successive redox potentials are separated by a large energy $U > E_g$, which restricts the accessible $4f^n$ configurations on a given rare earth cation to at most two, and two only if one happens to lie within $E_g$.

If a $4f^n/4f^{n+1}$ redox couple does lie within $E_g$, as illustrated in Fig. 20a, then it is possible to obtain mixed-valent compounds in which $E_F$ intersects the redox energy. In this case, metallic conduction could be expected were $b > \hbar\omega_R$, where $\omega_R$ is the frequency of a breathing-mode lattice vibration. However, $b$ is so small that the near neighbors have time to relax about a mobile electron, thereby trapping it in a local potential well. These lattice relaxations stabilize the occupied states at the expense of unoccupied states, as does the molecular reorganization in a liquid for a given redox couple. The mobile electrons of the mixed-valent state thus become "dressed" in a local lattice deformation, which introduces an activation energy into their mobility. These "dressed" electrons are called small polarons; they move diffusively, so **k** is no longer a good quantum number. The rare earth mixed-valent systems are not metallic.

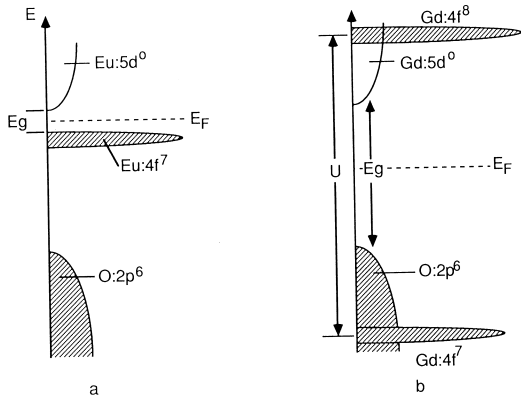



**FIGURE 20** Electron energies for (a) EuO and (b) $Gd_2O_3$. $E_F$ moves into the $4f^7$ energy level in $Eu_{1-\delta}O$ containing a $Eu^{3+/2+}$ mixed valence.

On the other hand, where a broad conduction band overlaps a $4f^n$ energy level and $E_F$ intersects the $4f^n$ energy level to give a $4f^{n+1}/4f^n$ mixed valence (in this case denoted an "intermediate valence"), hybridization of the $4f$ wave functions with the conduction-band wave functions may lead to "heavy-fermion" metallic behavior. Superconductivity has been observed in some heavy-fermion compounds (not oxides), but in all of them $T_c$ is low.

### 2. Outer *d* Electrons

*a. General considerations.* The $4f$ electrons of a rare earth ion in an oxide are only weakly perturbed from their free-ion behavior; the outer $s$ and $p$ electrons are so strongly perturbed that they are transformed into itinerant electrons. The perturbations of the $d$ wave functions in a transition-metal oxide are of intermediate strength. In this case also it is convenient to consider first the perturbations imposed by the nearest-neighbor metal–oxygen (M—O) interactions; these give rise to crystal-field orbitals containing the quantum-mechanical covalent mixing between overlapping cation and oxygen orbitals. Whether the crystal-field orbitals remain localized or are transformed into itinerant-electron band states depends on the relative strengths of the crystal-field intraatomic interactions $U$ and the bandwidth $W$ due to interatomic interactions between crystal-field wave functions on neighboring metal atoms M.

Covalent mixing between cation $d$ and oxygen $2s$ and $2p$ orbitals in a transition-metal oxide has two important consequences. First, the fivefold orbital degeneracy of the free ion is at least partially removed by a crystal-field splitting of the energies of the crystal-field orbitals. Second, mixing of oxygen wave functions with the $d$ wave functions spreads the crystal-field orbitals of $d$ wavefunction symmetry out over the oxygen atoms, which both reduces the intraatomic energy $U$ and increases the bandwidth $W$; it also allows M—O—M as well as M—M interactions.

For half-filled crystal-field orbitals, the addition of one more electron to a cation costs an intraatomic energy

$$U = U' + \Delta_{ex}, \tag{58}$$

where $U'$ is the energy required to add an electron to an empty orbital and $\Delta_{ex}$ is the additional electrostatic energy required to add it to a half-filled orbital. The term

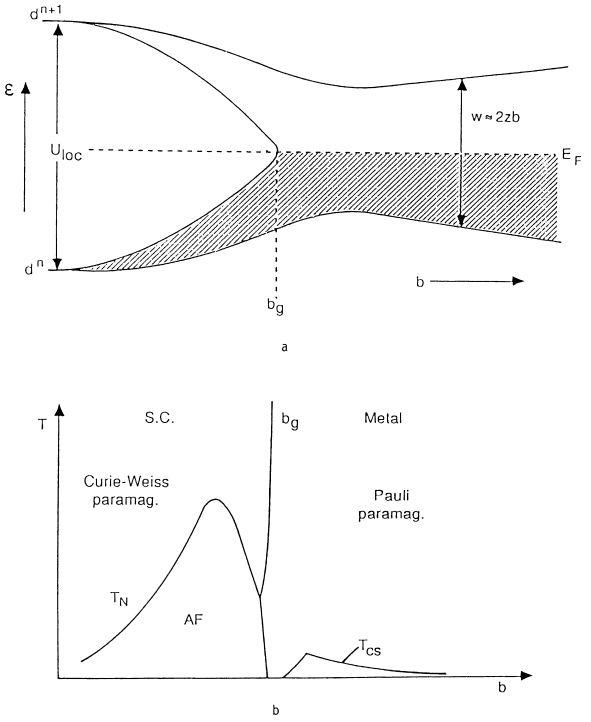



**FIGURE 21** (a) Evolution of electron energies versus resonance integral $b$ for interactions between nearest-neighbor like atoms on energetically equivalent sites. (b) Corresponding phase diagram for a single-valent system with a large separated atom $U$. (Half-filled band illustrated.)

$\Delta_{ex}$ enters $U$ wherever an electron is added to a half-filled manifold. Figure 21 shows the evolution of $U$ with increasing interatomic-interaction parameter $b$, where $W \approx 2zb$ in the tight-binding approximation with $W \gg U$. As $b$ increases, screening of the electrons of a given manifold by electrons on neighboring atoms causes $U$ to decrease rapidly with $b$ in the vicinity where $W \approx U$ causes a transition from semiconducting to metallic behavior. In the domain $W < U$, the crystal-field orbitals remain sufficiently localized to impart an atomic magnetic moment, whereas in the domain $W > U$ the compound not only is metallic, but also has no spontaneous atomic magnetic moment. Clearly a necessary criterion for metallic conductivity in a single-valent transition-metal compound is the condition

$$W > U. \tag{59}$$

If the initial $U$ at small $b$ is relatively small, as may occur where

$$U = U', \tag{60}$$

then the bands may be so narrow at $W \approx U$ that the occupied states become split from the unoccupied states (Fig. 22) by a displacement transition that changes the translational symmetry of the structure. Such displacement transitions exhibit atomic clustering: like-atom clustering where M—M or O—O interactions are important, M—O clustering in a disproportionation reaction where M—O—M interactions are dominant. Cooperative transitions are to static charge-density wave (CDW) states that compete with the superconductive state. This type of transition is not restricted to a single-valent situation; a longer wavelength CDW may be stabilized where like cations are present on energetically equivalent sites with a mixed valence.

For the case of a mixed valence on energetically equivalent lattice sites, a narrow band may also be split by small-polaron formation as discussed for the mixed-valent rare earth ions. Elimination of small polarons, a necessary criterion for metallic conductivity in mixed-valent systems, requires a bandwidth

$$W > \hbar\omega_R, \tag{61}$$

where $\omega_R$ is the frequency of the optical-mode vibration that traps the charge carrier in a local lattice deformation.

These several considerations are best illustrated by some specific examples.

*b. MnO, a single-valent compound with $W \ll U$.* Figure 23 illustrates the construction of an energy diagram for the antiferromagnetic insulator MnO. Comparison with Fig. 18 shows that it is similar to the construction

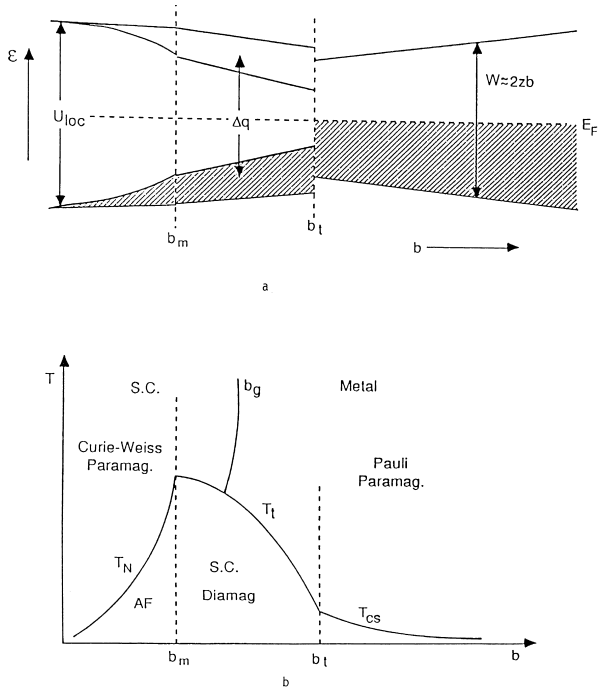



**FIGURE 22** Same as Fig. 21 for a small separated atom $U$.

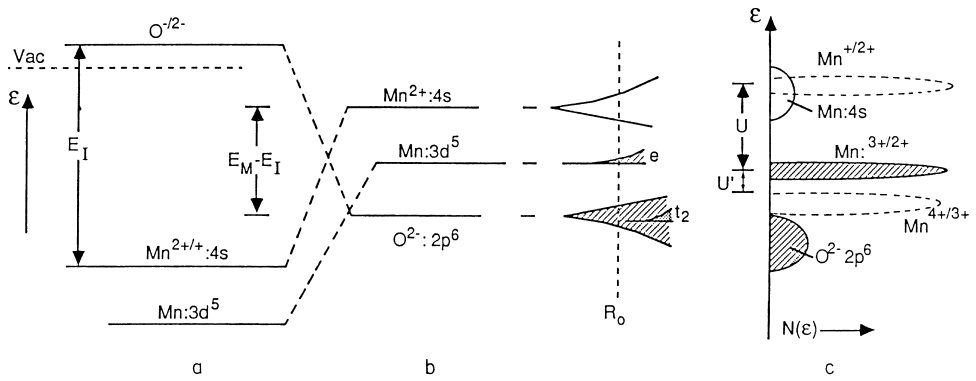



**FIGURE 23** Electron energies for MnO: (a) free ions; (b) point charge; (c) crystal.

for MgO except for the appearance of a filled crystal-field $d^5$ configuration within the large energy gap between the empty Mn:$4s^0$ and the filled O:$2p^6$ bands. Because $E_F$ lies above the top of the O:$2p^6$ band, the formal valence Mn(II) unambiguously assigns five $d$ electrons per Mn.

The Mn atom sits in an octahedral interstice of oxygen atoms in the rock-salt structure of MnO; in this configuration the fivefold $d$-orbital degeneracy is split in two by a cubic crystalline field. The wave functions for the two crystal-field manifolds have the form

$$\psi_e = N_\sigma[f_e - \lambda_s\phi_s - \lambda_\sigma\phi_\sigma], \tag{62}$$

$$\psi_t = N_\pi[f_t - \lambda_\pi\phi_\pi], \tag{63}$$

where, from Fig. 24, the twofold-degenerate $e$ orbitals of symmetry $x^2 - y^2$ and $[(z^2 - x^2) + (z^2 - y^2)]/\sqrt{2}\sigma$-bond with $2p_\sigma$ and $2s$ orbitals at neighboring oxygen atoms but are orthogonal to the O-$2p_\pi$ orbitals; the threefold-degenerate $t_2$ orbitals of symmetry $xy$, $yz$, $zx$ $\pi$-bond with the O-$2p_\pi$ orbitals but are orthogonal to the O-$2p_\sigma$ and O-$2s$ orbitals. The admixture wave functions $\phi_s$, $\phi_\sigma$, and $\phi_\pi$ are, respectively, linear combinations of nearest-neighbor O-$2S$, O-$2p_\sigma$, and O-$2p_\pi$ orbitals having the same symmetries as the atomic $f_e$ or $f_t$ orbitals with which they mix. The covalent-mixing parameters are defined as

$$\lambda_\sigma \equiv \left|b_\sigma^{\rm ca}\right|/\Delta E \quad \text{and} \quad \lambda_\pi \equiv \left|b_\pi^{\rm ca}\right|/\Delta E, \tag{64}$$

where $\Delta E = (E_d - E_p)$ is the energy required to transfer an electron from an O-$2p$ orbital to an empty $d$ orbital at the point-charge Mn(I) = $3d^6$ energy. Because the overlap integrals entering

$$b^{\rm ca} \equiv (\psi_{\rm cat}, H'\psi_{\rm anion}) \approx \varepsilon(\psi_{\rm cat}, \psi_{\rm anion}) \tag{65}$$

are larger for the $\sigma$-bonding orbitals, a $\lambda_\sigma > \lambda_\pi$ raises the energy of the antibonding crystal-field $e$ orbitals relative to that of the $t_2$ orbitals. The crystal-field splitting

$$10Dq \equiv \Delta_{\rm c} = \Delta_{\rm m} + \tfrac{1}{2}\left(\lambda_\sigma^2 - \lambda_\pi^2\right)(E_d - E_p) + \tfrac{1}{2}\lambda_s^2(E_d - E_s) \tag{66}$$

between $e$ and $t_2$ crystal-field energies contains only a relatively small electrostatic term $\Delta_{\rm m}$.

The energy difference between the Mn(I):$3d^6$ and the Mn(II):$3d^5$ manifolds is

$$U = U_t + \Delta_{\rm ex}, \tag{67}$$

where $U_t$ is the energy required to add an electron to an empty $t_2$ orbital. In MnO, a $\Delta_{\rm ex} \approx$ 3 eV makes both $U$ and $\Delta E$ large, and a large $\Delta E$ makes $\lambda_\sigma$ and $\lambda_\pi$ relatively small. Therefore a $W \ll U$ and a $\Delta_{\rm c} < \Delta_{\rm ex}$ stabilize a localized $t_2^3e^2$ configuration at a Mn(II) ion, and a direct exchange interaction between spins in orthogonal orbitals couples the spins parallel—in accordance with Hund's highest multiplicity rule for free ions—to give a localized Mn(II)-ion magnetic moment of $5\mu_{\rm B}$.

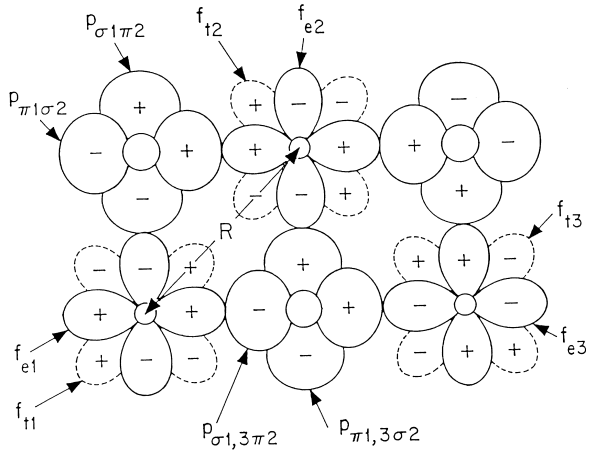



**FIGURE 24** Illustration of cation $d$ and anion $p$ orbitals in the (001) plane of the rock-salt structure.

In the presence of a $W \ll U$, the interactions between half-filled orbitals on like atoms are treated in second-order perturbation theory:

$$\Delta\varepsilon \approx |t|^2/U, \tag{68}$$

where the spin-dependent resonance integral is

$$t = b\sin(\Theta/2) \tag{69}$$

because the rotation of a spin through an angle $\Theta$ transforms as

$$\begin{aligned} \alpha &= \alpha'\cos(\Theta/2) + \beta'\sin(\Theta/2), \\ \beta &= -\alpha'\sin(\Theta/2) + \beta'\cos(\Theta/2), \end{aligned} \tag{70}$$

and the Pauli exclusion principle allows transfer only of an antiparallel spin to an orbital that is already half-filled. Substitution of Eq. (69) into Eq. (68) gives a spin–spin contribution to the interatomic interaction of the form

$$\begin{aligned} H_{\text{ex}} &= -2J_{ij}(\mathbf{s}_i \cdot \mathbf{s}_j), \\ J_{ij} &= 2b^2/U, \end{aligned} \tag{71}$$

for $s = \frac{1}{2}$. This superexchange interaction has the form of the Heisenberg spin–spin interatomic-exchange interaction. It is responsible for long-range antiferromagnetic order below a Néel temperature $T_{\text{N}}$.

In MnO, 180° Mn—O—Mn interactions compete with Mn—Mn interactions. The magnetic order and the exchange striction below $T_{\text{N}} = 118$ K demonstrate that Mn—O–Mn interactions are dominant in this compound. Placement of $E_{\text{F}}$ in a large energy gap between the Mn:$4s^0$ band and the Mn:$t_2^3e^2$ level makes MnO an antiferromagnetic insulator.

*c. Li[Mn₂]O₄, a mixed-valent compound with $W < \hbar\omega_R$.* Because the Mn(II):$t_2^3e^2$ energy lies above the O:$2p^6$ bands, it is possible to oxidize Mn(II) to Mn(III) by the removal of a single $e$ electron per Mn atom. The intraatomic electrostatic energy separating the Mn(II):$t_2^3e^2$ and Mn(III):$t_2^3e^1$ manifolds does not contain either $\Delta_{\text{ex}}$ or $\Delta_{\text{c}}$; it is

$$U = U_e, \tag{72}$$

which is small enough to retain the Mn(III) level above the top of the O:$2p^6$ band. Consequently it is also possible to remove the remaining $e$ electron at a Mn(III) ion to oxidize it to Mn(IV):$t_2^3e^0$. The Mn(IV):$t_2^3e^0$ level, on the other hand, lies well below the top of the O:$2p^6$ band because it is separated from the Mn(III):$t_2^3e^1$ level by the relatively large energy

$$U = U_e + \Delta_{\text{c}}. \tag{73}$$

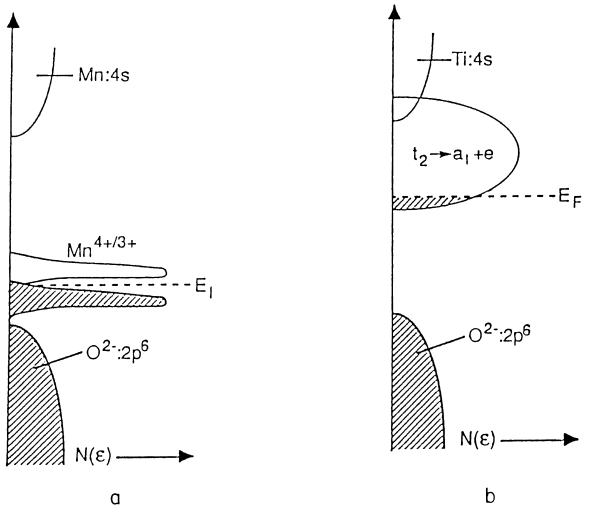



**FIGURE 25** Electron energies for two spinels: (a) $Li[Mn_2]O_4$ and (b) $Li[Ti_2]O_4$.

Therefore an octahedral-site Mn(V) valence is not stabilized in oxides.

The cubic spinel $Li[Mn_2]O_4$ illustrates a mixed Mn(III) + Mn(IV) valence configuration on energetically equivalent octahedral sites. In this compound, the Fermi energy intersects the Mn(III):$t_2^3e^1$ energy level (Fig. 25a), and electron transport can occur via the reaction

$$t_2^3e^1 + t_2^3e^0 = t_2^3e^0 + t_2^3e^1. \tag{74}$$

However, the bandwidth of the $t_2^3e^1$ level is so narrow that the time $\tau_{\text{h}} \approx \hbar/W$ for an electron to hop to a neighboring site is long compared to the period $\omega_{\text{R}}^{-1}$ of the optical-mode lattice vibration that traps it as a small polaron. Therefore, the Mn:$d^4$ level is split by a polaron energy $\varepsilon_{\text{p}}$ into occupied Mn(III):$t_2^3e^1$ and empty Mn(IV):$t_2^3e^0$ states in a manner analogous to the splitting of a $Mn^{4+/3+}$ redox couple in a liquid electrolyte. Small-polaron formation introduces an activation energy into the charge-carrier mobility, so $Li[Mn_2]O_4$ is a magnetic semiconductor; it is not a superconductor.

*d. Li[Ti₂]O₄, a mixed-valent compound with $W > \hbar\omega_R$.* In contrast to the manganese oxides, which generally have localized $d^n$ configurations, the titanium oxides generally have itinerant $d$ electrons. For example, the cubic spinel $Li[Ti_2]O_4$ is a superconductor with $T_{\text{c}} = 13.7$ K. In this compound, the Ti—Ti interactions are strong enough to make $\tau_{\text{h}} < \omega_{\text{R}}^{-1}$ (i.e., $W > \hbar\omega_{\text{R}}$) for the electron-transfer reaction

$$\text{Ti(III)}{:}t_2^1e^0 + \text{Ti(IV)}{:}t_2^0e^0 = \text{Ti(IV)}{:}t_2^0e^0 + \text{Ti(III)}{:}t_2^1e^0, \tag{75}$$

so band theory becomes applicable (Fig. 25b).

*e. $Ti_4O_7$ and the bipolaron.* The mixed-valent compound $Ti_4O_7$ is obtained by removing oxygen from $TiO_2$. The oxygen vacancies order into shear planes so as to leave a close-packed tetragonal oxide-ion array as in $TiO_2$, but with the Ti atoms arranged to give $TiO_2$ slabs connected every four Ti atoms along the $TiO_2$ $c$ axis by a shear plane across which Ti atoms share common octahedral-site faces rather than octahedral-site edges. A strong electrostatic repulsion between cations across a shear plane displaces the shear-plane Ti atoms away from each other. The possibility of a ferroelectric-type displacement of a Ti(IV) ion in an octahedral site stabilizes the shear-plane structure and locates the Ti–3$d$ electrons within the slabs.

At room temperature, both the Ti—Ti and the Ti—O—Ti interactions within a rutile slab satisfy the condition $W > \hbar\omega_R$, and $Ti_4O_7$ is metallic. However, a first-order semiconductor–metal transition occurs at a $T_t \approx 150$ K due to Ti—Ti dimerization within the slabs; the electrons are trapped out as pairs within specific Ti—Ti homopolar bonds. Such a transition would be typical of cation clustering except that in the mixed-valent compound $Ti_4O_7$ the homopolar bonds are mobile in the temperature range $130 < T < 150$ K; they become stationary only below 130 K. These mobile homopolar bonds represent strongly coupled, localized electron pairs that, like small polarons, move diffusively in the crystal. Such a mobile electron pair is called a bipolaron. Formation of spin-paired bipolarons causes a sharp drop in the paramagnetic susceptibility. But the compound becomes a semiconductor, not a superconductor, and there is no Meissner effect.

*f. TiO, a single-valent compound with $W > U$.* The titanium atom may be stabilized in single-valent oxides as Ti(II) in TiO, Ti(III) in $Ti_2O_3$, and Ti(IV) in $TiO_2$. This is possible, even though $TiO_2$ is an insulator with empty $d$ orbitals some 3 eV above the top of the O:$2p^6$ band, because the energy separating the Ti(III):$t_2^1e^0$ and Ti(II):$t_2^2e^0$ levels is a relatively small $U = U_t$. Moreover, the absence of antibonding $e$ electrons allows the Ti—O bond to be short relative to the radial extension of the $t_2$ crystal-field orbitals, so the overlap of $t_2$ orbitals on neighboring Ti atoms sharing a common octahedralsite edge is large enough to ensure a $W > U_t$ for Ti—Ti interactions and the overlap of Ti-$t_2$ and O-$2p_\pi$ orbitals is large enough to make $W \approx U_t$ for 180° Ti(III)—O—Ti(III) interactions.

The corundum structure of $Ti_2O_3$ contains, on a hexagonal basis, $c$-axis pairs of cations sharing a common octahedral-site face. Although $Ti_2O_3$ is a metal above room temperature, it becomes a semiconductor at lower temperatures because the $d$ electrons become trapped in homopolar Ti—Ti bonds within the $c$-axis pairs.

TiO, on the other hand, is metallic and a superconductor ($T_c = 1.5$ K), even though it contains about 15% cation and anion vacancies in its rock-salt structure that become ordered at lower temperatures. Here, also, the intraatomic energy $U = U_t$ contains no additional term $\Delta_c$ or $\Delta_{ex}$, and some hybridization of titanium 4$s$ and 3$d$ orbitals increases $W$.

*g. Oxides with only M—O—M interactions.* The perovskite and pyrochlore structures provide systems having 135 to 180° M—O—M interactions and no M—M interactions. A survey of the oxides with these structures shows that the condition $W > U$ may be fulfilled in single-valent oxides and the condition $W > \hbar\omega_R$ may be found in mixed-valent oxides with these structures.

The relevant bandwidth $W \cong 2zb$ arising from M—O—M interactions is proportional to either $\lambda_\pi^2$ or $\lambda_\sigma^2$. A large covalent-mixing parameter $\lambda_\pi$ or $\lambda_\sigma$ requires, according to Eq. (64), a small $\Delta E = (E_d - E_p)$ and/or a large $b^{ca}$. The energy $\Delta E$ decreases with increasing formal charge on the cation and, for a given charge, on going to the right in any long period, provided that $\Delta_c$ or $\Delta_{ex}$ is not introduced into $U$ on adding another $d$ electron to compensate for the increased nuclear charge. However, $b^{ca}$ also decreases on going to heavier atoms, but it increases on going from 3$d$ to 4$d$ to 5$d$ orbitals. The overlap integrals are also sensitive to the M—O—M angle and to the character of the countercation A in the $AMO_3$ perovskites and the $A_2M_2O_7$ pyrochlores.

Perovskite and pyrochlore oxides containing partially filled 5$d$ orbitals are generally metallic, whether stoichiometric as in $ReO_3$ or mixed-valent as in the $Na_xWO_3$ bronzes; but single-valent metallic conductivity, as in $ReO_3$, does not guarantee that the compounds are superconductors.

Perovskite and pyrochlore oxides containing partially filled 4$d$ orbitals are intermediate in character; the 4$d$ electrons are itinerant, but a $W \approx U$ may result in spontaneous magnetism.

The perovskite and pyrochlore oxides containing partially filled 3$d$ orbitals commonly contain localized $3d^n$ configurations; however, several of these oxides have a $W \geq U$ and are either metallic or stabilize itinerant-electron antiferromagnetic order. In the perovskites, the homogeneous electronic picture in Fig. 21 does not hold where $W \approx U$; a static CDW and/or spin-density wave (SDW) may be stabilized as indicated in Fig. 22. Recently, dynamic displacive phase segregations have been identified in the metallic systems $Sr_{1-x}Ca_xVO_3$, $La_{1-x}Nd_xCuO_3$, and $Ln_{1-x}Ca_xMnO_3$ and the $LnNiO_3$ family (Ln = lanthanide). These dynamic segregations occur because there is a first-order change in the equilibrium M—O bond length on going from localized 3$d$ electrons on the M atoms ($W < U$) to itinerant 3$d$ electrons ($W > U$); the equilibrium M—O bond is longer for

localized than for itinerant $3d$ electrons. In the $LnNiO_3$ family, the onset of a static CDW/SDW below a critical temperature $T_t$ has been shown to be an order–disorder transition of localized-electron fluctuations in an itinerant-electron host; it is not due to either Fermi-surface nesting in a new Brillouin zone created by a change in lattice symmetry or to the homogeneous Mott–Hubbard transition illustrated in Fig. 21. In the $Ln_{1-x}Ca_xMnO_3$ system, all the Mn atoms carry a localized $t^3$ configuration with spin $S = 3/2$, and a single electron per Mn(III) occupies a narrow ($W \approx U$) $\sigma^*$ band of $e$-orbital parentage. In this case, the transition from localized $e$ to itinerant $\sigma^*$ electrons is approached from the localized-electron side, and ordering of the twofold-degenerate $e$ orbitals promotes $e$-electron localization and charge ordering. In the absence of charge and orbital ordering, the system becomes ferromagnetic below a Curie temperature $T_c$, and a dynamic phase segregation in compositions with $W \approx U$ results in a "colossal magnetoresistance" (CMR) at temperatures $T \geq T_c$; a dynamically segregated, ferromagnetic, metallic phase of higher Curie temperature grows in an applied magnetic field at the expense of the host phase until it reaches a percolation threshold. This evidence of vibronic (hybridization of electronic and vibrational state) phenomena in oxides with perovskite-related structures is of great significance for our understanding of the high-temperature superconductivity in copper oxides.

*h. Summary.* A review of the known properties of transition-metal oxides reveals the following generalizations that apply to compounds having an $E_F$ above the top of the O:$2p^6$ bands.

1. Formal valences provide a count of the number of crystal-field $d$ electrons per transition-metal ion. Any ambiguity in the distribution of $d$ electrons among different transition-metal ions or between crystallographically inequivalent lattice sites can generally be resolved.
2. Single-valent oxides require a $W > U$ to be metallic, and $W$ is larger for $5d$ than for $4d$ or $3d$ electrons. Itinerant $3d$ and $4d$ electrons are found only where $U$ contains neither a $\Delta_c$ nor a $\Delta_{ex}$.
3. Mixed-valent oxides require a $W > \hbar\omega_R$ to be metallic. Metallic mixed-valent oxides are more commonly superconductors.
4. Single-valent oxides having $W \approx U$ are not superconducting; electron correlations that introduce an enhancement of the magnetic susceptibility, even if they do not induce magnetic order at low temperatures, compete with superconductivity. Moreover, where a small $U$ permits a large $N(E_F)$ compatible with $W > U$, static charge-density waves compete with superconductivity. In fact, CDWs may also compete in mixed-valent oxides.
5. At the crossover from localized to itinerant electronic behavior ($W \approx U$ or $W \approx \hbar\omega_R$), a first-order change in the equilibrium M—O bond length can give rise to vibronic phenomena in both single-valent and mixed-valent transition-metal arrays.

*i. Peculiarity of copper oxides.* Copper oxides are unusual in two respects. First, octahedral-site Cu(II):$t^6e^3$ contains a single $e$ hole in the $3d$ shell, which makes it orbitally degenerate and therefore a strong Jahn–Teller ion; consequently, Cu(II) ions normally occupy octahedral sites that are deformed to tetragonal ($c/a > 1$) symmetry by Jahn–Teller orbital ordering. However, in the absence of a cooperativity that stabilizes long-range orbital ordering, the electrons may couple locally to E-mode vibrations, forming vibronic states in a dynamic Jahn–Teller coupling. Second the Cu(II):$3d^9$ energy level lies below the top of the $O^{2-}$:$2p^6$ valence band in an ionic model; the introduction of covalent bonding creates states of $e$-orbital symmetry at the top of the $O^{2-}$:$2p^6$ bands that have a large O-$2p_\sigma$ component (see Fig. 26). Locally this O-$2p_\sigma$ component increases dramatically on oxidation of Cu(II) to Cu(III). The change in hybridization represents a polarization of the oxygen atoms that decreases the equilibrium Cu—O bond length, but the change in polarization is fast

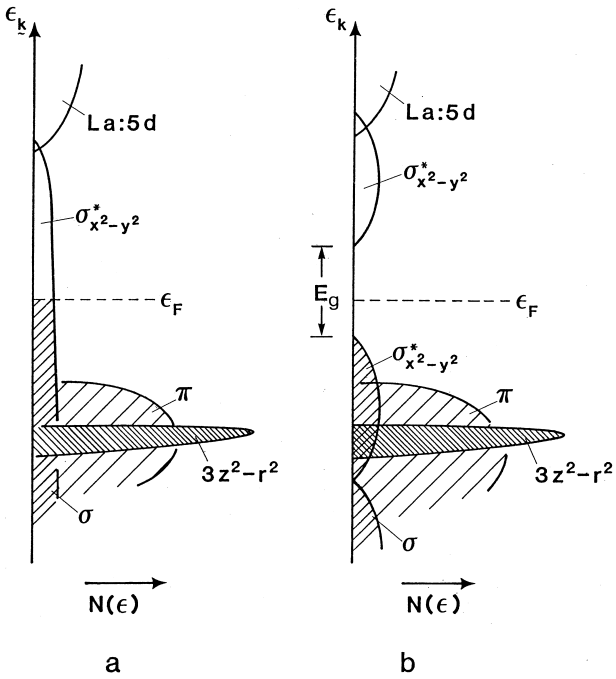



**FIGURE 26** Schematic energy density of one-electron states for $La_2CuO_4$ with bandwidth $x^2$–$y^2$. (a) $W_\sigma > U$ and (b) $W_\sigma < U$, $E_g = U - W_\sigma$.

relative to the motion of the oxygen nucleus. Therefore, a dynamic vibronic phenomenon may reflect coupling to the polarization cloud of the oxygen atoms rather than to significant oxygen-atom displacements. Nevertheless, hybridization with a polarization wave on the oxygen-atom array would significantly increase the effective mass $m^*$ of an itinerant electron.

The copper-oxide superconductors all contain $CuO_2$ sheets in which apical Cu—O bonds perpendicular to a sheet are significantly longer than the in-plane Cu—O bonds. This structural feature signals full occupancy of the $(3z^2–r^2)$ orbitals of an $e$-orbital pair. The parent compounds of the superconductive systems contain all Cu(II) in the $CuO_2$ sheets, which leaves the in-plane $(x^2–y^2)$ orbitals half-filled with a

$$U = U_e + \Delta_{ex}. \tag{76}$$

A $W < U$ results in localized $(x^2–y^2)$ electrons that interact with one another on nearest neighbors by superexchange to give antiferromagnetic order within a semiconductive $CuO_2$ sheet. On oxidation of the $CuO_2$ sheets, the system undergoes a crossover from localized to itinerant electronic behavior, and a thermodynamically distinguishable $p$-type superconductive phase is found at crossover with a hole concentration $x$ per Cu atom of the $CuO_2$ sheets in the range $0.14 \leq x \leq 0.22$. Superconductivity has also been observed on reduction of the $CuO_2$ sheets, but $n$-type superconductivity is more difficult to stabilize and has been studied much less.

## IV. HIGH-$T_c$ SUPERCONDUCTORS

### A. System $BaPb_{1-x}Bi_xO_3$

#### 1. Structure

The high-$T_c$ superconductors are oxides having structures related to the cubic perovskite. The ideal cubic perovskite has the composition $ABX_3$, where A is a large cation, B is a smaller cation, and X is an anion. As illustrated in Fig. 27, the $BX_3$ array consists of a framework of corner-shared octahedra, and the large A cation occupies the center of each "cage" of the framework. For the A cation to fit easily into the cage, the A—X and B—O bond lengths must satisfy the following relation among the "ionic radii" $R_A$, $R_B$, and $R_X$:

$$t \lesssim (R_A + R_X)/\sqrt{2}(R_B + R_X), \tag{77}$$

where $t$ is called the Goldschmitt tolerance factor. Since the A—X and B—X bond lengths can, at best, be optimized simultaneously only at a single temperature for a fixed pressure, it is common in $AMO_3$ perovskites to find

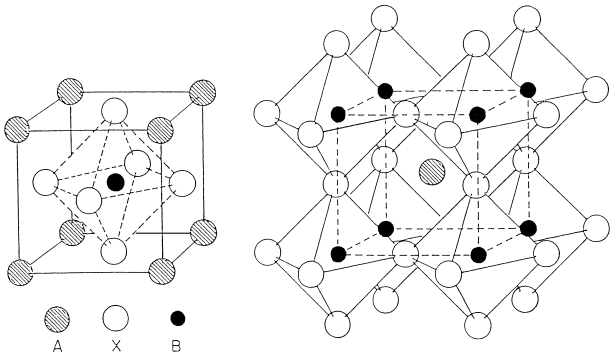



**FIGURE 27** Two views of the $ABX_3$ cubic perovskite structure.

that optimization of the A—O interactions induces a distortion of the cubic $MO_3$ cage; these distortions—to orthorhombic, rhombohedral, or tetragonal symmetry—are accomplished by a cooperative rotation of the $MO_{6/2}$ octahedra that somewhat reduces the M—O—M bond angles from 180°.

The "cubic" perovskite structure can sustain a wide range of compositional variations. Partial substitution of any of the ions is possible: $A_{1-x}A''_xMO_3$, $AM_{1-x}M'_xO_3$, and $AMO_{3-x}F_x$ are known, for example. Removal of A cations and anions is also possible: the cubic bronzes $Na_xWO_3$ contain an A-cation deficiency, and the $ReO_3$ structure consists of just the cubic $MO_3$ array. Small concentrations of oxygen deficiency may be disordered, as in the superconductor $SrTiO_{3-x}$; but a large electrostatic repulsion between oxygen vacancies tends to introduce short-range order, at least, and commonly a long-range order that defines a new structural type.

#### 2. Superconductive versus CDW State

Pure $BaPbO_3$ is a pseudocubic, metallic perovskite that is distorted to orthorhombic symmetry by a cooperative rotation of the $PbO_{6/2}$ octahedra; all the Pb(IV) ions are in energetically equivalent octahedral sites. Oxygen-deficient $BaPbO_{3-y}$ is an $n$-type metal and a conventional superconductor.

Pure $BaBiO_3$, on the other hand, is monoclinic, with two distinguishable bismuth octahedra obtained by a cooperative shifting of the oxygen atoms away from one near-neighbor bismuth toward the other so as to make the Bi—O bonds short and long in alternate octahedra. Such a "breathing-mode" oxygen displacement is indicative of a disproportionation reaction

$$2\text{Bi(IV)} \rightarrow \text{Bi(III)} + \text{Bi(V)} \tag{78}$$

in which the energy gained by stronger covalent mixing in a $Bi(V)O_6$ more than compensates for the electrostatic energy $U$ required to transfer an electron from one Bi(IV)

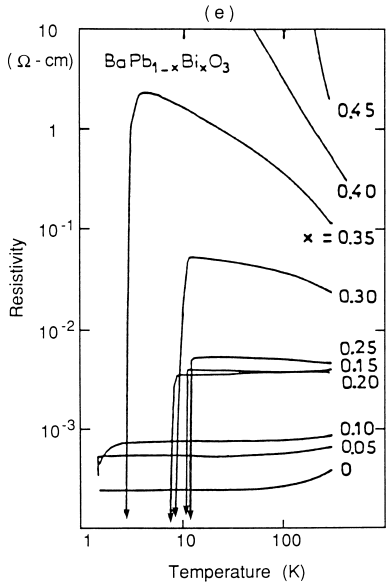



**FIGURE 28** Temperature variation of the resistivity for various values of $x$ in the system $BaPb_{1-x}Bi_xO_3$. [After Thanh, T. D., Koma, A., and Tanaka, S. (1980). *Appl. Phys.* **22,** 205.

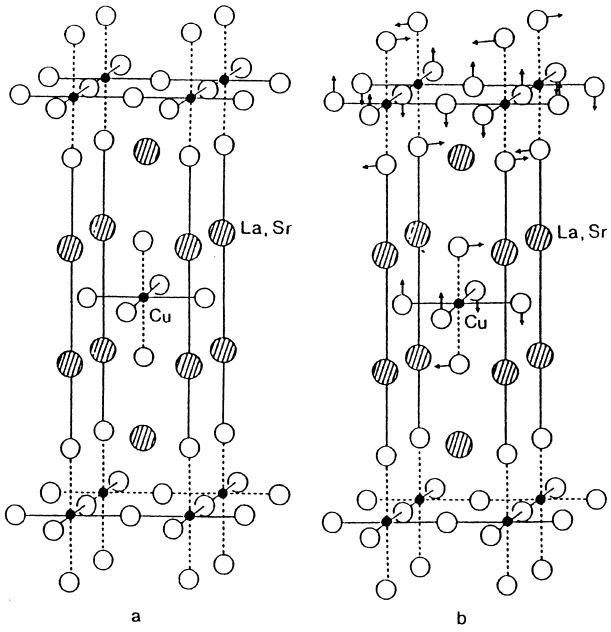



**FIGURE 29** Structure of $La_2CuO_4$: (a) tetragonal and (b) orthorhombic cooperative $CuO_{6/2}$ rotations.

to the other. Physicists refer to such a spontaneous disproportionation as a "negative $U$" reaction.

The system $BaPb_{1-x}Bi_xO_3$ is pseudocubic in the compositional range $0.05 \leq x \lesssim 0.3$, and it is an unconventional superconductor with a $T_c$ that increases with $x$ to about 13 K at the limiting composition of the pseudocubic single-phase field (Fig. 28).

Superconductivity was first discovered in the $BaPb_{1-x}Bi_xO_3$ system by Sleight of DuPont. Although this perovskite system reaches a maximum $T_c$ of only about 13 K, it is considered unconventional because such a $T_c$ requires a large $V_{BCS}N(E_F)$ product [Eq. (42)], and a small measured $N(E_F)$ then requires an exceptionally large pairing potential $V_{BCS}$. Therefore, the system has been examined for clues to the strong coupling mechanism operative in the higher-$T_c$ copper oxides (Fig. 1).

Superconductivity also appears on suppression of the static CDW of $BaBiO_3$ by substitution of more than 12% $K^+$ for $Ba^{2+}$ in $Ba_{1-x}K_xBiO_3$; a maximum $T_c = 32$ K is found near $x = 0.4$, where the system becomes cubic. For $x > 0.47$, the system behaves as a normal metal without any superconductivity. The effect on $T_c$ of substituting $^{18}O$ for $^{16}O$ gave a conventional isotope shift, $\alpha = 0.4$ to 0.5, which indicates that the BCS phonon-mediated pairing mechanism is operative in these systems. On the other hand, Kumal, Hall, and Goodrich have shown that the transition at $T_c$ is fourth-order, not second-order, in the Ehrenfest classification.

## B. Copper-Oxide Superconductors

### 1. Structure

Where the tolerance factor in Eq. (77) approaches unity, epitaxial (001) interfaces between an AX rock-salt layer and an $ABX_3$ perovskite layer are lattice matched. Nature recognizes this fact by stabilizing intergrowth structures $(AX)(ABX_3)_n$ in which perovskite layers alternate with rock-salt layers along an [001] axis. The $La_2CuO_4$ structure in Fig. 29, for example, is tetragonal at high temperatures, with LaO rock-salt layers alternating with $LaCuO_3$ perovskite layers on traversing the $c$-axis. Lattice matching requires a 45° rotation of the [100] axis of a rock-salt layer relative to that of a perovskite layer. As in the perovskite structure itself, the Goldschmitt tolerance factor of Eq. (77) is a measure of the mismatch of the equilibrium A—X and B—X bond lengths. Since the A—X and B—X bonds have different thermal expansions and compressibilities, matching ($t = 1$) of the bond lengths can be perfect only at a specific temperature for a given pressure, and the value of $t$ calculated from tabulated ionic radii corresponds to room temperature at ambient pressure. On cooling, $t$ decreases, and a $t < 1$ is compensated by a cooperative rotation of the $CuO_{6/1.5}$ octahedra about a tetragonal [110] axis (Fig. 29b), to lower the symmetry to orthorhombic. These rotations buckle the Cu–O–Cu bonds from 180° to $(180° - \phi)$, so the $CuO_2$ planes become $CuO_2$ sheets.

In the case of $La_2CuO_4$, the crystallographic $c/a$ ratio is anomalously high because the Cu(II) ions distort their

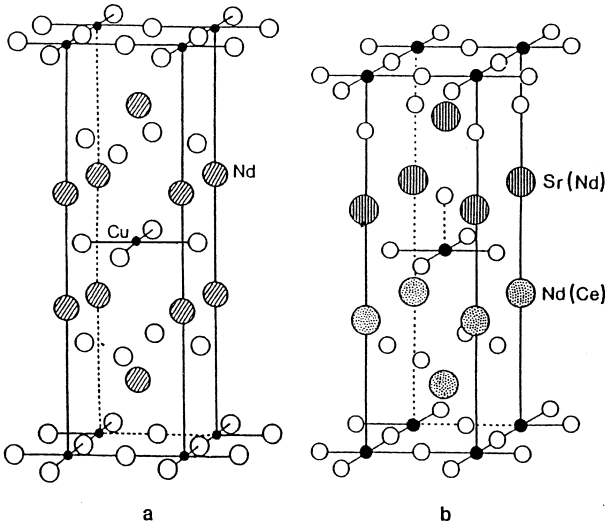



**FIGURE 30** (a) $T'$-tetragonal structure of $Nd_2CuO_4$. (b) $T^*$-tetragonal structure of $Nd_{2-x-y}Ce_ySr_xCuO_4$.

octahedra to tetragonal ($c/a > 1$) symmetry with their long apical Cu—O bond along the $c$-axis. This $c$-axis ordering of the filled ($3z^2–r^2$) orbitals has an important structural consequence: the apical oxygen atoms are not strongly bound to the Cu atoms and may be removed from the oxygen coordination at a Cu(II). Consequently, the Cu(II) ion may be found in six-, five-, or fourfold oxygen coordination, but the strong square-coplanar bonding within a $CuO_2$ sheet is always maintained.

Replacement of La with a smaller trivalent rare earth ion illustrates well the weak bonding of the apical oxygen. Substitution of a smaller A cation lowers $t$, and the structure accommodates the bond-length mismatch by displacing the apical oxygen to tetrahedral interstices of an (001) La bilayer to form a fluorite Ln—$O_2$—Ln layer (Ln = Pr to Gd) as illustrated in Fig. 30a. This structure is labeled T′-tetragonal to distinguish it from the T-tetragonal phase of high-temperature $La_2CuO_4$. An important consequence for the chemistry of these phases is that the displacement of the apical oxygen in the T′ phase places the $CuO_2$ planes under tension, whereas the $CuO_2$ sheets of $La_2CuO_4$ are under compression as a result of the bond-length mismatch. A tensile stress is relieved by adding antibonding ($x^2–y^2$) electrons to the $CuO_2$ planes; a compressive stress is relieved by removing antibonding ($x^2–y^2$) electrons from the $CuO_2$ sheets. As a result, the T′ phase can only be doped $n$-type to give $n$-type superconductivity, whereas $La_2CuO_4$ can only be doped $p$-type. In fact, care must be exercised in the preparation of $La_2CuO_4$, as it may accept interstitial oxygen in the tetrahedral sites of the rock-salt bilayers to give $La_2CuO_{4+\delta}$; this composition phase segregates below room temperature to give filamentary $p$-type superconductivity in the oxygen-rich phase. In the $p$-type system $La_{2-x}Sr_xCuO_4$, the larger $Sr^{2+}$ ion relieves the compressive stress, and oxygen stoichiometry is more easily achieved as $x$ increases.

In the T*-tetragonal structure of $Nd_{2-x-y}Ce_ySr_xCuO_4$ (Fig. 30b), the larger $Sr^{2+}$ ions order into alternate A-cation bilayers; the $Sr^{2+}$ ions stabilize rock-salt bilayers, whereas the alternate bilayers have the fluorite structure. As a result, the Cu(II) are fivefold coordinated. Whether the Cu(II) are six-, five-, or fourfold coordinated, superconductivity requires preservation of the translational symmetry within a $CuO_2$ plane or sheet and, therefore, the same nearest-neighbor oxygen coordination for every copper atom within a plane or sheet.

The variable oxygen coordination at a Cu(II) also makes it possible to remove the apical oxygen atoms from a perovskite multilayer if the A-site cations of the perovskite block are stable in eightfold oxygen coordination. In fact, all the copper-oxide superconductors that would contain perovskite multilayers contain eightfold coordination of the A' cations (A' = Ca, Y, or a trivalent lanthanide) to form an $A'_{m-1}(CuO_2)_m$ layer, with integral $m \geq 2$; these superconductive layers alternate with AO–Φ–AO layers (A = La, Sr, Ba) that have a rock-salt AO interface. The intralayer composition Φ has a variable oxygen content and may be quite varied, as illustrated in Figs. 31 to 35.

The nonsuperconductive layer may act as a charge reservoir for the holes in the $p$-type superconductive layers. This situation is found, for example, in the $YBa_2Cu_3O_{6+x}$ system, where the $0 \leq x < 1$ oxygen in the BaO—$CuO_x$—BaO layers order into Cu—O—Cu chains

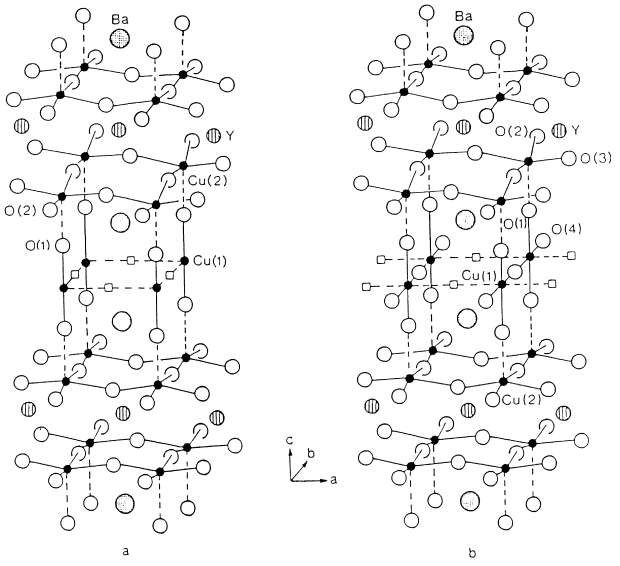



**FIGURE 31** Structures of (a) tetragonal $YBa_2Cu_3O_6$ and (b) orthorhombic, ideal $YBa_2Cu_3O_7$.

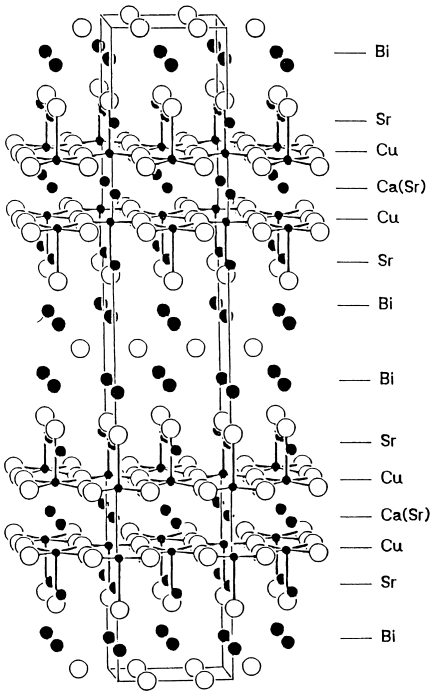



**FIGURE 32** Tetragonal subcell of $Bi_2Sr_{3-x}Ca_xCu_2O_{8+y}$ showing the $CuO_2$ layers. Metal atoms are shaded and only Cu–O bonds are indicated. Oxygen atom positions for the Bi layers are idealized. [After Subramanian, M. A., *et al.* (1988). *Science* **239,** 1015.]

for $x > 0.4$; the fully formed chains are more conductive than the superconductive $CuO_2$ sheets and they become superconductive with the $CuO_2$ sheets. Moreover, displacement of the apical oxygen regulates the distribution of holes between the chains and the sheets. On the other hand, the superconductive ($T_c = 16$ K) $RuSr_2GdCu_2O_8 = (CuO_2GdCuO_2)(SrORuO_2SrO)$; is also ferromagnetic, with a Curie temperature of 133 K. As in the ferromagnetic perovskite $SrRuO_3$, the Ru atoms carry a magnetic moment $\mu_{Ru} \approx 1\ \mu_B$. Clearly the nonsuperconductive, ferromagnetic layer in this compound must be electronically isolated from the superconductive layer, but the internal magnetic field lowers $T_c$.

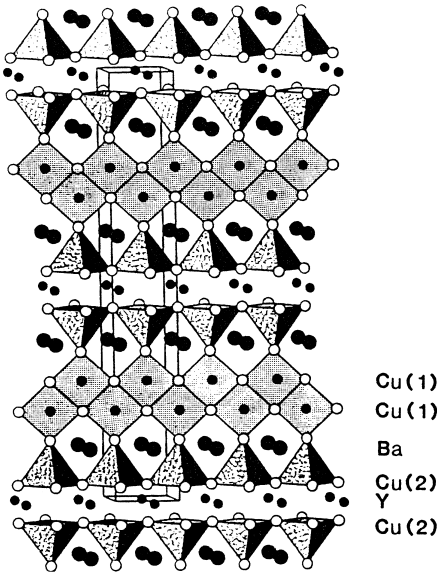



**FIGURE 33** Structure of $YBa_2Cu_4O_8$ showing double chains in the $BaO–Cu_2O_2–BaO$ layer.

## 2. System $La_{2-x}Sr_xCuO_4$

*a. Phase identification.* The intergrowth structure of $La_2CuO_4$ (Fig. 29) is the simplest that exhibits *p*-type superconductivity. Holes may be introduced into the $(x^2–y^2)$ band by creating A-site vacancies or interstitial oxygen; more useful is the substitution of an alkaline-earth ion $A^{2+}$ for $La^{3+}$ in oxygen-stoichiometric $La_{2-x}A_xCuO_4$. The $La_{2-x}A_xCuO_4$ system is of particular interest for two reasons: (1) the number $x$ of holes per formula unit is unambiguously introduced into the $(x^2–y^2)$ band of the $CuO_2$ sheets; and (2) the solid-solution range $0 \le x \le 0.3$ spans the entire range of superconductive compositions, as shown in the phase diagram in Fig. 36.

Crystallographically, there are two distinguishable phases in Fig. 36, a high-temperature tetragonal (HTT) and a low-temperature orthorhombic (LTO) phase, resulting from cooperative rotations of the $CuO_{6/1.5}$ octahedra. In the $La_{2-x}Ba_xCuO_4$ system, cooperative rotations about the [100] and [010] axes in alternate $CuO_2$ sheets produce a low-temperature tetragonal (LTT) phase below about 60 K in the range $0.12 \le x \le 0.15$. In Fig. 36, the LTO–HTT transition temperature $T_t$ is seen to drop with increasing $x$, crossing $T_c$ near $x = 0.22$.

The transport data distinguish three electronic phases below room temperature: an antiferromagnetic phase in the range $0 \le x \le 0.02$, a superconductive phase in the range $0.1 < x < 0.22$, and (3) an *n*-type metallic phase for $0.26 < x \le 0.30$. At $x \approx 0.125$, there is a weak suppression of $T_c$ vs $x$; in the $La_{2-x}Ba_xCuO_4$ system, superconductivity is completely suppressed in the range $0.12 \le x \le 0.13$ by the stabilization of a static CDW. Tranquada *et al.* have shown that the LTT phase of $La_{2-x}Ba_xCuO_4$ stabilizes at $x \approx 0.125$, a static CDW having the form of alternating hole-rich and antiferromagnetic stripes running parallel to the tetragonal [100] and [010] axes, respectively, in alternate $CuO_2$ sheets. Under hydrostatic pressure, superconductivity is restored to the compositions where it was suppressed by the static CDW. From X-ray absorption fine structure (XAFS), Bianconi *et al.* previously found evidence suggesting mobile stripes in superconductive samples, and an open question is whether the stripes are mobile in the superconductive phase or whether a related vibronic coupling characterizes the charge carriers.

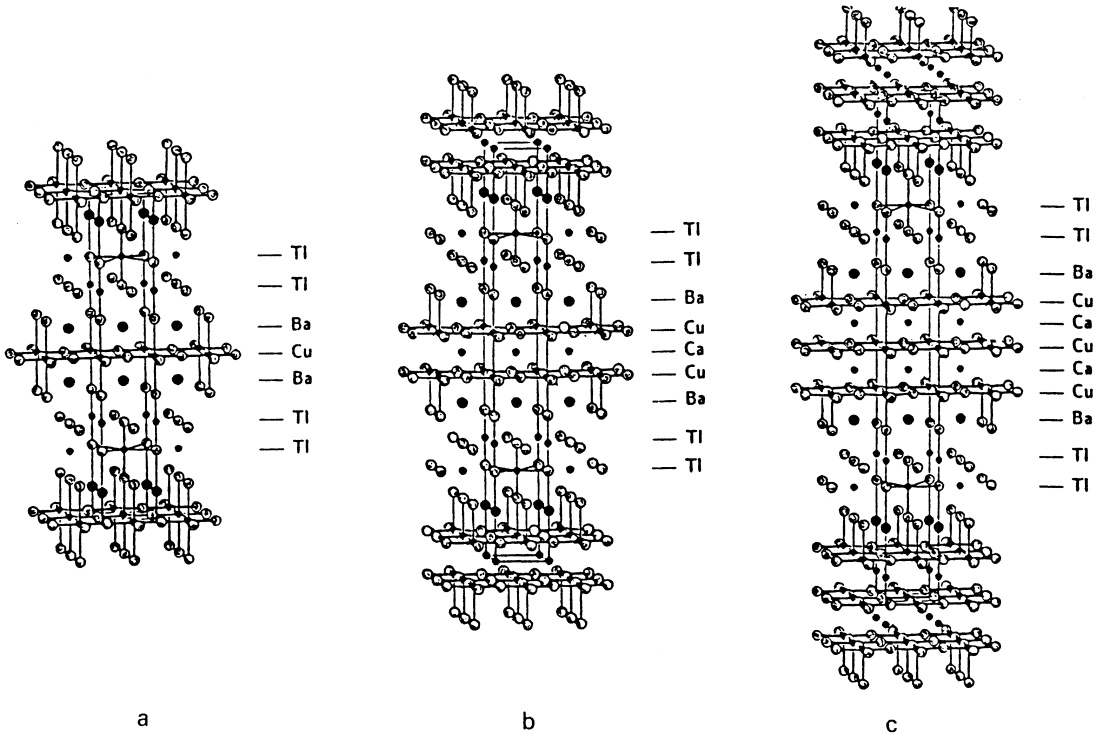



**FIGURE 34** Structures of (a) $Tl_2Ba_2CuO_6$, (b) $Tl_2Ba_2CaCu_2O_8$, and (c) $Tl_2Ba_2Ca_2Cu_3O_{10}$.

*b. Underdoped.* The "underdoped" region $0 < x < 0.1$ supports polaronic conduction, but the thermoelectric power indicates that the nonadiabatic polorans are not small, centered on one Cu atom, but embrace about five Cu centers. A pseudo Jahn–Teller deformation of the in-plane square-coplanar coordination at a Cu(III) would be resisted by an elastic energy, but this energy would be reduced by cooperative deformations over several Cu centers. Calculation has shown that the gain in elastic energy would result in a polaron that embraced five to seven Cu centers. However, some other vibronic mechanism may be responsible for preventing the polaron collapse to a single Cu center. Within the polarons, antiferromagnetic order is suppressed, which indicates that the hole occupies a molecular orbital that includes all the Cu centers of the polaron. In this respect, it represents a mobile metallic phase in the antiferromagnetic matrix. As the volume of this second electronic phase increases, it breaks up the long-range antiferromagnetic order of the parent phase, which causes $T_N$ to decrease precipitously with $x$ from 340 K at $x = 0$. However, localized spins in regions of short-range order persist into the superconducting compositions; they give rise to a maximum in the paramagnetic susceptibility at a $T_{max}$ that decreases with increasing $x$.

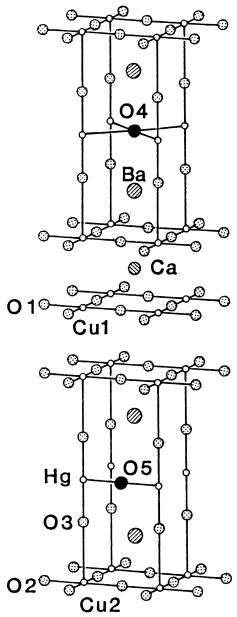



**FIGURE 35** Structure of $HgBa_2Ca_2Cu_3O_{8+\delta}$. The $O_4$ and $O_5$ are the $\delta$ interstitial oxygen.

The appearance of a superconductive $T_c$ that increases with $x$ for compositions $0.05 \leq x \leq 0.10$ indicates that the polarons condense at lower temperatures into superconductive filaments. The transition temperatures $T_F$ and $T_\rho$ in Fig. 36 mark anomalies in the temperature dependence of the transport properties; others have noted anomalies in

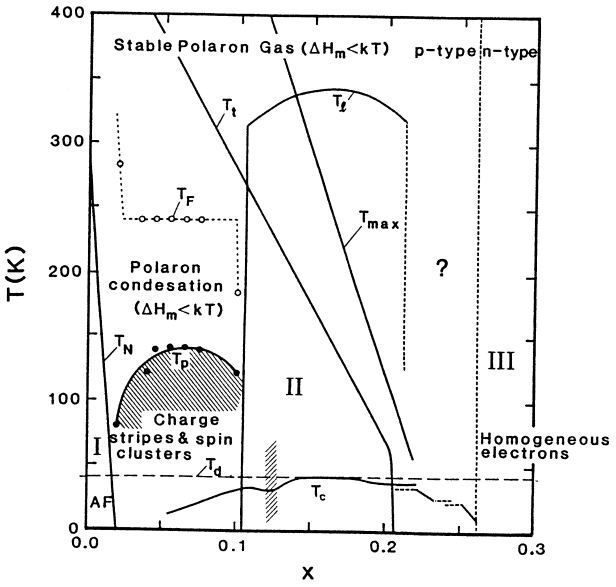



**FIGURE 36** Phase diagram of $La_{2-x}Sr_xCuO_4$.

NMR and thermodynamic measurements at similar temperatures $T^*$ (not shown in Fig. 36) that indicate the opening of a psuedogap, i.e., the lowering of the density of states, at $\varepsilon_F$. These anomalies appear to reflect interactions between polarons and their ordering into hole-rich stripes. Hunt *et al*. have used $^{63}Cu$ nuclear quadruple resonance (NQR) to reveal the presence of slowly fluctuating, quasitatic charge (i.e., hole-rich) stripes in the range $\frac{1}{16} \leq x \leq \frac{1}{8}$; the stripes become increasingly ordered on lowering the temperature, but the ordering temperature decreases with increasing $x$, vanishing at $x = \frac{1}{8}$. The polaron ordering apparently occurs within a parent phase that decreases in volume with increasing $x$; it is replaced by a single superconductive phase.

*c. Optimally doped.* The superconductive compositions $0.14 \leq x \leq 0.20$ exhibit a nearly temperature-independent thermoelectric power $\alpha(T)$ above $T_l$; there is no dramatic change in the evolution of $\alpha(T)$ with $x$ above $T_l$. However, below $T_l$ there is an abrupt change in the character of $\alpha(T)$ between $x = 0.10$ and $x = 0.15$. On cooling below $T_l$ in the range $0.15 \leq x \leq 0.22$, $\alpha(T)$ increases relatively steeply to a maximum value at about 140 K, too high a temperature to be due to phonon drag. Zhou and Goodenough have shown that this unusual features is present in all the single-phase copper-oxide superconductors, and only where there are superconductive $CuO_2$ sheets. This feature reflects the appearance of itinerant quasiparticles of momentum $\hbar\mathbf{k}$ that have an unusual dispersion $\varepsilon_{\mathbf{k}}(\mathbf{k})$ of their one-particle energies. Mihailovic *et al*. have used femtosecond time-domain spectroscopy to demonstrate a change from a polaronic to an itinerant character of the mobile holes on passing from the underdoped to the optimally doped compositions in the $YBa_2Cu_3O_{6+x}$ system, and the Fermi surface of the itinerant quasiparticles in optimally doped $CuO_2$ sheets has been mapped with photoemission spectroscopy (PES). Most significant, angle-resolved PES as a function of temperature by Norman *et al*. and Dessau *et al*. has revealed a massive transfer of spectral weight on cooling from the $\pi, \pi$ to the $\pi, 0$ directions within a $CuO_2$ sheet. These data indicate a progressive stabilization of the itinerant quasiparticles with $\mathbf{k}$ vectors along the Cu—O—Cu bonds of a $CuO_2$ sheet relative to those directed along a tetragonal [110] direction. The $\varepsilon_{\mathbf{k}}(\mathbf{k})$ dispersion becomes extremely flat at the Fermi energy $\varepsilon_F$ in the direction of the Cu—O—Cu bonds, indicating that the quasiparticles of the dominant population at $\varepsilon_F$ have an unusually heavy mass $m^*$.

The origin of the heave mass $m^*$ has not been resolved. $T_c$ is expected to reach a maximum at the crossover from Cooper pairing to Bose condensation of bipolarons. Alexandrov and Mott have explored the bipolaron option most thoroughly. Markiewicz has argued extensively for trapping of the Fermi energy in a van Hove singularity. Although there is considerable evidence that the cuprates are close to the Bose–Einstein condensation regime, the PES data show the charge carriers are itinerant in the superconductive phase. Goodenough and Zhou have suggested that the large $m^*$ is due to an unusual electron–lattice (or electron–polarization) interaction that gives rise to vibronic itinerant quasiparticles. The transfer of spectral weight in the angle-resolved PES spectra are consistent with the latter view, as are the data on the pressure dependence of $T_c$. A vibrational or polarization wave that is hybridized with a traveling-electron wave would be sensitive to changes in the bending angle $\phi$ of a $(180^\circ - \phi)$ Cu—O—Cu bond. The hydrostatic pressure $P$ decreases $\phi$, and the $T_c$ of the LTO phase increases with $P$, whereas a $dT_c/dP = 0$ is found for the tetragonal ($\phi = 0^\circ$) phase. Moreover, epitaxial $La_{1.85}Sr_{0.15}CuO_4$ films on $SrTiO_3$ have their $CuO_2$ sheets under tension and the $T_c$ is lowered; those on $LaSrAlO_4$ have their $CuO_2$ sheets under compression and the $T_c$ is raised. The compressive stress built into the films on $LaSrAlO_4$ allows an added hydrostatic pressure achievable in a Cu—Be pressure cell to access at low temperature the tetragonal phase of the optimally doped $La_{1.85}Sr_{0.15}CuO_4$; $T_c$ increased with $P$ to 47 K, where it became $P$-independent on going from the orthorhombic to the tetragonal phase.

Below $T_c$, NMR Knight shift and other measurements have established that superconductive particles consist of two spin-paired electrons as in a conventional superconductor. However, a short coherence length $\xi_\circ \approx 15$ Å means that the coulomb repulsion between paired electrons is much stronger in the copper oxides. In a conventional superconductor, weak coulomb interactions

result in pair wave functions with $s$-wave symmetry; the superconductive energy gap $2\Delta_0$ is finite over the entire Fermi surface. The pair wave functions in the $CuO_2$ sheets of the copper oxides have $(x^2-y^2)$ $d$-wave symmetry; the energy gap $2\Delta_0$ has nodes along the tetragonal [110] and [1$\bar{1}$0] axes. This symmetry reduces the coulomb repulsion between the paired electrons.

*d. Overdoped.* The overdoped compositions $x > 0.25$ are not superconductors, and a change from $p$-type to $n$-type conduction signals a transfer of spectral eight from the lower and upper Hubbard bands of the $x = 0$ parent compound to Fermi-liquid states in the gap $(U-W)$ of the parent. Nevertheless, the metallic resistivity remains high with an anomalous temperature dependence, which indicates that the transition from vibronic to Fermi-liquid states may not be complete by $x = 0.3$.

The decrease in $T_c$ with increasing $x$ in the range $0.22 < x < 0.26$ is not smooth; it is characterized by a series of steps typical of phase segregation. Since $T_t$ crosses $T_c$ near this compositional range, these steps may reflect segregation of orthorhombic and tetragonal phases.

### 3. System $YBa_2Cu_3O_{6+x}$

The possibility of practical superconductive devices operating at the boiling point of liquid nitrogen (77 K) captured the imagination of the technical community on the discovery of a superconductive critical temperature of 90 K in $YaBa_2Cu_3O_{6.95}$. Although other superconductors with a higher $T_c$, a greater chemical stability, and cleavage planes that simplify fabrication into tapes and wires have since been discovered, the $YBa_2Cu_3O_{6+x}$, $0 \le x < 1$, system continues to be of technical importance because, so far at least, films of $YaBa_2Cu_3O_{6.95}$ have been able to sustain the highest critical currents.

The structure, shown in Fig. 31, contains $CuO_2$—Y—$CuO_2$ layers and BaO—$CuO_x$—BaO layers. The oxygen atoms of the BaO buckled planes are $c$-axis apical oxygen atoms of the Cu in the $CuO_2$—Y—$CuO_2$ layers; these Cu all have fivefold oxygen coordination. The Cu of the BaO—$CuO_x$—BaO layers bridge the apical oxygen atoms with 180° O—Cu—O bonds oriented parallel to the $c$-axis. The $x$ interstitial oxygen atoms in the BaO—$CuO_x$—BaO layers are mobile above 300°C, and their equilibrium concentration depends on the temperature and atomsphere. An air anneal at 400°C is used to obtain the optimally doped $YaBa_2Cu_3O_{6.95}$ composition. Since the apical oxygen atoms participate in the interstitial oxygen diffusion, it is important to ensure that the thermal history does not leave apical oxygen vacancies, which perturb the periodic potential of the superconductive $CuO_2$—Y—$CuO_2$ layers and suppress $T_c$. It is interesting that the $Y^{3+}$ ion may be substituted by any trivalent rare earth ion (with the exception of Pr) without influencing $T_c$ significantly. Only in the case of Pr is there an important interaction between the lanthanide $4f$ orbitals and the $(x^2-y^2)$ band of the $CuO_2$ sheets.

$YBa_2Cu_3O_6$ is an antiferromagnetic insulator with Cu(II) in the $CuO_2$—Y—$CuO_2$ layers and Cu(I) in the BaO—Cu—BaO layers; antiferromagnetic order between Cu(II) ions sets in at a $T_N > 500$ K. The initial interstitial oxygen atoms enter the BaO—Cu—BaO layers randomly and oxidize the neighboring Cu(I) to Cu(II). However, a threefold-coordinated Cu(II) attracts a second interstitial oxygen atom to form square-coplanar coordination, which initiates the formation of a chain segment. The twofold-coordinated Cu remain Cu(I), so the formation of chain segments initiates oxidation of the $CuO_2$—Y—$CuO_2$ layers. For $x < 0.3$, the chain segments remain disordered, so the crystallographic symmetry is tetragonal; but $T_N$ drops precipitously with increasing oxidation of the $CuO_2$—Y—$CuO_2$ layers in the interval $0.1 < x < 0.25$ (see Fig. 37). For $x \ge 0.4$, the crystallographic symmetry is

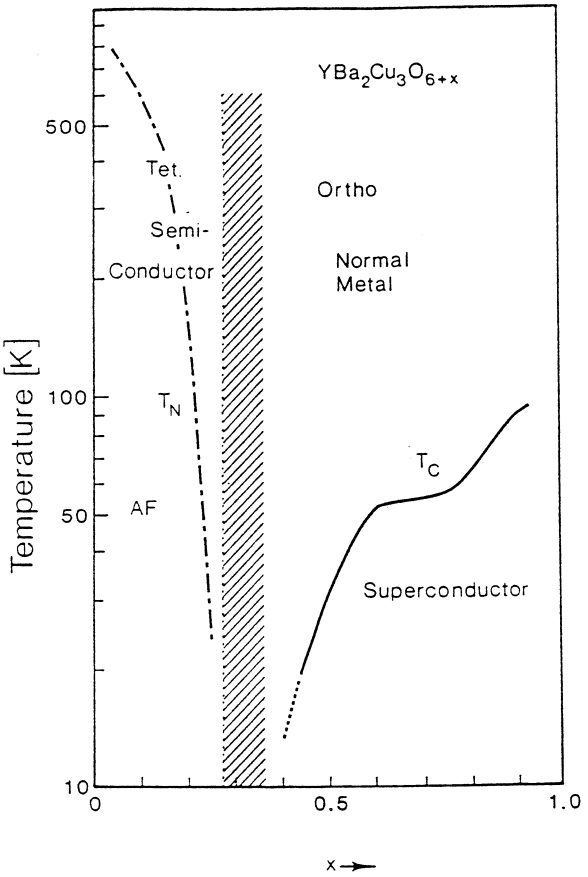



**FIGURE 37** Phase diagram for the system $YBa_2Cu_3O_{6+x}$.

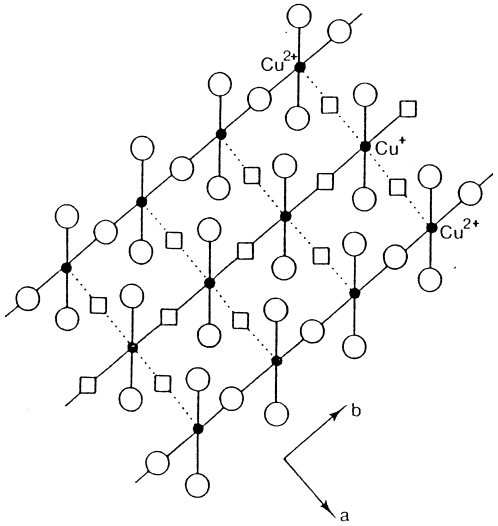



**FIGURE 38** Ideal oxygen ordering of chains in $YBa_2Cu_3O_{6.5}$.

orthorhombic as a result of an alignment of the chains along the orthorhombic $b$-axis. The orthorhombic compositions are superconductors with a $T_c$ that increases with $x$ in two steps, a $T_c \approx 60$ K plateau appearing in the interval $0.6 < x < 0.8$. At $x = 0.5$, the chains order, alternating with Cu(I) $b$-axis rows as illustrated in Fig. 38. If the fully formed chains contained all Cu(II), the number of holes per Cu atom in the $CuO_2$—Y—$CuO_2$ layers would be 0.15 at $x = 0.5$, close to the optimal doping. However, the fully formed chains are oxidized beyond Cu(II), which reduces the number of holes per Cu atom in the $CuO_2$—Y—$CuO_2$ layers and makes the chains metallic conductors. As first pointed out by Cava, the chains act as charge reservoirs for the superconductive $CuO_2$—Y—$CuO_2$ layers. In $YBa_2Cu_3O_{6.95}$, the $CuO_2$—Y—$CuO_2$ layers are optimally doped with about 0.18 hole/Cu atom and the chains in the BaO—$CuO_{0.95}$—BaO layers are also superconductive.

Orthorhombic symmetry and superconductive chains are not determinants of the superconductivity of the $CuO_2$—Y—$CuO_2$ layers. By doping an equal amount of Ca for Y and La for Ba, the total hole concentration is kept constant. By 40% doping of Ca and La, the La in the BaO—$CuO_x$—BaO layers breaks up the chains into randomly oriented chain segments, changing the symmetry to tetragonal and suppressing superconductivity in the chain segments. As a result, the mobility of $H_2O$ or $CO_2$ species in the nonsuperconductive layers is reduced, which suppresses chemical degradation at room temperature on exposure to the atmosphere, but the superconductive transition temperature is reduced only from 90 to 78 K.

Figure 39 shows the phase diagram of applied magnetic field $H$ vs temperature $T$ for clean $YBa_2Cu_3O_{6.95}$ with $H$ parallel to the $c$-axis. The copper-oxide superconductors are all strongly Type II, and $H_{c1}(T)$ marks the

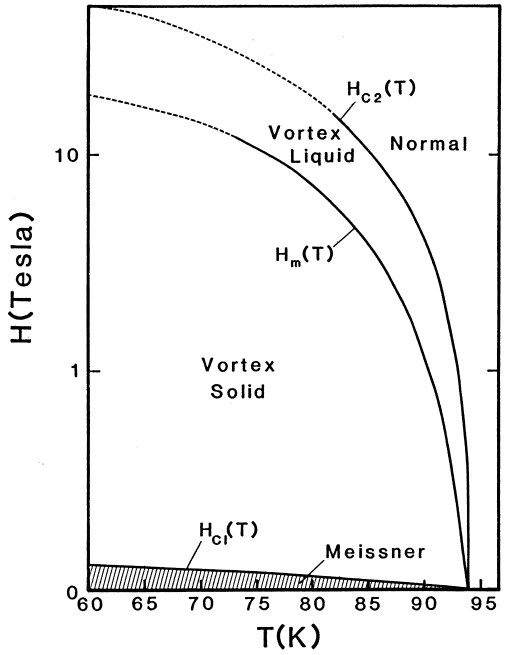



**FIGURE 39** Upper and lower critical magnetic fields $H_{c2}(T)$ and $H_{c1}(T)$ and critical field for vortex melting $H_m(T)$ for $YBa_2Cu_3O_{6.95}$.

transition between the Meissner phase and the vortex state. The vortex solid consists of an array of stationary vortices. The ability to grow crystals with a good surface quality has allowed imaging of the vortex lattice with scanning tunneling microscopy (STM). The vortex lattice does not show long-range order into the hexagonal close-packed structure; it represents a glassy state rather than a regular lattice. However, locally the flux lines are arranged in an oblique lattice, with approximately equal primitive lattice vectors forming an angle between them of $77 \pm 5°$. Moreover, the shape of the vortex cores is elliptical, not circular, with the long axis along an in-layer orthorhombic axis. These features reflect the anisotropy within an $a$–$b$ plane that is induced by orientation of the chains of the BaO—$CuO_x$—BaO layers along the orthorhombic $b$-axis.

The strength of the pinning of the vortex solid in copper-oxide superconductors depends on the coupling between layers. With weak coupling, the vortex lattice of one layer may be displaced relative to that of an adjacent layer, thereby bending the flux trajectory through the vortex cores. The interlayer vortex coupling is relatively strong in $YaBa_2Cu_3O_{6.95}$. Nevertheless, the vortex solid melts at an $H_m(T) < H_{c2}(T)$; the melting transition is weakly first-order. The vortices of the vortex liquid are mobile; moving vortices dissipate energy and introduce a finite resistance. Associated with $H_m(T)$ is an irreversibility line $H_{irr}(T) < H_m(T)$. Both $H_{irr}(T)$ and $H_m(T)$ decrease

sharply with loss of oxygen from the BaO—$CuO_x$—BaO layers.

Pinning of the vortex solid is an extrinsic phenomenon, and considerable effort has been given to finding ways to increase the pinning. In films, surface roughness and the deposition of a protective $(Y_{1-x}Ca_x)(Ba_{2-x}La_x)Cu_3O_{7-\delta}$ overlayer increase $H_m(T)$ and therefore the critical current.

The brittleness of the ceramics and the two-dimensional superconductive layers makes fabrication of flexible tapes or wires a formidable challenge. Alignment of the layers from grain to grain is a critical requirement that is most easily achieved by deposition of films on a flexible metallic tape.

### 4. Layers with Three $CuO_2$ Sheets

The highest values of $T_c$ have been obtained with structures containing layers with three $CuO_2$ sheets. Figure 35 shows the structure of $HgBa_2Ca_2Cu_3O_{8+\delta}$, which has a $T_c = 135$ K at ambient pressure and a $T_c = 164$ K under a quasi-hydrostatic pressure of 30 Gpa. The lattice oxygen atoms of the BaO—$HgO_\delta$—BaO nonsuperconductive layers supply the apical oxygen atoms of the two outer sheets of the $CuO_2$—Ca—$CuO_2$—Ca—$CuO_2$ superconductive layers; the Cu atoms of the inner $CuO_2$ plane have square-coplanar oxygen coordination. The Cu—O—Cu bond angles within the outer $CuO_2$ sheets approach the optimal 180°. The value of $T_c$ at ambient pressure decreases from 135 K to 94 K as the number of $O_4$ interstitial oxygen in the $HgO_\delta$ planes decreases from 0.18 to 0.10 per Hg atom. The $O_4$ oxygen atoms oxidize the superconductive layers, the $O_5$ interstitial oxygen of the $HgO_\delta$ planes do not. This result implies near-optimal doping with 0.36 hole per formula unit in the superconductive layers. These holes would be distributed predominantly in the outer sheets with Cu in fivefold oxygen coordination, which would give a maximum hole concentration of 0.18/Cu atom in these sheets. Optimal doping is thus seen to correspond well with that in other $p$-type copper-oxide superconductors. Why pressure increases the $T_c$ in $HgBa_2Ca_2Cu_3O_{8+\delta}$ is not known, but it is reasonable to assume that a redistribution of holes within the superconductive layers increases the coupling between the outer $CuO_2$ sheets.

### 5. Electron Superconductors

The equilibrium Cu—O bond length for Cu(II) ions in square-coplanar oxygen coordination is about 1.93 Å. To dope $n$-type a $CuO_2$ plane without apical oxygen atoms by reducing Cu(II) to Cu(I), it has been necessary to place the $CuO_2$ planes under tension so as to make the Cu—O bond of the compound larger than 1.93 Å. This feature is illustrated by the two copper-oxide structures that exhibit $n$-type superconductivity.

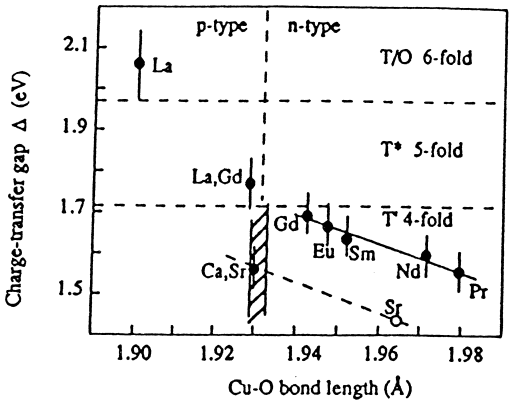



**FIGURE 40** Charge-transfer gap vs Cu—O bond length for $Ln_2CuO_4$ oxides.

*a. $T'$-$Ln_{2-x}Ce_xCuO_4$.* The parent $T'$ phases $Ln_2CuO_4$ have been prepared at atmospheric pressure for Ln = Pr,. . . , Gd; they have the structure of Fig. 30a, which has isolated $CuO_2$ planes having no apical oxygen. However, care must be taken to order the oxygen in the fluorite layers, as fivefold oxygen coordination at a few Cu atoms would perturb the periodic potential. Sensitivity to perturbations of the periodic potential is another indicator that the charge carriers are itinerant in both the $p$-type and the $n$-type superconductors.

The parent compounds contain only Cu(II) and are antiferromagnetic insulators such as the parent $La_2CuO_4$ compound with the T/O structure. Figure 40 shows, for room temperature, the magnitude of the energy gap $E_g = U - W$ vs the Cu—O bond length, which remains longer than the equilibrium bond length over the entire series of $n$-type superconductors. In each $T'$ system $Ln_{2-x}M_xCuO_4$, M = Ce or Th, $n$-type superconductivity is found only in a narrow compositional range $0.10 \le x \le 0.18$, and at larger $x$ a nonsuperconductive metallic state persists to lowest temperatures. Thus the $n$-type superconductors, like the $p$-type superconductors, appear as a distinguishable thermodynamic phase at a crossover from an antiferromagnetic-insulator to a metallic phase (Fig. 41). However, there are also significant differences between the $p$-type and $n$-type superconductors. For example, the charge carriers in the underdoped $T'$ systems are conventional small polarons and $T_N$ decreases only slowly with increasing $x$ in a manner typical of a simple dilution with nonmagnetic Cu(I) ions. Moreover, the transition from antiferromagnetic semiconductor to superconductor appears to be a conventional first-order phase change occurring at a critical charge-carrier concentration $x_c$. With decreasing Cu—O bond length, there is a systematic increase in $x_c$

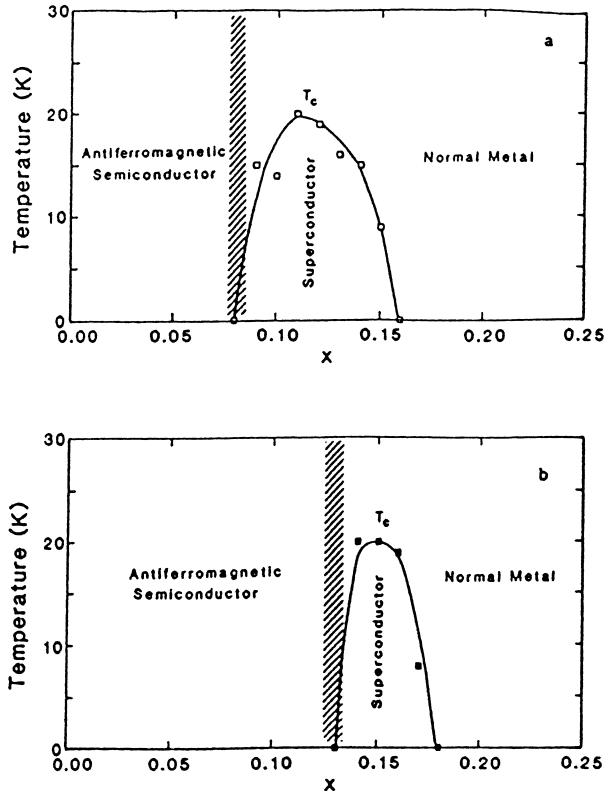



**FIGURE 41** Variations of $T_c$ with $x$ for (a) $LaNd_{1-x}Ce_xCuO_4$ and (b) $Nd_{2-x}Ce_xCuO_4$. Shaded areas refer to two-phase regions.

and a decrease in the Ce solubility limit $x_l$ that results in a narrowing of the superconductive phase field until, in $Gd_2CuO_4$, it disappears altogether. The appearance of $n$-type superconductivity is restricted not only to compounds with Cu in fourfold, square-coplanar coordination, but also to those where the Cu—O bonds of the $CuO_2$ planes are stretched sufficiently beyond 1.93 Å to give an $x_c < x_l$.

*b. Infinite layers.* The infinite-layer structure in Fig. 42 was first stabilized in the compound $Ca_{0.86}Sr_{0.14}CuO_2$; it has an equilibrium Cu—O bond length of 1.93 Å at room temperature. Synthesis at atmospheric pressure allows little variation in the Ca/Sr ratio, and attempts to dope the compound either $p$-type or $n$-type were unsuccessful. On the other hand, $SrCuO_2$ can be prepared under high pressure; it has a Cu—O bond length stretched to 1.965 Å, which satisfies the criterion for $n$-type doping. Therefore, $Sr_{1-x}Ln_xCuO_2$ (Ln = La, Pr, Nd) were prepared under high pressure; they proved to be $n$-type superconductors with a $T_c \approx 30$ K.

## 6. Mechanism: An Open Question

The pairing mechanism in the copper-oxide superconductors remains an open question; a consensus on the character of the charge carriers in the normal state has yet to be reached. Most theorists have investigated the role of spin–spin exchange interactions without consideration of electron coupling to the lattice or to the oxygen polarization. Since the Cu(III) are diamagnetic, these efforts have been able to justify the separation of holes into charge stripes, but the charge separation can be achieved by other forces. To date, a convincing description of the high-$T_c$ phenomenon has yet to emerge. Nevertheless, experiment has shown that spin fluctuations persist into the superconductive phase, and inelastic neutron scattering has revealed a commensurate $(\pi, \pi)$ resonance peak in the spectrum of the antiferromagnetic susceptibility $\chi(q, \omega)$ that has a half-width in momentum space that varies linearly with $T_c$. These measurements define a characteristic velocity that is lower than a typical electron velocity at the Fermi energy and an order of magnitude smaller than a spin-wave velocity. These spin fluctuations could be associated with either mobile stripes or with slowly moving electron-density fluctuations as in a vibronic state. The data do not reveal what is the driving force for the formation of these density fluctuations.

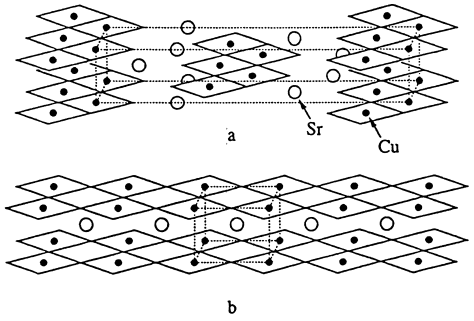



**FIGURE 42** Comparison of the (a) atmospheric-pressure and (b) high-pressure forms of $SrCuO_2$.

## ACKNOWLEDGMENT

Support of this work by the R. A. Welch Foundation, Houston, Texas, is gratefully acknowledged.

## SEE ALSO THE FOLLOWING ARTICLES

BONDING AND STRUCTURE IN SOLIDS • ELECTRONS IN SOLIDS • SUPERCONDUCTING DEVICES • SUPERCONDUCTIVITY

## BIBLIOGRAPHY

Alexandrov, A. S., and Mott, N. F. (1996). "Polarons and Bipolarons," World Scientific, Singapore.

Blatt, J. M. (1964). "Theory of Superconductivity," Academic Press, New York.
de Gennes, P. G. (1966). "Superconductivity of Metals and Alloys," Benjamin, New York.
Ginsberg, D. M. (ed.) (1996). "Physical Properties of High-Temperature Superconductors," Vol. 5, World Scientific, Singapore.
Ginzburg, V. L., and Kirzhmits, D. A. (eds.) (1982). "High Temperature Superconductivity" (A. K. Agyei, transl.; J. L. Birman, transl. ed.), Pergamon, Oxford.
Goodenough, J. B. (1972). *Prog. Solid State Chem.* **5,** 145.
Goodenough, J. B., and Longo, J. M. (1970). "Crystallographic and Magnetic Properties of Perovskite and Perovskite Related Compounds, in Landolt–Bornstein Tabellen," New Series Group III/4a, No. 126, Springer-Verlag, Berlin, New York.
Kaldis, E. (ed.) (1994). "Materials and Crystallographic Aspects of $HT_c$—Superconductivity," NATO ASI, Series E: Applied Sciences, Vol. 263, Kluwer Academic, Dordrecht.
Kittel, C. (1976). "Introduction to Solid State Physics," 5th ed., Wiley, New York.
Kulik, I. O., and Yanson, I. K. (1972). "Josephson Effect in Superconductive Tunneling Structures," Halsted, New York.
Kuper, C. G. (1968). "Introduction to the Theory of Superconductivity," Oxford University Press, London, New York.
London, F. (1950). "Superfluids," Vol. I, Wiley, New York.
Lynton, E. A. (1971). "Superconductivity," 3rd ed., Halsted, New York.
Markiewicz, R. S. (1997). *J. Phys. Chem. Solids* **58,** 1179.
McMillan and Rowell (1969). *In* "Superconductivity" (R. D. Parks, ed.), p. 561, Dekker, New York.
Manousakis, E. (1991). *Rev. Mod. Phys.* **63**(1), 1–62.
Mendelssohn, K. (1966). "Quest for Absolute Zero," McGraw–Hill, New York.
Newhouse, V. L. (ed.) (1975). "Applied Superconductivity," Academic Press, New York.
Rickayzen, G. (1965). "Theory of Superconductivity," Wiley (Interscience), New York.
Saint-James, D., Sarma, G., and Thomas, E. J. (1969). "Type II Superconductivity," Pergamon, Oxford.
Scalapino, D. J. (1995). *Phys. Rev.* **250,** 329.
Schrieffer, J. R. (1964). "Theory of Superconductivity," Benjamin, New York.
Solymar, L. (1972). "Superconductive Tunnelling and Applications," Halsted, New York.
Taylor, A. W. B. (1970). "Superconductivity," Wykeham, London, Winchester.
Tinkham, M. (1975). "Introduction to Superconductivity," McGraw–Hill, New York.
Wallace, P. R. (ed.) (1969). "Superconductivity," Gordon and Breach, New York.
Williams, J. E. C. (1970). "Superconductivity and Its Applications," Arrowsmith, Bristol, England.
Ziman, J. M. (1972). "Principles of the Theory of Solids," 2nd ed., Cambridge University Press, London, New York.



# Thermoelectricity

**Timothy P. Hogan**
*Michigan State University*

I. Introduction and Basic Thermoelectric Effects
II. Thermodynamic Relationships
III. Thermodynamics of an Irreversible Process
IV. Statistical Relationships
V. Applications
VI. Summary

## GLOSSARY

**Boltzmann equation** An equation based on the Fermi distribution equation under nonequilibrium conditions. The Boltzmann equation describes the rate of change of the distribution function due to forces, concentration gradients, and carrier scattering.

**Fermi distribution** A function describing the probability of occupancy of a given energy state for a system of particles based on the Pauli exclusion principle.

**Fermi level** The energy level which exhibits a 50% probability of being occupied.

**Joule heating** Heating due to $I^2R$ losses.

**Onsager relations** A set of simultaneous equations that describe the macroscopic interactions between "forces" and "flows" within a thermoelectric system.

**Peltier effect** Absorption or evolution of thermal energy at a junction between dissimilar materials through which current flows.

**Seebeck effect** Open-circuit voltage generated by a circuit consisting of at least two dissimilar conductors when a temperature gradient exists within the circuit between the measuring and the reference junctions.

**Thermocouple** A pair of dissimilar conductors joined at one set of ends to form a measuring junction.

**Thermoelectric cooler** A heat pump designed from thermoelectric materials typically configured in an array as a series of thermocouples with the junction exposed.

**Thermopower** This is defined here as the absolute Seebeck coefficient and corresponds to the rate of change of the thermoelectric voltage with respect to the temperature of a single conductor with a temperature gradient between the ends.

**Thompson effect** The absorption or evolution of thermal energy from a single homogeneous conductor through which electric current flows in the presence of a temperature gradient along the conductor.

**THE FIELD** of thermoelectricity involves the study of characteristics resulting from electrical phenomena occurring in conjunction with a flow of heat. It includes flows of electrical current and thermal current and the interactions between them.

## I. INTRODUCTION AND BASIC THERMOELECTRIC EFFECTS

In 1822, Seebeck reported on "the magnetic polarization of metals and ores produced by a temperature difference" (Joffe, 1957). By placing two conductors in the configuration shown in Fig. 1, Seebeck observed a deflection of the magnetic needle in his measurement apparatus (Gray, 1960). The deflection was dependent on the temperature difference between junctions and the materials used for the conductors. Shortly after this, Oersted discovered the interaction between an electric current and a magnetic needle. Many scientists subsequently researched the relationship between electric currents and magnetic fields including Ampère, Biot, Savart, Laplace, and others. It was then suggested that the observation by Seebeck was not caused by a magnetic polarization, but due to a thermoelectric current flowing in the closed-loop circuit.

Seebeck did not accept this explanation, and in an attempt to refute it, he reported measurements on a number of solid and liquid metals, alloys, minerals, and semiconductors. The magnetic polarization hypothesis was incorrect as can be seen in the open-circuit configuration of his experiment.

Experimentally, a voltage ($\Delta V$) at the open-circuit terminals is measured when a temperature gradient exists between junctions such that

$$\Delta V = \int_{T_1}^{T_2} S_{AB}\, dT, \tag{1}$$

where $S_{AB}$ is the Seebeck coefficient for the two conductors, which is defined as being positive when a positive voltage is measured for $T_1 < T_2$. The voltage is measured across terminals maintained at a constant temperature $T_0$. For this voltage to appear in the open-circuit configuration (Fig. 2), there must exist a current which flows in the closed-circuit configuration. Furthermore, in the open-circuit configuration, Seebeck would no longer observe a deflection of the magnetic needle, which is not expected if a magnetic polarization is taking effect.

The diligence of his measurements was vertified by the confirmation of his values years later by Justi and Meisner

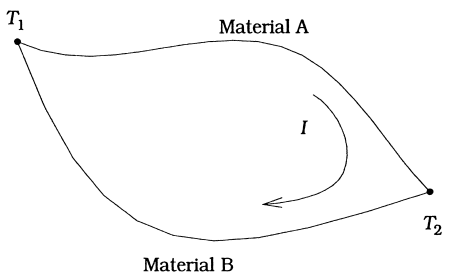



**FIGURE 1** Closed-circuit Seebeck effect.

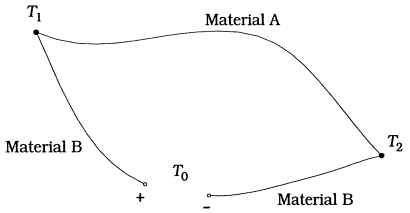



**FIGURE 2** The open-circuit Seebeck effect.

as well as by Telkes, who showed, 125 years after Seebeck's measurements, that the best couple for energy conversion was formed using ZnSb and PbS, which were two materials examined by Seebeck.

Twelve years after Seebeck's discovery, a scientist and watchmaker named Jean Peltier reported a temperature anomaly at the junction of two dissimilar materials as a current was passed through the junction. It was unclear what caused this anomaly, and while Peltier attempted to explain it on the basis of the conductivities and/or hardness of the two materials, Lenz removed all doubt in 1838 with one simple experiment. By placing a droplet of water in a dimple at the junction between rods of bismuth and antimony, Lenz was able to freeze the water and subsequently melt the ice by changing the direction of current through the junction. In a way, Lenz had made the first thermoelectric cooler. The rate of heat ($\dot{Q}$) absorbed or liberated from the junction was later found to be proportional to the current, or

$$\dot{Q} = \Pi \cdot I, \tag{2}$$

where the proportionality constant ($\Pi$) was named the Peltier coefficient.

Near this time, the field of electromagnetics was being formed and captured much attention in the scientific community. Therefore, another 16 years passed before Thomson (later called Lord Kelvin) reasoned that if the current through the two junctions in Fig. 1 produced only Peltier heating, then the Peltier voltage must equal the Seebeck voltage and both must be linearly proportional to the temperature. Since this was not observed experimentally, he reasoned that there must be a third reversible process occurring. This third process is the evolution or absorption of heat whenever current is passed through a single homogeneous conductor along which a temperature gradient exists, or in equation form,

$$\dot{Q} = \Im I \frac{dT}{dx}, \tag{3}$$

where $\dot{Q}$ is the rate of heat absorbed or liberated along the conductor, $\Im$ is the Thomson coefficient, $I$ is the current through the conductor and $dT/dx$ is the temperature gradient maintained along the length of the conductor.

Thomson then applied the first and second laws of thermodynamics to the Seebeck, Peltier, and Thomson effects to find the Kelvin relationships

$$\Pi = S_{AB}T, \tag{4}$$

$$\frac{dS_{AB}}{dT} = \frac{\Im_A - \Im_B}{T}, \tag{5}$$

where the subscripts $A$ and $B$ correspond to the two materials in Fig. 2. The second Kelvin relation suggests that the Seebeck coefficient for two materials forming a junction can be represented as the difference between quantities based on the properties of the individual materials making up the junction. Integration of the second Kelvin relation gives

$$\int dS_{AB} = \int \frac{\Im_A - \Im_B}{T}\,dT = \int \frac{\Im_A}{T}\,dT - \int \frac{\Im_B}{T}\,dT \tag{6}$$

or

$$S_{AB} = \int \frac{\Im_A}{T}\,dT - \int \frac{\Im_B}{T}\,dT. \tag{7}$$

Defining the first term on the right-hand side as the "absolute" Seebeck coefficient of material $A$ and the second term as the "absolute" Seebeck coefficient of material $B$, we find that the Seebeck coefficient for a junction is equal to the difference in "absolute" Seebeck coefficients of the individual materials making the junction. This is a very significant result, as measurements of the individual materials can be used to predict how junctions formed from various combinations of materials will behave, thus removing the need to measure every possible combination of materials. The "absolute" Seebeck coefficient or thermoelectric power of a material, hereafter referred to simply as the thermopower of the material, can be found for material $A$ if the thermopower of material $B$ is known or if the thermopower of material $B$ is zero. A material in the superconducting state has a thermopower of zero, and once a material is calibrated against a superconductor, it can then be used as a reference material to measure more materials. This has been done for several pure materials such as lead, gold, and silver (Roberts, 1977; Wendling *et al.*, 1993).

Further understanding of the basic thermoelectric properties and the relationships between them can be found through comparisons of macroscopic and microscopic derivations. The Onsager relations formulate various flows (consisting of matter or energy) as functions of the forces that drive them, thus describing macroscopic observations of materials. Another useful technique for understanding these basic thermoelectric properties utilizes semiclassical statistical mechanics to describe the microscopic processes. Comparisons between the macroscopic and the microscopic analyses can be used in deriving many useful formulas for calculating thermoelectric properties of various materials. The following sections are dedicated to developing the macroscopic and microscopic analyses.

## II. THERMODYNAMIC RELATIONSHIPS

As shown by the Seebeck effect, when a temperature gradient is placed over the length of a sample, carrier flow will be predominantly from the hot side to the cold side. This indicates that a temperature gradient, $\Delta T$, is a *force* that can cause a *flow* of carriers. It is well known that applying the *force* of an electric potential gradient, $\Delta V$, can also induce carrier *flow*. In 1931, Onsager developed a method of relating the flows of matter or energy within a system to the forces present. In this method the forces are assumed to be sufficiently small so that a linear relationship between the forces, $\mathbf{X}_i$, and the corresponding flows, $\mathbf{J}_i$, can be written.

$$\begin{aligned} \mathbf{J}_1 &= L_{11}\mathbf{X}_1 + L_{12}\mathbf{X}_2 + \cdots L_{1n}\mathbf{X}_n, \\ \mathbf{J}_2 &= L_{21}\mathbf{X}_1 + L_{22}\mathbf{X}_2 + \cdots L_{2n}\mathbf{X}_n, \\ \mathbf{J}_3 &= L_{31}\mathbf{X}_1 + L_{32}\mathbf{X}_2 + \cdots L_{3n}\mathbf{X}_n, \end{aligned} \tag{8}$$

or

$$\mathbf{J}_i = \sum_{m=1}^{n} L_{im}\mathbf{X}_m \qquad (i = 1, 2, 3, \ldots, n). \tag{9}$$

For carrier and heat flow as described above, the Onsager relationships can be written

$$\begin{aligned} \mathbf{J} &= L_{11}\nabla V + L_{12}\nabla T, \\ \mathbf{J}_Q &= L_{21}\nabla V + L_{22}\nabla T, \end{aligned} \tag{10}$$

where $\mathbf{J}$ is the current density (electric charge flow), and $\mathbf{J}_Q$ is the heat flux density (heat flow). Without a temperature gradient ($\Delta T = 0$), a heat flux of zero would be expected, contrary to what Eq. (10) would indicate. It is, therefore, important to understand further the primary coefficients $L_{ii}$ and interaction coefficients $L_{ij}$ $(i \neq j)$ linking these equations. To do so requires a consideration of the thermodynamics of an irreversible process (one in which the change in entropy is greater than zero $\Delta \bar{S} > 0$).

## III. THERMODYNAMICS OF AN IRREVERSIBLE PROCESS

The general application of the Onsager relationship was derived by Harman and Honig (1967) and is summarized here. For a constant electric potential, $V$, throughout the sample,

$$dQ = T\,d\bar{S} = dU + P\,d\mathcal{V} - \sum_i \mu_i\,dn_i, \qquad (11)$$

where $Q$ is the heat energy density, $\bar{S}$ is the entropy density, $U$ is the internal energy density, $P$ is the pressure, $\mu_i$ is the chemical potential of the particle species, and $n_i$ is the particle density. The magnitude of the differential volume, $d\mathcal{V}$, is zero since each quantity has been specified per unit volume. This can be combined with the total energy density, $E$, given by

$$E = U + V\sum_i Z_i q n_i, \qquad (12)$$

where $q$ is the magnitude of the electronic charge $[1.602 \times 10^{-19}(\mathrm{C})]$, $V$ is an externally applied bias, and $Z_i$ is the number and sign of the charges on the $i$th particle species. For example, an electron would have charge $Z_e q$, where $Z_e = -1$. The time derivative of Eq. (12) gives

$$\frac{\partial E}{\partial t} = \frac{\partial U}{\partial t} + \frac{\partial V}{\partial t}\sum_i Z_i q n_i + V\sum_i Z_i q\frac{\partial n_i}{\partial t}. \qquad (13)$$

From Eq. (11) with $d\mathcal{V} = 0$,

$$dU = T\,d\bar{S} + \sum_i \mu_i\,dn_i. \qquad (14)$$

Taking the time derivative of (14) gives

$$\frac{\partial U}{\partial t} = T\frac{\partial \bar{S}}{\partial t} + \bar{S}\frac{\partial T}{\partial t} + \sum_i \mu_i\frac{\partial n_i}{\partial t} + \sum_i n_i\frac{\partial \mu_i}{\partial t}. \qquad (15)$$

Using (15) in (13) yields

$$\frac{\partial E}{\partial t} = T\frac{\partial \bar{S}}{\partial t} + \bar{S}\frac{\partial T}{\partial t} + \sum_i \mu_i\frac{\partial n_i}{\partial t} + \sum_i n_i\frac{\partial \mu_i}{\partial t} + \frac{\partial V}{\partial t}\sum_i Z_i q n_i + V\sum_i Z_i q\frac{\partial n_i}{\partial t}. \qquad (16)$$

This can be simplified by considering the Gibbs–Duhem relation (Guggenheim, 1957),

$$\bar{S}\frac{\partial T}{\partial t} + \sum_i n_i\frac{\partial \mu_i}{\partial t} = 0 \qquad (17)$$

and

$$\bar{\mu}_i = \mu_i + Z_i q V, \qquad (18)$$

where $\bar{\mu}_i$ is the electrochemical potential, $\mu_i$ is the chemical potential, and $Z_i q V$ is the electrostatic potential energy. The relationship among the chemical potential, $\mu$, the electrochemical potential, $\bar{\mu}$, and the temperature for electrons is shown in Fig. 3, where the right side of the sample is at a potential of $-V_1$ relative to the left.

Equation (16) then reduces to

$$\frac{\partial E}{\partial t} = T\frac{\partial \bar{S}}{\partial t} + \sum_i \bar{\mu}_i\frac{\partial n_i}{\partial t} + \frac{\partial V}{\partial t}\sum_i Z_i q n_i. \qquad (19)$$

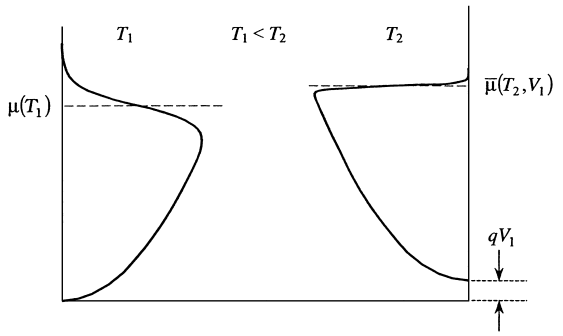



**FIGURE 3** The density of states for a metal at a temperature $T_1 > 0$ K on the left and at a lower temperature, $T_2 < T_1$, on the right.

The rate of change in the particle density $n_i$, is governed by the *equation of continuity*,

$$\frac{\partial n_i}{\partial t} = \left(\frac{\partial n_i}{\partial t}\right)_s - \nabla\cdot\mathbf{J}_i, \qquad (20)$$

which states that the total rate of change in $n_i$ is equal to the local particle generation rate, or source rate, minus the transport of the $i$th species across the boundary of the differential volume (or local system) of interest. The first term on the right-hand side of the equation is the source term and represents the particle generation (or capture) rate through chemical reactions, for example. The last term is found using Gauss's theorem,

$$\iint \mathbf{J}_i\cdot\hat{\mathbf{n}}\,dA = \iiint \nabla\cdot\mathbf{J}_i\,d\mathcal{V}, \qquad (21)$$

where $\mathbf{J}_i$ is the flux vector equal to the number of particles of type $i$ moving past a unit cross section per unit time in the direction of $\mathbf{J}_i$, and $\hat{\mathbf{n}}$ represents a unit vector outward normal from an element of area $dA$ on the boundary surface. This represents the total outward flux of the $i$th particle species from the differential volume of interest. The particle species over which the summation in Eq. (19) is evaluated includes core species, $L$, which form the host lattice; neutral donors, $D$; ionized donors, $D^+$; neutral acceptors, $A$; ionized acceptors, $A^-$; electrons in the conduction band, $n$; and holes in the valence band, $p$. Therefore,

$$\frac{1}{T}\sum_i \bar{\mu}_i\frac{\partial n_i}{\partial t} = \frac{1}{T}\left(\bar{\mu}_L\frac{\partial n_L}{\partial t} + \bar{\mu}_D\frac{\partial n_D}{\partial t} + \bar{\mu}_{D^+}\frac{\partial n_{D^+}}{\partial t} + \bar{\mu}_A\frac{\partial n_A}{\partial t} + \bar{\mu}_{A^-}\frac{\partial n_{A^-}}{\partial t} + \bar{\mu}_n\frac{\partial n_n}{\partial t} + \bar{\mu}_p\frac{\partial n_p}{\partial t}\right). \qquad (22)$$

Equation (20) can now be used for each term on the right-hand side of (22). Some simplification can be readily

made, however, when the species $L$, $D$, $D^+$, $A$, and $A^-$ are assumed to be immobile such that

$$\mathbf{J}_L = \mathbf{J}_D = \mathbf{J}_{D^+} = \mathbf{J}_A = \mathbf{J}_{A^-} = 0. \tag{23}$$

Furthermore, the lattice will not be affected by local transformations, and

$$\left(\frac{\partial n_L}{\partial t}\right)_s = \frac{\partial n_L}{\partial t} = 0. \tag{24}$$

Additional relationships can be found to simplify (22) further by identifying the different mechanisms for generation of electrons, $n$, or holes, $p$, as follows:

$$D \Rightarrow D^+ + n, \tag{25}$$

$$A \Rightarrow A^- + p, \tag{26}$$

$$\Rightarrow n + p. \tag{27}$$

These reactions are reversible such that the time rate of change of ionized and unionized donors and acceptors must be considered in (22). Identifying the reactions in (25), (26), and (27) as I, II, and III, respectively, the following reaction velocities can be written

$$\left.\begin{aligned} -\left(\frac{\partial n_D}{\partial t}\right)_s &= \left(\frac{\partial n_{D^+}}{\partial t}\right)_s = \left(\frac{\partial n_n}{\partial t}\right)_{\mathrm{I}} = \nu_{\mathrm{I}}, \\ -\left(\frac{\partial n_A}{\partial t}\right)_s &= \left(\frac{\partial n_{A^-}}{\partial t}\right)_s = \left(\frac{\partial n_p}{\partial t}\right)_{\mathrm{II}} = \nu_{\mathrm{II}}, \\ \left(\frac{\partial n_n}{\partial t}\right)_s &= \left(\frac{\partial n_p}{\partial t}\right)_{\mathrm{III}} = \nu_{\mathrm{III}}. \end{aligned}\right\} \tag{28}$$

Therefore, (22) becomes

$$\begin{aligned} \frac{1}{T}\sum_i \bar{\mu}_i \frac{\partial n_i}{\partial t} = \frac{1}{T}\Bigg\{ & \bar{\mu}_D\left(\frac{\partial n_D}{\partial t}\right)_s + \bar{\mu}_{D^+}\left(\frac{\partial n_{D^+}}{\partial t}\right)_s \\ & + \bar{\mu}_A\left(\frac{\partial n_A}{\partial t}\right)_s + \bar{\mu}_{A^-}\left(\frac{\partial n_{A^-}}{\partial t}\right)_s \\ & + \bar{\mu}_n\left[\left(\frac{\partial n_n}{\partial t}\right)_{\mathrm{I}} + \left(\frac{\partial n_n}{\partial t}\right)_{\mathrm{III}} - \nabla\cdot\mathbf{J}_n\right] \\ & + \bar{\mu}_p\left[\left(\frac{\partial n_p}{\partial t}\right)_{\mathrm{II}} + \left(\frac{\partial n_p}{\partial t}\right)_{\mathrm{III}} - \nabla\cdot\mathbf{J}_p\right]\Bigg\}. \end{aligned} \tag{29}$$

From the relations in (28), Eq. (29) can be written in terms of the reaction velocities, $\nu_{\mathrm{I}}$, $\nu_{\mathrm{II}}$, and $\nu_{\mathrm{III}}$ as follows:

$$\begin{aligned} \frac{1}{T}\sum_i \bar{\mu}_i \frac{\partial n_i}{\partial t} = \frac{1}{T}\{ & (-\bar{\mu}_D + \bar{\mu}_{D^+} + \bar{\mu}_n)\nu_{\mathrm{I}} + (-\bar{\mu}_A \\ & + \bar{\mu}_{A^-} + \bar{\mu}_p)\nu_{\mathrm{II}} + (\bar{\mu}_n + \bar{\mu}_p)\nu_{\mathrm{III}} \\ & - \bar{\mu}_n\nabla\cdot\mathbf{J}_n - \bar{\mu}_p\nabla\cdot\mathbf{J}_n\} \end{aligned} \tag{30}$$

This can be substituted into (19) to give

$$\begin{aligned} \frac{\partial E}{\partial t} = T\frac{\partial \bar{S}}{\partial t} + \frac{1}{T}\{ & A_{\mathrm{I}}\nu_{\mathrm{I}} + A_{\mathrm{II}}\nu_{\mathrm{II}} + A_{\mathrm{III}}\nu_{\mathrm{III}} - \bar{\mu}_n\nabla\cdot\mathbf{J}_n \\ & - \bar{\mu}_p\nabla\cdot\mathbf{J}_n\} + \frac{\partial V}{\partial t}\sum_i Z_i q n_i, \end{aligned} \tag{31}$$

where $A_{\mathrm{I}}$, $A_{\mathrm{II}}$, and $A_{\mathrm{III}}$ *affinities* are defined as

$$\begin{aligned} A_{\mathrm{I}} &\equiv -\bar{\mu}_D + \bar{\mu}_{D^+} + \bar{\mu}_n, \qquad A_{\mathrm{II}} \equiv -\bar{\mu}_A + \overline{\mu}_{A^-} + \bar{\mu}_p, \\ A_{\mathrm{III}} &\equiv \bar{\mu}_n + \bar{\mu}_p. \end{aligned} \tag{32}$$

These general derivations can now be applied to more specific cases by solving for the energy flux term on the left-hand side of the equation using the appropriate approximations for the material under consideration.

## A. Metals

In metals, the energy density term $\partial E/\partial t$ can be viewed as composed of four contributions.

- The rate at which an externally applied field delivers energy to the local system.
- Two terms arise from the rate of change in the electrostatic energy either due to a change in the charge concentration or due to a change in the potential, $V$.
- Electrons in the higher-energy states [the energies above $\mu(T_1)$ in Fig. 3] can transition to the available lower-energy states by giving up this excess energy to the lattice, resulting in a heat flux, $\mathbf{J}_Q$.

As an externally applied electric field accelerates charged carriers, they do not continue to increase in velocity as they would in free space, but attain some average drift velocity. Therefore, an internal force must exist to counterbalance the external force. This internal force is caused mainly by collisions of the carriers with the lattice, thus providing a mechanism of energy transfer from the applied electric field to the lattice. The first contribution is given by

$$\begin{aligned} \left(\frac{\partial E}{\partial t}\right)_{\mathrm{I}} &= \mathcal{E}\cdot(-n_n q\mathbf{v}_n + n_p q\mathbf{v}_p) = \mathbf{J}\cdot\mathcal{E} = -\mathbf{J}\cdot\nabla V \\ &= q(\mathbf{J}_n - \mathbf{J}_p)\cdot\nabla V, \end{aligned} \tag{33}$$

where $\mathbf{J}_n$ and $\mathbf{J}_p$ represent the particle flux densities, while $\mathbf{J}$ represents the current density such that

$$\mathbf{J} = q(\mathbf{J}_p - \mathbf{J}_n). \tag{34}$$

The electrostatic energy density is given by $V\sum_i Z_i q n_i$. The time rate of change of the electrostatic energy density is

$$\frac{\partial}{\partial t}\left(V\sum_i Z_i q n_i\right) = \frac{\partial V}{\partial t}\sum_i Z_i q n_i + V\sum_i Z_i q\frac{\partial n_i}{\partial t}. \tag{35}$$

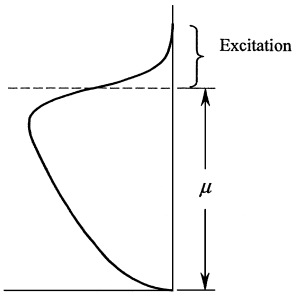



**FIGURE 4** The density of states at a finite temperature. Only the excitation energy can be transferred to the lattice.

This gives the second and third contributions to the total rate of change in energy density,

$$\left(\frac{\partial E}{\partial t}\right)_{\text{II}} = \frac{\partial V}{\partial t}\sum_i Z_i q n_i, \tag{36}$$

$$\left(\frac{\partial E}{\partial t}\right)_{\text{III}} = V\sum_i Z_i q\frac{\partial n_i}{\partial t} = -V(\nabla\cdot\mathbf{J}) = qV[\nabla\cdot(\mathbf{J}_n-\mathbf{J}_p)], \tag{37}$$

where Eq. (20) was used, with the assumption of no generative sources. The fourth contribution comes from the excitation energy depicted in Fig. 4, which gives rise to a heat flux, $\mathbf{J}_Q$. Relative to the bottom of the conduction band, the total heat flux density, $\mathbf{J}_u$, is

$$\mathbf{J}_u = \mathbf{J}_Q - \frac{\mu}{q}\mathbf{J}, \tag{38}$$

thus giving the fourth contribution to energy flow through Fourier's law of heat conduction,

$$\left(\frac{\partial E}{\partial t}\right)_{\text{IV}} = -\nabla\cdot\mathbf{J}_u = -\nabla\cdot\left(\mathbf{J}_Q - \frac{\mu}{q}\mathbf{J}\right). \tag{39}$$

Summing contributions I through IV gives the total energy density rate of change as

$$\frac{\partial E}{\partial t} = -\mathbf{J}\cdot\nabla V - V(\nabla\cdot\mathbf{J}) + \frac{\partial V}{\partial t}\sum_i Z_i q n_i - \nabla\cdot\mathbf{J}_u = -\nabla\cdot V\mathbf{J} + \frac{\partial V}{\partial t}\sum_i Z_i q n_i - \nabla\cdot\mathbf{J}_u = \frac{\partial V}{\partial t}\sum_i Z_i q n_i - \nabla\cdot\mathbf{J}_E, \tag{40}$$

where

$$\mathbf{J}_E = \mathbf{J}_u + V\mathbf{J} \tag{41}$$

is the total energy flux density. Substituting (40) into (19) gives

$$\frac{\partial V}{\partial t}\sum_i Z_i q n_i - \nabla\cdot\mathbf{J}_E = T\frac{\partial\bar{S}}{\partial t} + \sum_i \bar{\mu}_i\frac{\partial n_i}{\partial t} + \frac{\partial V}{\partial t}\sum_i Z_i q n_i, \tag{42}$$

or, after cancellation and using (20), assuming no generative sources,

$$-\nabla\cdot\mathbf{J}_E = T\frac{\partial\bar{S}}{\partial t} + \frac{\bar{\mu}}{q}\nabla\cdot\mathbf{J} = T\frac{\partial\bar{S}}{\partial t} + \bar{\mu}\nabla\cdot\mathbf{J}_q. \tag{43}$$

The rate of change of entropy can, therefore, be written

$$\frac{\partial\bar{S}}{\partial t} = \frac{-\nabla\cdot\mathbf{J}_E}{T} - \frac{\bar{\mu}}{qT}\nabla\cdot\mathbf{J} = -\nabla\cdot\left(\frac{\mathbf{J}_E}{T}\right) - \nabla\cdot\left(\frac{\mathbf{J}\bar{\mu}}{qT}\right) + \mathbf{J}_E\cdot\nabla\left(\frac{1}{T}\right) + \mathbf{J}\cdot\nabla\left(\frac{\bar{\mu}}{qT}\right), \tag{44}$$

or using an entropy flux, $\mathbf{J}_{\bar{s}}$, defined as

$$T\mathbf{J}_{\bar{s}} = \mathbf{J}_E + \frac{\bar{\mu}}{q}\mathbf{J}, \tag{45}$$

gives

$$\frac{\partial\bar{S}}{\partial t} = \frac{\partial\bar{S}_0}{\partial t} + \frac{\partial\bar{S}_s}{\partial t} = -\nabla\cdot\mathbf{J}_{\bar{s}} + \mathbf{J}_E\cdot\nabla\left(\frac{1}{T}\right) + \mathbf{J}\cdot\nabla\left(\frac{\bar{\mu}}{qT}\right), \tag{46}$$

where the total entropy is given by the sum of the equilibrium entropy plus additional entropy sources, or $\bar{S} = \bar{S}_0 + \bar{S}_s$. The irreversible process for which $\Delta\bar{S} = (\bar{S} - \bar{S}_0) = \bar{S}_s > 0$ then consists of the last two terms in the above equation such that

$$\frac{\partial\bar{S}_s}{\partial t} = \mathbf{J}_E\cdot\nabla\left(\frac{1}{T}\right) + \mathbf{J}\cdot\nabla\left(\frac{\bar{\mu}}{qT}\right). \tag{47}$$

Using Eq. (45) to substitute for $\mathbf{J}_E$, (47) becomes.

$$\frac{\partial\bar{S}_s}{\partial t} = \frac{-\mathbf{J}_s}{T}\cdot\nabla T + \frac{\mathbf{J}}{qT}\cdot\nabla\bar{\mu}, \tag{48}$$

or using $\bar{\mu} = \mu + qV$ along with (41) and (45) to give

$$T\mathbf{J}_{\bar{s}} = \mathbf{J}_Q, \tag{49}$$

then Eq. (48) becomes

$$\frac{\partial\bar{S}_s}{\partial t} = \frac{-\mathbf{J}_Q}{T^2}\cdot\nabla T + \frac{\mathbf{J}}{qT}\cdot\nabla\bar{\mu}. \tag{50}$$

These three equations, (47), (48), and (50), could each be written in the general form of

$$\frac{\partial\bar{S}_s}{\partial t} = \sum_i \mathbf{J}_i\cdot\mathbf{X}_i. \tag{51}$$

This is a necessary condition for using the Onsager reciprocity relation that $L_{12} = L_{21}$ in Eq. (10). Three sets of

Onsager relations can then be written, by extracting the forces, $\mathbf{X}_i$, from Eqs. (47), (48), and (50).

$$\left.\begin{aligned}\mathbf{J} &= \frac{\mathbf{Z}_{11}}{q}\nabla\left(\frac{\bar{\mu}}{T}\right) + \mathbf{Z}_{12}\nabla\left(\frac{1}{T}\right),\\ \mathbf{J}_E &= \frac{\mathbf{Z}_{21}}{q}\nabla\left(\frac{\bar{\mu}}{T}\right) + \mathbf{Z}_{22}\nabla\left(\frac{1}{T}\right),\end{aligned}\right\} \tag{52}$$

$$\left.\begin{aligned}\mathbf{J} &= \frac{\mathbf{B}_{11}}{qT}\nabla\bar{\mu} - \frac{\mathbf{B}_{12}}{T}\nabla T,\\ \mathbf{J}_{\bar{s}} &= \frac{\mathbf{B}_{21}}{qT}\nabla\bar{\mu} - \frac{\mathbf{B}_{22}}{T}\nabla T,\end{aligned}\right\} \tag{53}$$

$$\left.\begin{aligned}\mathbf{J} &= \frac{\mathbf{L}_{11}}{qT}\nabla\bar{\mu} - \frac{\mathbf{L}_{12}}{T^2}\nabla T,\\ \mathbf{J}_Q &= \frac{\mathbf{L}_{21}}{qT}\nabla\bar{\mu} - \frac{\mathbf{L}_{22}}{T^2}\nabla T,\end{aligned}\right\} \tag{54}$$

thus relating the electrical current density, $\mathbf{J}$, to the energy flux density, $\mathbf{J}_E$, the entropy flux density, $\mathbf{J}_{\bar{s}}$, and the heat flux density, $\mathbf{J}_Q$. Equations (52), (53), and (54) can now be used to identify various thermoelectric properties.

## IV. STATISTICAL RELATIONSHIPS

Within crystalline materials, electron behavior can be described by the wave nature of electrons and Schrödinger's equation,

$$\nabla^2\Psi + \frac{2m}{\hbar^2}(E - V)\Psi = \frac{-\hbar}{j}\frac{\partial E}{\partial t}, \tag{55}$$

where $\Psi$ is the electron wave function, $E$ is the total energy, and $V$ is the potential energy of the electrons. The solution to this equation is

$$\Psi(\mathbf{r}, t) = \psi(\mathbf{r})e^{-j\omega t}, \tag{56}$$

where $\Psi$ is the time-independent solution to Schrödinger's equation. This solution forms a wave packet with a group velocity, $\mathbf{v}$, equal to the average velocity of the particle it describes, such that

$$\mathbf{v} = \nabla_k\omega = \frac{\partial\omega}{\partial\mathbf{k}} = \frac{1}{\hbar}\nabla_k E = \frac{1}{\hbar}\frac{\partial E}{\partial\mathbf{k}}, \tag{57}$$

where the use of Planck's relationship, $E = h\nu = \hbar\omega$, was made. Force times distance is equal to energy, or with a time derivative,

$$\mathbf{v}\cdot\mathbf{F} = \frac{\partial E}{\partial t} = \frac{1}{\hbar}\frac{\partial E}{\partial\mathbf{k}}\cdot\hbar\frac{\partial\mathbf{k}}{\partial t}, \tag{58}$$

giving

$$\mathbf{F} = \hbar\frac{\partial\mathbf{k}}{\partial t}. \tag{59}$$

The electron wave function, $\Psi(\mathbf{r}, t)$, itself does not have physical meaning, however, the product of $\Psi^*(\mathbf{r}, t)\Psi(\mathbf{r}, t)$ represents the probability of finding an electron at position $\mathbf{r}$ and time $t$. As a probability implies, there is a factor of uncertainty, which was quantified in 1927 by Heisenberg.

$$\Delta p_x\Delta x \geq h, \qquad \Delta p_y\Delta y \geq h, \qquad \Delta p_z\Delta z \geq h, \tag{60}$$

where $\Delta p_x$, $\Delta p_y$, and $\Delta p_z$ are the momentum uncertainties in the $x$, $y$, and $z$ directions, respectively. The positional uncertainties in the three directions are given by $\Delta x$, $\Delta y$, and $\Delta z$. It is possible to utilize these uncertainties to define the smallest volume (in real space, or momentum space) that represents a discrete electronic state. Within a cube of material with dimensions $L \times L \times L$, the maximum positional uncertainty for a given electron would be $\Delta x = \Delta y = \Delta z = L$, since the electron must be located somewhere within the cube. This would correspond to the minimum $\Delta p_x$, $\Delta p_y$, and $\Delta p_z$ given by

$$\begin{aligned}\Delta p_{x_{\min}} &= \frac{h}{\Delta x} = \frac{h}{L}, \qquad \Delta p_{y_{\min}} = \frac{h}{\Delta y} = \frac{h}{L},\\ \Delta p_{z_{\min}} &= \frac{h}{\Delta z} = \frac{h}{L}.\end{aligned} \tag{61}$$

Thus the product

$$\Delta p_{x_{\min}}\Delta p_{y_{\min}}\Delta p_{z_{\min}} = \frac{h^3}{L^3} \tag{62}$$

gives the minimum elemental volume in momentum space to represent two discrete electronic states (one for spin-up and one for spin-down). The number of states, $dg$, per unit volume in an element $dp_x\,dp_y\,dp_z$ of momentum space can be written

$$dg = \frac{1}{h^3}\,dp_x\,dp_y\,dp_z. \tag{63}$$

Schrödinger's time-independent equation for a free electron ($V = 0$) is

$$\nabla^2\psi + \frac{2m}{\hbar^2}E\psi = 0, \tag{64}$$

which has the solution

$$\psi = Ae^{j\mathbf{k}\cdot\mathbf{r}}. \tag{65}$$

Substituting back into (64) gives

$$E = \frac{\hbar^2}{2m}k^2 = \frac{p^2}{2m}, \tag{66}$$

where $p^2 = \hbar k$ was used. Within a crystal, a similar formula can be found when the concept of effective mass, $m^*$, is utilized to account for internal forces on the electrons due to the ion cores at each lattice point. Electrons with energies below some value $E$ are then defined by a sphere in momentum space with radius $p = \sqrt{2m^*E}$. The number

of electronic states within the material cube ($L \times L \times L$) is found by dividing the total momentum space volume by the volume per state, or

$$N = 2\left[\frac{4/3\pi r^3}{h^3/L^3}\right] = 2\left[\frac{4/3\pi(\sqrt{2m^*E})^3}{h^3/L^3}\right] = \frac{8\pi(2m^*E)^{3/2}}{3h^3}L^3. \quad (67)$$

The factor of 2 is included to account for electrons of both spin-up and spin-down.

The density of states is defined as the number of states per unit energy per unit volume, or

$$g(E) = \frac{(dN/dE)}{L^3} = \frac{4\pi(2m^*)^{3/2}}{h^3}E^{1/2} = \frac{1}{2\pi^2}\left(\frac{2m^*}{\hbar^2}\right)^{3/2}E^{1/2}, \quad (68)$$

where $\hbar = h/2\pi$ is Plank's reduced constant. This equation describes the number of available states for electrons to go into, but it does not describe the way the electrons fill those available states.

## A. The Fermi Distribution

To determine the number of electrons in a given band, it is necessary to find the probability of a given state being occupied by an electron and then integrate over all available states. A more realistic result for metals, which does not assume spherical constant energy surfaces in **k**-space, thus allowing for the electron energy to deviate from $E = \hbar^2k^2/2m^*$ would be found using the density of states from (63).

Within a crystalline material, charge carriers are known to follow the Pauli exclusion principle, which states that only one electrons can occupy a given energy state. The probability that an electron occupies an energy state can be found by considering a simple statistical exercise. If a system is defined to have three allowed energy levels ($E_1$, $E_2$, and $E_3$), two electrons, and a total energy of 4 eV as shown in Fig. 5, with the three energy levels defined as $E_1 = 1$ eV, $E_2 = 2$ eV, and $E_3 = 3$ eV, it would be expected that 80% of the time, a distribution of one electron in energy level $E_1$, zero electrons in $E_2$, and one electron in $E_3$, or a distribution of (1, 0, 1), would occur. The entropy of a system is related to the most probable arrangement, $W_m$, of the particles through Boltzmann's definition,

$$\bar{S} = k \ln W_m. \quad (69)$$

When only electrons are considered, the entropy is related to the internal energy of the system, $U$, the total number of electrons, $N$, and the volume, $\mathcal{V}$, of the system through Euler's equation

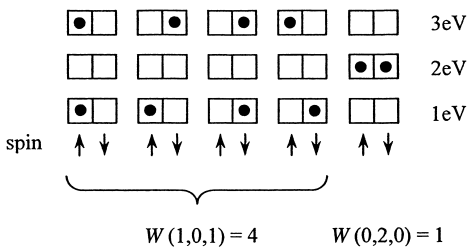



**FIGURE 5** The number of ways, *W*, two electrons can be distributed in three energy levels to obtain a total energy of 4 eV.

$$U = T\bar{S} - P\mathcal{V} + \mu N - qVN, \quad (70)$$

where $V$ represents the internal electrostatic potential. For a simple system with just two available energy states (energy = 0 or energy = $E$), the probability of finding the system with energy $E$ to that of finding it with energy 0 is

$$\frac{\mathcal{P}(E)}{\mathcal{P}(0)} = \frac{W(U_0 - E)}{W(U_0)} = \frac{e^{\bar{S}(U_0-E)/k}}{e^{\bar{S}(U_0)/k}}. \quad (71)$$

Using the approximation $\bar{S}(U_0 - E) \approx \bar{S}(U_0) - E(\frac{\partial \bar{S}}{\partial U_0})$ and

$$\frac{\partial \bar{S}}{\partial U} = \frac{1}{T} = \frac{\partial}{\partial U}\left(\frac{U + P\mathcal{V} - \mu N + qVN}{T}\right) \quad (72)$$

simplifies (71) to

$$\frac{\mathcal{P}(E)}{\mathcal{P}(0)} = \frac{e^{(\bar{S}(U_0)/k)-(E/kT)}}{e^{\bar{S}(U_0)/k}} = e^{-E/kT}. \quad (73)$$

To determine the probability of a system in energy state $E$, and that the state is occupied by an electron, then the influence of the total number of electrons, $N$, must also be taken into consideration. Then the ratio of the probability that the system is occupied by one electron at energy $E$ to the probability that the system is unoccupied with energy 0 is

$$\frac{\mathcal{P}(1, E)}{\mathcal{P}(0, 0)} = \frac{W[(U_0 - E), (N_0 - 1)]}{W[U_0, N_0]} = \frac{e^{\bar{S}[(U_0-E),(N_0-1)]/k}}{e^{\bar{S}[U_0,N_0]/k}}. \quad (74)$$

Using $\bar{S}[(U_0 - E), (N_0 - 1)] \approx \bar{S}[U_0, N_0] - E(\partial\bar{S}/\partial U_0) - (\partial\bar{S}/\partial N_0)$ yields

$$\frac{\mathcal{P}(1, E)}{\mathcal{P}(0, 0)} = \frac{e^{(\bar{S}[U_0,N_0]/k)-(E/kT)+((\mu-q\xi)/kT)}}{e^{\bar{S}[U_0,N_0]/k}} = e^{(E_F-E)/kT}, \quad (75)$$

where the Fermi level is defined as $E_F = \mu - qV$. Since $\mathcal{P}(1, E) + \mathcal{P}(0, 0) = 1$,

$$\mathcal{P}(1, E) = f(E) = \frac{1}{1 + e^{((E_F-E)/kT)}}. \quad (76)$$

This is the Fermi–Dirac distribution and represents the probability of occupancy of an energy state in equilibrium.

## B. Carrier Concentrations

In a free-electron approximation, the total number of electrons in a given energy band can then be found by integrating the product of the density of states and the probability of occupancy of the state for spherical energy surfaces:

$$n = \int_{E_{\text{bottom}}}^{E_{\text{top}}} g(E) f(E)\, dE = \frac{(2mkT)^{3/2}}{2\pi^2\hbar^3} F_{1/2}(\eta), \quad (77)$$

where $\eta = E_F/kT$, and $F_\nu(\eta)$ is the Fermi–Dirac function,

$$F_\nu(\eta) = \int_0^\infty \frac{\eta^\nu}{1 + e^{x-\eta}}\, dx. \quad (78)$$

Here, the bottom of the energy band was taken to be zero energy corresponding to $\mathbf{k} = 0$, and the integration was allowed to extend to $\infty$ since the Fermi–Dirac distribution falls to zero at high energy levels. In the degenerate limit, when $E_F \gg kT$, a series expansion of (78) leads to the following approximations for metals:

$$E_F \approx E_{F_0}\left[1 - \frac{\pi^2}{12}\left(\frac{kT}{E_{F_0}}\right)^2 + \cdots\right],$$
$$n \approx \frac{(2m)^{2/3}}{3\pi^2\hbar^3}\left(\frac{E_F}{kT}\right)^{3/2}\left[1 + \frac{\pi^2}{8}\left(\frac{kT}{E_F}\right)^2 + \cdots\right]. \quad (79)$$

At $T = 0$ K,

$$E_F = E_{F_0} = \frac{\pi^2\hbar^2}{2m}\left(\frac{3n}{\pi}\right)^{2/3}. \quad (80)$$

For nonspherical energy surfaces, the number of electrons per unit volume within an element of momentum space, $dp_x\, dp_y\, dp_z$, is found using (63) and the relation $\mathbf{p} = \hbar\mathbf{k} = \frac{h}{2\pi}\mathbf{k}$,

$$dn = \frac{2}{h^3} f(\mathbf{p}, \mathbf{r})\, dp_x\, dp_y\, dp_z = \frac{1}{4\pi^3} f(\mathbf{k}, \mathbf{r})\, dk_x\, dk_y\, dk_z, \quad (81)$$

where the factor of 2 accounts for two electrons of opposite spin. The total electron density is then found by integration.

## C. The Boltzmann Function

If the material is disturbed from equilibrium, then the distribution will vary, in general, as a function of wavevector, $\mathbf{k}$, position, $\mathbf{r}$, and time, $t$, or $f(\mathbf{k}, \mathbf{r}, t)$. At a time $t + dt$, the probability that a state with wavevector $\mathbf{k} + d\mathbf{k}$ is occupied by an electron at position $\mathbf{r} + d\mathbf{r}$ can be found, using Eq. (59), to be

$$f(\mathbf{k} + d\mathbf{k}, \mathbf{r} + d\mathbf{r}, t + dt) = f\left(\mathbf{k} + \frac{1}{\hbar}\mathbf{F}_t \cdot \nabla_k\, dt, \mathbf{r} + \mathbf{v}\, dt, t + dt\right). \quad (82)$$

The total rate of change of the distribution function near $\mathbf{r}$ is then

$$\frac{df}{dt} = \frac{1}{\hbar}\mathbf{F}_t \cdot \nabla_k f + \mathbf{v} \cdot \nabla_r f + \frac{\partial f}{\partial t}, \quad (83)$$

which is Boltzmann's transport equation. The first term on the right-hand side of this equation accounts for contributions from forces, $\mathbf{F}_t$, including externally applied forces, $\mathbf{F}$, and collision forces, $\mathbf{F}_c$. The middle term adds the contributions from concentration gradients, and the last term is the local changes in the distribution function about the point $\mathbf{r}$. Equation (83) is equal to zero since the total number of states in the crystal is constant, thus

$$\begin{aligned}\frac{\partial f}{\partial t} &= \frac{-1}{\hbar}\mathbf{F}_t \cdot \nabla_k f - \nu \cdot \nabla_r f \\ &= \frac{-1}{\hbar}\mathbf{F}_c \cdot \nabla_k f - \frac{1}{\hbar}\mathbf{F} \cdot \nabla_k f - \nu \cdot \nabla_r f \\ &= \left(\frac{\partial f}{\partial t}\right)_c - \frac{1}{\hbar}\mathbf{F} \cdot \nabla_k f - \nu \cdot \nabla_r f. \end{aligned} \quad (84)$$

With external forces applied, the distribution function, $f$, will be disturbed from the equilibrium value, $f_0$. Upon the removal of those external forces, equilibrium will be reestablished through collisions, $(\partial f/\partial t)_c$. Calculation of this collision term is a formidable task dependent largely on the scattering mechanisms for the material investigated. For small disturbances, however, a relaxation-time approximation is often used which assumes that

$$\left(\frac{\partial f}{\partial t}\right)_c = \frac{-(f - f_0)}{\tau_k} = \frac{-f_1}{\tau_k}, \quad (85)$$

where $\tau_k$ is the momentum relaxation time. In steady state, $\partial f/\partial t = 0$ and Eq. (84) becomes

$$\left.\begin{aligned} 0 &= \frac{-f_1}{\tau_k} - \frac{1}{\hbar}\mathbf{F} \cdot \nabla_k f - \mathbf{v} \cdot \nabla_r f \\ \text{or} \qquad f_1 &= -\frac{\tau_k}{\hbar}\mathbf{F} \cdot \nabla_k f - \tau_k \mathbf{v} \cdot \nabla_r f. \end{aligned}\right\} \quad (86)$$

The electric and heat current densities are given by

$$\mathbf{J} = -q\mathbf{v}n = -q\int \mathbf{v}\, dn = -\frac{q}{4\pi^3}\int f_1(\mathbf{k})\, d\mathbf{k},$$
$$\mathbf{J}_Q = \frac{1}{4\pi^3}\int \mathbf{v}(E - E_F) f_1(\mathbf{k})\, d\mathbf{k}. \quad (87)$$

Substituting Eq. (86) into Eqs. (87) starting with the electric current density, $\mathbf{J}$, gives

$$\mathbf{J} = -\frac{q}{4\pi^3}\int\left(\frac{-\tau_k}{\hbar}\mathbf{v}\mathbf{F} \cdot \nabla_k f - \tau_k \mathbf{v}\mathbf{v} \cdot \nabla_r f\right) d\mathbf{k}. \quad (88)$$

Assuming parabolic bands, the gradient of the distribution function in $\mathbf{k}$ space can be written

$$\nabla_k f = \frac{\partial f}{\partial E}\nabla_k E = \frac{\partial f}{\partial E}\hbar\mathbf{v}. \quad (89)$$

Also, the following can be shown by direct substitution of the Fermi–Dirac distribution (76):

$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial E} T \left[ E \frac{\partial}{\partial x}\left(\frac{1}{T}\right) - \frac{\partial}{\partial x}\left(\frac{E_F}{T}\right)\right], \tag{90}$$

with similar results in the $y$ and $z$ directions. Substituting these results into Eq. (88) gives

$$\mathbf{J} = -\frac{q}{4\pi^3}\int \left( \frac{-\tau_k}{\hbar}\mathbf{v}\mathbf{F}\frac{\partial f}{\partial E}\hbar\mathbf{v} - \mathbf{v}\mathbf{v}\tau_k \frac{\partial f}{\partial E} T \times \left[ E\nabla_r\left(\frac{1}{T}\right) - \nabla_r\left(\frac{E_F}{T}\right)\right]\right) d\mathbf{k} \tag{91}$$

or

$$\mathbf{J} = \frac{q}{4\pi^3}\int \tau_k \mathbf{v}\mathbf{v}\mathbf{F}\frac{\partial f}{\partial E}\, d\mathbf{k} + q\int \tau_k \mathbf{v}\mathbf{v}\frac{\partial f}{\partial E} T(E - E_F)\nabla_r\left(\frac{1}{T}\right) d\mathbf{k}. \tag{92}$$

For example, the applied force might include a contribution from an external electric field ($-q\mathcal{E}$), plus a contribution caused by a temperature gradient (see Fig. 3), or in general as $\nabla\bar{\mu} = \nabla(\mu + qV) = \nabla\mu - q\mathcal{E}$. Then (92) would be

$$\mathbf{J} = \frac{q}{4\pi^3}\int \tau_k \mathbf{v}\mathbf{v}\frac{\partial f}{\partial E}(\nabla\mu - q\mathcal{E})\, d\mathbf{k} - q\int \tau_k \mathbf{v}\mathbf{v}(E - E_F)\frac{\partial f}{\partial E}\frac{1}{T}\nabla_r T\, d\mathbf{k}, \tag{93}$$

where $\nabla_r(1/T) = -(1/T^2)\nabla_r T$ was used. The electrical current density can be simplified and put into a format similar to the Onsager relations as shown in (54) by using transport integrals defined as

$$\mathbf{K}_n = -\frac{1}{4\pi^3}\int \tau_k \mathbf{v}\mathbf{v}(E - E_F)^n \frac{\partial f_0}{\partial E}\, d\mathbf{k}, \tag{94}$$

where it is assumed that the deviations from equilibrium are small, such that $\partial f/\partial E$ in Eq. (93) may be replaced with $\partial f_0/\partial E$. This leads to an electrical current density of

$$\mathbf{J} = -q\mathbf{K}_0\nabla\bar{\mu} + \frac{q}{T}\mathbf{K}_1\nabla T. \tag{95}$$

Similarly, the heat current density, $\mathbf{J}_Q$, follows the same derivation to arrive at

$$\mathbf{J}_Q = \frac{1}{4\pi^3}\int \tau_k \mathbf{v}\mathbf{v}\frac{\partial f}{\partial E}(\nabla\mu - q\mathcal{E})(E - E_F)\, d\mathbf{k} - \frac{1}{4\pi^3}\int \tau_k \mathbf{v}\mathbf{v}(E - E_F)^2\frac{\partial f}{\partial E}\frac{1}{T}\nabla_r T\, d\mathbf{k} \tag{96}$$

or

$$\mathbf{J}_Q = -\mathbf{K}_1\nabla\bar{\mu} + \frac{1}{T}\mathbf{K}_2\nabla T. \tag{97}$$

These derivatives form the link between the macroscopic Onsager equations and the atomistic derivations from Boltzmann's equation. Comparison of Eqs. (95) and (97) with Eq. (54) shows the following relations:

$$\left.\begin{aligned} \frac{\mathbf{L}_{11}}{qT} &= -q\mathbf{K}_0, \\ \frac{-\mathbf{L}_{12}}{T^2} &= \frac{q}{T}\mathbf{K}_1, \\ \frac{\mathbf{L}_{21}}{qT} &= -\mathbf{K}_1, \\ \frac{-\mathbf{L}_{22}}{T^2} &= \frac{1}{T}\mathbf{K}_2, \end{aligned}\right\} \quad \text{or} \quad \left\{\begin{aligned} \mathbf{L}_{11} &= -q^2T\mathbf{K}_0, \\ \mathbf{L}_{12} &= -qT\mathbf{K}_1, \\ \mathbf{L}_{21} &= -qT\mathbf{K}_1, \\ \mathbf{L}_{22} &= -T\mathbf{K}_2. \end{aligned}\right. \tag{98}$$

This shows the Onsager reciprocity relation, in that $\mathbf{L}_{12} = \mathbf{L}_{21}$. The thermoelectric properties can now be determined through an evaluation of the transport integrals and the appropriate boundary conditions of isothermal ($\nabla T = 0$), isoelectric ($\nabla V = -\mathcal{E} = 0$), static ($\mathbf{J} = 0$), or adiabatic ($\mathbf{J}_Q = 0$). For example, under isothermal conditions, where $\nabla T = 0$ and thus $\nabla\mu = 0$ (for a homogeneous metal),

$$\mathbf{J} = q^2\mathbf{K}_0\mathcal{E} = \boldsymbol{\sigma}\mathcal{E}, \tag{99}$$

and the electrical conductivity is

$$\boldsymbol{\sigma} = q^2\mathbf{K}_0. \tag{100}$$

The electronic contribution to the thermal conductivity is defined for static conditions as $\mathbf{J}_E|_{J=0} = -\boldsymbol{\kappa}_e\nabla T$, or when $\mathbf{J}_E = \mathbf{J}_Q$ as in a one-band material, $\mathbf{J}_Q = \boldsymbol{\kappa}_e\nabla T$, where (95) becomes

$$0 = -q\mathbf{K}_0\nabla\bar{\mu} + \frac{q}{T}\mathbf{K}_1\nabla T. \tag{101}$$

Solving for $\nabla\bar{\mu}$ and substituting into (97) gives

$$\mathbf{J}_Q = -\mathbf{K}_1\left(\frac{1}{T}\frac{\mathbf{K}_1}{\mathbf{K}_0}\nabla T\right) + \frac{1}{T}\mathbf{K}_2\nabla T = \frac{1}{T}\left(\mathbf{K}_2 - \frac{\mathbf{K}_1\mathbf{K}_1}{\mathbf{K}_0}\right)\nabla T \tag{102}$$

or

$$\boldsymbol{\kappa} = \frac{1}{T}\left(\mathbf{K}_2 - \frac{\mathbf{K}_1\mathbf{K}_1}{\mathbf{K}_0}\right). \tag{103}$$

The absolute Seebeck coefficient, or thermopower, $\mathbf{S}$, can also be found from the static condition, where the use of Eq. (101) gives

$$\mathbf{S} = \frac{1}{q}\frac{\nabla\bar{\mu}}{\nabla T} = \frac{1}{qT}\frac{\mathbf{K}_1}{\mathbf{K}_0}. \tag{104}$$

The Peltier coefficient, $\Pi$, can be found by evaluating the heat current density, Eq. (97), for isothermal conditions:

$$\mathbf{J}_Q = q\mathbf{K}_1\mathcal{E}. \tag{105}$$

**TABLE I Combined Results from Macroscopic and Atomistic Analysis**

| Thermoelectric property | Transport integral | Onsager coefficient |
|---|---|---|
| $\boldsymbol{\sigma} = q^2\mathbf{K}_0$ | $\mathbf{K}_0 = \dfrac{\boldsymbol{\sigma}}{q^2}$ | $\mathbf{L}_{11} = -\boldsymbol{\sigma}T$ |
| $\mathbf{S} = \dfrac{1}{qT}\dfrac{\mathbf{K}_1}{\mathbf{K}_0}$ | $\mathbf{K}_1 = T\dfrac{\boldsymbol{\sigma}}{q}\mathbf{S}$ | $\mathbf{L}_{12} = \mathbf{L}_{21} = -T^2\boldsymbol{\sigma}\mathbf{S}$ |
| $\boldsymbol{\kappa}_e = \dfrac{1}{T}\left(\mathbf{K}_2 - \dfrac{\mathbf{K}_1\mathbf{K}_1}{\mathbf{K}_0}\right)$ | $\mathbf{K}_2 = \boldsymbol{\kappa}_e T + T^2\boldsymbol{\sigma}\mathbf{S}^2$ | $\mathbf{L}_{22} = -T^2\boldsymbol{\kappa}_e - T^3\boldsymbol{\sigma}\mathbf{S}^2$ |

Substituting for the electric field, $\mathcal{E}$, from Eq. (99) gives the direct relationship between heat current density and electric current density,

$$\mathbf{J}_Q = \frac{\mathbf{K}_1}{q\mathbf{K}_0}\mathbf{J} = \boldsymbol{\Pi}\mathbf{J}, \tag{106}$$

where the proportionality constant is simply the Peltier coefficient, $\Pi$. Comparing the Peltier coefficient (106) with the thermopower (104) leads to Kelvin's second relation:

$$\boldsymbol{\Pi} = T\mathbf{S}. \tag{107}$$

Results of the transport integrals are summarized in Table I. Thus Eqs. (54) can be rewritten in terms of the thermoelectric properties as

$$\left.\begin{aligned}\mathbf{J} &= \frac{-\boldsymbol{\sigma}}{q}\nabla\bar{\mu} + \boldsymbol{\sigma}\mathbf{S}\nabla T,\\ \mathbf{J}_Q &= \frac{-T}{q}\boldsymbol{\sigma}\mathbf{S}\nabla\bar{\mu} + \left(\boldsymbol{\kappa}_e + T\boldsymbol{\sigma}\mathbf{S}^2\right)\nabla T.\end{aligned}\right\} \tag{108}$$

Substitution of the transport integrals can be used to evaluate further the thermoelectric properties. Estimations can be made through a series expansion of the transport integrals using a Sommerfeld expansion,

$$\begin{aligned}\mathbf{K}_n &= -\int_0^\infty \phi_n(E)\frac{\partial f_0}{\partial E}\,dE\\ &= \phi_n(E_F) + \frac{\pi^2}{6}(kT)^2\frac{d^2}{dE_F^2}\phi_n(E_F) + \cdots.\end{aligned} \tag{109}$$

For the electrical conductivity,

$$\boldsymbol{\sigma} = q^2\mathbf{K}_0 = -\frac{q^2}{4\pi^3}\int \tau_k\mathbf{v}\mathbf{v}\frac{\partial f_0}{\partial E}\,d\mathbf{k}. \tag{110}$$

In its simplest form for cubic symmetry, this reduces to

$$\sigma = \frac{nq^2\tau_k}{m^*}, \tag{111}$$

where $n$ is the electron density with energies near $E_F$, and $m^*$ is the effective mass of the electrons. Both the electron density near the Fermi level and the relaxation time are functions of energy, such that the electrical conductivity can be approximated as $\sigma = \text{const}\cdot E^\xi$, where $\xi$ is some number.

A relationship between the electrical conductivity and the thermopower can be found by series expansion $\mathbf{K}_1$, which gives

$$\mathbf{K}_1 = -\frac{\pi^2}{3q^2}(kT)^2\frac{\partial\sigma}{\partial E}\bigg|_{E=E_F}, \tag{112}$$

along with $\boldsymbol{\sigma} = q^2\mathbf{K}_0$ and substituting into (104). This leads to the Mott–Jones equation (Barnard, 1972):

$$S_d = \frac{-\pi^2}{3}\frac{k^2T}{q}\left(\frac{\partial\ln\sigma}{\partial E}\right)_{E_F}. \tag{113}$$

A distinction of the diffusion thermopower, $S_d$, has been made here to separate it from a low-temperature effect that has not been considered above. The low-temperature effect typically appears as a peak in the measured thermopower (near 60 K for monovalent noble metals) and is the result of an increased electron–phonon interaction. When a temperature gradient exists across a crystal, heat will flow from the hot side to the cold side through lattice vibrations (phonons) and through electron flow. Various interactions among phonons, lattice defects, and electrons can be described by scattering times for each type of interaction. At high temperatures, phonon–phonon interactions are more frequent than electron–phonon interactions ($\tau_{p,p} < \tau_{p,e}$). At these high temperatures (above the Debye temperature, $T > \theta_D$), $\tau_{p,e}$ is approximately temperature independent, while $\tau_{p,p} \propto 1/T$. Under these conditions, the total thermopower is dominated by the diffusion thermopower as given in Eq. (113). At low temperatures ($T < \theta_D$), $\tau_{p,e} \propto 1/T$ and $\tau_{p,p} \propto e^{\theta_D/2T}$, therefore, as the temperature drops, $\tau_{p,p}$ increases more rapidly than $\tau_{p,e}$. When this occurs, $\tau_{p,p} > \tau_{p,e}$ and electron–phonon interactions will occur more frequently, causing electrons to "dragged" along with the phonons. This gives rise to a larger gradient of carrier concentration across the sample and is additive to the diffusion thermopower such that $S = S_d + S_g$, where $S_g$ is the phonon-drag component of the thermopower described above. At still lower temperatures, phonon-impurity interactions can dominate, causing the magnitude of the thermopower to decrease toward zero. For the remainder of this chapter, the

temperature is assumed to be much higher than the Debye temperature, such that $S \approx S_d$ and the diffusion subscript is dropped.

When the electrical conductivity can be written $\sigma = \text{const} \cdot E^{\xi}$, this can be used in the Mott–Jones equation to give

$$S = \frac{-\pi^2}{3}\frac{k^2 T}{qE_F}\xi = -0.0245\frac{T}{E_F}\xi\left(\frac{\mu\text{V}}{\text{K}}\right). \tag{114}$$

### 1. Normal Metals

In monovalent noble metals (Cu, Ag, and Au), $\xi \approx -\frac{3}{2}$ has been measured, giving the positive quantity

$$S = 0.03675\frac{T}{E_F}\left(\frac{\mu\text{V}}{\text{K}}\right). \tag{115}$$

It is instructive to compare the thermopower of noble metals to the electronic heat capacity, $C_{\text{el}}$, which is dependent on the density of states, $g(E_F)$, evaluated at the Fermi level. Substituting $\xi \approx -\frac{3}{2}$ into Eq. (114) gives

$$S = \frac{\pi^2}{2}\frac{k^2 T}{qE_F} \tag{116}$$

and

$$C_{\text{el}} = \frac{\pi^2}{3}g(E_F)k^2 T = \frac{\pi^2 N}{2}\frac{k^2 T}{E_F}, \tag{117}$$

where $N$ is the total number of carriers. Then it can be seen that the electronic heat capacity per carrier is simply the electronic charge times the thermopower,

$$\frac{C_{\text{el}}}{N} = qS. \tag{118}$$

### 2. Transition Elements

The electronic properties of transition metals are usually considered to have contributions from two bands that overlap at the Fermi level: the $s$-band, from the $s$ levels of the individual atoms, and the $d$-band, consisting of five individual overlapping bands. The $s$-band is broad and typically approximated as free electron-like, while the $d$-band is narrow, with a high density of states and high effective mass, thus the $s$ electrons carry most of the current. The relaxation time is, however, greatly affected by the high density of states of the $d$-band. This comes about through the inverse proportionality of the relaxation time to the probability of scattering from one wavevector, $\mathbf{k}$, to another, $\mathbf{k}'$. The occupancy and availability of each of these wavevectors are, in turn, proportional to the density of states at the Fermi level. This leads to the relationship of the inverse proportionality of the relaxation time to the density of states:

$$\frac{1}{\tau} \propto g(E)|_{E=E_F}. \tag{119}$$

Due to the relatively high density of states in the $d$-band, the relaxation time of the highly responsive $s$-band electrons is dominated by $s$–$d$ transitions, or

$$\frac{1}{\tau_s} \approx \frac{1}{\tau_{s-d}} \propto g_d(E)|_{E=E_F}. \tag{120}$$

Neglecting the $d$-band contribution to the electrical conductivity and rewriting Eq. (110) in terms of the density of states gives

$$\begin{aligned}\sigma &= \frac{2}{3}q^2 v_s^2 \tau_s g_s(E)|_{E=E_F} \\ &= \text{const} \cdot v_s^2 \frac{g_s(E)}{g_d(E)}\bigg|_{E=E_F}.\end{aligned} \tag{121}$$

Defining the bottom of the $s$-band as zero energy, and the partially filled $d$-band in terms of the holes in the band so it can be referenced to the top of the $d$-band, such that $E_0$ is the energy at the top of the $d$-band, and $g_d(E) = \text{const} \cdot (E_0 - E_F)^{1/2}$, then approximating the $s$-band electrons as free electrons gives

$$\frac{\partial \ln \sigma}{\partial E}\bigg|_{E=E_F} = \frac{3}{2E_F} + \frac{1}{2(E_0 - E_F)}. \tag{122}$$

Typically $E_F \gg (E_0 - E_F)$, such that approximating the above equation as the second term on the right-hand side and using this in the Mott–Jones equation (113) gives

$$S = \frac{-\pi^2}{6}\frac{k^2 T}{q(E_0 - E_F)}. \tag{123}$$

Again, the electronic heat capacity can be compared to find

$$C_{\text{el}} = \frac{\pi^2}{6}\frac{Nk^2 T}{(E_0 - E_F)}, \tag{124}$$

and the relationship between the magnitude of the electronic heat capacity and the thermopower remains

$$\frac{C_{\text{el}}}{N} = qS. \tag{125}$$

### 3. Semimetals

The petavalent elements of As, Sb, and Bi are semimetals with rhombohedral crystal structures. This leads to nonspherical Fermi surfaces and anisotropic scattering such that $\tau \propto \mathbf{k}_x^s$ for a given crystallographic direction, where $s$ accounts for the anisotropy. Likewise, the density of states $g(E) \propto \mathbf{k}_x^3$, and $\mathbf{k}_x \propto (E_0 - E_F)^{1/2}$. Using the density of states and the relaxation time for the electrical conductivity in an equation similar to (121) gives

$$\sigma = \text{const} \cdot (E_0 - E_F)^{(3+s)/2} \tag{126}$$

or, in (113),

$$S \cong \frac{-\pi^2}{6} \frac{k^2 T}{q(E_0 - E_F)}(3+s), \tag{127}$$

where $(3+s) < 0$. There is an exception to (127) in bismuth, which shows the expected anisotropic thermopower, but an unexpected negative thermopower ($S_\perp \approx -50\,\mu$V/K, and $S_{||} \approx -100\,\mu$V/K at 273 K). For bismuth, a value of $\xi = (3+s)/2$ should be used in (114) describing electron conduction, instead of using Eq. (127), which is for conduction by holes.

### 4. Alloys

Matthiessen's rule states that the total resistivity of an alloy formed by two metals can be found by

$$\rho = \frac{1}{\sigma} = \rho_i + \rho_j, \tag{128}$$

where $\rho_i$ is the resistivity of the pure solvent metal due to scattering of carriers by thermal vibrations, and $\rho_j$ represents scattering of carriers from impurities. This rule is often used for approximations but is not widely applicable since many cases exhibit anisotropic scattering of carriers, causing a large deviation from (128). Assuming the validity of Matthiessen's rule, (113) can be written

$$S = \frac{\pi^2}{3} \frac{k^2 T}{q} \left( \frac{\partial \ln(\rho_i + \rho_j)}{\partial E} \right)_{E_F}, \tag{129}$$

which can be written in terms of the difference between $S$ for the alloy and the thermopower of the pure solvent metal, $S_i$, or $\Delta S = S - S_i$ leads to

$$\frac{\Delta S}{S} = -\frac{1 - (x_j/x_i)}{1 + (\rho_i/\rho_j)}, \tag{130}$$

where

$$x_i = -\left( \frac{\partial \ln \rho_i}{\partial E} \right)_{E_F} \quad \text{and} \quad x_j = -\left( \frac{\partial \ln \rho_j}{\partial E} \right)_{E_F}. \tag{131}$$

Using the Gorter–Nordheim relation for the impurity component of Mattheissen's rule, $\rho_i = CX(1-X)$, where $C$ is the Nordheim coefficient and $X$ is the atomic fraction of the solute atoms in a solid solution, yields a more useful relationship:

$$S = S_j + \frac{\rho_i}{\rho}(S_i - S_j), \tag{132}$$

where $S_j$ is the thermopower for the impurity.

The third thermoelectric parameter listed in Table I is thermal conductivity. This can likewise be determined using (109) and substituting the transport integrals into (103). Series expansion of $\mathbf{K}_2$ gives

$$\mathbf{K}_2 = -\frac{\pi^2}{3} \frac{k^2 T^2}{q^2} \boldsymbol{\sigma}. \tag{133}$$

The thermal conductivity is given by (103), repeated here for convenience:

$$\boldsymbol{\kappa}_e = \frac{1}{T} \left( \mathbf{K}_2 - \frac{\mathbf{K}_1 \mathbf{K}_1}{\mathbf{K}_0} \right). \tag{103}$$

In metals, $(\partial/\partial E)\sigma(E)|_{E=E_F} \approx (\sigma/E_F)$, thus $\mathbf{K}_1 \approx -(\pi^2/3q^2)(kT)^2(\sigma/E_F)$, or

$$\frac{\mathbf{K}_1 \mathbf{K}_1}{\mathbf{K}_0} \approx \frac{\left[(\pi^2/3q^2)(kT)^2(\sigma/E_F)\right]^2}{\sigma/q^2} = \left[ \frac{\pi^2}{3q^2}(kT)^2 \sigma \right] \left[ \frac{\pi^2}{3} \frac{(kT)^2}{E_F^2} \right], \tag{134}$$

giving

$$\boldsymbol{\kappa}_e \approx \frac{\pi^2 k^2 T}{3q^2} \sigma \left( 1 + \frac{\pi^2}{3} \frac{(kT)^2}{E_F^2} \right) \approx \frac{\pi^2 k^2 T}{3q^2} \sigma, \tag{135}$$

where the approximation of $(\pi^2/3)((kT)^2/E_F^2) \ll 1$ was used, thus arriving at the Wiedemann–Franz law, or $\boldsymbol{\kappa}_e/\sigma T = 2.443 \times 10^{-8}\,((\text{W} \cdot \Omega)/\text{K}^2)$.

The total thermal conductivity, $\kappa$, must also include a lattice contribution, $\kappa_L$, such that

$$\boldsymbol{\kappa} = \boldsymbol{\kappa}_L + \boldsymbol{\kappa}_e. \tag{136}$$

The lattice thermal conductivity for metals is generally much lower than the electronic contribution.

## D. Semiconductors

The above analysis is applicable to normal metals, where it is assumed that the carriers are electrons and $\nabla\mu$ is a function of temperature only. Furthermore, the Onsager relations were developed using four contributions, (33), (36), (37), and (39), to the energy density rate of change, however, two additional contributions exist for semiconductors. These contributions account for transitions of electrons across the bandgap, or the rate of change in carrier concentrations in each band, and for positional gradients of the band edges (valence band and conduction band). The last contribution could arise from temperature gradients and/or compositional variations, for example. These additional contributions have the form

$$\left( \frac{\partial E}{\partial t} \right)_{\text{V}} = -q\Phi_C(-\nabla \cdot \mathbf{J}_n) + q\Phi_V(-\nabla \cdot \mathbf{J}_p) \tag{137}$$

and

$$\left( \frac{\partial E}{\partial t} \right)_{\text{VI}} = \mathbf{J}_n \cdot \nabla q\Phi_C - \mathbf{J}_p \cdot \nabla q\Phi_V, \tag{138}$$

where $-q\Phi_C$ and $q\Phi_V$ represent the internal potential energies of the electrons and holes at the bottom of the conduction band and the top of the valence band, respectively. This leads to the Onsager relations for a two-band model, where, in a steady-state condition (Harman and Honig, 1967),

$$\begin{aligned}\mathbf{J}_Q &= L_{11}\mathbf{X}_Q + L_{12}\mathbf{X}_n + L_{13}\mathbf{X}_p,\\ \mathbf{J}_- &= L_{21}\mathbf{X}_Q + L_{22}\mathbf{X}_n + L_{23}\mathbf{X}_p,\\ \mathbf{J}_+ &= L_{31}\mathbf{X}_Q + L_{32}\mathbf{X}_n + L_{33}\mathbf{X}_p,\end{aligned} \tag{139}$$

where

$$\begin{aligned}\mathbf{X}_Q &= -\frac{1}{T^2}\nabla T,\\ \mathbf{X}_n &= -\frac{1}{T}\nabla\varphi_C + \nabla\frac{\bar{\mu}_C}{qT},\\ \mathbf{X}_p &= -\frac{1}{T}\nabla\varphi_V - \nabla\frac{\bar{\mu}_V}{qT},\end{aligned} \tag{140}$$

Also, $\bar{\mu}_C$ and $\bar{\mu}_V$ represent the difference between the chemical potential energy and the internal potential energy of the carriers in the two bands. The total potential energy of the carriers in an applied field for a semiconductor must include the potential energy from the field as well as the internal potential energies $-q\Phi_C$ and $q\Phi_V$, from the band edges. Contributions to the electrical current density come from electrons, $\mathbf{J}_- = -q\mathbf{J}_n$, and from holes, $\mathbf{J}_+ = q\mathbf{J}_p$, for the total current density given by $\mathbf{J} = \mathbf{J}_- + \mathbf{J}_+$. Applying the same procedure for this case as followed for metals above, with the additional consideration of the relative potential energies of the band edges using $\bar{\mu}_V = -(E_F + E_V)$ and $\bar{\mu}_C = E_F - E_C$, gives the following formula for a two-band semiconductor:

$$\begin{aligned}\sigma &= \sigma_n + \sigma_p,\\ S &= \frac{S_n\sigma_n + S_p\sigma_p}{\sigma_n + \sigma_p},\\ \kappa &= \kappa_L + \kappa_n + \kappa_p + \frac{\sigma_n\sigma_p}{\sigma T q^2}\left[\frac{K_1^n}{K_0^n} + \frac{K_1^p}{K_0^p} + (E_C - E_V)\right]^2.\end{aligned} \tag{141}$$

Of course, as a semiconductor is doped $n$-type or $p$-type, the corresponding contributions, subscripted $n$ or $p$, respectively, above will dominate. The last term in the thermal conductivity formula, when multiplied by $-\nabla$T, would relate to the transport of bandgap energy along the negative temperature gradient and is defined as an ambipolar transport mechanism.

# V. APPLICATIONS

## A. Thermocouples

Thermocouples are the most common application of thermoelectric materials. Application of the Seebeck coefficient (1), along with the Thompson relation (7), allows one to determine the open-circuit potential for a circuit containing temperature gradients by integrating over temperature as one traverses through the circuit from one terminal of the open circuit to the other. For example, in the circuit shown in Fig. 6 the open-circuit voltage can be written

$$\begin{aligned}\Delta V &= \int_{T_0}^{T_1} S_A\,dT + \int_{T_1}^{T_2} S_B\,dT + \int_{T_2}^{T_3} S_C\,dT\\ &\quad + \int_{T_3}^{T_4} S_C\,dT + \int_{T_4}^{T_5} S_D\,dT + \int_{T_5}^{T_0} S_A\,dT\\ &= \int_{T_5}^{T_1} S_A\,dT + \int_{T_1}^{T_2} S_B\,dT + \int_{T_2}^{T_4} S_C\,dT\\ &\quad + \int_{T_4}^{T_5} S_D\,dT.\end{aligned} \tag{142}$$

When measuring this potential difference, care must be taken to include the contribution from the leads of the meter. This can be minimized by assuring that the thermocouple-circuit open terminals (in Fig. 6) are at a constant temperature $T_0$ and the terminals on the voltage meter are also at a constant temperature (not necessarily $T_0$).

## B. Generators and Coolers

Lenz first demonstrated a thermoelectric cooler by freezing water at the junction between two conductors formed by rods of bismuth and antimony; however, a more common configuration for a thermoelectric cooler is shown in

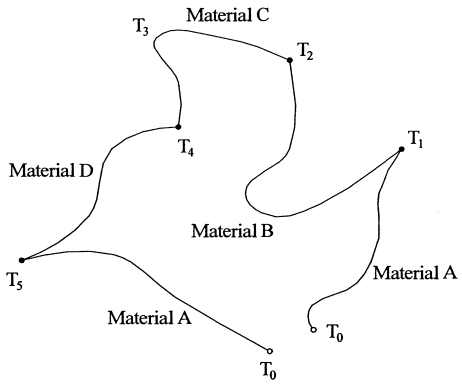



**FIGURE 6** Thermocouple circuit.

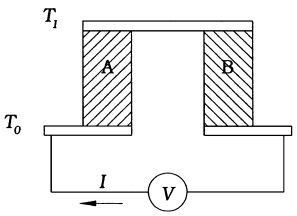



**FIGURE 7** Thermoelectric cooler.

Fig. 7. Here the cooling (or warming) junction is made more accessible for device cooling (or heating).

Since current is defined as positive in the direction of positive carrier flow (hole flow), and likewise for the direction opposite to negative carrier flow (electron flow), by using one $p$-type leg and one $n$-type leg to the cooler, the highest efficiency can be achieved (Fig. 8). In this situation, all carriers flow in the same physical direction (either top to bottom or bottom to top) in both legs. Since charge carriers also carry heat as shown through the Onsager relations, heat will flow through the device in the direction of the carriers. Although the configuration shows a *pn* junction, these devices do not behave as diodes and electrical current is reversible. This is due to the fact that each of the legs is doped to degeneracy, or near-degeneracy, such that ohmic contacts with the metals are exhibited.

The goal in making a thermoelectric cooler is to maximize the coefficient of performance, $\varphi$, of the device, defined as

$$\varphi = \frac{\dot{Q}_0}{W}, \tag{143}$$

where $\dot{Q}_0$ is the rate of heat absorbed from the object being cooled over the amount of power, $W$, it takes to drive the cooler. Assuming that the thermopower of materials A and B in Fig. 7 do not vary significantly over the temperature range $T_0$ to $T_1$, then the Thompson heat may be neglected, and

$$\dot{Q}_0 = \dot{Q}_\Pi - \dot{Q}_\mathrm{T}, \tag{144}$$

where $\dot{Q}_\Pi$ is the Peltier heat absorbed at the cold junction and $\dot{Q}_\mathrm{T}$ is the thermal losses down the arms of the cooler. The Peltier heat absorbed is $\dot{Q}_\Pi = \Pi \cdot I$, and the thermal losses down the arms consist of thermal conduction losses, $K(T_0 - T_1)$, where $K$ is the thermal conductance of the arms, and Joule heating losses, $\frac{1}{2}I^2R$. A factor of $\frac{1}{2}$ on the Joule heating losses is due to half of this heat flowing to the cold end and half flowing to the warm end of the cooler. Substituting gives

$$\dot{Q}_0 = \Pi \cdot I - \tfrac{1}{2}I^2R - K(T_0 - T_1). \tag{145}$$

Maximizing $\dot{Q}_0$ with respect to current yields $\Pi = I \cdot R$, or $I_\mathrm{max} = \Pi/R$. Using the Kelvin relations,

$$I_\mathrm{max} = \frac{(S_A - S_B)T_1}{R}. \tag{146}$$

In steady state, $\dot{Q}_0 = 0$, and the maximum temperature gradient $\Delta T_\mathrm{max} = (T_0 - T_1)$ is

$$\Delta T_\mathrm{max} = \frac{1}{2}\frac{(S_A - S_B)^2}{RK}T_1^2 = \frac{1}{2}ZT_1^2, \tag{147}$$

where $Z$ is defined as the figure of merit for the cooler. Equation (147) clearly shows that the maximum temperature gradient is increased by choosing materials with the largest difference in thermopower values. Therefore, the logical choice is to use one $n$-type and one $p$-type material as mentioned previously.

Continuing with the evaluation of the coefficient of performance for the cooler, the power absorbed by the device is simply the product of the current and voltage supplied to the cooler, or

$$W = IV = I\{IR + (S_A - S_B)(T_0 - T_1)\}, \tag{148}$$

where the voltage across the device includes the resistive and thermoelectric voltage drops. Dividing this into $\dot{Q}_0$ yields the coefficient of performance,

$$\varphi = \frac{\Pi I - \frac{1}{2}I^2R - K(T_0 - T_1)}{I^2R + (S_A - S_B)(T_0 - T_1)I}. \tag{149}$$

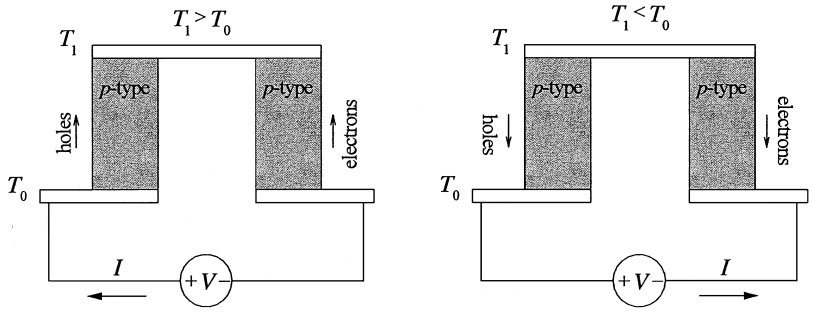



**FIGURE 8** Thermoelectric cooler current flow.

Taking the derivative with respect to current and setting it equal to zero gives

$$\frac{d\varphi}{dI} = 0 = \frac{(\Pi - IR)[S\Delta T \cdot I + I^2R] - [S\Delta T + 2IR]\left[\Pi I - \frac{1}{2}I^2R - K(T_0 - T_1)\right]}{[S\Delta T \cdot I + I^2R]^2}, \tag{150}$$

where the substitutions $S = (S_A - S_B)$ and $\Delta T = (T_0 - T_1)$ were used. After expansion and cancellation in the numerator,

$$\frac{d\varphi}{dI} = 0 = \frac{I^2\left[-R\Pi - \frac{1}{2}R \cdot S\Delta T\right] + I[2KR(T_0 - T_1)] + K \cdot S\Delta T^2}{[S\Delta T \cdot I + I^2R]^2}. \tag{151}$$

Substituting $\Pi = (S_A - S_B)T_1$ for the Peltier heat removed at the cold junction gives

$$0 = I^2\left\{-R(S_A - S_B)\left[T_1 + \tfrac{1}{2}(T_0 - T_1)\right]\right\} + I[2KR(T_0 - T_1)] + K(S_A - S_B)(T_0 - T_1)^2. \tag{152}$$

Solving this quadratic equation yields the maximum coefficient of performance at the optimum current,

$$I_{\text{opt}} = \frac{(S_A - S_B)(T_0 - T_1)}{R(\sqrt{1 + Z\bar{T}} - 1)}, \tag{153}$$

where $\bar{T}$ is the average temperature $\frac{1}{2}(T_0 + T_1)$. Using this in Eq. (143) yields

$$\varphi_{\text{opt}} = \frac{T_1}{(T_0 - T_1)} \frac{\sqrt{1 + Z\bar{T}} - (T_0/T_1)}{\sqrt{1 + Z\bar{T}} + 1}, \tag{154}$$

where the first term represents the coefficient of performance for an ideal heat pump. This shows that both $\varphi$ and $\Delta T$ are directly dependent on the figure of merit, $Z$. Thus maximizing the figure of merit for the individual materials,

$$Z = \frac{S^2}{\rho\kappa} = \frac{S^2\sigma}{\kappa}, \tag{155}$$

maximizes the efficiency of the cooler. Desirable materials have large-magnitude thermopowers, $S$ (one $n$-type and one $p$-type), and low electrical resistivities, $\rho$, or, equivalently, high electrical conductivities, $\sigma$, and low thermal conductivities, $\kappa$. Since the figure of merit has units of $K^{-1}$, the unitless quantity of $ZT$ is often reported.

It should also be noted that the Peltier heat, $\dot{Q}_\Pi = \Pi \cdot I$, is either absorbed or liberated based on the current direction. Therefore, the same configuration can be used as either a thermoelectric cooler or a heater.

For comparison, and to illuminate the present challenge, the coefficient of performance for standard Freon-based refrigeration systems is 1.2 to 1.4, for a refrigerator operating at a cold temperature of 263 K while the outside (hot temperature) is at 323 K. Freon-based cooling systems have coefficients of performance that would correspond to a thermoelectric device with $ZT$ between 3 and 4. Also shown is the COP for the present value of $ZT \sim 1$. The advantages of thermoelectric devices includes size scalability without loss of efficiency, robustness, low maintenance, a relatively small electromagnetic signature, and the ability both to heat and to cool from a single device, and they are environmentally cleaner than conventional CFC-based coolers. Many thermoelectric companies presently exist, indicating an existing market such that any increase in $ZT$ through a new material and/or configuration could have a direct impact; however, a significant increase in the market is anticipated for an increase in $ZT$ to 2. This, therefore, represents the current goal in Fig. 9.

These devices are heat pumps, in that it is also possible to remove the electrical power source, and force a temperature gradient across the thermoelectric device, by contacting one end of it to an external heat source. With a load connected to the device instead of the electrical power source, it then functions as a thermoelectric generator. Thus, the application of an electrical potential gradient causes the generation of a temperature gradient (thermoelectric cooler) and the application of a temperature gradient causes the generation of electrical power (thermoelectric generator).

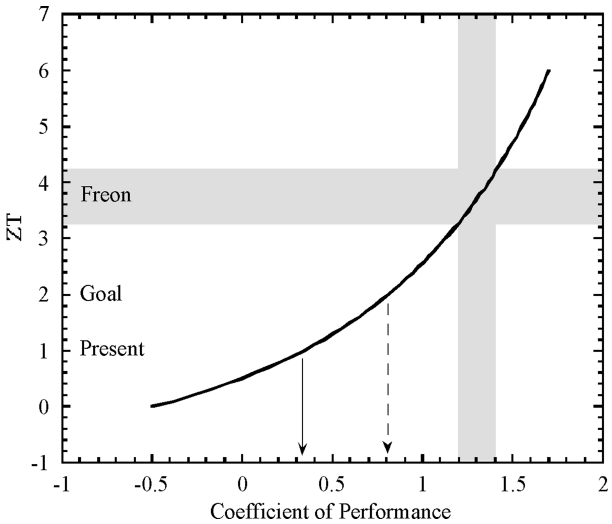



**FIGURE 9** The figure of merit versus the coefficient of performance.

In the case of a generator, the efficiency, $\eta$, of the device is defined as the ratio of the power supplied to the load to the heat absorbed at the hot junction:

$$\eta = \frac{T_H - T_C}{T_H} \frac{\sqrt{1 + Z\bar{T}} - 1}{\sqrt{1 + Z\bar{T}} + (T_C/T_H)}. \qquad (156)$$

This is again dependent on the figure of merit of the device. Through Thompson's relations we can split the figure of merit for the device into a figure of merit for each of the two legs. When each of these has been maximized individually, then the total device figure of merit will also be maximized assuming that one leg is $n$-type and one $p$-type.

## C. New Directions

Traditional materials used in thermoelectric devices are listed in Table II. Near-room-temperature devices have been designed largely for cooling applications, while higher-temperature materials have been generally used in electrical power generation. Research on thermoelectrics was highly active during the decade following 1954, with the United States showing a great interest in high-temperature power generation applications, such as the Si–Ge-based generators used on the satellites Voyager I and II. Recently there has been a resurgence of interest in thermoelectrics, spurred on partly by predictions of the high $ZT$s possible in quantum confined structures (Hicks and Dresselhaus, 1993). It was predicted that in such structures, both the electrical conductivity and the thermopower could be simultaneously increased due to the sharpening of the density of states as confinement increases from 3D $\rightarrow$ 2D $\rightarrow$ 1D $\rightarrow$ 0D (Broido and Reinecke, 1995). The influence of such sharpening can be seen clearly within the Mott–Jones equation for thermopower (113). An indication of the effect from a rapidly varying density of states comes from mixed-valent compounds such as $CePd_3$ and $YbAl_3$, which have shown the largest power factor, $\sigma S^2$, among all known materials. Unfortunately, the high thermal conductivity in these materials prevents them from having a correspondingly high figure of merit. An additional increase in $ZT$ for quantum confined materials comes from a decrease in the thermal conductivity due to confinement barrier scattering.

Another avenue for investigating thermoelectric materials has been coined the "phonon glass electron crystal"

**TABLE II The Most Widely Used TE Materials**

| | $Z_{max}$ ($K^{-1}$) | Useful range (K) | $T_{max}$ (K) |
|---|---|---|---|
| $Bi_2Te_3$ | $3 \times 10^{-3}$ | <500 | 300 |
| PbTe | $1.7 \times 10^{-3}$ | <900 | 650 |
| Si–Ge | $1 \times 10^{-3}$ | <1300 | 1100 |

**TABLE III Desirable Material Properties for Thermoelectric Applications (Kanatzidis, 2001)**

1. Many valley bands near the Fermi level, but located away from the Brillouin zone boundaries.
2. Large atomic number elements with large spin–orbit coupling.
3. Compositions with two or more elements such as ternaries and quaternaries.
4. Low average electronegativity differences between elements.
5. Large unit cells.
6. Energy gaps near 10 kT.

(PGEC) method (Slack, 1995), in which short phonon mean free paths and long electron mean free paths are simultaneously sought in a material. A suggested way for a material to exhibit PGEC behavior is making a material that incorporates cages and/or tunnels in its crystal structure large enough to accommodate an atom. The caged atom provides strong phonon scattering by rattling within the cage. Electrons would not be significantly scattered by such "rattlers" since the main crystal structure would remain intact.

Additional guidance has been provided by identifying a $B$ parameter defined as

$$B = \gamma \frac{1}{3\pi^2} \left( \frac{2kT}{h^2} \right)^{3/2} \sqrt{m_x m_y m_z} \frac{k^2}{q\kappa_L} \mu_x, \qquad (157)$$

where $\gamma$ is the degeneracy parameter (Hicks and Dresselhaus, 1993). This function should be maximized for optimal $ZT$. With a large number of valleys within a band, fewer carriers can exist in each valley, thus increasing the contribution to the thermopower from that valley. At the same time, the total number of carriers can be maintained for a high electrical conductivity. High degeneracy parameters are generally found in highly symmetric crystal systems. Large effective masses, or large effective mass components in the axes perpendicular to the current flow, allow for a high electrical conductivity in the direction of interest while maintaining a high $B$ parameter. Equation (157) also indicates that high mobilities, $\mu_x$, in the transport direction, and a low lattice thermal conductivity are also desirable. It has recently been shown that semiconductors with bandgaps of approximately 10 kT best satisfy these criteria (Mahan, 1998). Six properties of thermoelectric materials that give the best results are listed in Table III.

# VI. SUMMARY

Macroscopic and atomistic derivations of the thermoelectric properties of electrical conductivity, thermoelectric power, and thermal conductivity have been presented, with approximations for various material systems. The derivations outlined have considered external forces of electric

fields and temperature gradients. Additional effects are realized when more forces such as magnetic fields are included. These include Hall, magnetoresistance, Nernst, Ettingshausen, and Righi–Leduc effects. These magnetic fields also affect the operation of thermoelectric coolers, with significant enhancements of the efficiency possible under strong fields.

## ACKNOWLEDGMENT

I wish to thank Sangeeta Lal, of Bihar University, for her helpful review of the manuscript.

## SEE ALSO THE FOLLOWING ARTICLES

ELECTROMAGNETICS • ELECTRONS IN SOLIDS • SEMICONDUCTOR ALLOYS • SUPERCONDUCTIVITY • THERMODYNAMICS • THERMOMETRY

## BIBLIOGRAPHY

Barnard, R. D. (1972). "Thermoelectricity in Metals and Alloys," Halsted Press (Division of John Wiley & Sons), New York.

Broido, D. A., and Reinecke, T. L. (1995). "Thermoelectric figure of merit of quantum wire superlattices," *Appl. Phys. Lett.* **67**(1), 100–102.

Gray, P. E. (1960). "The Dynamic Behavior of Thermoelectric Devices," Technology Press of the Massachusetts Institute of Technology/John Wiley & Sons, New York.

Guggenheim, E. A. (1957). "Thermodynamics," 3rd ed., North-Holland, Amsterdam.

Harman, T. C., and Honig, J. M. (1967). "Thermoelectric and Thermomagnetic Effects and Applications," McGraw–Hill, New York.

Hicks, L. D., and Dresselhaus, M. S. (1993). "Effect of quantum-well structures on the thermoelectric figure of merit," *Phys. Rev. B* **47**(19), 12 727–12 731.

Ioffe, A. F. (1957). "Semiconductor Thermoelements and Thermoelectric Cooling," Infosearch, London.

Kanatzidis, M. G. (2001). The role of solid-state chemistry in the discovery of new thermoelectric materials. *In* "Solid State Physics" (H. Ehrenreich and F. Spaepen, eds.), Vol. 69, pp. 51–100, Academic Press, New York.

Mahan, G. D. (1998). Good thermoelectrics. *In* "Solid State Physics" (H. Ehrenreich and F. Spaepen, eds.), Vol. 51, pp. 82–157, Academic Press, New York.

Roberts, R. B. (1977). "The absolute scale of thermoelectricity," *Philos. Mag.* **36**(1), 91–107.

Slack, G. A. (1995). New materials and performance limits for thermoelectric cooling. *In* "CRC Handbook of Thermoelectrics" (D. M. Rowe, ed.), CRC Press, New York.

Wendling, N., Chaussy, J., and Mazuer, J. (1993). "Thin gold wires as reference for thermoelectric power measurements of small samples from 1.3 K to 350 K," *J. Appl. Phys.* **73**(6), 2878–2881.



# Transmission Electron Microscopy

**S. Amelinckx**
**J. Van Landuyt**
*University of Antwerp*

I. Bragg's Law
II. Reciprocal Lattice—Ewald Construction
III. Convergent-Beam Electron Diffraction
IV. Kinematical Diffraction Theory
V. Dynamical Theory
VI. Multiple-Beam Diffraction
VII. Diffraction Contrast Images
VIII. Column Approximation
IX. Moiré Fringes
X. Dislocation Contrast
XI. Weak-Beam Imaging
XII. Computer Simulation of Dislocation Images
XIII. Diffraction Contrast at Planar Interfaces
XIV. Image Formation in an Ideal Microscope
XV. Image Formation in a Real Microscope
XVI. Image Formation of a Weak-Phase Object
XVII. Optimum Defocus Images
XVIII. Lattice Images
XIX. Imaging Modes
XX. Scanning Electron Microscopes
XXI. High-Voltage Electron Microscopy
XXII. Analytical Electron Microscopy
XXIII. Specimen Preparation for Transmission Electron Microscopy
XXIV. Examples of Applications

## GLOSSARY

**Absorption contrast** Image formation mechanism resulting from local differences in density or absorption of the material.

**Anomalous absorption** Orientation-dependent absorption of electrons described by a complex extinction distance.

**Bormann effect** Maximum in the rocking curve of the transmitted beam for a slightly positive excitation error. The phenomenon was first discovered in X-ray diffraction.

**Column approximation** Approximation made with the assumption that electrons propagate through the specimen along narrow columns. These columns can be considered as the "picture elements" of the diffraction contrast image.

**Diffraction contrast** Image formation mechanism in transmission electron microscopy resulting from local differences in orientation, thickness, or structure

factor. The image consists of the intensity distribution in a single diffraction spot. If this spot is the transmitted beam, the image is called a bright-field image. If it is a diffracted beam, the resulting image is a dark-field image. A dark-field image in a reflection for which the excitation error $s$ has a large value is called a weak-beam image.

**Dynamical diffraction theory** Diffraction theory that takes into account multiple scattering events.

**Ewald construction** Geometrical construction that allows the directions of scattered beams to be derived based on the concept of reciprocal lattice.

**Ewald sphere** Sphere in reciprocal space with a radius of $1/\lambda$ used in the Ewald construction ($\lambda$ is the electron wavelength).

**Excitation error** Distance in reciprocal space from a reciprocal lattice node to the Ewald sphere, describing the deviation from the exact Bragg condition.

**High-resolution electron microscopy** The image is formed by the interference between a large number of diffraction beams; with modern instruments, this allows the resolution of well-separated atom columns in crystal structures.

**Kinematical diffraction theory** Diffraction theory that takes into account only single scattering events. It is a reasonable approximation for neutron and X-ray scattering but not for electron scattering.

**Pendellösung effect** Periodic transfer of intensity from the incident beam to the scattered beam in a two-beam situation. At the exact Bragg condition, the depth period of this transfer is the extinction distance, which depends on the reflection.

**Reciprocal lattice** Lattice of which the base vectors $\mathbf{A}_1$, $\mathbf{A}_2$, and $\mathbf{A}_3$ are related to the base vectors $\mathbf{a}_1$, $\mathbf{a}_2$, and $\mathbf{a}_3$ of the direct lattice by the relation

$$\mathbf{A}_1 = \frac{(\mathbf{a}_2 \times \mathbf{a}_3)}{V}, \qquad \mathbf{A}_2 = \frac{(\mathbf{a}_3 \times \mathbf{a}_1)}{V},$$
$$\mathbf{A}_3 = \frac{(\mathbf{a}_1 \times \mathbf{a}_2)}{V}.$$

**Rocking curve** Functional dependence of the transmitted and scattered intensity on the excitation error (i.e., on the specimen orientation).

**Scherzer focus** Focusing condition leading to optimal image resolution in high-resolution transmission electron microscopy. The Scherzer focus does not coincide with the Gaussian focus considered in geometrical optics.

**TRANSMISSION** electron microscopy combined with electron diffraction has become an important tool in the study of the structural geometry of solids, which in many respects is complementary to other diffraction techniques such as neutrons and X rays. Electron microscopy has been especially successful in the study of crystal defects (e.g., dislocations, stacking faults, antiphase boundaries, domain boundaries, inversion boundaries, and discommensurations) as well as long-period superstructures resulting from the periodic arrangement of such defects (i.e., modulated structures, long-period antiphase boundary structures, and polytypes).

The existence of quasi-crystals, that is, objects without translation symmetry but with noncrystallographic rotation symmetry elements, such as 5-fold and 10-fold axes, was first demonstrated in 1984 by means of electron diffraction and electron microscopy.

Transmission electron microscopy finds applications in a wide variety of disciplines (e.g., solid-state physics, solid-state chemistry, physical metallurgy, mineralogy, and geology); for instance, the mineralogy of moon rocks was studied mainly by means of electron microscopy.

In recent years the application of high-resolution electron microscopy to problems of crystal chemistry has been expanding very rapidly, and recently it has, in particular, been applied with considerable success to the study of high-critical temperature ($T_c$) superconductors. The study of the atomic structure of surfaces is another very recent application.

In 1986 the Nobel prize for physics was attributed to the inventors of the electron microscope. Although the invention dates back to 1937, it was not until 1958 that transmission electron microscopy was used in the study of solids. Rapid progress, following closely the progress in electron optics, has been made since then; during the 1980s atomic resolution became possible in many materials.

## I. BRAGG'S LAW

The contrast of image produced in transmission electron microscopy of crystalline solids is due mainly to diffraction effects and, to a smaller extent, to absorption. In direct space the diffraction of electrons by a crystal lattice can be described in terms of Bragg's law (Fig. 1c):

$$2d_H \sin\theta_n = n\lambda, \tag{1}$$

where $d_H$ is the interplanar spacing of the lattice planes with Miller indices $H(hkl)$, $\theta_n$ the Bragg angle, $n$ an integer (i.e., the order of the reflection), and $\lambda$ the de Broglie wavelength of the monokinetic electrons. This relation defines the angles $\theta_n$ (i.e., the orientations of the crystal for which the waves scattered by successive lattice planes are in phase and produce a peak in scattered intensity). The peaks are very sharp because many unit cells contribute to the interference phenomenon.

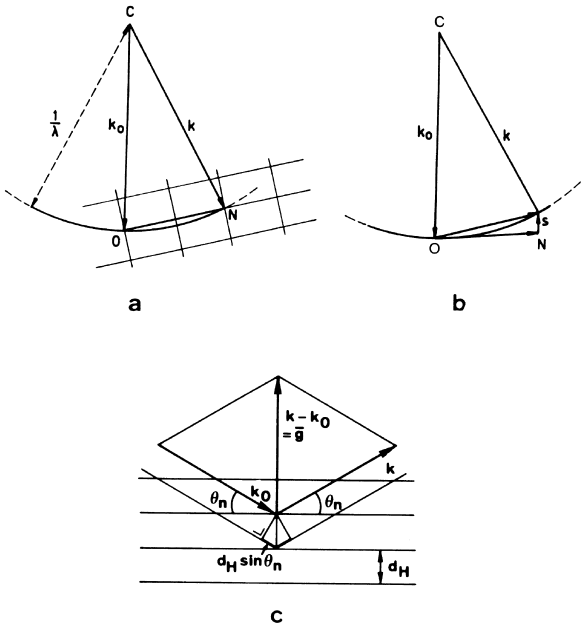



**FIGURE 1** (a) Ewald construction representing the diffraction conditions in reciprocal space. (b) A small deviation from the exact diffraction conditions is present. (c) Bragg's law representing diffraction in real space as "reflection" against a set of lattice planes. Constructive interference occurs if $2d_H \sin\theta_n = n\lambda$. [From Amelinckx, S. (1964). "The Direct Observation of Dislocation," Solid State Physics, Suppl. 6, Academic Press, New York.]

The electron wavelength $\lambda$ is related to the accelerating voltage $V$ by the relation

$$\lambda = \frac{1}{\left[2m_0V_e\left(1 + eV/2m_0c^2\right)\right]^{1/2}}, \qquad (2)$$

which takes into account relativistic effects, where $h$ is the Planck constant, $m_0$ the electron rest mass, $e$ the electron charge, and $c$ the velocity of light in vacuum. For a typical value $V = 100$ kV the wavelength is $\lambda = 0.0037$ nm. Since $d_H$ is of the order of 0.1 nm, the Bragg angles as deduced from Eq. (1) are very small, of the order of $10^{-2}$ to $10^{-3}$ rad.

## II. RECIPROCAL LATTICE—EWALD CONSTRUCTION

An alternative description makes use of the notion of reciprocal lattice. The basic vectors $\mathbf{A}_1$, $\mathbf{A}_2$, and $\mathbf{A}_3$ of the reciprocal lattice are given in terms of the basic vectors $\mathbf{a}_1$, $\mathbf{a}_2$, and $\mathbf{a}_3$ of the direct lattice by the expressions

$$\mathbf{A}_1 = \frac{(\mathbf{a}_2 \times \mathbf{a}_3)}{V}, \qquad \mathbf{A}_2 = \frac{(\mathbf{a}_3 \times \mathbf{a}_1)}{V},$$

$$\mathbf{A}_3 = \frac{(\mathbf{a}_1 \times \mathbf{a}_2)}{V},$$

where $V = (\mathbf{a}_1 \times \mathbf{a}_2)\,\mathbf{a}_3$, is the volume of the unit cell of the lattice.

The geometry of the diffraction phenomenon is now described by the Ewald construction (Fig. 1a), which has the same physical contents as Bragg's law. If $\mathbf{k}_0 = (1/\lambda)\mathbf{e}_0$ ($\mathbf{e}_0$ is an unit vector) is the wave vector of the incident beam, and $\mathbf{g}$ is a reciprocal lattice vector such that $d = |1/g|$, the diffraction condition is given by

$$\mathbf{k} = \mathbf{k}_0 + \mathbf{g}, \qquad (3)$$

where $\mathbf{k}$ is the wave vector of the scattered wave $\mathbf{k} = (1/\lambda)\mathbf{e}_s$, with $\mathbf{e}_s$ the unit vector along the scattered beam. This leads to the geometrical construction represented in Fig. la, where 0 is the origin of the reciprocal lattice. Diffraction maxima will occur in the direction of $\mathbf{k}$ each time a node point of the reciprocal lattice touches a sphere with radius $1/\lambda$, the Ewald sphere, and when the center $C$ is situated in $-\mathbf{k}_0$ drawn from the origin 0.

In electron diffraction the wavelength $\lambda \approx 0.0037$ nm is small compared to the unit cell edge ($\approx$0.2 nm), and as a result, the radius of the Ewald sphere $(1/\lambda)$ is large compared to the unit cell edges $(1/g)$ of the reciprocal lattice. The Ewald sphere can be approximated by a plane normal to the wave vector $\mathbf{k}_0$ of the incident beam.

Since electrons are absorbed in solid matter, it is necessary to use very thin foil specimens (<200 nm). The diffraction conditions are then significantly relaxed since only a small number of unit cells along the normal to the foil can contribute to the diffraction phenomena. However, in directions parallel to the foil the peaks remain sharp. This relaxation of the diffraction conditions is represented in reciprocal space by letting the reciprocal lattice nodes become thin rods perpendicular to the foil plane, instead of points (Fig. 1b). It is now sufficient that the Ewald sphere intersects such a rod to produce a diffracted beam of which the direction is obtained by joining the center $C$ of the Ewald sphere with this intersection point. As a result, many beams can be excited simultaneously, especially in thin foils, and the diffraction pattern along a simple zone axis becomes a two-dimensional array of spots, which reflects the crystal symmetry and can, in fact, be considered a planar section of the reciprocal lattice (Fig. 2).

The relaxation of the diffraction conditions makes it possible to have a diffracted beam even though the relation in Eq. (1) is not exactly fulfilled. The intensity is then a function of the deviation from the exact Bragg condition. In reciprocal space this deviation can be described by the vector $\mathbf{s}$ leading from the reciprocal lattice node point to the intersection point of the rod with the Ewald sphere. Conventionally $s$ will be positive when the reciprocal lattice node point is inside the Ewald sphere and negative when it is outside the sphere (Fig. 1b).

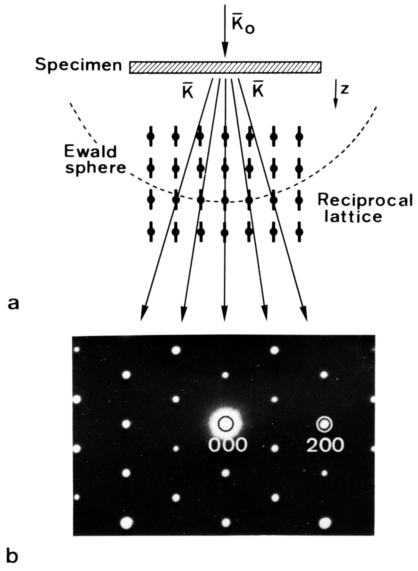



**FIGURE 2** Electron diffraction pattern of a thin foil reflecting the crystal symmetry along the zone axis. The diffraction pattern is a planar section of the reciprocal lattice. (a) Reciprocal lattice construction and (b) corresponding observed diffraction pattern.

Apart from sharp diffraction spots produced by very thin foils, diffraction patterns consisting of lines (Kikuchi lines) are produced in thicker specimens (Fig. 3). These lines form along the intersection lines of a flat double cone with the photographic plate. They are, in fact, hyperbola, which can be approximated by their asymptotes. Each double cone is formed by the geometrical locus of the diffracted beams due to one family of lattice planes produced by randomly and inelastically scattered electrons in the entrance part of the specimens. The semiapex angle of the cones is $90^\circ - \theta$ (i.e., they are, thus, very flat since $\theta$ is small) (Fig. 4). They occur in pairs, one produced by each blade of the double cone. The one due

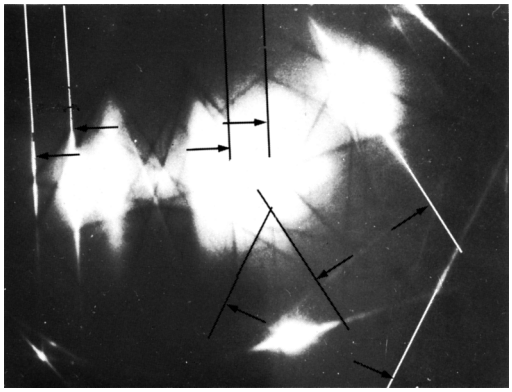


**FIGURE 3** The Kikuchi line pattern is a relatively thick foil. Note the pairs of bright and dark lines emphasized by the drawn lines.

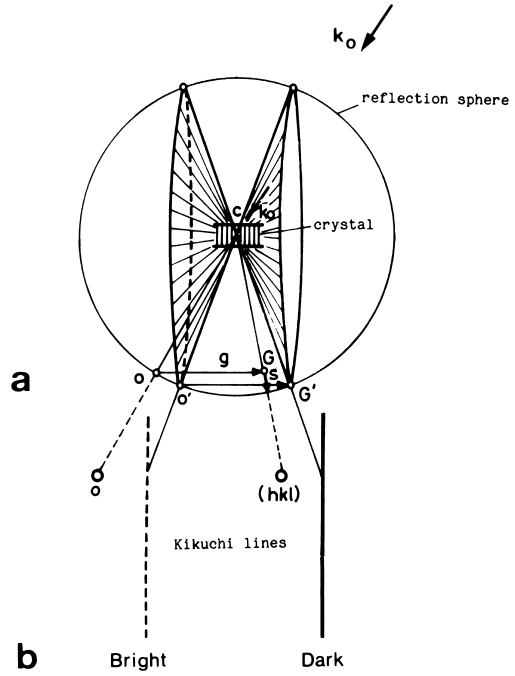



**FIGURE 4** (a) Formation of Kikuchi lines as the intersection lines of a double cone of Bragg reflected beams with the photographic plate. (b) Relation between the positions of Kikuchi lines and Bragg spots.

to forward scattering is brighter than the background; the other one is darker. Their angular separation is $2\theta$. Since each cone pair is strictly bound to one set of lattice planes, rotating the specimen over an angle $\alpha$ causes a parallel displacement of the lines over the same angular range. The Kikuchi lines are of practical importance since they allow the determination of the magnitude and the sense of a **s** from the relative positions of the sharp Bragg spot and the corresponding Kikuchi line (see Fig. 4).

## III. CONVERGENT-BEAM ELECTRON DIFFRACTION

All diffraction techniques described so far make use of a parallel incident beam of electrons. Recently it has been realized that more information, which is more accurate, can be obtained from convergent beam electron diffraction patterns. A convergent beam of electrons is focused on the specimen, and the crossover is situated in the specimen, which is flooded by electrons incident within a wide solid angle of orientations. Bragg scattering then takes place simultaneously with many sets of lattice planes, and is much like the case of Kikuchi lines, but where Kikuchi lines are formed by inelastically scattered electrons, these are due to elastic scattering.

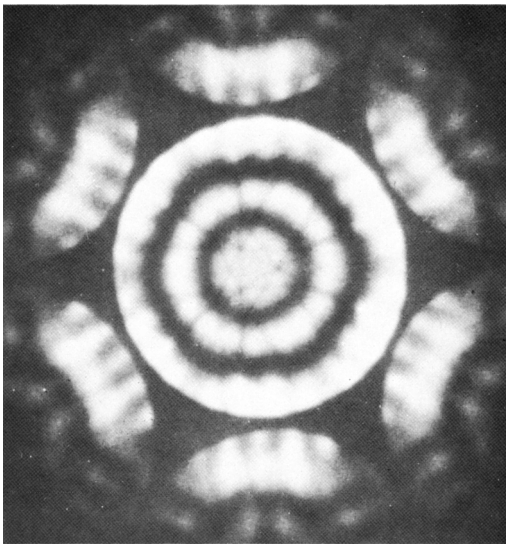


**FIGURE 5** Convergent-beam electron diffraction pattern; the pattern indicates threefold symmetry. [From Tanaka, M. (1978). *JEOL News* **16E,** 13.]

Convergent-beam patterns yield information on the space group of the crystal, including the presence or absence of a center of symmetry, and allow structure amplitude measurements and study of the lattice potential. Furthermore, they allow precise lattice parameter measurements of very small areas.

Convergent-beam electron diffraction patterns are the electron analogue of Kossel lines in X-ray diffraction (Fig. 5).

## IV. KINEMATICAL DIFFRACTION THEORY

According to the kinematical theory of diffraction, the incident radiation is scattered only once, and although the scattered beam is in a position to produce Bragg scattering with the same set of lattice planes, such processes are neglected. Moreover, it is assumed that the incident beam is not depleted so each scatterer sees the same intensity of incident radiation. This approximation is acceptable in X-ray and neutron diffraction, where the probability of multiple scattering is relatively low. But it is only a poor approximation in electron diffraction, where the interaction between the atoms and the incident electrons is very strong. Nevertheless, this theory allows a qualitative understanding of a wide variety of phenomena, and moreover, the geometrical features of the diffraction patterns are, in general, correctly predicted. Also, other features of the diffraction phenomena are qualitatively well described provided **s** is not very small. However, the intensities of diffraction spots predicted by the kinematical theory are usually not in agreement with the observations due to the absence of a simple relation between intensity and structure factor in electron scattering.

According to this theory the amplitude of the scattered beam is given by the expression

$$A_S = F \sin \pi st / \pi s, \tag{4}$$

where $s$ is the deviation parameter, $t$ the foil thickness, and $F$ the structure amplitude. It is obvious from this expression that the amplitude of the scattered beam depends on $s$ in the oscillatory way represented in Fig. 6. This curve is called the rocking curve since it gives the scattered intensity as a function of the angle of incidence. This curve can be visualized directly in the electron microscope by making a dark-field image in a diffracted beam produced by a cylindrically bent crystal; a wide range of angles of incidence (i.e., of $s$ values) is then present simultaneously. The geometrical loci of the points with equal inclination (i.e., equal $s$ values) are lines parallel with the axis of the cylinder. The separation of the zeros is $1/t$ and the central peak is $2/t$ wide (Figs. 6 and 7a).

A rocking curve for the transmitted beam can be obtained by subtracting $I_S$ from unity, $I_T = 1 - I_S$, since absorption is neglected.

Similarly, by making a dark-field image of a wedge-shaped crystal, the intensity scattered in a fixed direction (i.e., in a given reflection) can be imaged as a function of thickness. The geometrical loci of the points of equal thickness (i.e., of equal intensity) are lines parallel with the wedge edge (Fig. 7b). The depth period is given by $1/s$ and becomes infinite for $s \to 0$ . However, for $s = 0$ the kinematical theory is no longer valid and the dynamical theory has to be used.

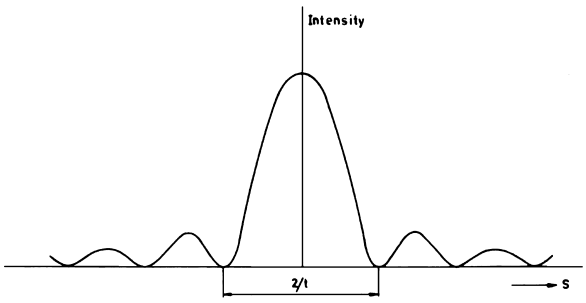



**FIGURE 6** Scattered intensity as a function of the excitation error (rocking curve) according to the kinematical theory, valid in very thin crystals only. [From Amelinckx, S. (1964). "The Direct Observation of Dislocation," Solid State Physics, Suppl. 6, Academic Press, New York.]

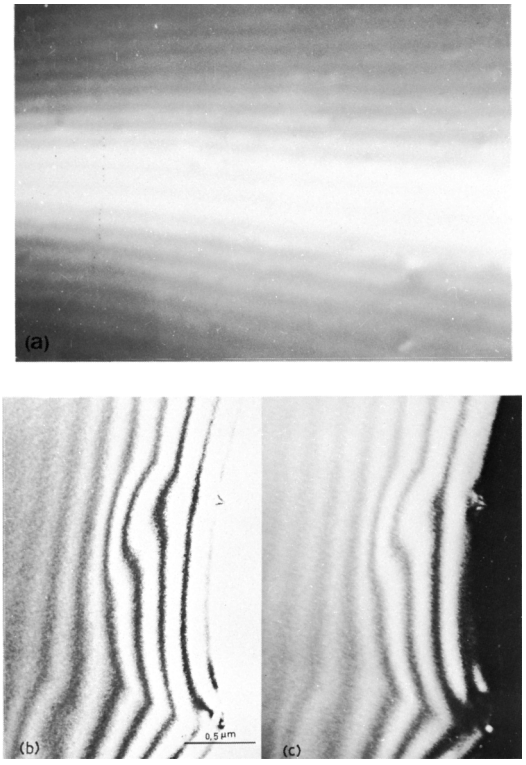



**FIGURE 7** Extinction contours: (a) equi-inclination extinction contours in a cylindrically bent graphite foil of constant thickness; (b) thickness extinction contours in a wedge-shaped crystal (bright-field image); (c) dark-field image corresponding to b. [From Amelinckx, S. (1964). "The Direct Observation of Dislocation," Solid State Physics, Suppl. 6, Academic Press, New York.]

## V. DYNAMICAL THEORY

We summarize the main results for the case of a perfect crystal and assume that only one beam, apart from the incident beam, is strongly excited (i.e., has a $s$ value close to 0); this is called the two-beam case. For quantitative studies of defects, working conditions which closely approximate a two-beam case are practically required.

According to the two-beam dynamical theory of electron diffraction, in a perfect crystal there is a constant interplay between incident and diffracted beams. This can be described by a set of two coupled differential equations, one for each beam, and which is similar to the equations describing the motion of two coupled pendulums (sometimes called Pendellösung). The simplest form of the equations is

$$\begin{aligned}(dT/dz) + \pi i s T &= (\pi i/t_g)S,\\ (dS/dz) - \pi i s S &= (\pi i/t_g)T,\end{aligned} \tag{5}$$

where $T$ and $S$ are the amplitudes of the transmitted and scattered beam, respectively; $t_g$ is a parameter depending on the material and on the reflection $g$ with dimensions of a length, which is inversely proportional to the strength of the considered reflection **g** (i.e., to its structure factor); and $z$ is the distance in the foil measured from the entrance face.

So-called anomalous absorption (i.e., orientation-dependent absorption) can be taken into account phenomenologically by assuming the extinction distance to be complex, replacing $1/t_g$ with $1/t_g + i/\tau_g$, where $\tau_g$ is the absorption length belonging to the reflection $g$. This is a standard procedure also used in other branches of physics to account for absorption. The normal absorption further leads to an exponential damping factor depending only on thickness, and not on orientation. These equations further imply the column approximation, which is discussed later.

The solution of these equations leads to the rocking curve for the transmitted and scattered beam represented in Fig. 8. The oscillating character of Fig. 6 is strongly attenuated, mainly as a result of anomalous absorption. Moreover, the symmetry with respect to $s$ of the curves for the transmitted beam is lost (A in Fig. 8b). The strong transmission for $s > 0$ relative to that for $s < 0$ is called the Bormann effect, which also occurs in X-ray diffraction. For the scattered beam the symmetry, with respect to $s$, is still conserved (B in Fig. 8b).

The depth period of the scattered and transmitted beams is now

$$t_g' = t_g/\sqrt{1 + (st_g)^2}; \tag{6}$$

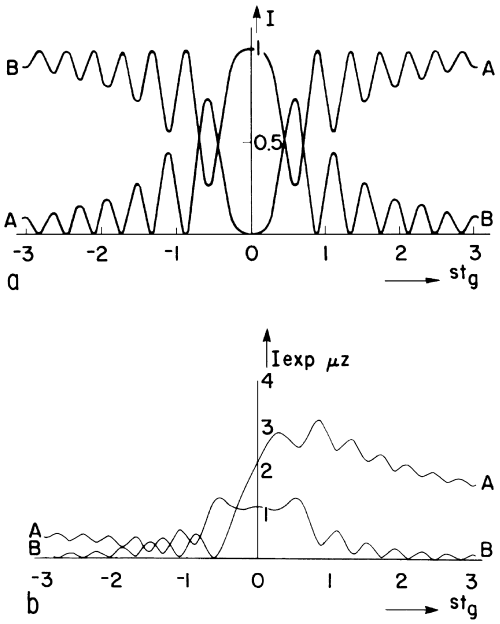



**FIGURE 8** Scattered (B) and transmitted (A) intensity as a function of the excitation error $s$ (rocking curve) according to the dynamical theory (a) without absorption and (b) taking into account anomalous absorption.

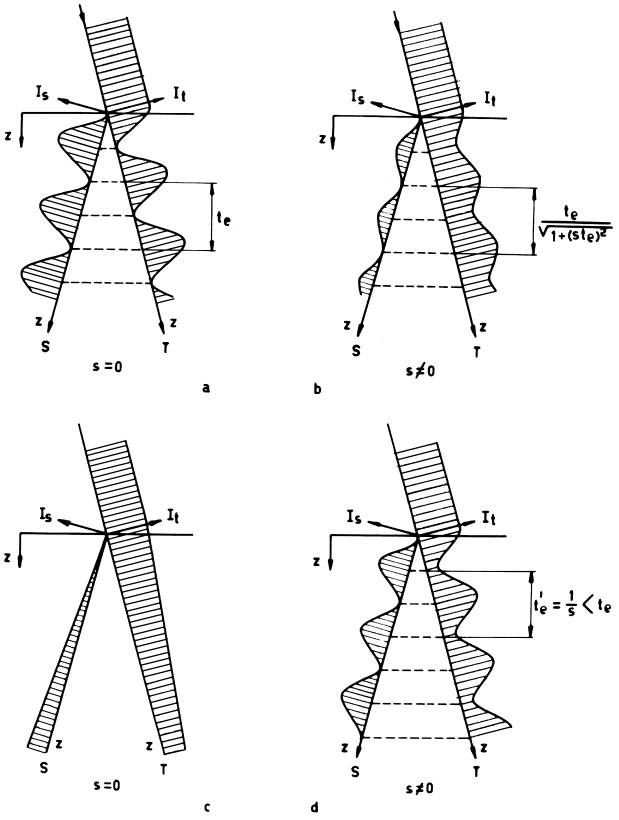



**FIGURE 9** Intensity of scattered and transmitted beams as a function of specimen thickness (Pendellösung effect) according to the dynamical (a, b) and kinematical (c, d) theory. Periodicity is lost in c. [From Amelinckx, S. (1964). "The Direct Observation of Dislocation," Solid State Physics, Suppl. 6, Academic Press, New York.]

in particular, for $s = 0$ it reduces to $t_g$. This gives a simple physical meaning to the extinction distance $t_g$. After having covered a distance $t_g$ in the foil for $s = 0$, the incident-beam intensity is completely transferred into the scattered beam, and vice versa. This repeats periodically with period $t_g$ (Fig. 9). This effect is called the Pendellösung effect and it is responsible for the thickness extinction contours in wedge crystals, as described previously. The dynamical theory (Figs. 9a and b) gives a better description of the observations than does the kinematical theory (Figs. 9c and d). In particular, for $s = 0$ the thickness extinction contours are periodic, whereas the kinematical theory predicts a loss of periodicity.

## VI. MULTIPLE-BEAM DIFFRACTION

Under most conditions, a large number of beams is excited simultaneously; the beams with the smallest $s$ values are most strongly excited. The diffraction pattern is a two-dimensional section of the reciprocal lattice. The beams all interact one with another; there is a continuous transfer of electrons from one beam to all other beams. The simple system of Eq. (5) must now be generalized into a system of $N$ coupled linear differential equations between $N$ wave amplitudes,

$$d\psi/dz = 2\pi i s z S \psi, \tag{7}$$

which is derived from Schrödinger's equation for a periodic potential taking the column approximation into account. Here $\psi$ is a column vector representing the amplitudes of the transmitted and scattered beams, and $S$ a square matrix of which the diagonal elements are the excitation errors $s$ for the different beams. The off-diagonal elements are associated with pairs of beams; they are related to transition probabilities from beam $n$ to beam $m$ and are of the form $1/t_{nm}$ $(n \neq m)$, where $t_{nm}$ is a length (i.e., extinction distance). This set of equations reduces in the two-beam case to Eq. (5). The simple quasi-periodic character for the intensity as a function of depth (4 or 6) is lost. Analytical solutions are possible only in certain symmetrical situations and for a small number of beams. The equations have to be solved numerically using a procedure whereby the crystal is dissected into slices parallel to the foil plane, and the propagation of electrons followed from slice to slice. Efficient computer programs have been developed to perform such calculations in either reciprocal or direct space.

## VII. DIFFRACTION CONTRAST IMAGES

Diffraction contrast results from the spatial dependence of scattered or transmitted intensity along a foil. A single beam produces a diffraction spot to which different parts of the foil contribute differently; however, this is not apparent in an overfocused diffraction spot. A single beam can be selected by means of an aperture in the back focal plane of the objective lens and then magnified by the electron optical system. In this way a highly magnified map of the intensity distribution along the exit face of the foil is obtained in a given diffraction spot (i.e., along a given direction). This map is what we call an image.

Local variations of the orientation of the crystal lattice are revealed in this map as intensity differences according to the rocking curve. A bent perfect foil produces equi-inclination extinction contours or bent contours (see Section IV). Also, local changes in thickness give rise to local intensity variations (i.e., produce contrast). In particular, a wedge-shaped perfect crystal produces thickness extinction contours.

Local composition changes lead to local structure factor differences and, hence, to brightness differences in the image; this is called structure factor contrast.

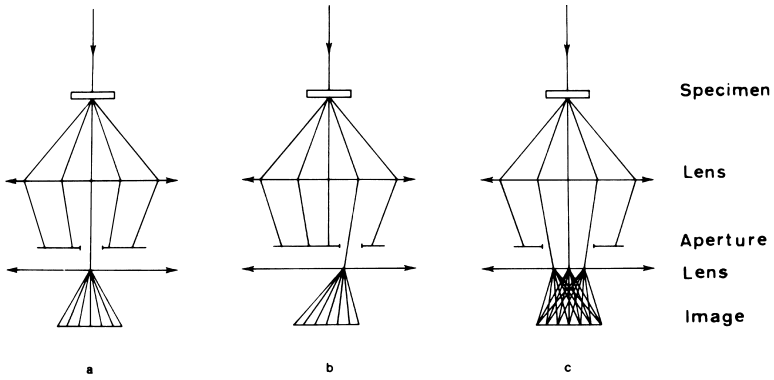



**FIGURE 10** Schematic beam path in an electron microscope illustrating different imaging modes: (a) diffraction contrast bright-field image, (b) diffraction contrast dark-field image, and (c) lattice image mode. [From Van Dyck, D. (1978). In "Diffraction and Imaging Techniques in Material Science," p. 355, North-Holland, Amsterdam.]

Defects, such as dislocations, produce a strain field that causes local lattice curvature and, hence, produce images by strain contrast.

The relative displacement of two crystal parts along a planar interface introduces phase shifts between the electron beams diffracted by the two crystal parts. The interference between these electron beams produces quasi-periodic intensity variations in the scattered as well as in the transmitted beam. In a foil with constant thickness, the intensities of the emerging beams vary with the depth position of the interface in the foil in a quasi-periodic fashion; the quasi-period being the extinction distance $t_g$ (see Sections V and VI).

The selected beam can be either a transmitted or a scattered beam. In the first case the image is called a bright-field image (Fig. 10a) and in the second a dark-field image in the selected reflection (Fig. 10b).

## VIII. COLUMN APPROXIMATION

The possibility to obtain a relatively undistorted image that is in fact a highly magnified part of the diffraction spot is directly related to the small magnitude of the Bragg angles in electron diffraction. As a consequence of this, it is a good approximation to assume that electrons travel through the foil along narrow columns, the lateral dimensions of which are determined by the foil thickness $t$, by the Bragg angles $\theta$, and the upper limit of the lateral spread $w$ given by $w = \theta t$.

It is, therefore, justified to associate a local intensity with the exit point in the back surface of a column, which has sampled a columnar volume of the foil. In a sense the columns can be considered as picture elements (pixels) of the image, which diffract quasi-independently. This procedure is known as the column approximation and its main importance is in computing and interpreting the images of defects.

## IX. MOIRÉ FRINGES

Superposed crystals that differ slightly in either orientation or lattice parameter (or both) produce moiré fringes. The interference between the beam diffracted first by crystal 1 and subsequently transmitted through crystal 2 and the beam transmitted first through crystal 1 and subsequently diffracted by crystal 2 produces dark-field moiré fringes. The bright-field moiré fringes result from the interference between the doubly transmitted beam and the doubly scattered beam.

Let the active diffraction vectors in crystal 1 and 2 be, respectively, $\mathbf{g}_1$ and $\mathbf{g}_2$, which are equal in magnitude but differ slightly in orientation. The moiré fringes are then perpendicular to $\mathbf{\Delta g} = \mathbf{g}_1 - \mathbf{g}_2$ (i.e., roughly parallel to $\mathbf{g}_1 \simeq \mathbf{g}_2$), and their spacing is $\Lambda = 1/|\mathbf{\Delta g}|$; $\mathbf{\Delta g}$ can be considered as the wave vector of the moiré fringes.

The formation of moiré fringes can be simulated by means of the optical analogues in Fig. 11, in which two identical parallel line systems are superposed with a small angular difference. It is clear that much wider coincidence fringes are generated, which are the rotation moiré fringes (Fig. 11b).

The parallel superposition of two line gratings with slightly different spacings, $d_1 = 1/|\mathbf{g}_1|$ and $d_2 = 1/|\mathbf{g}_2|$, also produces coincidence fringes perpendicular to $\mathbf{g}_1 \simeq \mathbf{g}_2$ (i.e., to $\mathbf{\Delta g}$) and with a spacing $\Delta = d_1 d_2/(d_1 - d_2)$. The formation of parallel moiré fringes is demonstrated by means of the optical analogue in Fig. 11a.

Such fringes can be realized in the electron microscope using sandwiches of two thin metal layers grown epitaxially one onto the other. It is clear that if $\mathbf{\Delta g}$ is

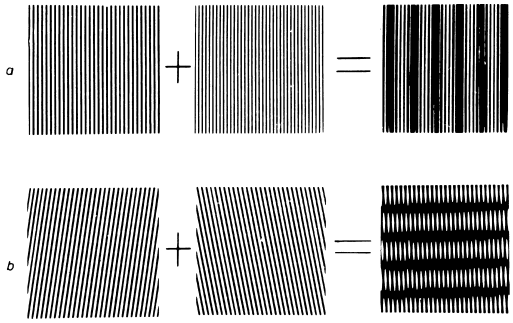



**FIGURE 11** Optical analogues for the production of moiré fringes in the electron microscope: (a) spacing difference moiré fringes and (b) rotation moiré fringes.

small or if $d_1 \approx d_2$, the Moiré spacing becomes large and can easily be resolved in the microscope.

If one of the two components of such a sandwich contains an edge dislocation perpendicular to the foil, the moiré pattern also exhibits a dislocationlike configuration (i.e., one or more supplementary half-fringes) (Fig. 12). The number of supplementary fringes is given by $\mathbf{g} \cdot \mathbf{b}$, where $\mathbf{b}$ is the Burger vector of the dislocation. In a rotation moiré pattern the supplementary half-plane in the moiré pattern is perpendicular to that in the actual dislocation. In a parallel moiré pattern both supplementary half-planes are parallel, and the sign of the moiré dislocation is the same as that of the crystal dislocation if the perfect crystal has a larger lattice parameter than the dislocated one. The sign of the moiré dislocation is the opposite of that of the crystal dislocation if the perfect foil has the smaller lattice parameter.

A stacking fault in one of the components of the sandwich causes a fractional shift of the moiré fringes, which is given by $\mathbf{g} \cdot \mathbf{R}$ ($\mathbf{R}$ is the displacement vector).

Moiré fringes can be considered as a means of producing a geometrical magnification of the crystal lattice and of the defects contained in it. They can be used to derive small differences in lattice spacing of superposed crystals (e.g., of a precipitate particle relative to the surrounding matrix).

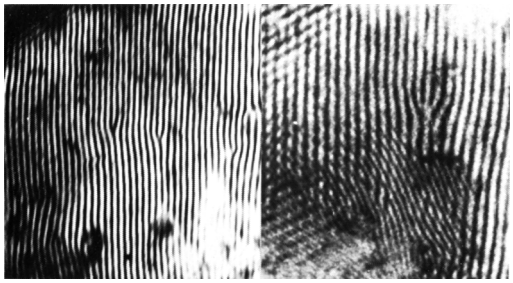


**FIGURE 12** Moiré images of dislocations in metal sandwiches. [Courtesy of P. Delavignette.]

## X. DISLOCATION CONTRAST

The following intuitive reasoning explains the origin of contrast at dislocations and allows us to determine the image side. Let the foil contain an edge dislocation in $E$. In the perfect part of the foil the orientation is such that the transmitted and the scattered beams are comparable in intensity for reflection against the lattice planes shown in Fig. 13. The line width is a rough measurement for the intensity of the beams. The $s$ value of the active diffraction vector that points to the right is positive.

At the left side of the dislocation in $E_1$ the considered lattice planes are sloping in such a way that the Bragg condition is locally better satisfied than in the perfect part of the foil. As a result, locally more electrons will be scattered into the diffracted beam than in the perfect part, and a lack of electrons will be noted all along a line slightly to the left of the dislocation (i.e., a dark line will be observed in the bright-field image). In this approximation a bright line would be observed in the dark-field image. The opposite would apply to the crystal part slightly right of the dislocation, that is, in $E_2$ since the lattice planes are sloping in the opposite sense compared to the orientation in the perfect part. A similar reasoning is applicable to screw dislocations since lattice planes are also inclined in opposite senses left and right of the dislocation line.

The diffraction contrast at dislocations is thus one-sided for $s \neq 0$ and not too small $s$ values, that is, dislocations are observed as dark lines situated slightly on one side of the actual position of the dislocation core. The side of the dislocation on which the image line occurs, referred to here as the image side, is determined by the sign of the

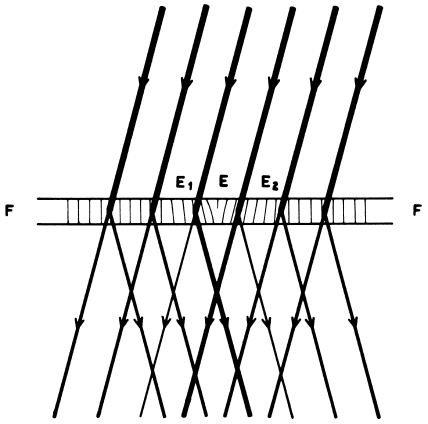



**FIGURE 13** Diffraction contrast at a dislocation in a thin foil (intuitive picture). The thickness of the lines is a measurement of their intensity. [From Amelinckx, S. (1964). "The Direct Observation of Dislocation," Solid State Physics, Suppl. 6, Academic Press, New York.]

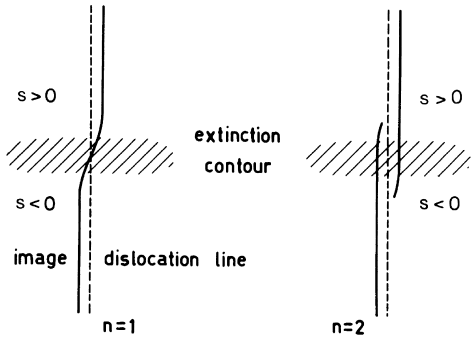



**FIGURE 14** Schematic representation of a dislocation image crossing an extinction contour; note the change of image side for $s > 0$ and $s < 0$. [From Amelinckx, S. (1964). "The Direct Observation of Dislocation," Solid State Physics, Suppl. 6, Academic Press, New York.]

product $p = (\mathbf{g} \cdot \mathbf{b})s$. Accepting the FS/RH convention[1] for the definition of the Burger vector **b** and the previous definition of $s$ and arranging a normal positive print (i.e., one viewed from below the specimen) so as to define a left and right side of the dislocation, then for a dislocation line direction defined from bottom to top on the print, the image will be on the right for $p > 0$ and on the left for $p < 0$. Similarly, for a closed dislocation loop where the sense of the dislocation line is defined by a clockwise circuit, the image will be inside the loop for $p > 0$ and outside for $p < 0$.

Since the actual position of the dislocation line cannot be observed, the image side can be determined only by producing successively a left and a right image. There are two means of achieving this: (1) by producing an image with $s > 0$ and another with $s < 0$ for the same **g** or (2) by producing images for **g** and $-\mathbf{g}$, both with an $s$ value of the same sign (Fig. 14).

The same intuitive reasoning also demonstrates that Bragg reflection from the family of lattice planes that are left undeformed by the dislocation will not reveal the presence of the dislocation (Figs. 15a and b). Thus, the condition for extinction of the image is $\mathbf{g} \cdot \mathbf{b} = 0$. This is an approximation, however, and the lattice planes parallel with the glide plane of an edge dislocation, for instance, satisfy the extinction criterion, but some contrast is observed, which is due to the slight deformation of such planes. In the case of an edge dislocation this slight displacement is perpendicular to the glide plane. This effect is clearly visible for pure edge prismatic loops observed with **g** in the plane of the loop. In the latter case the extinction will be complete only along those parts of the loop where the radial displacement is perpendicular to **g** (Fig. 16), that is, there will be a line of no contrast only perpendicular to **g**.

[1]FS/RH or finish–start/right-hand convention refers to the way in which a Burger circuit is defined around a dislocation line.

From the one-sided nature of the contrast at dislocations it is possible to determine the sign of dislocations. Intuitive reasoning easily demonstrates that changing the sign of the dislocation (i.e., of the Burger vector) changes the image side; the same is true for edge and screw dislocations. This result follows from the change in sign of $p$ as **b** changes sign (Fig. 17).

The problem in determining whether a prismatic loop is due to the precipitation of vacancies or interstitials is equivalent to determining the sign of the dislocation bordering the loop. Several practical methods, based on determining the sign of $(\mathbf{g} \cdot \mathbf{b})s$, have been described. Such methods have been used extensively in the study of radiation damage, of quench defects in metals and alloys, and of loops due to nonstoichiometry.

For small $s$ values, dislocations that are not parallel to the foil plane produce an image that changes perodically with the level in the foil, the period being the effective extinction distance. This is called oscillating or dotted

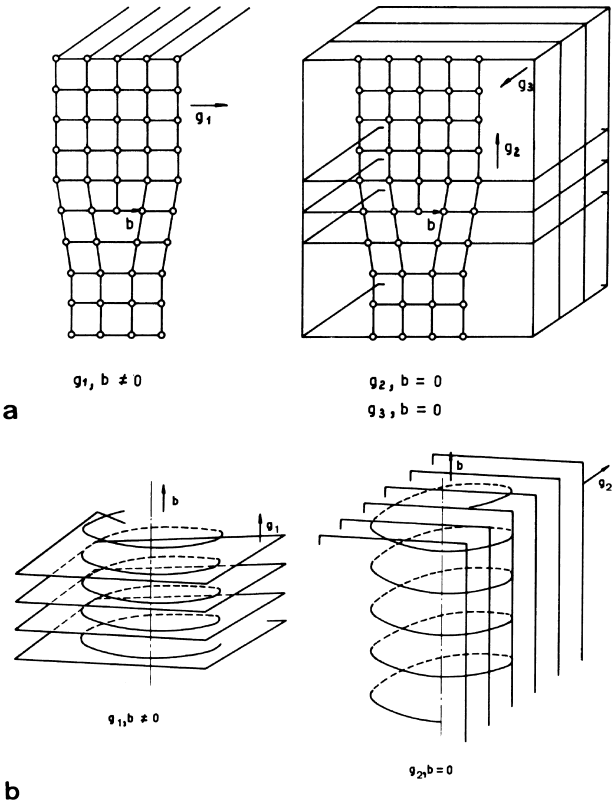



**FIGURE 15** Extinction conditions for dislocations. (a) Edge dislocations: planes perpendicular to $\mathbf{g}_1$ are deformed and planes perpendicular to $\mathbf{g}_3$ remain flat. (b) Screw dislocations: planes perpendicular to $\mathbf{g}_1$ acquire screw shapes and planes perpendicular to $\mathbf{g}_2$ remain flat. [From Amelinckx, S. (1964). "The Direct Observation of Dislocation," Solid State Physics, Suppl. 6, Academic Press, New York.]

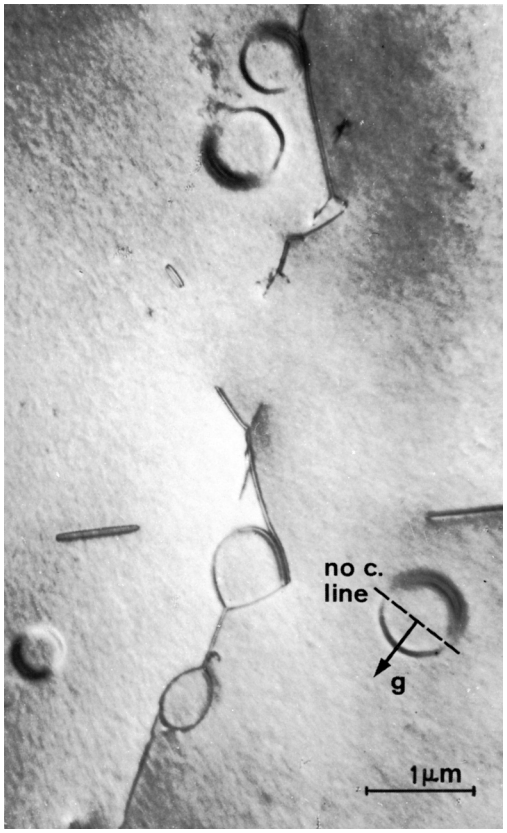



**FIGURE 16** Diffraction contrast of prismatic dislocation loops in the (0001) plane of zinc. The line of no contrast perpendicular to **g** is indicated. [Courtesy of A. Fourdeux.]

contrast (Fig. 18). In sufficiently thick foils where anomalous absorption is important, the oscillations in bright- and dark-field images are similar (i.e., in phase) at the top of the foil and complementary at the bottom of the foil (i.e., in antiphase).

Small dislocation loops or very small precipitates may produce black or bright dots as images. Dislocations seen end-on produce characteristic contrast effects that are due to a large extent to surface relaxation along their emergence points. In cases where the dislocation image is extinct (Figs. 19a and b), the emergence points may still produce contrast (Fig. 19c).

## XI. WEAK-BEAM IMAGING

The image width of defects decreases with increasing $s$. This is a consequence of the fact that the effective extinction distance becomes very small for large $s$ values, as follows from Eq. (6). As a result, the image width of dislocations becomes much narrower. It is therefore possible to increase the resolution of defect images by using the so-called weak-beam method, which consists essentially of making a dark-field image in a weakly excited reflection. The exposure time is correspondingly longer of course. Under these conditions (i.e., large $s$ values), the kinematical theory is a reasonable approximation.

Weak-beam images are used mainly to study the fine structure of dislocations, that is, to study the splitting of perfect dislocations into multiribbons of partials (Fig. 20). The separation of partials is determined, among other factors, by the magnitude of the stacking fault energy. Weak-beam images therefore offer a unique method for the quantitative determination of stacking fault energies.

For large $s$ values the effective extinction distance becomes smaller. Under such conditions planar interfaces produce many fringes, which are sometimes useful, such as for the study of antiphase boundaries in ordered alloys. The extinction distances corresponding to the superstructure reflections, which are needed to image antiphase boundaries, are, in general, large, thus producing only a small number of fringes, unless large $s$ values are used.

## XII. COMPUTER SIMULATION OF DISLOCATION IMAGES

Computer programs have been developed to simulate two-beam dislocation images. Identification of the

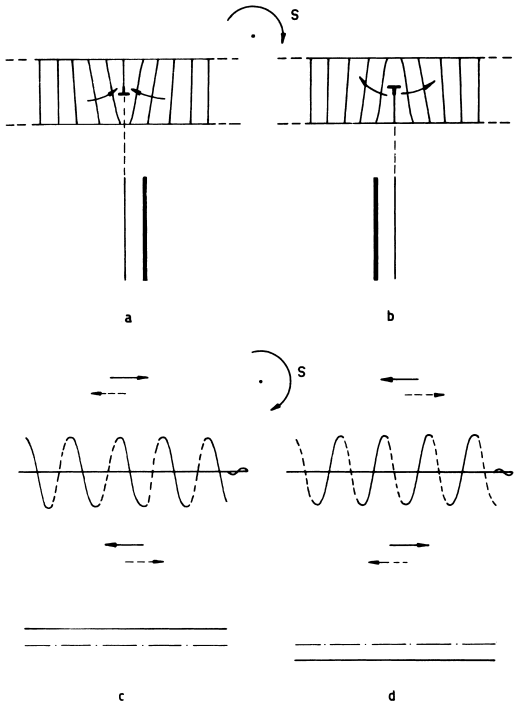



**FIGURE 17** Image side and sign of dislocation. When the sign of the dislocation changes, the image side changes: (a, b) edge dislocations; (c, d) screw dislocations.

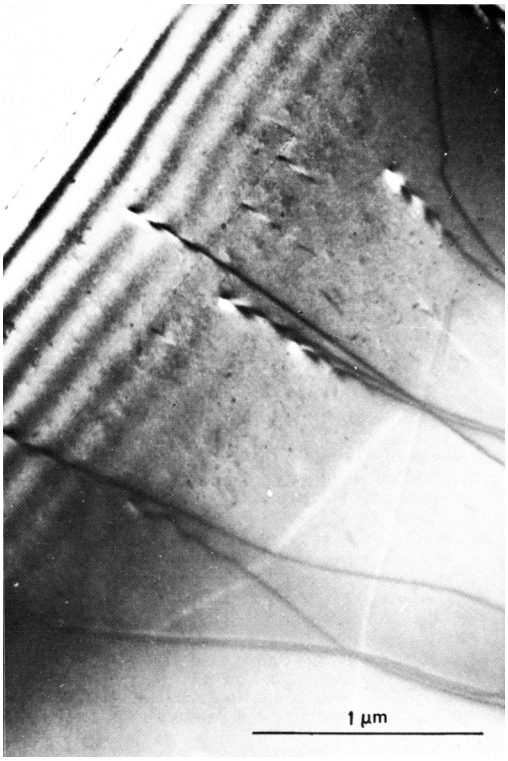



**FIGURE 18** Oscillating contrast at dislocations that are inclined with respect to the foil plane and finally emerge in the surface. [Courtesy of P. Delavignette.]

characteristics of dislocations proceeds through comparison of observed and simulated images. The "strength" of dislocation contrast depends on $n = \mathbf{g} \cdot \mathbf{b}$, which is a small integer for perfect dislocations ($n = 1, 2, 3$), and on the diffraction variables. For partial dislocation $n$ adopts nonintegral values. For $n = \frac{2}{3}$ the dislocation will usually still be visible, whereas for $n = \frac{1}{3}$ visibility becomes questionable.

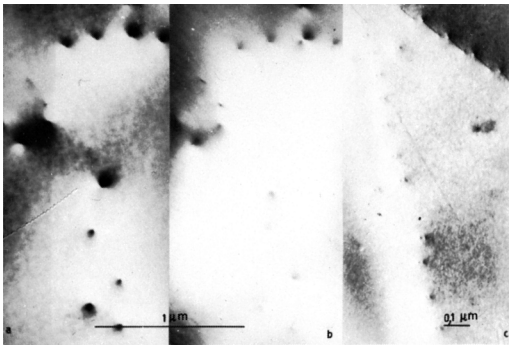



**FIGURE 19** Contrast due to surface relaxation at the emergence points of dislocations in a platinum foil. The dislocations themselves are out of contrast in c; only the surface relaxation is visible.

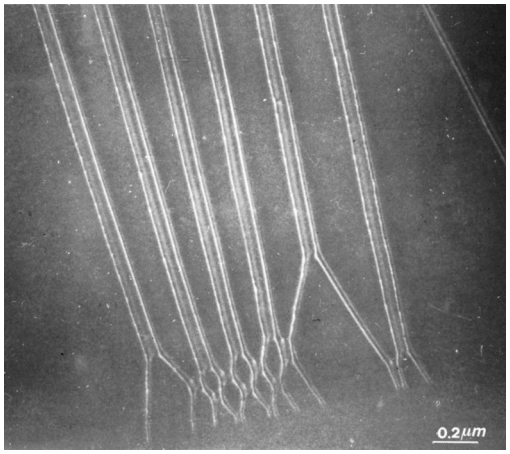



**FIGURE 20** Weak-beam images of ribbons of partial dislocations in the basal plane of $RhSe_2$. [From Amelinckx, S., and Van Landuyt, J. (1976). In "Electron Microscopy in Mineralogy" (H. R. Wenk, ed.), pp. 68–112, Springer-Verlag, Berlin.]

It is possible to determine the magnitude of the Burger vector from the knowledge of $n$, $\mathbf{g}$, and the direction of $\mathbf{b}$.

The images of close-neighboring parallel dislocations with the same Burger vector may be quite different; this is due to the fact that the combined displacement field of the two dislocations produces the contrast. Therefore the image is not the superposition of the two images that would be produced by two isolated single dislocations. This "vicinity" effect is especially striking in layer structures where it extends far along the layer planes as a result of the elastic anisotropy of such materials. In graphite, for instance, triple ribbons containing three partial dislocations with the same Burger vector are frequently observed. Nevertheless, the three partials produce quite different images, and the image of these ribbons is, furthermore, strongly dependent on the sign of $s$.

## XIII. DIFFRACTION CONTRAST AT PLANAR INTERFACES

Different classes of planar interfaces can be distinguished from characteristic image features under two-beam diffraction contrast conditions.

### A. Translation Interfaces

The two parts of the crystal are related by a pure translation described by the constant displacement vector $\mathbf{R}$. If $\mathbf{R}$ is not a lattice vector, such a defect is called a stacking fault (Fig. 21a). If the displacement vector $\mathbf{R}$ is a lattice vector, but not a superlattice vector (e.g., in an ordered alloy) (Fig. 21b), the interface is called an antiphase boundary,

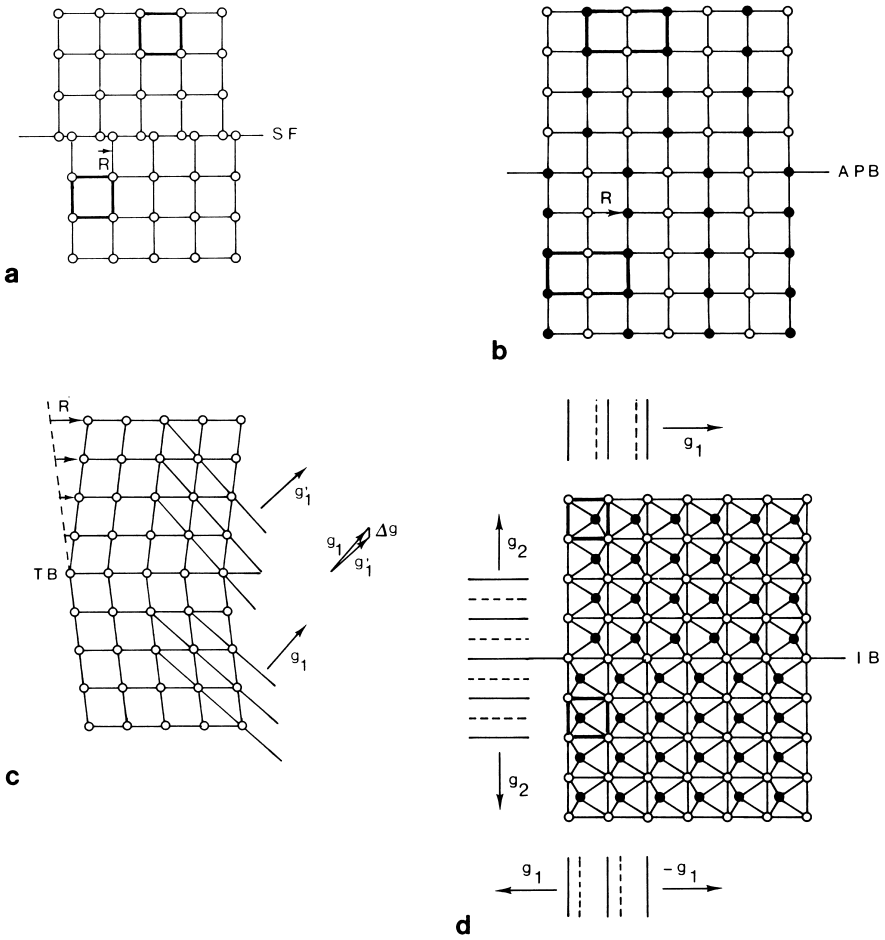



**FIGURE 21** Schematic representation of different planar defects: (a) stacking fault (SF) in elementary crystal, (b) antiphase boundary (APB) in an ordered alloy, (c) coherent twin boundary (TB), and (d) inversion boundary (IB). [From Amelinckx, S., and Van Landuyt, J. (1976). In "Electron Microscopy in Mineralogy" (H. R. Wenk, ed.), pp. 68–112, Springer-Verlag, Berlin.]

but strictly speaking only if **R** is one-half of a lattice vector. If this is not the case, the same terminology may be used although out-of-phase boundary would be more correct.

The lattice planes with diffraction vector **g** in one part of the crystal are shifted with respect to those in the second part of the crystal over a fraction of the interplanar distance given by $\mathbf{g} \cdot \mathbf{R}$. As a result, the electrons diffracted by the second part of the crystal undergo a phase shift over $\alpha = 2\pi\mathbf{g} \cdot \mathbf{R}$ with respect to those diffracted by the first part of the crystal. The phase of the transmitted beam is not affected by this shift. Interference between (1) the beam transmitted by the first and, again, by the second part of the foil $T_1T_2$ and (2) the beam scattered by the first part and, again, scattered in the incident direction by the second part, $S_1S_2^-$, gives rise to periodic variations of the final transmitted intensity with the position of the interface within the foil (Fig. 22). The period is the extinction distance, corrected for deviations from the exact Bragg conditions, $t'_g$ [i.e., as given by Eq. (6)].

The final scattered beam results from the interference between (1) the beam transmitted through the first part and scattered by the second part $T_1S_2$ and (2) the beam scattered by the first part and transmitted through the second $S_1T_2^-$. Also, this beam periodically varies in intensity with the position of the interface in the foil, the period being the same as for the transmitted beam (Fig. 22).

If the interface is inclined with respect to the foil surfaces, one observes a set of fringes with a depth period equal to $t'_g$, in the bright-field image as well as in the dark-field image, and the projected period depends on the inclination of the interface. We now discuss, in some detail, the properties of the images of faults for which $\alpha = \pm\frac{2}{3}\pi$, which occur in close-packed structures.

In foils sufficiently thick for absorption effects to be important, the bright-field image is symmetrical with respect

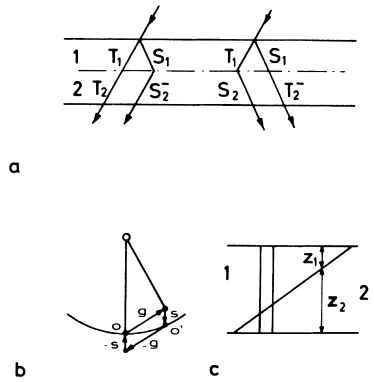



**FIGURE 22** (a) Schematic beam path occurring upon diffraction by a planar interface between crystal parts 1 and 2. $T$ and $S$ are the amplitudes of transmitted ($T$) and scattered ($S$) beams. (The minus sign in the superscript indicates that the sign of $s$ should be changed in the corresponding expression.) (b) Reciprocal space construction illustrating the reversal of the sign of the excitation error $s$ upon reversal of the sense of the diffraction vector **g**. (c) Geometry for an inclined interface; $(z_1 + z_2)$ equals the specimen thickness.

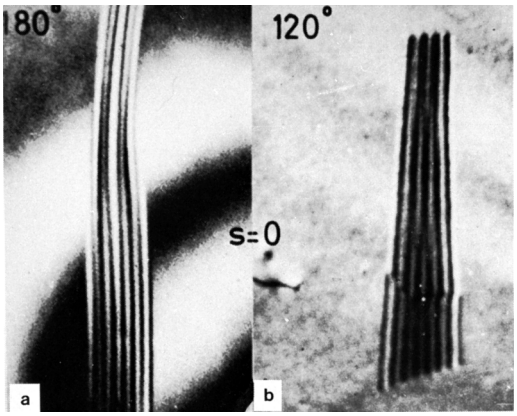



**FIGURE 23** Fringe patterns at planar stacking faults with (a) $\alpha = 180^\circ$ and (b) $\alpha = 120^\circ$.

to the projection of the foil center, but asymmetrical in the dark-field image. In the bright-field image the first fringe at the entrance face is bright if $\sin \alpha > 0$ and dark if $\sin \alpha < 0$; the same is true for the dark-field image. On the other hand, at the exit face the last fringes are opposite in nature in bright- and dark-field images. Table I summarizes these characteristics.

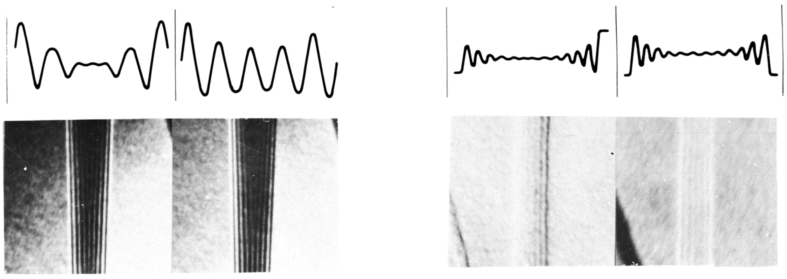
In wedge-shaped foils the fringes are parallel with the intersection lines of the interface and the nearest foil surface. As the foil thickens additional fringes are formed in the center of the foil (Fig. 23b).

Fringes associated with antiphase boundaries, for which $\alpha = \pi$, have somewhat different properties. Bright- and dark-field images are now complementary. The central fringe is bright in the bright-field image and dark in the dark-field image. The fringes are parallel to the foil center rather than with the foil surface. As a result, in a wedge-shaped crystal, new fringes are created at the surfaces of the foil (Fig. 23a). The contrast within the domains on either side of a translation interface is always the same since the lattices are parallel.

It is clear that for these reflections **g**, for which $\mathbf{g} \cdot \mathbf{R}$ is an integer, no fringes will be produced since the lattice planes with diffraction vector **g**, in both crystal parts, are then in register again. As a result, in face-centered crystals where $\mathbf{R} = \frac{1}{6}[11\bar{2}]$ (or the equivalent, $\frac{1}{3}[111]$), reflections

TABLE I. Properties of $\alpha$ and $\delta$ Fringes

| $\alpha = 2\pi \mathbf{g} \cdot \mathbf{R}$ | Bright Field | | Dark Field | | $\delta = s_1 t g_1 - s_2 t g_2$ | Bright Field | | Dark Field | |
|---|---|---|---|---|---|---|---|---|---|
| | First | Last | First | Last | | First | Last | First | Last |
| $\sin \alpha > 0$ | Bright | Bright | Bright | Dark | $\delta > 0$ | Bright | Dark | Bright | Bright |
| $\alpha \neq \pi$ | — | — | — | — | — | — | — | — | — |
| $\sin \alpha < 0$ | Dark | Dark | Dark | Bright | $\delta < 0$ | Dark | Bright | Dark | Dark |



for which $h+k+l=3n$ (i.e., the sum of the indices is a threefold) will not produce any fringes at stacking faults; for the other reflections $\alpha = \pm\frac{2}{3}\pi$. For antiphase boundaries in ordered alloys, $\alpha = \pm\pi$ for superstructure reflections, and $\alpha = 0 \pmod{2\pi}$ for basic reflections. The extinction distances associated with superstructure reflections are mostly large. As a result, antiphase boundaries are usually imaged by a smaller number of fringes than stacking faults in the same foil thickness.

Sometimes weak fringes are visible along stacking faults, and antiphase boundaries for reflections for which $\mathbf{g}\cdot\mathbf{R}$ is an integer. This is a consequence of the fact that the displacement vector is not a simple vector but differs from this by a small vector $\varepsilon$ as a result of relaxation along the interfaces. It is possible to deduce $\varepsilon$ from contrast effects.

### B. Coherent Twin Interfaces

After a phase transition, single crystals are usually broken up in domains. The structures within the domains are related by symmetry operations lost during the transformation. Coherent twin boundaries (Fig. 21c) often result after a displacive transformation. The two parts of the crystal on either side of such a twin boundary are then related by a mirror operation or by a 180° rotation.

In many cases one can alternatively derive one part of the crystal from the other part by means of the displacement field shown in Fig. 21c, that is, the displacement vector has a constant direction and sense but increases in magnitude with distance away from the interface. If the displacement per atom plane is a small fraction of the interatomic distance, the difference $\Delta\mathbf{g}$ of simultaneously active diffraction vectors $\mathbf{g}_1$ and $\mathbf{g}_2$ in the two crystal parts is a small vector (i.e., $\Delta g \ll g$), which is perpendicular to the interface. The two crystal parts diffract simultaneously under two-beam conditions, although with different excitation errors $s_1$ and $s_2$. Also, the extinction distances are, in general, different in the two crystal parts for simultaneously excited reflections. However, in diffraction contrast such interfaces are also imaged as fringe patterns of which the properties are different from those produced by translation interfaces.

The contrast is now determined by $\delta = s_1 t_{g_1} - s_2 t_{g_2}$. The nature, bright or dark, of the outer fringes is given in Table I. The depth period may now be different close to front and back surfaces if the extinction distances in the two crystal parts are significantly different. If $t_{g_1} = t_{g_2}$, the fringe pattern is symmetrical in the dark-field image provided $|\mathbf{s}_1| = |\mathbf{s}_2|$. In general, the domains on either side of the boundary have different contrast. No fringes are produced if $\delta = 0$ (i.e., for reflections for which $\Delta g = 0$).

### C. Inversion Boundaries

Noncentrosymmetrical crystals often contain domains built on a common lattice, but of which the structures are related by an inversion operation (Fig. 21d). Such domains can be made to produce a different brightness in a dark-field image made in a multiple-beam situation along a zone that does not produce a center of symmetry in projection along the zone axis. In the bright-field image the domains always have the same brightness. The method is based on the violation of Friedel's law in multiple-beam situations. The interfaces can also be imaged as fringe patterns of the same nature as those produced by translation interfaces.

### D. Permutation Twins

In certain crystals the symmetry of the structures is lower than that of the lattice. The lattice for an orthorhombic structure may, for instance, be tetragonal (e.g., $\delta - \mathrm{NiMo}$) or a rhombohedral structure may be based on a hexagonal lattice (e.g., $\alpha$-quartz). In such cases the crystal axis may be "permuted" in adjacent regions, domains can then be revealed by structure factor contrast. Although the two domains produce reflections in the same directions, since they are built on the same lattice, the intensities of certain reflections may be significantly different for the two domains. The interfaces are again imaged as fringe patterns, the origin of the contrast being the difference in phase of simultaneously active reflections.

## XIV. IMAGE FORMATION IN AN IDEAL MICROSCOPE

Let the incident electron beam be described by a plane wave of amplitude 1. Diffraction occurs in the object, and electrons emerge from the exit face. The object is characterized by a two-dimensional transmission function $q(x, y)$, which describes the amplitude and phase of the emerging beams at each point $(x, y)$ (Fig. 24). The back surface of the object can be considered as a planar assembly of point sources of spherical wavelets in the sense of Huyghens. The interference between these wavelets generates the diffracted beams in the case of a crystalline specimen and produces a diffraction pattern in the back focal plane of the objective lens. This diffraction pattern can, to a good approximation, be described by Fraunhofer diffraction, because of the relative dimension of the lenses and the electron wavelength and because of the paraxial nature of most of the diffracted beams. This is a consequence of the fact that in electron diffraction, Bragg angles are very small, as already mentioned. Thus the diffraction

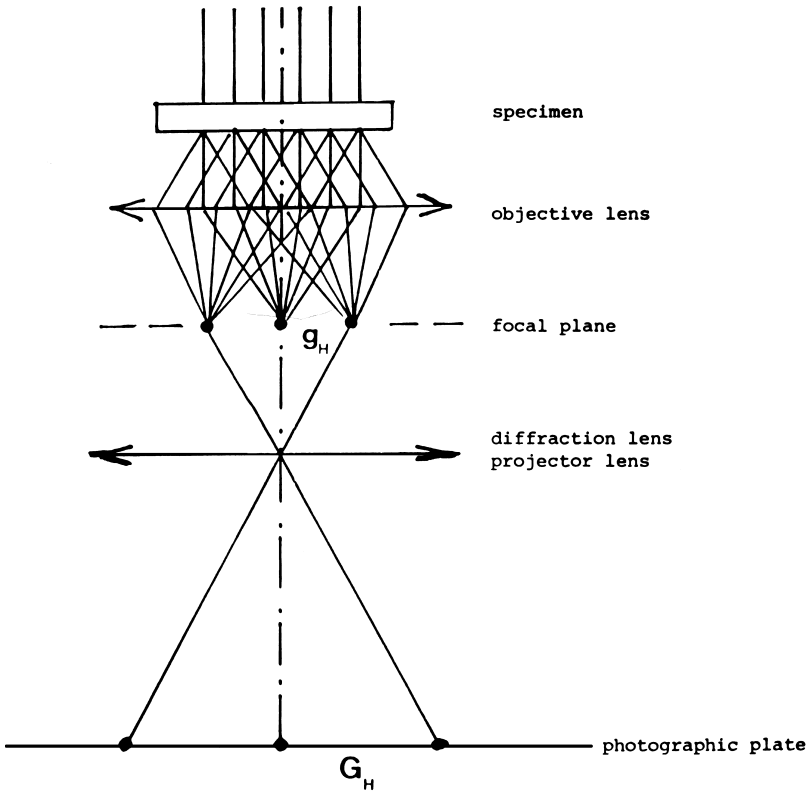



**FIGURE 24** Image formation in an ideal microscope. The diffraction pattern is the Fourier transform of the object and the image is the inverse Fourier transform of the diffraction pattern.

amplitude is the Fourier transform of the function $q(x, y)$. In turn, the diffraction pattern in this back focal plane acts as a source of Huyghens spherical wavelets, which interfere to produce an enlarged image of the transmission function. This image is, again, the Fourier transform of the diffraction pattern. We can, therefore, conclude that the ideal microscope acts as an analogue computer and performs a double Fourier transformation apart from a linear magnification, and, thus, reproduces the object. Unfortunately ideal microscopes are not available and the actual situation is somewhat more complicated.

## XV. IMAGE FORMATION IN A REAL MICROSCOPE

Real microscopes are subject to a number of limitations that induce deviations from the ideal imaging conditions just described.

### A. Spherical Aberration

In real magnetic lenses the paraxial approximation, which leads to point-to-point representation in the Gaussian focal plane (i.e., the focal plane considered in geometrical optics), breaks down. This is due to the fact that the value of sin $\beta$, which enters into the expression for the Lorentz force on a moving charge, can no longer be approximated by the angle $\beta$. This is analogous to approximating sin $\beta$ by $\beta$ in Snell's law for paraxial rays in ordinary optics, in which higher-order terms, up to the third power in $\beta$, are required.

The radius of the disk of confusion in object space resulting from this lens aberration is then given by

$$\rho_S = C_S \beta^3, \tag{8}$$

where $C_S$ is the spherical aberration constant, which has a value between 1 and 10 mm. A typical high-resolution microscope operating at 200 kV has a value of $C_S = 1.2$ mm. As a result of spherical aberration, electron beams inclined at an angle with the optical axis suffer a phase shift $\chi_S$ with respect to the central beam ($\beta = 0$), which is given by

$$\chi_S = 2\pi(\Delta/\lambda),$$

where $\Delta$ is the path difference caused by the beam that does not pass along the axis. From Fig. 25a it can be concluded that $\Delta = \rho_S \sin\beta \simeq \rho_S \beta$ and hence $d\Delta = \rho\, d\beta$

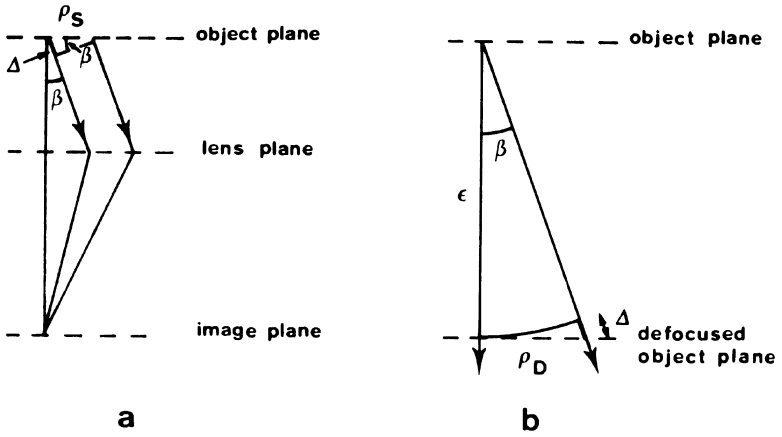



**FIGURE 25** Microscope aberrations. Phase shifts $\chi$ due to (a) spherical aberration and (b) defocusing. [From Amelinckx, S. (1986). In "Examining the Submicron World," p. 71, Plenum, New York.]

and $d\chi_S = 2\pi\rho_S\, d\beta/\lambda = 2\pi C_S\beta^3\, d\beta/\lambda$. After integration from 0 up to the angle $\beta$,

$$\chi_S = (1/2)\pi C_S\beta^4/\lambda. \tag{9}$$

## B. Aperture

The microscope contains an objective aperture that eliminates beams that enclose an angle $\beta$ with the optical axis exceeding an angle $\beta_A$ to reduce the spherical aberration. This imposes a limit to the theoretically achievable resolution called the Abbe limit. A geometrical point is imaged as a circle (the disk of confusion) with the radius

$$\rho_A = 0.61\lambda/\beta_A, \tag{10}$$

which means, in practice, that only points separated by at least this distance in the object can be observed as separate points in the final image.

## C. Defocus

Most high-resolution images are automatically made under conditions where visual contrast is best. It turns out that in the exact Gaussian focal plane the contrast is smallest, at least, for a phase object, that is, an object that changes only the phase of the incident beam. One therefore usually works under somewhat defocused conditions. Also, defocusing causes phase shift and a disk of confusion, which we can estimate with reference to Fig. 25b.

When defocusing the electron microscope by an amount $\varepsilon$, leaving the plane of observation unchanged, the object being situated near the first focal plane, results in an apparent displacement $\varepsilon$ of the object plane (Fig. 25b). One clearly has

$$\rho_D = \varepsilon \sin\beta \approx \varepsilon\beta,$$

and for the value of $\Delta$,

$$\Delta = \varepsilon/(\cos\beta) - \varepsilon \approx 1/2\beta^2\varepsilon,$$

and hence,

$$\chi_D = 2\pi\Delta/\lambda = \pi\varepsilon\beta^2/\lambda. \tag{11}$$

$\varepsilon > 0$ means lens strengthening and $\varepsilon < 0$ means lens weakening.

## D. Chromatic Aberration

As a result of high-voltage instabilities of the microscope, the incident electron beams exhibit a wavelength spread, since $\lambda$ is related to the acceleration potential $V$ by the nonrelativistic approximate relation [Eq. (2)]:

$$\lambda = h(2meV)^{-1/2}. \tag{12}$$

Moreover, variations in the lens currents $\Delta I/I$ also cause aberrations which are of the same nature.

A third origin of aberration is the inelastic scattering in the specimen, which is equivalent to a change in energy of the electrons entering the lens system $\Delta E/E$. The net effect of all these phenomena on the image formation is a spread $\Delta f$ on the focal distance $f$ of the objective lens. The latter is proportional to $EI^{-2}$, assuming that $\Delta E$ and $\Delta I$ are uncorrelated; $\Delta f$ is given by

$$\Delta f = C_c[(\Delta E/E)^2 + 4(\Delta I/I)^2]^{1/2}.$$

The corresponding disk of confusion in object space has a radius

$$\rho_c = \beta\Delta f.$$

The constant $C_c$ is called the chromatic aberration constant. Since instabilities at high voltages and lens currents can be reduced to smaller than $10^{-6}$, $\Delta f$ takes a typical

value of 10 nm, which corresponds with a $C_c$ value smaller than 10 mm.

### E. Beam Divergence

Because of the finite dimensions of the electron source and the condensor lense aperture, the incident beam is somewhat divergent. Under the intense illumination conditions used in high-resolution imaging, the apex angle of the illumination cone may reach a value of the order of $\sim 10^{-3}$ rad.

The influence of incoherent beam divergence on the image can be described as being due to the superposition of independent images (i.e., intensities) corresponding to different incident directions within the divergence cone.

### F. Ultimate Resolution

Apart from the lens imperfections discussed previously, which lead to image blurring and phase shifts, a number of other imperfections occur, but these are unimportant compared to those discussed. Furthermore, resolution also depends on mechanical stability (e.g., vibration and drift); these effects can be eliminated to a large extent by a proper microscope design.

The ultimate resolution is limited mainly by three, up to the present, inevitable phenomena: finite aperture, spherical aberration, and chromatic aberration.

The final disk of confusion has a radius given by

$$\rho = \left(\rho_A^2 + \rho_s^2 + \rho_c^2\right)^{1/2}.$$

Because of the difference in angular dependence of the different aberrations, it turns out that in present-day high-resolution electron microscopes, chromatic aberration has only a relatively small influence. The limiting factor at small angles ($\beta < 5 \times 10^{-3}$ rad) is the aperture, whereas in the range $\beta > 5 \times 10^{-3}$ rad, the limiting factor is the spherical aberration. (This is true for $E = 100$ kV, $C_s = 8.2$ mm, and $C_c = 3.9$ mm.)

The curves $\rho_A$ and $\rho_s$ versus $\beta$ [i.e., Eqs. (8) and (10)] have opposite slopes. There is therefore a minimum value for $\rho$, which occurs for $\partial\rho/\partial\beta = 0$, where $\rho = (\rho_A^2 + \rho_s^2)^{1/2}$. This minimum, which corresponds to the optimum compromise $\beta_0$ between spherical aberration and aperture effects, occurs for $\beta_0 = (0.61\lambda/C_s\sqrt{3})^{1/4}$. The corresponding radius of the confusion disk is then $\rho_0 = 0.9\lambda^{3/4}C_s^{1/4}$. Representative values are $\beta_0 = 5 \times 10^{-3}$ rad and $\rho_0 \approx 0.5$ nm. This expression makes it clear that more can be gained in resolution by decreasing the wavelength (i.e., by increasing the accelerating voltage) than by decreasing $C_s$.

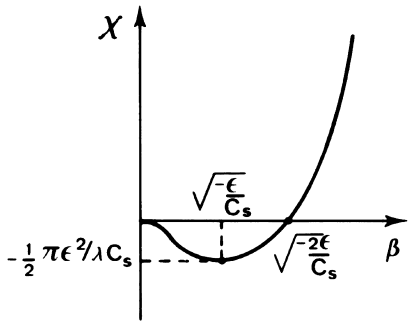



**FIGURE 26** Dependence of the phase shift $\chi$ on the angle $\beta$. [From Amelinckx, S. (1986). In "Examining the Submicron World," p. 71, Plenum, New York.]

### G. Phase Shift

We have pointed out that spherical aberration and defocus cause phase shifts of the nonaxial electron beams with respect to the axial beam. These phase shifts depend on $\beta$ in the following manner:

$$\chi(\beta) = \left(\tfrac{1}{2}\pi C_s\beta^4 + \pi\varepsilon\beta^2\right)/\lambda. \qquad (13)$$

Note that $\chi = 0$ for the defocus value

$$\varepsilon = -\tfrac{1}{2}C_s\beta^2.$$

This value still depends on $\beta$ can be approximately satisfied in only a limited range of $\beta$ values. It is therefore necessary to take the phase shifts into account when performing image calculations. The general aspect of the curve is as presented in Fig. 26.

## XVI. IMAGE FORMATION OF A WEAK-PHASE OBJECT

The amplitude distribution in the back focal plane of the objective lens is given by the Fourier transform ($F$) of the object function. In the case of a crystalline specimen the object function is the electron wave function at the exit face of the thin foil. The amplitude distribution in the diffraction pattern is the Fourier transform of the wave function. The final image amplitude is the Fourier transform of the diffraction amplitude. However, the electrons are now moving in a lens system and therefore undergo the phase shifts $\chi(\beta)$ discussed previously. Moreover, an aperture is limiting the number of beams transmitted through the system. This can be taken care of by introducing an aperture function in the plane of the diffraction pattern. This function is 1 over the surface of the aperture and 0 outside of this.

The main features of the image formation can be illustrated in the simplest case, where the specimen can be assimilated with a weak-phase grating. Since the wavelength of the electron is different in vacuum, and in the specimen the passage of the electron beam through the specimen causes a phase shift that can be written $\chi(x, y) = \sigma\phi(x, y)$, where $\phi$ is the projected lattice potential along the propagation direction of the electrons, $x$ and $y$ are coordinates in the specimen plane, and $\sigma = \pi/\lambda E$ ($\lambda$ is the wavelength of the electrons in vacuum, and $E$ the accelerating potential). The object function is then

$$q(x, y) = e^{i\sigma\phi(x,y)} \simeq 1 + i\sigma\phi(x, y). \qquad (14)$$

The image amplitude $U(x, y)$ obtained by Fourier transformation followed by the inverse Fourier transformation then becomes of the form

$$U(x, y) = 1 + \sigma\phi(x, y)\sin\chi + i\sigma\phi(x, y)\cos\chi, \quad (15)$$

where $\chi$ is supposed to be a constant. In reality, $\chi$ depends on $\beta$ (i.e., on $x$ and $y$), but we shall show that the imaging conditions require that $\sin\chi$ is at least approximately constant to obtain a directly interpretable image.

In the particular case where $\sin\chi = -1$ (and, thus, $\cos\chi = 0$), the intensity distribution (i.e., the image) becomes

$$I = UU^* \qquad \text{or} \qquad I(x, y) = 1 - 2\sigma\phi(x, y), \quad (16)$$

which clearly has a direct relationship with the object represented by its projected potential $\phi(x, y)$. The image contrast, defined as $(I - I_0)/I_0 \simeq 2\sigma\phi(x, y)$, turns out to be directly proportional to the projected potential.

If one could make $\sin\chi = +1$, one would obtain

$$I(x, y) = 1 + 2\sigma\phi(x, y). \qquad (17)$$

The intensity is now larger for a larger projected potential. The situation is somewhat like positive and negative phase contrast. The lenses have introduced phase shifts of $\pi/2$, similar to what the quarter-wavelength ring does in optical phase contrast microscopy. The lens aberrations are exploited to produce phase contrast that would be absent in a perfect microscope.

## XVII. OPTIMUM DEFOCUS IMAGES

If the image is to be a "faithful" representation of the projected potential, then $\sin\chi \approx \pm 1$, not just for a single beam but for as many diffracted beams contributing to the image as possible.

The value of $\sin\chi$ will not vary rapidly in the vicinity of a stationary point of $\chi$ (i.e., a minimum or a maximum). We have found that $\chi$ adopts a stationary value $\chi = -\pi\varepsilon^2/2\lambda C_s$ for $\beta = (-\varepsilon/C_s)^{1/2}$ (Fig. 26). Since this value of $\chi$ is essentially negative, we cannot satisfy simultaneously the requirement $\sin\chi = +1$ (i.e., $\chi = \pi/2$), but we can do so for $\sin\chi = -1$ (i.e., $\chi = -\pi/2$). It is sufficient to choose the defocus $\varepsilon$ in such a way that $-\pi\varepsilon^2/2\lambda C_s = -\pi/2$ [i.e., $\varepsilon_s = -(\lambda C_s)^{1/2}$]. Since the $\sin\chi$ function is stationary around $\chi = \pm\pi/2$, the $\sin\chi$-versus-$\beta$ curve will present a flat part in the region of $\beta = (-\varepsilon/C_s)^{1/2}$ provided $\varepsilon = \varepsilon_s$. The defocus value $\varepsilon_s = -(\lambda C_s)^{1/2}$, which corresponds to the optimum imaging conditions of a phase grating, is called the Scherzer defocus, and the quantity $(\lambda C_s)^{1/2}$ is used as a unit of defocus. A more complete expression is $\varepsilon_s = -\frac{4}{3}[\lambda C_s]^{1/2}$.

The dependence of $\sin\chi$ on the diffraction angle $\beta$ is represented in Fig. 27 for a typical situation close to the Scherzer defocus. It is a rapidly oscillating function and therefore it is not possible to fulfill the required condition for all $\beta$ values. The curve in Fig. 27, which was drawn for the Scherzer defocus value of $-210$ nm in this particular case, does exhibit a region where $\sin\chi$ is approximately $-1$ as required. Beams that are diffracted in this angular range give rise to an image that is a direct representation of the object. If not enough beams can be passed through this "window" in the $\sin\chi$ curve, the image may be rudimentary in the sense that it can give true detail only up to some maximum spatial frequency. This limiting frequency corresponds roughly with the $\beta$ angle for which $\sin\chi$ goes the first time through zero at the Scherzer focus, that is, $\beta_{max} = (-2\varepsilon/C_s)^{1/2}$; for $\chi(\beta_{max}) = 0$ with $\varepsilon = \varepsilon_s$, this becomes $\beta_{max} = 2(\lambda/3C_s)^{1/4}$, and the radius of the corresponding disk of confusion is $\rho = (\lambda/2)^{3/4}C_s^{1/4}$, to be compared with the expression just given. Figures 27a and b show two image transfer functions for two instruments; the advantage of a high voltage becomes quite apparent from the width of the window.

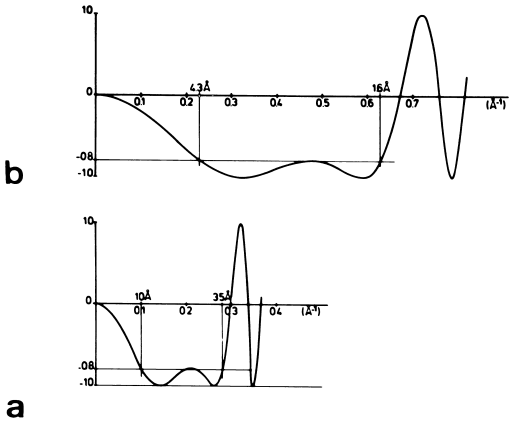



**FIGURE 27** Image transfer function $\sin\chi(\beta)$ for the Scherzer defocus. Accelerating voltage: (a) 100 kV ($C_s = 4.2$ mm, $\Delta f = 720$ Å) and (b) 1000 kV ($C_s = 1.4$ mm, $\Delta f = 860$ Å). [From Amelinckx, S. (1986). In "Examining the Submicron World," p. 71, Plenum, New York.]

It is worth noting that the image (i.e., the positive or the negative deviations from the background) is proportional to the projected potential $\phi(x, y)$.

## XVIII. LATTICE IMAGES

In images made using the diffraction contrast mode, only one beam is used. A quite different type of image is obtained when admitting more than one beam through the selector aperture. In the simplest situation, two beams, usually the transmitted beam and one scattered beam, are selected and then made to interfere. The resulting interference pattern behind the exit surface of the crystal consists of straight sinusoidal fringes with a spacing equal to the interplanar spacing corresponding to the chosen reflection. Their formation is illustrated in Fig. 28, where the transmitted and a single scattered beam are represented, assuming the Bragg condition to be satisfied exactly. The successive planar wave fronts (spacing $\lambda$) of maximum elongation associated with the two beams ovelap in the space behind the foil and produce maxima (and minima) in a set of parallel planes. Although the plane waves are propagating, these parallel planes of maximum (or minimum) amplitude form a stationary pattern that can be observed after magnification by the electron optical system. Elementary geometrical considerations show that the spacing between the planes in this pattern is equal to the spacing of the lattice planes for which the Bragg condition is satisfied. The direction of the fringes is perpendicular to the acting diffraction vector and thus parallel with the lattice planes; their intensity distribution is sinusoidal since they are two beam interference fringes. The observed fringes represent admittedly rudimentary images of lattice planes, in terms of an intensity variation of the electron beam. They can also be considered to represent one Fourier component of the lattice potential.

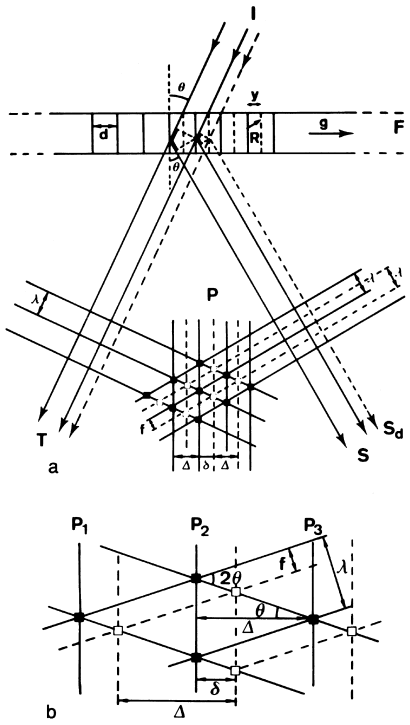



**FIGURE 28** Intuitive model for the formation of lattice fringes in (a) a perfect crystal and in (b) a crystal containing a stacking fault. [From Amelinckx, S. (1986). *J. Electron Microsc. Tech.* **3,** 131.]

In a sense, the crystal foil acts in this simple case on the incident electron beam in very much the same way that certain optical devices act on a light wave to produce interference fringes by beam splitting (the Fresnel biprism or mirror experiment).

If the crystal contains a planar defect (such as a stacking fault, an antiphase boundary, or a discommensuration wall with displacement vector $\mathbf{R}$), an intersecting set of lattice fringes undergoes a lateral fractional shift along the intersection line, given by $\mathbf{g}\cdot\mathbf{R}$, where $\mathbf{g}$ is the acting diffraction vector. The model allows a simple understanding of this effect. Suppose that in the second part of the crystal the lattice planes are displaced over a distance $y$ with respect to the position in the first part, so as to occupy the dotted position (Fig. 28). This shift will not affect the phase of the transmitted beam but it will cause a phase shift of the scattered beam; the path difference $f$ is given by $f = 2y \sin\theta$. This quantity also determines the relative phase shift of the wave fronts diffracted by the displaced part with respect to that diffracted by the undisplaced part. The stationary interference pattern formed by $T$ and $S_\mathrm{d}$ (dotted wave front) will now be displaced sideways over a distance $\delta$ (into the dotted position). From simple geometrical considerations one can conclude the $\delta/\Delta = (\mathbf{g}\cdot\mathbf{R})/d$, which is the relation used in practice to determine the displacement vector $\mathbf{R}$ of stacking faults and of antiphase boundaries from the fringe shifts along the trace of the interface.

Likewise every experimental parameter or instrumental factor (e.g., defocusing and beam tilt) that influences the relative phase of $T$ and $S$ will also shift the fringe positions. It is, in general, not possible to associate a fringe position with a plane in the crystal structure. The fringe spacing $1/g$ has a structural significance—it is directly related to the interplanar spacing in the crystal.

If the foil is oriented exactly perpendicular to the incident beam, the reflection $+\mathbf{g}$ and $-\mathbf{g}$ are both excited to the same extent (the same $s$ value). Lattice images can now be obtained by selecting the three beams, $-\mathbf{g}$, 0, and $+\mathbf{g}$, which exhibit fringes with a spacing $1/g$ as well as $\frac{1}{2}g$ with a different brightness (Fig. 29). The latter fringes (the second harmonic) arise as a result of the interference between

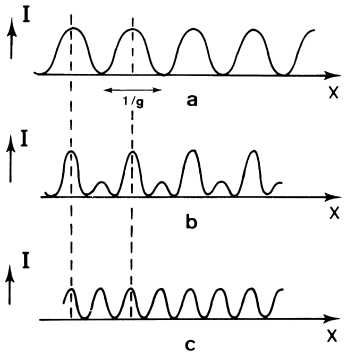



**FIGURE 29** (a) Two-dimensional interference fringes due to $(0, g)$. (b) Three beam fringes, due to $-g, 0$, and $+g$ (asymmetrical incidence). (c) Symmetrical incidence. [From Amelinckx, S. (1986). In "Examining the Submicron World," p. 71, Plenum, New York.]

the beams $+\mathbf{g}$ and $-\mathbf{g}$. These fringes are remarkable in that the angle-dependent aberrations of the microscope do not disturb the image formation since the interfering beams $+\mathbf{g}$ and $-\mathbf{g}$ enclose the same angle with the optical axis of the instrument. Fringes of this type are sometimes used by microscope manufacturers to demonstrate the capabilities of their instruments, but they must be evaluated having in mind this latest remark.

In cases where a number of reflections in a linear array $(\ldots -2g, -g, 0, +g, +2g \ldots)$ are excited and selected, the fringes are still straight but their profile becomes more and more complicated since higher harmonics now play a role.

If several beams are admitted through the selector aperture, they each contribute one (or more) Fourier components (i.e., planar sinusoidal image waves or intensity waves). The electron microscopic image can be considered as being due to the superposition of these different "image waves," one corresponding to each diffraction vector.

The wave vectors of these waves are given by the vectors joining the diffraction spots, corresponding to the different admitted interfering beams, also among each other as a consequence of multiple scattering. An image wave has an amplitude proportional to the intensity of the diffraction spot, which corresponds to the considered wave vector. In this simple case where only the central beam 0, one first-order $\mathbf{g}$, and one second-order $2\mathbf{g}$ reflection in a linear arrangement are admitted, one obtains straight fringes with a periodic intensity distribution containing two Fourier terms, one with period $1/g$ and the second with period $1/2g$, with corresponding vectors $\mathbf{g} = 0\bar{A} = A\bar{B}$ and $2\mathbf{g} = 0B$. The more beams used in the linear arrangement, the more Fourier components contribute, and the more detail can be imaged.

The generalization to two dimensions is obvious. For the simple diffraction pattern where, apart from the direct beam, only four scattered beams forming a square are admitted, one obtains the superposition of four beams (i.e., four sets of fringes with wave vectors $\mathbf{g}_1$, $\mathbf{g}_2$, $\mathbf{g}_3$, and $\mathbf{g}_4$) (Fig. 30). Moreover, also the higher harmonics (i.e., $\mathbf{g}_5 = \overline{AC}$ and $\mathbf{g}_6 = \overline{BD}$) are inevitably present, and, finally, $\mathbf{g}_7 = \overline{AB}$, $\mathbf{g}_8 = \overline{BC}$, $\mathbf{g}_9 = \overline{CD}$, and $\mathbf{g}_{10} = \overline{DA}$ are produced. The superposition of all these waves (i.e., their interference) is sufficient to produce a rudimentary image revealing the lattice without structural details on a subunit cell level, unless the structure is very simple (e.g., a face-centered-cubic element). In the particular case represented in Fig. 30 the number of beams is sufficient to represent the structure. As more beams are admitted, the structural detail revealed in the image becomes finer. There is, of course, a limit to the detail that can be represented, which is imposed by the width of the window through which the beams can be transmitted in the correct phase relationship, and this in turn is determined by the resolution of the microscope.

This consideration implicitly assumed that all considered waves interfere "in phase" (i.e., in the correct phase relationship) with the incident beam and among themselves.

Unfortunately, with increasing order of the Fourier components, corresponding beams enclose angles of increasing magnitude with the optical axis. As long as we make use of only beams that pass through the "window" or "plateau" in the image transfer function (i.e., the sin $\chi$-versus-$\beta$ curve), the different components interfere with the correct phase relationship and hence produce a directly interpretable image for a properly chosen defocus value (the Scherzer defocus). For high-resolution studies it is important to have an instrument with a wide plateau in the image transfer function and to eliminate the beams outside of it by an aperture. From Fig. 27 the advantages of using a high voltage become apparent in this respect.

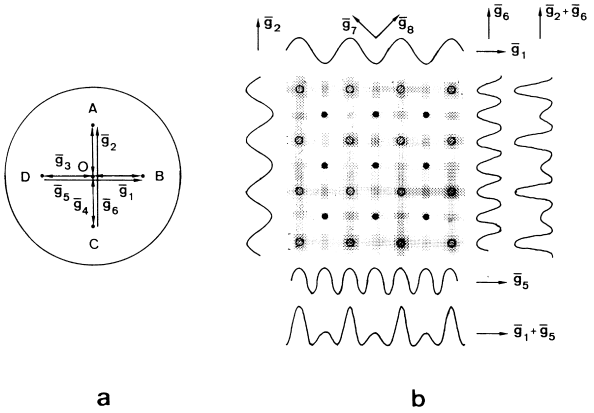



**FIGURE 30** Two-dimensional lattice image formation. (a) Reflections used in imaging along a centered square zone and (b) Fourier components giving rise to the image.

One can also use beams corresponding to $\beta$ angles outside of the window, but then the resulting image is no longer directly interpretable except in very special cases, such as in the aberration-free focus method. For simple crystals (e.g., Si) with a small unit cell it is possible that a number of reflections outside of the plateau still keep the correct phase by using a small number of beams and by choosing defocus and an appropriate $C_s$ value. An image is then obtained that represents the true structure with a resolution that may exceed the point resolution.

The interpretation usually proceeds by the trial-and-error method, which consists of comparing the calculated image for a given structural model to the observed image. In this calculation the phase shifts introduced by the lens system must be taken into account properly. Usually the calculation is made for different defocus values and for different specimen thicknesses, since we know that these parameters affect the phases of the different Fourier components and, hence, the final image.

A number of computational methods are in use. Examples of the application of this method are reproduced in Fig. 31. The calculated images have been plotted on a cathode-ray screen, which simulates images of the same nature as the ones observed in the microscope.

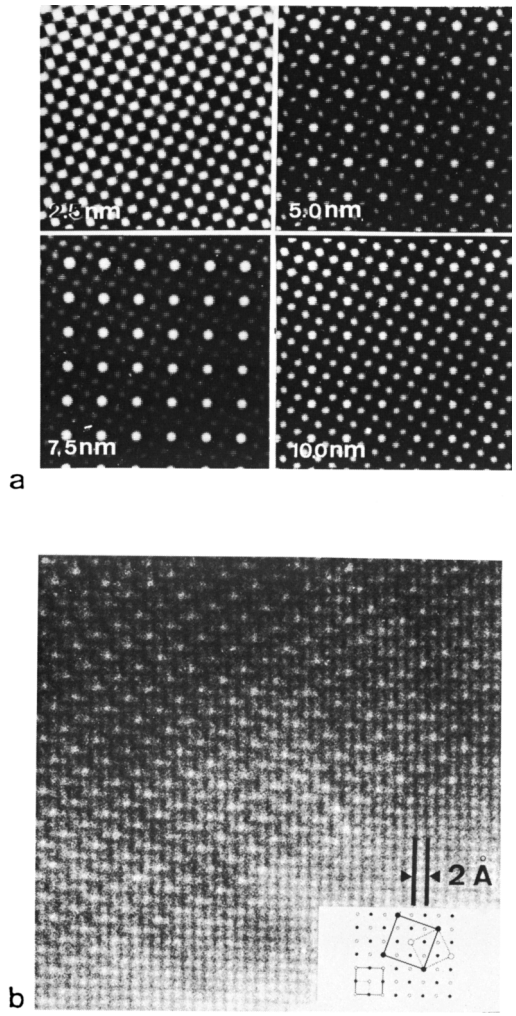



**FIGURE 31** Computed and corresponding experimental structure image of $Au_4Mn$. (a) Four computed images at optimum defocus but for different thicknesses. The left-bottom corner was occupied by a manganese column. The brightest dots are thus located at manganese columns. (b) Experimental image: Note that in the thick part only manganese columns produced bright dots, whereas in the thinnest part all atom columns produce bright dots. The inset shows a model of the tetragonal structure. [From Van Dyck, D., Van Tendeloo, G., and Amelinckx, S. (1982). *Ultramicroscopy* **10,** 263–280.]

## XIX. IMAGING MODES

### A. One-Dimensional Images

If only a one-dimensional representation of the structure is required because the structure is a long-period one-dimensional superstructure, one can use the following imaging techniques (Fig. 32a).

#### 1. Mode 1

One can select two neighboring superstructure reflections belonging to the same basic spot in a row of spots passing through the origin (i.e., in a central row). One uses only one Fourier component, and consequently the image reveals only the long spacing.

#### 2. Mode 2

If all superstructure reflections belonging to a central row are used, excluding the basic reflections in the row, one obtains the distribution of long spacings.

#### 3. Modes 3 and 4

If one selects basic reflections as well as superlattice reflections in a central row (circle 3) or a noncentral row (circle 4), one images also the set of lattice planes of the basic structure that is parallel with the periodic interfaces that produce the superstructure. Any variability of the long spacing is now imaged in terms of the spacing of the basic lattice.

#### 4. Mode 5

One can also use a sequence of superlattice reflections from a noncentral row. This is a useful mode if one wants to reveal polysynthetic subunit cell twinning.

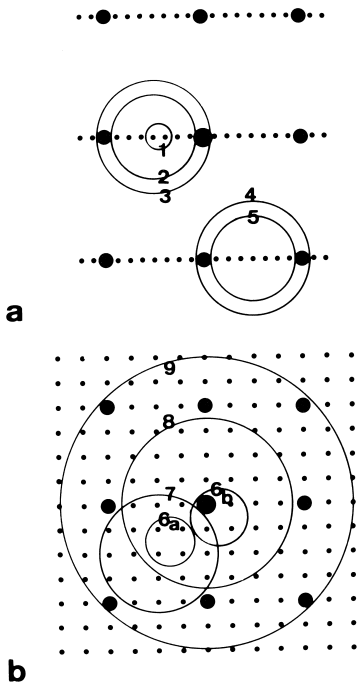



**FIGURE 32** Imaging modes: The figure represents the diffraction spots admitted through the selector aperture. (a) One-dimensional lattice fringes and (b) two-dimensional lattice fringes. [From Amelinckx, S. (1986). In "Examining the Submicron World," p. 71, Plenum, New York.]

## B. Two-Dimensional Images

For Modes 6–9 see Fig. 32b.

### 1. Mode 6

If only the superlattice needs to be imaged, it is sufficient to include only pairs of neighboring superlattice spots in two directions. This can be done either in the dark field (6a) or in the bright field (6b). In the latter case the image is formed by the direct beam and the first shell of superlattice reflections around it; the contrast is usually lower than for 6a.

### 2. Mode 7

All superlattice reflections are selected that are present within one mesh of the reciprocal lattice of the basic structure, excluding the direct beam and the basic reflections. This is clearly a multiple-beam dark-field image. The Fourier components contributing to the image are, in general, just sufficient to locate the positions of the minority atoms.

### 3. Mode 8

Except for the direct beam, all other basic reflections are excluded, but as many shells of superlattice reflections as feasible are admitted. The number of Fourier components required to image the columns of minority atoms is now redundant, however, such columns will therefore be imaged as sharper dots than in mode 7.

### 4. Mode 9

All beams originating from the basic as well as from the superstructure reflections are selected, provided they do not correspond to spacings that are smaller than the instrumental resolving power of the microscope. For most current instruments this means that up to the first or, possibly, up to the second shell of reflections, due to an FCC matrix, can usefully be included. The face-centered-cubic matrix will now be prominently revealed in the thin parts of the specimen, whereas in the thicker part the superstructure will be revealed.

Examples of high-resolution images are reproduced in Section XXIV.

## C. High-Resolution Imaging Interpretation and Simulation

Image interpretation consists in relating an atomic structure model to an HRTEM image. A very thin foil acts as a two-dimensional phase grating; in such a foil the local image brightness is directly related to the local projected lattice potential and hence to the structure mentioned in Section XVI. Unfortunately this in true only for extremely thin specimens in which multiple diffraction is not taking place. In real specimens the image is the result of the dynamical interaction between the numerous diffracted beams, whose amplitudes also depend on the foil thickness. Moreover, the microscope introduces angle-dependent phase shifts between these beams, which depend on the focus and on instrumental parameters ($C_s$, $C_c$, beam divergence, etc.) as discussed above.

Image interpretation proceeds mostly by "trial and error," in much the same way as structure determination by X-ray diffraction in its early days. The image of a structure is computer simulated, then compared with the observed image, and the model is refined until the correspondence between the observed and the simulated image is judged satisfactory. This similarity can be quantified by a "goodness-of-fit" criterion similar to that used in X-ray diffraction.

The most frequently used simulation method is the "multislice" method. The specimen is dissected into slices limited by planes parallel to the foil plane that are thin enough so that each slice can be considered a pure phase grating. The structure in each slice is represented by its projected potential; the slice then acts as a two-dimensional phase grating. The specimen now consists of

a succession of parallel two-dimensional phase gratings, separated by layers of vacuum with a thickness equal to the slice thickness. Scattering of the electrons is followed slice after slice by computing the electron wave amplitude, alternatingly taking into account the phase shift due to the propagation between two gratings, followed by computing the diffracted amplitude by the successive gratings. The latter operation amounts to a Fourier transformation as shown above. In practice, the slices can be taken to be one unit cell thick. For each slice the exit waves of the previous slice act as the input waves. The electron waves emerging from the back surface of the specimen are subsequently assumed to suffer the angle-dependent phase shifts introduced by the microscope before interfering to form the final image. Suitable computer programs using the multislice algorithm are available commercially.

More recently developed methods proceed by the direct retrieval of the projected structure. The main problem with direct retrieval is similar to the phase problem in X-ray diffraction; it is a consequence of the fact that only intensities are recorded in the image (and in the diffraction pattern), not amplitudes, unless use is made of holographic methods.

A recent retrieval method relies on the use of a sequence of images taken at closely spaced foci (focus variation method). This allows us to eliminate the effects of the microscopic optics and, thus, to obtain the corrected wave function at the exit face of the specimen. Using an analytical approach, based on the channeling model, then allows us to obtain the projected structure.

## XX. SCANNING ELECTRON MICROSCOPES

So far we have discussed conventional transmission electron microscopy (CTEM), in which the electron probe is a parallel stationary beam incident along a fixed direction with respect to the specimen and in which the transmitted or scattered beam(s) produces the image on a recording medium (film, channel plate-CCD camera, etc.).

In a second class of electron optical instruments (so-called scanning electron microscopes), a fine electron probe is scanned across the specimen and the signal of interest, produced locally by the probe, is selected, detected, amplified, and displayed by modulating the intensity of the electron beam of a TV monitor which is scanned synchronously with the probe. The signal can be observed either in backscattering, as in conventional scanning electron microscopy (CSEM or simple SEM), or in transmission (STEM). In this chapter we consider only the latter case and compare it to CTEM (Fig. 33).

In a scanning transmission electron microscope (STEM) the signal often consists of the transmitted (or scattered) electron beam; however, other signals (e.g., X rays) can also be detected, even in parallel with the electrons, provided that adequate detectors are available on the instrument.

In CTEM the achievable resolution is determined mainly by the quality of the imaging optics *behind* the specimen; in STEM the resolution limit is determined mainly by the probe size (which is usually of the order of 1 nm or less), i.e., by the probe forming optics *ahead* of the specimen.

In STEM the magnification is purely geometrical, i.e., it is given by the ratio of the area on the monitor scanned by the electron beam to the corresponding specimen area scanned by the probe. Whereas CTEM instruments are conceptually closely related to the classic light microscope, STEM instruments are not unlike TVs; they are essentially "mapping" or "plotting" devices of the spatial variation of the various signals captured by adequate detectors. The analog signals can further be digitized and images can thus be electronically treated and, for instance, magnetically stored.

The relationship between STEM and CTEM operating modes is represented schematically in Fig. 33. The diagram should be read from right to left in the case of

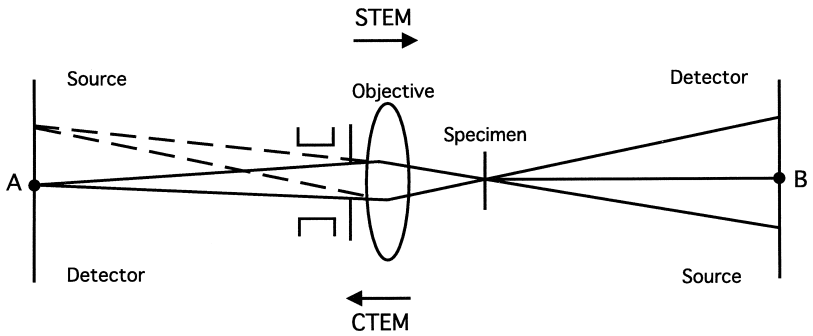



**FIGURE 33** Schematic ray paths of electron beams in an electron microscope in two modes: conventional electron microscopy (CTEM) and scanning transmission electron microscopy (STEM).

CTEM and from left to right in the case of STEM, i.e., the electrons are traveling in opposite directions in the two cases.

In the STEM the source in A is often a field emission gun, since this has a high brilliance. The objective lens is a demagnifying lens system producing a slightly convergent electron beam that becomes the probe, which is scanned over the specimen by means of the deflector following a two-dimensional raster. Behind the objective lens a convergent-beam electron diffraction pattern is formed. Part of this pattern is selected, and the signal collected, amplified, and displayed on a TV monitor. In this way bright- and dark-field images from elastically scattered electrons can be produced; moreover, a variety of other modes of operation is possible. Collecting many beams allows us to image atom columns in crystals as a dot pattern.

In the $Z$-contrast mode an annular detector, capturing the large-angle scattered beams only, allows us to obtain an image formed predominantly by incoherently scattered electrons. High-resolution images made in large-angle incoherently scattered electrons have the important property that the bright dot image does not suffer contrast reversal with changing defocus, i.e., the columns are always imaged as bright dots, irrespective of the amount of defocus or foil thickness, whereas in CTEM the same atom column can be imaged either as a bright or as a dark dot, depending on the thickness and focus. This difference in behavior is a consequence of the difference in shape of the image contrast transfer function (CTF). Under incoherent imaging conditions the CTF decreases monotonously with the spatial frequency, whereas in the coherent case it depends on the spatial frequency in an oscillatory manner. The dot brightness, furthermore, increases with the average $Z$ value of the atoms along the column. Chemically different columns, are thus imaged as dots of different brightnesses.

$Z$-contrast high-resolution images can be interpreted intuitively; they are thus particularly well suited for imaging of the geometry of defect configurations. Often an electron energy loss spectrometer (EELS) is fitted to the outcoming electron beam, which makes it possible to produce images using only electrons having suffered no or a characteristic energy loss, thus allowing chemical mapping.

## XXI. HIGH-VOLTAGE ELECTRON MICROSCOPY

Electron microscopes with an accelerating voltage significantly higher than the conventional 100 kV have come into use over the last decade. They offer the possibility of greater penetrating power and the use of thicker specimens that are more representative of the bulk (Fig. 34). Due to the shorter wavelength of the electrons used, they allow a better instrumental resolution to be achieved. Moreover, the contrast transfer function of the lens system can be designed so it produces roughly the same phase shift for a larger angular range of beams, and hence more faithful representation of crystal structures, than with 100-kV microscopes. The use of high-resolution, high-voltage microscopy offers, at present, perspectives for the direct study of crystal structures. However, the displacement and ionization damage produced by high-energy electrons constitute an intrinsic limitation that restricts the observation time as well as the resolution. Medium-voltage (∼300- to 400-kV) electron microscopes may turn out to be the best compromise for a number of applications (a 400-kV image is shown in Fig. 51).

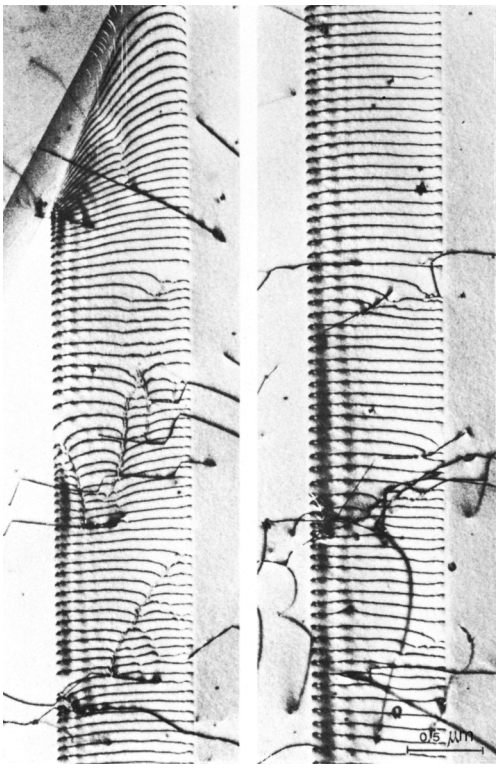



**FIGURE 34** Array of dislocations in stainless steel situated in a glide plane observed by means of high-voltage electron microscopy. [From Dupouy, G., Perrier, F., and Durieu, L. (1970). *J. Microsc.* **9,** 575.]

## XXII. ANALYTICAL ELECTRON MICROSCOPY

Apart from elastic scattering, which is responsible for electron diffraction, inelastic scattering events also occur as electrons pass through the foil. Inelastic processes

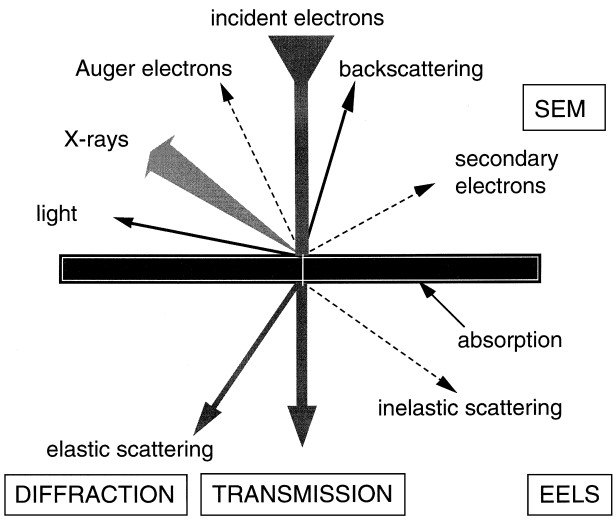



**FIGURE 35** Various types of interactions of electrons with a specimen in an electron microscope, giving rise to signals (X rays, secondary electrons, elastic and inelastic scattered electrons, etc.) that can be used for analytical electron microscopy.

can be caused by (1) single-electron excitations, such as X-ray and Auger electron production, and (2) collective excitations, such as volume and surface plasma oscillations (plasmons) and phonons. Collective excitation can be revealed indirectly as characteristic energy losses of the incident electrons. These various types of electron–specimen interactions can be used to turn an electron microscope into a versatile *in situ* microanalytical instrument where atomic-scale imaging can be coupled with chemical spectroscopy. The most important interactions are summarized in Fig. 35. The emerging signals carry, because of their origin, which lies within the atomic structure of the elements, chemical information on the irradiated area. The use of this information for chemical analysis on a microscale was first suggested by Castaing. Instrumental improvements and the advent of scanning techniques have resulted in applications whereby the various types of interactions can be usefully exploited for obtaining chemical information down to the subnanometer scale. The major physical processes used in electron microscopes to obtain chemical information are characteristic X rays, characteristic energy losses, and Auger electron production.

## A. Energy Loss Spectroscopy and Imaging

Energy loss is characteristic for an element, and if it can be detected by energy analysis of the transmitted electrons, elemental analysis is possible. This is the case, however, only for sufficiently thin specimens, where other energy loss processes do not overwhelm the characteristic spectrum. Plasmons have energies in the range of 10–20 eV, whereas phonon energies are of the order of 10 meV. Individual quantized plasmons are thus much easier to detect than phonons. Since the positions of the plasmon loss peaks are characteristic of the materials, they can be used as analytical tools for aluminium and magnesium, for example.

However, for chemical analysis, the absorption edges in EELS curves are more important; they reflect the absorption phenomena leading to X-ray production and exhibit a fine structure on the high-energy side. This fine structure is referred to as EXELFS (extended energy loss fine structure), which is the analogue of EXAFS (extended X-ray absorption fine structure). The steep rise on the low-energy side of the absorption edge is due to the excitation of inner-shell electrons and characterizes the element.

The fine structure is produced by the electron wave originating from an inner shell that is partially back-reflected by the surrounding atoms, which leads to a modulation of the excitation probability of inner-shell electrons. This fine structure can therefore provide information not only on the chemical nature of the absorbing atom, but also on the number of nearest neighbors and their distances. A typical spectrum of an electron beam analyzed for energy loss after interaction with a TiC specimen is shown in Fig. 36.

Nowadays dispersion by a wedge magnet is used to analyze the characteristic losses. For the latter application of electron loss spectroscopy, in particular, the brightness of the primary electron source is of great importance since only a narrow energy band corresponding to the loss peak is filtered and used for the imaging. The slit-filtered characteristic loss beam then enters quadrupole lens configurations allowing HREM imaging, thus enabling subnanoscale imaging related to the presence of

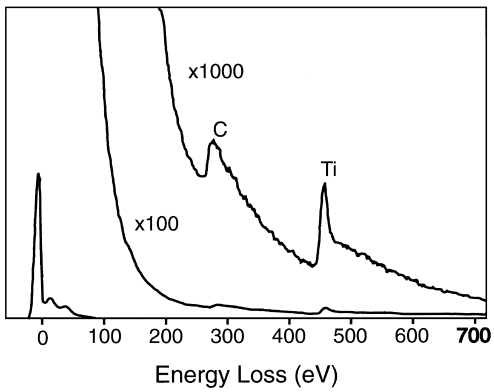



**FIGURE 36** Example of an energy loss spectrum. The intensity of the electrons is reproduced versus energy loss (in eV). Characteristic peaks for the elements C and Ti are clearly distinguished in the 300- to 500-eV energy loss range.

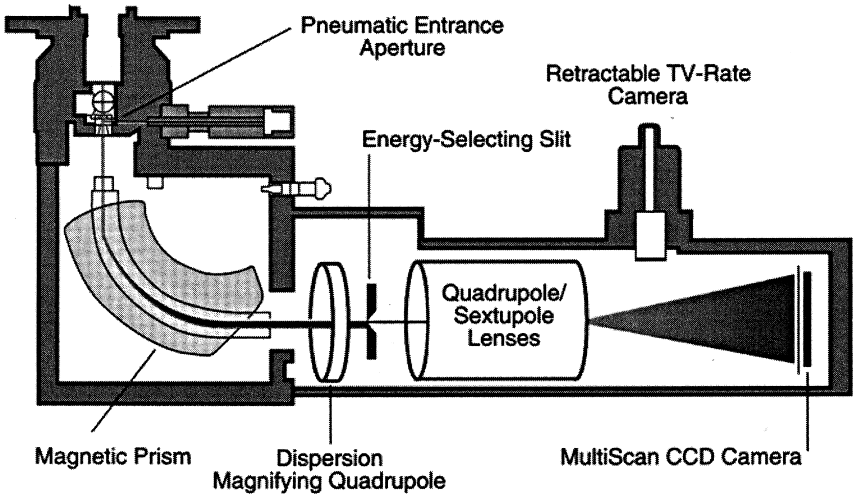



**FIGURE 37** Imaging filter (GIF) as an attachment to an electron microscope. It produces energy-filtered electron images and diffraction patterns, also known as electron spectroscopic images and diffraction patterns. It also produces electron energy loss spectra.

particular elements. A scheme of the experimental setup (Gatan imaging filter) is shown in Fig. 37.

### B. X-Ray Microanalysis

The higher energy state of the ionized atom can be reduced by an electron of an outer shell, e.g., $L_{||}$, falling into the K shell. This process results in the radiative emission of a photon with an energy equal to the difference between the two excited states:

$$E_{\text{ph}} = E_{\text{K}} - E_{\text{L}}.$$

The probability of these characteristic X-rays being emitted is called the X-ray fluorescence yield. It increases with atomic number and is larger for K-line emissions, than for L-line emissions.

For X-ray microanalysis energy-dispersive crystal detectors have nowadays replaced the curved crystal wavelength-dispersive systems based on Bragg diffraction. This technique employs a lithium-drifted silicon detector crystal, which has the advantage of covering a wide range of energies simultaneously. The convenience of display and spectrum processing and acceptable resolution have made this technique very popular. An extensive account of the instrumental aspects can be found in the Bibliography.

It is clear that X-ray microanalysis equipment mounted on a TEM constitutes a powerful tool enabling observation of the substructure of interest at high magnifications and *in situ* analysis of the elemental composition. This facility becomes even more powerful in a scanning transmission assembly (STEM) which has a very small electron probe size available, enabling analysis of particles as small as a few nanometers. Furthermore, use can be made of scanning optics to make images in the signal of, for example, one particular line, thus mapping the distribution of certain elements as in a micrograph, but on a much smaller scale.

### C. Auger Electron Emission

In contrast with the radiative method of deexciting the ion by emission of characteristic X rays, the energy liberated by an electron falling into the inner shell can be used to expel an electron from one of the higher shells. These have characteristic energies for each element and are called Auger electrons. These energies are rather low, resulting in easy absorption and the limitation of the use of these electrons to surface analysis only. Because of the necessity of an ultrahigh vacuum for reducing inelastic collisions, this technique is not commonly used in conjunction with electron microscopes and we do not elaborate on it further.

Both electron energy loss spectroscopy (EELS) and energy-dispersive X-ray spectroscopy (EDXS) enable (within their limits) convenient qualitative and quantitative elemental analysis. Whereas EELS is more specifically suited for light elements, EDXS is not, however, it can detect a wide spectrum of elements, from beryllium to uranium or from sodium to uranium, depending on the kind of spectrometer used (wavelength dispersive or energy dispersive). The former, using curved crystal geometry, is more suitable for quantitative analysis. The detection limit with EELS for an element in a matrix amounts to approximately $10^{-19}$ g, which means about 1000–100

atoms. For EDXS a sensitivity of 1% can be expected for an element detectable in the presence of another one.

Of the three techniques applicable in a TEM, X-ray microanalysis appears to be the most widely used. It is limited mainly in the low-$Z$ element side and in the resolution of the imaging capabilities. Energy loss spectroscopy has advantages in both these aspects, and its potential to provide nanoscale information on the elemental beams is most promising. Surface Auger spectroscopy has only limited applications except in dedicated STEM instruments, where instrumental conditions such as UHV and a high gun brightness are available. The three techniques, having in common their origin of inner shell excitations by the primary electron beam, appear to be complementary rather than competitive.

## XXIII. SPECIMEN PREPARATION FOR TRANSMISSION ELECTRON MICROSCOPY

One of the major drawbacks of electron microscopy techniques is the necessity of thinning the sample down to extremely small thicknesses, depending on the accelerating voltage used and on the resolution [diffraction contrast (100 Å) or HREM (1 to 5 Å) desired. The first is due mainly to the higher penetration with increased voltage, and the second to increased inelastic energy loss, giving rise to spread and chromatic aberration, with a consequent loss of resolution. For 100-kV microscopes a thickness of 1000 to 2000 Å is typically required for the diffraction contrast mode, whereas for HREM a thickness of 100 Å or less is required. For 1000-kV microscopes thicknesses of up to 1 to 5 $\mu$m of Si are acceptable for diffraction contrast observations.

These thickness requirements and the preparation of samples are undoubtedly destructive with respect to the bulk original material, which is indeed one of the disadvantages of TEM techniques. Fortunately appropriate techniques for specimen thinning have been developed for specific types of materials whereby the destructive aspects are kept under control and very valuable information on defect and structure characterization can be obtained.

### A. Thinning Methods

The main procedures for thinning various types of materials are summarized below. Sawing, slicing, and grinding or cold work (milling) are usually required as preparatory thinning procedures to obtain a starting thickness of 100 $\mu$m, from which the final thinning method proceeds.

Metals and alloys are generally thinned by electropolishing. Trepanned disks are mounted in special inert holders to yield self-supporting disks that can be mounted directly in the holder of the microscope.

Ceramics and semiconducting Si materials and like alloys are successfully thinned by ion milling. Ar ions (5 kV) bombard a rotating 3-mm disk under grazing incidence until perforation occurs, usually in the center of the disk. Chemical thinning is also used for these materials, but ion milling is more universal and reproducible.

For high-resolution observations a number of materials which exhibit a conchoïdal fracture habit can be thinned for HREM by mechanical crushing. The fragments are dispersed on holey carbon grids, and after appropriate tilting in the goniometer stage of the object holder, the thin fragments provide acceptable HREM observation conditions.

Layered crystals can usually be sliced down to useful thicknesses by repeated cleavage, and for some materials thin films can be prepared directly by one of the various types of deposition techniques, e.g., evaporation and chemical vapor deposition, on a substrate which is subsequently removed by dissolution. Multilayered-device materials are often prepared by this type of deposition or by in-depth doping or chemical treatment. For this type of material it is often of major importance to know the succession of the layers, their thicknesses, and detailed information on the interfaces. Therefore specimen preparation in "cross section" is required. Figure 38 shows the successive steps in sample preparation in comparison to "plan-view" specimen preparation as can be performed with the above-mentioned thinning methods.

Nowadays highly sophisticated instrumentation for ion milling semiconductor device samples allows viewing and thinning in the same instrument particular areas in a device

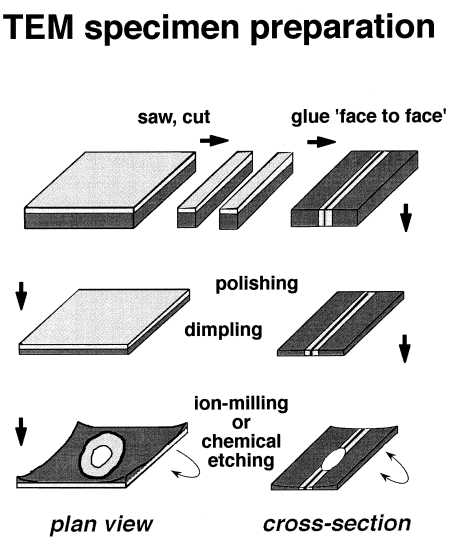



**FIGURE 38** Scheme illustrating the successive steps to prepare thin transmission electron microscope specimens in the plan-view mode (left) and in the cross-section mode (right).

to diagnose and or characterize device or interface configurations *in loco* down to a micrometer accuracy. This technique consists in using a focused ion beam for thinning and for imaging and is called the FIB method.

## XXIV. EXAMPLES OF APPLICATIONS

### A. Dislocations

The first application of transmission electron microscopy (TEM) was the study, by means of diffraction contrast, of dislocation configurations in materials heat treated in different ways; TEM allowed the first moving dislocations to be observed *in situ*. In Fig. 34 we reproduce an array of dislocations confined to a well-defined glide plane in stainless steel. Since the dislocations in this material are dissociated in Shockley partial dislocations, their movement is restricted to a well-defined glide plane, as cross-glide is difficult. This micrograph was made in a high-voltage electron microscope, which allowed a rather thick foil to be observed. The thickness can be deduced from the number of oscillations in the dislocation image.

In Fig. 39 a network of dissociated dislocations in graphite is shown. The configuration is situated in the (0001) plane, which is a glide plane and a cleavage plane. The cleaved specimen is also limited by (0001) planes. In Fig. 39a the extended dislocation nodes, which contain a stacking fault, exhibit a dark contrast. Also, one set of triple-ribbon dislocations is visible. In Figs. 42b and c, the configuration is imaged under three two-beam conditions.

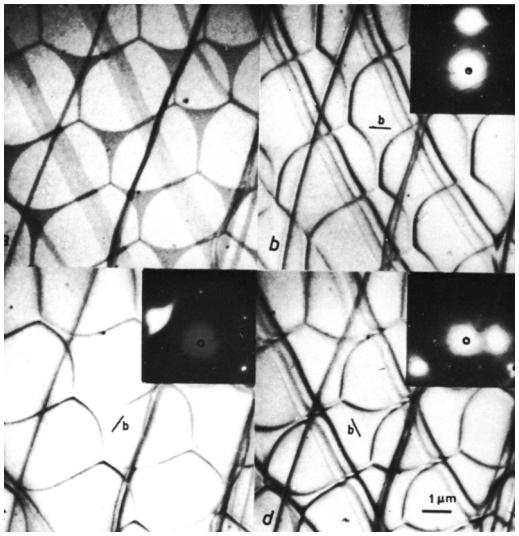



**FIGURE 39** Hexagonal network of dissociated dislocations in the basal plane of graphite: (a) stacking fault contrast; (b–d) different line contrasts. [From Amelinckx, S., and Delavignette, P. (1960). *J. Appl. Phys.* **31**, 2126].

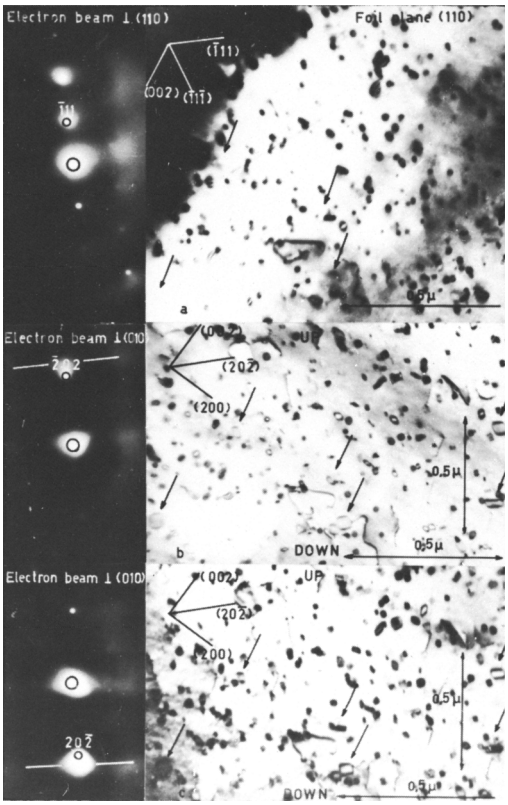



**FIGURE 40** Dislocation loops in irradiated platinum. A contrast experiment allows identification of the loop as being due to vacancies. [Courtesy of E. Ruedl.]

In each of the three images one set of dislocations is out of contrast; the Burger vector of the nonimaged dislocations is perpendicular to the active diffraction vector.

Under irradiation, by means of neutrons, point defects are formed that, with the appropriate heat treatment, agglomerate into small dislocation loops. Vacancies and interstitials may precipitate into separate disks, which, after collapsing, give rise to dislocation loops of two types. It is possible to distinguish these two types of dislocation loops by means of contrast experiments since the Burger vectors of the bordering dislocations have opposite signs (Fig. 40).

On quenching, similar dislocation loops, resulting from the agglomeration of vacancies, are observed in a number of metals (Fig. 41). In metals with a low stacking fault energy, such as gold and cobalt, the vacancies introduced by quenching are found to aggregate in stacking fault tetrahedra, a defect first discovered by means of TEM. The edges of such a tetrahedron are formed by stair rod dislocations, whereas the faces are stacking faults (Fig. 42). High-resolution electron microscopy allows one to distinguish between tetrahedra due to vacancies and tetrahedra due to interstitials.

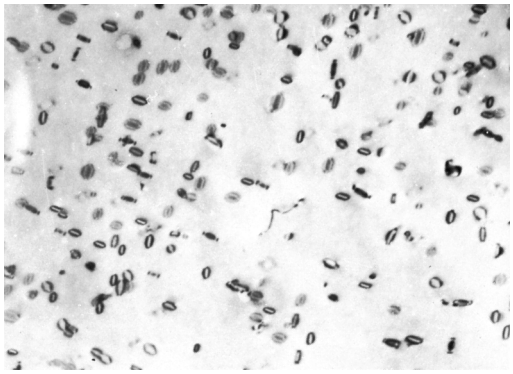


**FIGURE 41** Dislocation loops introduced by quenching in aluminum. [From Hirsch, P. B., Silcox, J., Smallman, R. E., and Westmacott, K. H. (1958). *Phil. Mag.* **3,** 897.]

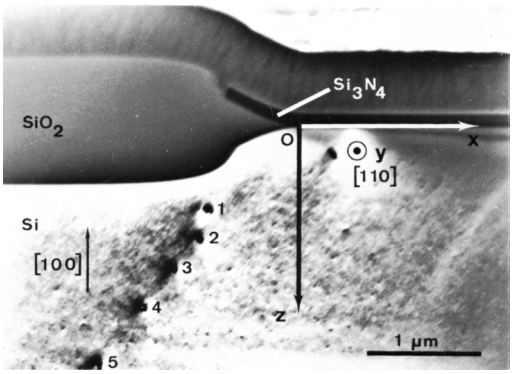



**FIGURE 43** Procession of dislocations in a silicon substrate formed during the fabrication process of a field effect transistor. [From Vanhellemont, J., Amelinckx, S., and Claeys, C. (1987). *J. Appl. Phys.* **61,** 2176.]

Figure 43 shows an application in the field of microelectronics. The sequence of dislocations results from the stresses caused by the fabrication process of field-effect devices. From the geometry of this dislocation array, the magnitude of the stress can be estimated.

The unambiguous interpretation of dislocation images should be based on a comparison of computer-generated images, using dynamical diffraction theory with experimental images.

The atomic arrangement around the core of edge dislocations can be imaged by means of high-resolution electron microscopy when viewed along a direction parallel to the dislocation line. Figure 44 shows a 60° dislocation in silicon.

## B. Planar Interfaces

As explained, stacking faults are images of different shade or fringe patterns. The first case arises when the fault plane is parallel to the foil plane (see, e.g., Fig. 39). Fringes are produced when the fault plane is inclined with respect to the foil plane. Figure 45 shows a bright- and a dark-field image of the same stacking fault, limited to Shockley partials in a face-centered cubic copper–aluminum alloy. From the nature of the outer fringes in the dark- and bright-field images, one can deduce whether the stacking fault is intrinsic (i.e., of the type ABCA–CABC) or extrinsic (i.e., of the type ABCABA–CABC). The displacement vector **R** of an interface, which relates the two crystal parts, can be determined by making images in different two-beam situations and the condition for the absence of contrast (i.e., so-called extinction) when $\mathbf{g} \cdot \mathbf{R}$ is an integer. The vector **R** describes the displacement of the crystal part last met by the electrons with respect to the front part. The problem of determining the nature of the fault consists in determining the displacement vector **R**.

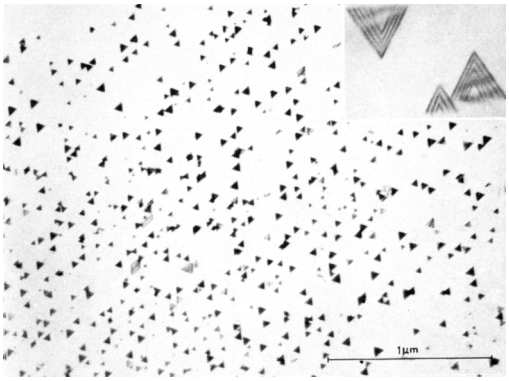



**FIGURE 42** Stacking fault tetrahedra in gold introduced by quenching. [From Hirsch, P. B., Cotterill, R. M. J., and Jones, M. W. (1962). *Proc. Int. Conf. Electron. Microsc., 5th, Philadelphia,* **1,** F-3.]

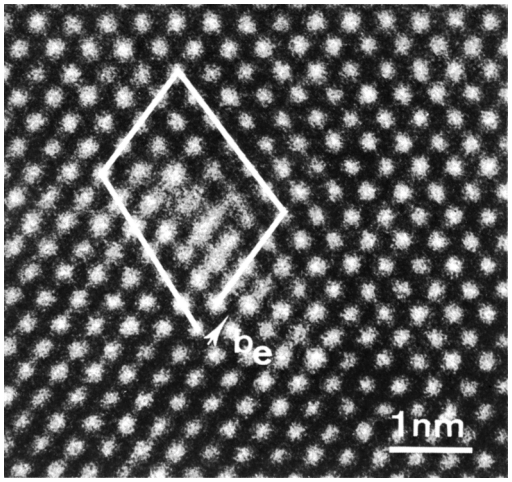



**FIGURE 44** High-resolution image of a 60° dislocation in silicon as viewed along the dislocation line. The edge component of the Burger vector is indicated. [Courtesy of H. Bender.]

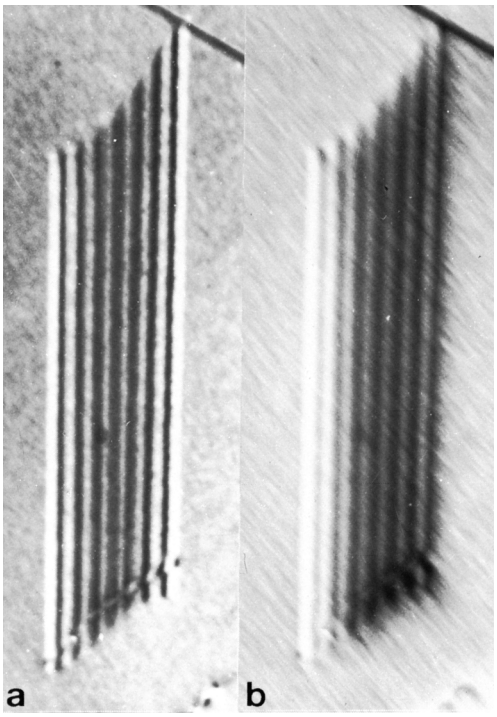



**FIGURE 45** Stacking fault fringes in a low-stacking fault energy alloy. (a) Bright-field image: the outer fringes are both bright. (b) Dark-field image: the outer fringes are opposite in nature. [Courtesy of A. Art.]

In face-centered-cubic crystals this displacement vector $\mathbf{R}$ can be written either $\frac{1}{6}[11\bar{2}]$ or $\frac{1}{3}[111]$ and the fault plane is called (111). In the particular case in Fig. 45 the fault is intrinsic.

The high-resolution image in Fig. 46 shows a stacking fault in zinc sulfide; the crystal is viewed along the close-packed directions of the structure. The bright dots can be

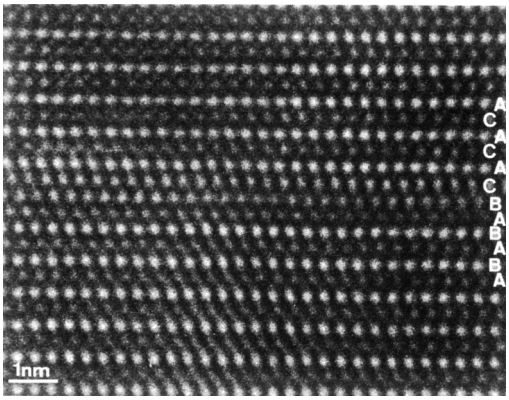



**FIGURE 46** High-resolution image of stacking faults in zinc sulfide as viewed along the close-packed rows of atoms. [From Amelinckx, S. (1986). In "Examining the Submicron World," p. 71, Plenum, New York.]

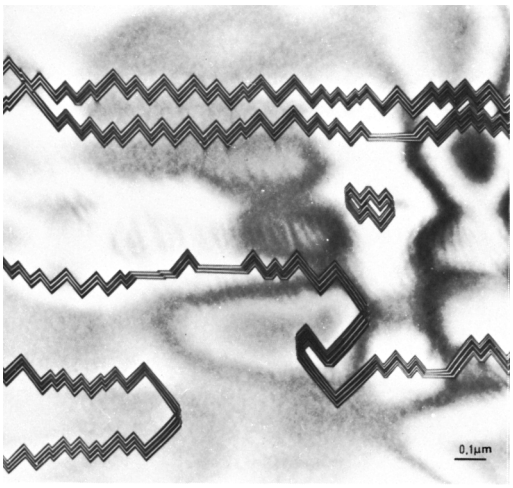



**FIGURE 47** Fringes associated with shear planes in nonstoichiometric rutile ($TiO_2$). [From Amelinckx, S., and Van Landuyt, J. (1978). In "Diffraction and Imaging Techniques in Material Science," p. 107, North-Holland, Amsterdam.]

assimilated with atom columns. The stacking sequence of the layers can be read off from the image. Stacking faults are conservative interfaces, that is, their presence does not change the chemical composition of the crystal. This is no longer the case for crystallographic shear planes that occur in nonstoichiometric oxides and accommodate deviations from the ideal stoichiometry. Figure 47 shows shear planes in rutile; they accommodate a slight deficiency of oxygen with respect to $TiO_2$. Similar shear planes occur in many nonstoichiometric oxides. Their contrast effects are similar to those of stacking faults, except that now the displacement vector is not necessarily a simple fraction of a lattice vector, but may be more complicated due to relaxation effects along the shear planes. As a result, the simple extinction criterion $\mathbf{g}\cdot\mathbf{R} = \text{integer}$ is no longer valid; weak residual fringes are observed for most $\mathbf{g}$ vectors.

Antiphase, or out-of phase, boundaries in ordered alloys are planar interfaces with a displacement vector that is a lattice vector but not a superlattice vector. In diffraction contrast such interfaces produce contrast features that are similar to those of stacking faults, except that the depth period of the fringes is much larger than in the case of stacking faults. As a result usually only one fringe can be formed in the foil thickness. Antiphase boundaries in $Cu_3Pd$ are imaged in the diffraction contrast mode in Fig. 48.

Periodic antiphase boundaries in $Au_4Mn$, giving rise to a one-dimensional long-period structure as imaged in the high-resolution mode, are reproduced in Fig. 49 next to a crystal area exhibiting the normal $Au_4Mn$ structure.

Periodic stacking faults lead to the formation of polytypes, that is, crystals with the same chemical composition but different stacking sequences of the same close-packed

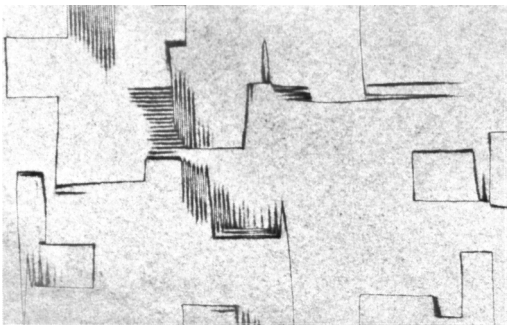


**FIGURE 48** Diffraction contrast image of antiphase boundaries in $Cu_3Pd$. [Courtesy of D. Broddin.]

layers. Depending on the imaging mode used, different details are revealed. In Fig. 50a, which refers to the 15R polytype of SiC, only two successive spots along the central row of superstructure spots were used. The image contains only one-dimensional fringes revealing only the long period. If two successive basic spots as well as the superstructure spots situated between them are used, the image still exhibits one-dimensional fringes, now revealing the elementary atomic layers but modulated in contrast with the superperiod (Fig. 50b). If at least two neighboring rows of basic spots are made to interfere, the stacking mode of the layers can be imaged (Fig. 50c) and the structure is revealed.

## C. Crystal Structures

In present-day microscopes the resolution reveals structural details at the subunit cell level. This has obvious application in the study of "microphases" (i.e., phases present in small volumes only and which would be difficult to detect by means of other diffraction techniques).

Figure 51 shows a two-dimensional superstructure that was found in $Au_{4-x}Mn$ next to the normal structure and to the one-dimensional long-period structure by means of electron microscopy (see also Fig. 49).

High-resolution electron microscopy has been applied extensively to the study of high-$T_c$ superconductors of

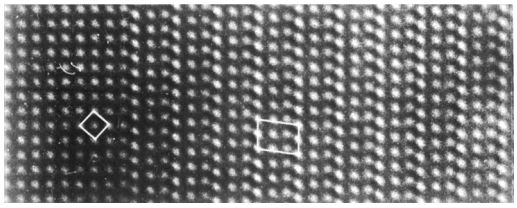


**FIGURE 49** Periodic antiphase boundaries in $Au_4Mn$ revealed by high-resolution electron microscopy. The bright dots represent manganese columns. The basic $Au_4Mn$ structure is imaged in the left part. [From Van Tendeloo, G., and Amelinckx, S. (1981). *Phys. Status Solidi A* **65,** 431.]

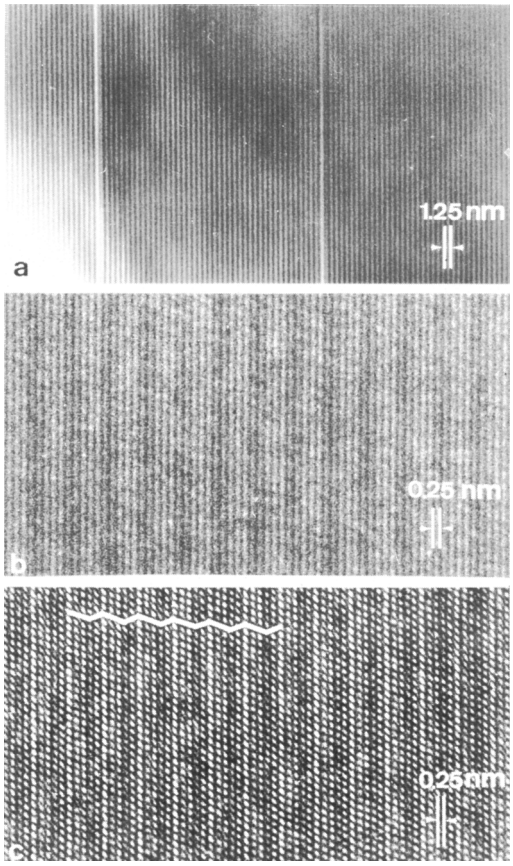



**FIGURE 50** High-resolution image of SiC polytype using different imaging modes. (a) Two successive spots along the central row are used. (b) Two basic spots and the intermediate superstructure spots are selected. (c) Spots on neighboring rows of basic spots are used. [From Amelinckx, S. (1986). In "Examining the Submicron World," p. 71, Plenum, New York.]

the perovskite type. Figure 52 shows a [100] zone image of $GdBa_2Cu_3O_{7-\delta}$. The superperiod containing three perovskite cubes can be clearly recognized; the relationship of the image with the structure is shown in the inset.

By using diffraction contrast in the structure factor contrast mode, an incommensurate modulated structure was discovered in quartz. It occurs in a narrow temperature range around 573°C and constitutes an intermediate phase between the $\alpha$ and the $\beta$ phase. It consists of a regular arrangement of small $\alpha$-type domains related by the Dauphiné twin law (Fig. 53). The domain size decreases with increasing temperature (i.e., the modulation wave vector increases with increasing temperature).

## D. Single Defects

High-resolution images allow the study of atomic arrangement along planar interfaces. Figure 54 shows the fine structure of a nonconservative antiphase boundary in

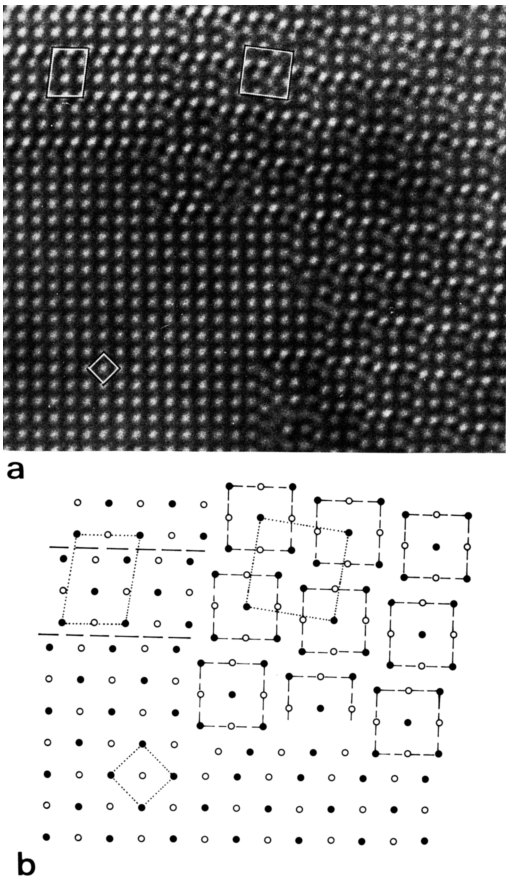



**FIGURE 51** Two-dimensional periodic antiphase boundary superstructures discovered in $Au_4Mn$. (a) The bright dots represent managanese columns. In the left-bottom part the basic $Au_4Mn$ structure is visible. (b) Schematic representation. [From Van Tendeloo, G., and Amelinckx, S. (1981). *Phys. Status Solidi A* **65,** 431.]

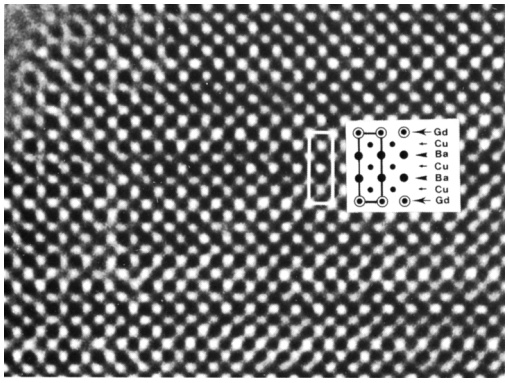



**FIGURE 52** High-$T_c$ superstructure $GdBa_2Cu_3O_{7-\delta}$ imaged along the [100] zone. Note the correspondence with the structure model. [Courtesy of G. Van Tendeloo.]

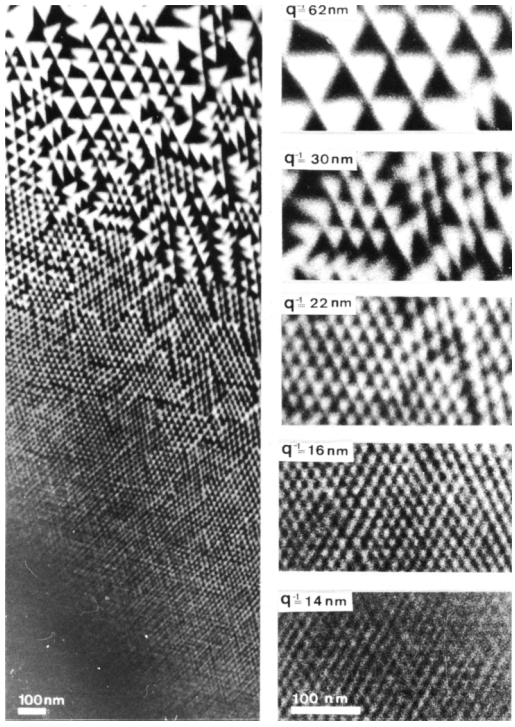



**FIGURE 53** Diffraction contrast images of the incommensurate phase in quartz ($SiO_2$) as observed along the *c* axis. The domain size decreases with increasing temperature from top to bottom. [From Van Landuyt, J., Van Tendeloo, G., and Amelinckx, S. (1986). *Phys. Rev. B* **34,** 2004.]

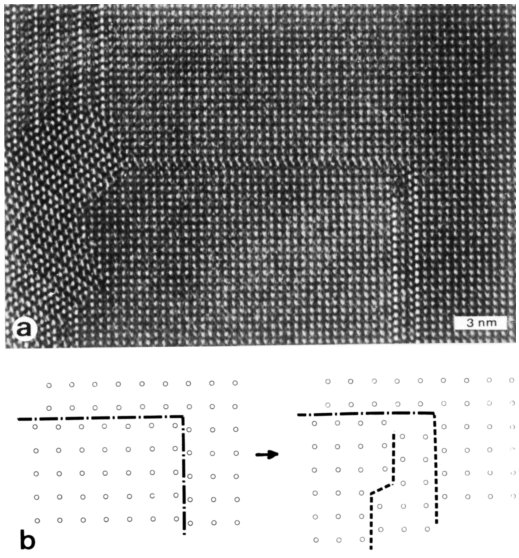



**FIGURE 54** (a) Fine structure of a nonconservative antiphase boundary in $Au_4Mn$. (b) The dissociation into the lower-energy configuration is illustrated schematically. [From Amelinckx, S. (1986). In "Examining the Submicron World," p. 71, Plenum, New York.]

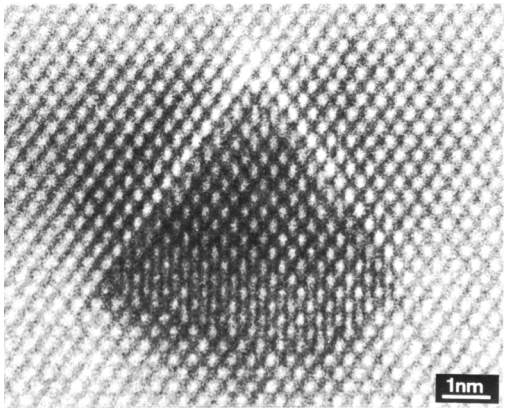



**FIGURE 55** High-resolution image of a stacking fault tetrahedron due to vacancies in silicon. [From Coene, W., Bender, H., and Amelinckx, S. (1985). *Phil. Mag. A* **52,** 369–381.]

$Au_4Mn$. On changing its orientation by 90°, it dissociates into two components, decreasing the free energy.

Figure 55 shows the high-resolution image of a stacking fault tetrahedron in silicon, which was first ion implanted and subsequently annealed. The tetrahedron is due to vacancy agglomeration (see also Fig. 42).

### E. *In Situ* Studies

The specimen chamber of an electron microscope can be transformed into a small laboratory in which specimens can be exposed to different environments and external parameters by the use of the appropriate specimen holders. Table II gives a survey of existing possibilities and some applications of the special specimen holders developed for these purposes.

*In situ* studies of electron radiation damage can be performed in a high-voltage electron microscope; the electrons create the damage and simultaneously form an image. *In situ* chemical analysis can be performed by X-ray microanalysis using the X rays excited by the electron beam or by analysis of characteristic electron energy loss peaks.

**TABLE I Microscope Holders and Their Applications**

| Holder | Application(s) |
|---|---|
| Heating holder | Phase transitions |
| | Chemical reactions |
| | Order–disorder phenomena |
| | Melting phenomena |
| | Annealing phenomena—grain growth |
| Cooling holder | Phase transitions |
| | Order–disorder phenomena |
| Environmental cell | Chemical reactions |
| | Crystal growth |
| Straining holder (possibly combined with heating and cooling) | Plastic deformation dislocation reactions |

### F. Surface Studies

Quite recently transmission electron microscopy was applied to the study of the atomic structure of surfaces. An excellent vacuum and clean surfaces are required for such studies. The reconstruction of surfaces has been demonstrated in this way, as well as the migration of steps on crystal surfaces during growth.

## SEE ALSO THE FOLLOWING ARTICLES

AUGER ELECTRON SPECTROSCOPY • CRYSTALLOGRAPHY • INCOMMENSURATE CRYSTALS AND QUASICRYSTALS • MICROSCOPY • OPTICAL DIFFRACTION • POSITRON MICROSCOPY • SCANNING ELECTRON MICROSCOPY • SCANNING PROBE MICROSCOPY • X-RAY PHOTOELECTRON SPECTROSCOPY

## BIBLIOGRAPHY

Amelinckx, S. (1964). "The Direct Observation of Dislocations," Academic Press, New York.

Amelinckx, S., and Nabarro, F. R. N. (eds.) (1979). "Dislocations in Crystals," North-Holland, Amsterdam.

Amelinckx, S., Gevers, R., and Van Landuyt, J. (eds.) (1978). North-Holland, Amsterdam.

Amelinckx, S., Van Dyck, D., Van Landuyt, J., and Van Tendeloo, G. (eds.) (1997a). "Handbook of Microscopy. Methods I," VCH, Weinheim.

Amelinckx, S., Van Dyck, D., Van Landuyt, J., and Van Tendeloo, G. (eds.) (1997b). "Handbook of Microscopy. Methods II," VCH, Weinheim.

Amelinckx, S., Van Dyck, D., Van Landuyt, J., and Van Tendeloo, G. (eds.) (1997c). "Handbook of Microscopy. Applications," VCH, Weinheim.

Amelinckx, S., Van Dyck, D., Van Landuyt, J., and Van Tendeloo, G. (eds.) (1997d). "Electron Microscopy, Principles and Fundamentals," VCH, Weinheim.

Bethge, H., and Heydenreich, J. (eds.) (1982). "Elektronenmikroskopie in der Festkörperphysik," VEB Deutscher Verlag der Wissenschaften, Berlin.

Cowley, J. M. (1975). "Diffraction Physics," North-Holland, Amsterdam.

Edington, J. W. (1977). "Monographs in Practical Electron Microscopy in Materials Science," Mcmillan, New York.

Glauert, A. M. (ed.) (1981). "Practical Methods: Electron Microscopy," North-Holland, Amsterdam.

Hawkes, P. W. (1972). "Electron Optics and Electron Microscopy," Taylor & Francis, London.

Head, A. K., Humble, P., Clarebrough, L. M., Morton, A. J., and Forwood, C. T. (1973). "Computed Electron Micrographs and Defect Identification," North-Holland, Amsterdam.

Jouffrey, B. (ed.) (1972). "Méthodes et techniques nouvelles d'observation en métallurgie physique," SFME, Paris.

Murr, L. E. (1970). "Electron Optical Applications in Materials Science," McGraw–Hill, New York.

Spence, J. C. H. (1981). "Experimental High-Resolution Electron Microscopy," Oxford University Press (Clarendon), London and New York.

Wenk, H.-R. (ed.) (1976). "Electron Microscopy in Mineralogy," Springer-Verlag, Berlin and New York.

Williams, D. B., and Carter, C. B. (1996a). "Transmission Electron Microscopy. I. Basics," Plenum, New York.

Williams, D. B., and Carter, C. B. (1996b). "Transmission Electron Microscopy. II. Diffraction from Crystals," Plenum, New York.

Williams, D. B., and Carter, C. B. (1996c). "Transmission Electron Microscopy. III. Imaging," Plenum, New York.

Williams, D. B., and Carter, C. B. (1996d). "Transmission Electron Microscopy. IV. Spectrometry," Plenum, New York.



# X-Ray, Synchrotron Radiation, and Neutron Diffraction

**P. Suortti**
*University of Helsinki and European Synchrotron Radiation Facility*

I. Elastic Scattering of X-Rays and Neutrons
II. Basic Formulas for Diffraction
III. Kinematical Theory of Diffraction by Crystals
IV. Perfect Crystal: Dynamical Theory of Diffraction
V. Real Crystal and Extinction, Powder Diffraction
VI. Synchrotron Radiation Sources and Properties of Radiation
VII. X-Ray Optics for Synchrotron Radiation Beamlines
VIII. X-Ray Diffraction Methods and Applications of Synchrotron Radiation
IX. Neutron Sources and Neutron Optics
X. Neutron Diffraction, Methods, and Applications
XI. Future Developments

## GLOSSARY

**Atomic form factor** Amplitude scattered by the electrons of an atom in units of the electron scattering length $r_e$ (for X-rays) or Bohr magneton $\mu_B$ (for neutrons).

**Autocorrelation function** Convolution of a function by itself; called Patterson function in the case of electron density of a crystal. The maxima correspond to interatomic distances.

**Convolution theorem** Fourier transform of a product of functions is the convolution of the Fourier transforms of these functions.

**Dynamical diffraction theory** Interaction between forward and diffracted beams is included in calculation of the diffracted intensity.

**Emittance** Area of the storage ring electron (positron) beam in position-angle phase space.

**Ewald's construction** Reciprocal lattice points on the surface of the Ewald's sphere fulfill the Laue equations. The radius of the sphere is $1/\lambda$, and its center is on the

line defined by the incident beam passing through the origin of reciprocal space.

**Extinction** Loss of diffracted intensity due to coherent coupling of forward and diffracted beams (primary extinction), or due to extra absorption arising from diffraction (secondary extinction).

**Interference function** Function giving the structure of diffracted intensity in reciprocal space. Contains the small-angle scattering term and the Fourier transform of the Patterson function.

**Kinematical diffraction** Interaction between forward and diffracted beams is negligible.

**Laue equations** Diffraction conditions in reciprocal space.

**Opening angle** Angular width of the radiation cone from a single relativistic electron.

**Phase problem** Phase angles of the structure factors are not obtained from diffracted intensity.

**Powder sample** Large number of randomly oriented crystallites or grains in small particles.

**Reciprocal lattice vectors** Span the reciprocal unit cell, perpendicular to the planes defined by crystal unit cell vectors **a**, **b**, **c**. The volume of the reciprocal unit cell is $V_c^* = 1/V_c$.

**Refractive index** Complex quantity, where the real part is slightly less than unity for X-rays and neutrons, and the imaginary part corresponds to absorption.

**Rietveld method** Calculated diffraction pattern from a structural model and peak shape functions is fitted to the experimental powder diffraction pattern.

**Spallation source** Neutron source where neutrons are stripped from nuclei by a high-energy proton beam.

**Unit cell** Basic building block of a crystal. The unit cell vectors **a**, **b**, **c** span a parallelepipedon of volume $V_c$.

**Wigglers and undulators** Periodic magnetic structures placed in the straight sections of storage rings for enhanced production of synchrotron radiation.

**THE DISCOVERY** of X-rays by W. C. Röntgen in 1895 came in the middle of great changes in physics. Much of the research in the last decades of the 19th century concentrated on the nature and propagation of radiation. New forms of radiation, radio waves, and cathode rays were produced, and the old argument of particles vs waves was rekindled. The prevailing ether theory was subjected to serious blows, and the need for reformulation of emission and absorption laws of radiation was becoming evident. X-rays propagated like visible light, were not affected by electric and magnetic fields, but were not refracted and penetrated through light materials, such as soft tissues of human body. This last property made X-rays instantly famous, and led also to medical applications. Röntgen himself suggested that X-rays might propagate as longitudinal vibrations of ether, and their true nature was revealed only gradually. The existence of short-wavelength electromagnetic radiation could be inferred from Maxwell's theory, but the wavelength of X-rays is more than 1000 times smaller than the wavelength of visible light, and the properties of X-rays could not be extrapolated from earlier observations. Röntgen received the first Nobel Prize in Physics in 1901, and in the subsequent years several Nobel Prizes were awarded to scientists working in this or closely related fields.

The wave nature of X-rays was established in the beginning of the 20th century, and their wavelength could be estimated. In 1912 M. von Laue had the idea that X-rays would be diffracted by crystals, which were known to be three-dimensional regular arrays of atoms, resembling one- or two-dimensional optical diffraction gratings. The fascinating story of the first X-ray diffraction experiment, which took place in Munich, Germany, has been told by P. P. Ewald (1962). The Laue-diffraction pattern was recorded on the photographic plate, and the intensity maxima could be correlated with atomic structure of the crystal. Soon after, father and son W. H. Bragg and W. L. Bragg observed reflection of X-rays from cleavage surfaces of crystals, and they established the Bragg law, which gives the relation between the diffraction angle, X-ray wavelength, and spacing of atomic planes. The Braggs solved the first crystal structures, those of cubic ZnS and alkali halides. The results were quite revealing, because they showed that at least the crystals of inorganic compounds are not composed of units of molecules, but of atoms making up a three-dimensional framework. This new concept was difficult to accept for some chemists, who considered the molecules the basic units of compounds. However, structure determination by X-ray diffraction developed rapidly, and it formed the foundation for understanding the nature of solid matter. The importance of this work was recognized immediately; von Laue received the Nobel Prize in Physics in 1914 and the Braggs the following year.

The neutron was discovered by Chadwick in 1932. By that time, the wave/particle dualism of radiation was generally accepted, and diffraction of electrons by crystals had been demonstrated. It was suggested that neutrons would be diffracted as well, and the first experimental evidence was obtained in 1936. However, the neutron beams from radioactive sources were too weak for any quantitative neutron diffraction experiments, and it was only after 1945, with the advent of nuclear reactors, that neutron diffraction became an important tool in study of condensed matter. The neutron wavelength is

$$\lambda = h/mv \tag{1}$$

where $h$ is Planck constant, $m$ the mass of the neutron, and $v$ its velocity. The fast neutrons from the reactor fuel elements are slowed down by successive collisions in the moderator, where they come to thermal equilibrium at the reactor temperature $T$. The neutrons have a Maxwellian distribution of velocities, and the root-mean-square velocity is given by

$$mv^2/2 = 3k_B T/2 \tag{2}$$

where $k_B$ is the Boltzmann constant. It is a fortunate circumstance that the rms velocities of thermal neutrons correspond to wavelengths best suited for diffraction studies. For instance, $T = 100°C$ corresponds to an rms velocity of 3000 m/s, $\lambda = 1.3$ Å $= 0.13$ nm, or an kinetic energy of 48 meV. Recently, neutrons were also produced by the so-called *spallation* sources, where neutrons are stripped from the target nuclei by a high-energy proton beam, and then moderated to thermal velocities. It will be seen that neutron diffraction and X-ray diffraction are complementary methods in studies of condensed matter. However, it took a long time before the pioneering work by B. N. Brockhouse and C. G. Shull was recognized by the Nobel Prize in Physics in 1994.

Production of synchrotron radiation (SR) in particle accelerators was predicted by several scientists in the 1940s, and it was first observed at a General Electric laboratory in 1947. SR wavelengths that are needed for diffraction studies are produced only in synchrotrons or storage rings where light particles (electrons or positrons) are accelerated to energies of several GeV. Such accelerators were built for particle physics research, and from the early 1960s SR was used for condensed matter research. The sources were optimized for collision experiments, not for production of SR, and many compromises had to be made. The first storage rings dedicated as SR sources were built in 1970s, and the present-day facilities are called the third-generation sources. Synchrotron radiation is X-rays, but the radiation is collimated to a narrow cone, its intensity is many orders of magnitude larger than that available from X-ray tubes, radiation is polarized, and it is emitted in short pulses. These properties have revolutionized the methods of X-ray scattering, diffraction included, and many new types of experiments have become possible. The wavelength distribution of SR is essentially continuous, and this has emphasized the complementary nature of SR and neutron scattering studies. Also, the mode of operation of an SR laboratory is very similar to that of a research reactor, and quite often the same scientists utilize both facilities.

# I. ELASTIC SCATTERING OF X-RAYS AND NEUTRONS

## A. X-Ray Scattering

Scattering of X-rays has been treated in many textbooks (James, 1962; Warren, 1969). There are several levels of sophistication, but it is important that a simple classical model of elastic scattering of an electromagnetic (EM) wave by an electron is fairly accurate. Descriptively, the EM wave makes the electron "dance" in the direction perpendicular to the propagation direction of the wave, and as an accelerated charge the electron radiates with the frequency of the incident wave. The scattering cross section is

$$d\sigma/d\Omega = r_e^2 K_{pol} \tag{3}$$

where $\Omega$ is the solid angle, $r_e = e^2/mc^2 = 2.82 \times 10^{-13}$ cm is the classical electron radius, and $K_{pol}$ is the polarization factor. For unpolarized radiation $K_{pol} = (1 + \cos^2\phi)/2$, where $\phi$ is the scattering angle.

The scattering units are actually rather atoms than free electrons. The electrons are bound to the atoms more or less strongly, except the outer electrons in metals, and the classical equivalent of scattering from bound electrons is that of reradiation by electrons in forced harmonic motion. The scattering amplitude of one electron in units of the free electron amplitude is

$$f = \omega^2 / \left(\omega^2 - \omega_s^2 - i\Gamma\omega\right) = f' + if'' \tag{4}$$

where $\omega$ is the X-ray frequency, $\omega_s$ the natural frequency of the oscillating electron, and $\Gamma$ a damping factor. The scattering factor is complex, and it depends strongly on $\omega$ near the resonant frequency $\omega_s$. The *refractive index* $n$ of X-rays can be calculated by considering the sum wave of the incident and scattered radiation,

$$n = 1 - N(\lambda^2/2\pi) r_e f = 1 - \delta - i\beta \tag{5}$$

Here $N$ is the number of electrons per unit volume and $\lambda$ the X-ray wavelength. It is seen that the real part $1 - \delta < 1$, when $\omega > \omega_s$, corresponding to a change of phase of the transmitted wave. The numerical value of $\delta$ is small, typically $10^{-5}$. The imaginary part of $n$ corresponds to absorption of the wave,

$$\mu_o = 2\omega\beta/c = 2N\lambda r_e f'' \tag{6}$$

where $\mu_o$ is the linear absorption coefficient, and $c$ the velocity of light. For an atom the scattering from electrons is calculated as the vector sum of amplitudes from the electron density distribution $\rho(\mathbf{r}) = |\psi|^2$, where $\psi$ is the wave function of the electrons (see Fig. 1). The amplitude or *atomic form factor* is

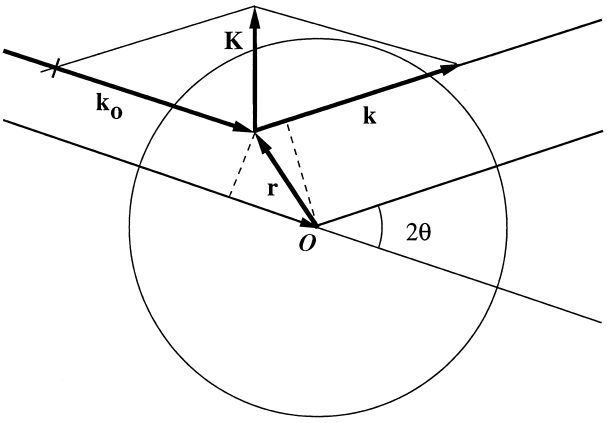



**FIGURE 1** Elastic scattering by an atom. The scattering amplitude from a volume element is proportional to the electron density $\rho(\mathbf{r})$, and the phase angle is $(\mathbf{k}-\mathbf{k_0})\cdot\mathbf{r}=\mathbf{K}\cdot\mathbf{r}$.

$$f(\mathbf{K}) = \int \rho(\mathbf{r}) \exp(i\mathbf{K}\cdot\mathbf{r})\, d^3\mathbf{r} \tag{7}$$

where the scattering vector is the difference between the wavevectors of the scattered and incident waves; $\mathbf{K} = \mathbf{k} - \mathbf{k}_o$. It is seen from Fig. 1 that $K = 2\sin\theta\, k_o$, where $2\theta$ is the scattering angle. The distribution $\rho(\mathbf{r})$ is almost isotropic even when the atoms are bound in a solid, so that

$$f(K) = \int 4\pi r^2 \rho(r)(\sin Kr/Kr)\, dr \tag{8}$$

Because of normalization of $\psi$, $f(0) = Z$, i.e., the number of electrons of the atom.

Several important features are seen already in this simple formulation of X-ray scattering. First, the scattering amplitude is complex, and it changes strongly near the resonant frequencies, which correspond to the $K$, $L$, ... absorption edges of the atom. This is utilized in the methods based on "anomalous dispersion," which are becoming more and more important in X-ray scattering studies. Second, when $\mathrm{Re}(n) = 1 - \delta < 1$, total external reflection takes place at grazing incidence. This has made possible diffraction studies from atomic layers on surfaces, and it is the basis for many X-ray optical components. The most important result is that of Eq. (7), which shows that the scattering amplitude is the Fourier transform of the electron density $\rho(\mathbf{r})$. This is true for any distribution under certain conditions, and it is the basis of all X-ray diffraction studies.

The preceding formalism, which is based on classical concepts, is retained in quantum-mechanical treatment of scattering. The real and imaginary parts of the atomic scattering amplitude are interpreted in terms of electron wave functions and transition probabilities, and they can be calculated with great precision. Also, the inelastic or Compton scattering of X-rays from electrons is included in the quantum-mechanical treatment. A thorough discussion on the correspondence of classical and quantum-mechanical formulations is given by James, Ch. IV (1962).

The electrical dipole (Thomson) scattering, which was described earlier, is the dominant process, but the EM wave interacts also with the magnetic moment of electrons, due either to the electron spin or the net orbital moment of atoms. The ratio of the spin-scattering amplitude to the Thomson scattering amplitude is

$$R \cong (E/mc^2)\sin\theta \tag{9}$$

where $E$ is the energy of the X-ray photon. $R$ is appreciable only at high photon energies, because the electron rest energy $mc^2 = 511$ keV, but near the absorption edges of the atom there is a resonant enhancement of the scattering amplitude. This has made magnetic scattering of X-rays a powerful tool for studies of magnetic structures. Previously these were studied by neutron diffraction only, but X-ray diffraction provides complementary information. Magnetic scattering of X-rays has been discussed by Brunel and de Bergevin (1991), and by Lovesey and Collins (1996).

## B. Neutron Scattering

The scattering amplitude of neutron is a more complicated quantity than the X-ray scattering amplitude. There are two main contributions: scattering of neutron from the nucleus and scattering from the magnetic moments of electrons. These will be discussed separately in the following.

In its simplest form, nuclear scattering of neutrons is understood as formation of a compound nucleus and reemission of a neutron. The energy level structure of the unstable compound nucleus determines the cross sections of possible nuclear reactions, which may give rise to scattering or absorption of the incident neutrons. In the scale of the neutron wavelengths used in the diffraction studies the nuclei are point scatterers. Therefore, the scattering amplitude $b$ is independent of the scattering angle, unlike in X-ray scattering [cf. Eq. (8)], and the increase of $b$ with the nuclear charge is weak and not systematic. To make a distinction with the angle-dependent scattering factor $f$, $b$ is usually called "scattering length." In the same way as in X-ray scattering the excitation energies of the compound nucleus make the scattering length complex. In the case of a single resonance energy $E_r$,

$$b = \xi + C/[(E - E_r) + i\Gamma_r/2] \tag{10}$$

where $E$ is the energy of the incident neutron, $\Gamma_r$ the width of the resonance, and $C$ a constant. The first term corresponds to "potential" scattering, and it is equal to the nuclear radius, and the second term corresponds to

"resonance" scattering. Under certain conditions the resonance term is negative and large enough to make the real part of $b$ negative.

If the scattering nucleus has a spin $I$, the compound nucleus has a spin $I + \frac{1}{2}$ or $I - \frac{1}{2}$. These nuclei have different scattering lengths $b_+$ and $b_-$, and the total cross section is

$$\sigma = 4\pi b^2 = 4\pi\left(w_+ b_+^2 + w_- b_-^2\right) \tag{11}$$

Here $w_+ = (I+1)/(2I+1)$ and $w_- = I/(2I+1)$ are the weight factors of the two possible compound nuclear states. The cross section is split in two parts, those of coherent scattering, $\mathcal{S}$, and incoherent scattering, $s$,

$$\sigma = \mathcal{S} + s = 4\pi(w_+ b_+ + w_- b_-)^2 + 4\pi\, w_+ w_-(b_+ - b_-)^2 \tag{12}$$

Only the first term can produce interference, which is observed in the diffraction pattern.

There is even more diversity in $b$ due to the existence of different isotopes. In the general case, each of these has its characteristic $b_+$ and $b_-$ . The different isotopes of an element will be distributed at random among the atomic positions, and in the case of a crystal, scattering can be divided into scattering from the average structure (ordered scattering) and scattering from fluctuations of the structure (disorder scattering). This division will be discussed in the section where a general formulation for diffraction from a crystal is given. The different factors affecting the nuclear scattering length complicate the interpretation of neutron diffraction patterns, but on the other hand they allow a great variety of studies. Experimental determination of $b$ is the basis of all quantitative scattering studies, and one of the methods used is determination of the critical angle of the total external reflection. In the same way as in the case of X-rays the refractive index for neutrons is a complex quantity,

$$n = 1 - N\,(\lambda^2/2\pi)\,\langle b\rangle \tag{13}$$

where $\langle b\rangle$ is the average value of the bound coherent scattering amplitude of $N$ nuclei in unit volume. For most isotopes, $\langle b\rangle$ is positive, and $1 - n \cong 10^{-6}$ for thermal neutrons. The critical angle is small, typically of the order of 0.1 mrad, which makes possible optical components and experiments similar to those with X-rays.

In addition to nuclear scattering there is scattering due to the interaction between the neutron magnetic moment and that of the atom. The bulk of knowledge of the magnetic structures of solids is based on neutron diffraction studies. There are two important groups of atoms with net magnetic moments, namely the first transition elements with incomplete 3d shells, and rare-earth elements with incomplete 4f shells. The scattering cross section is

$$d\sigma_{\mathrm{pm}}/d\Omega = (2/3)S(S+1)\gamma^2 r_{\mathrm{e}}^2 f_{\mathrm{m}}^2 \tag{14}$$

where $S$ is the electron spin quantum number, $\gamma$ the neutron magnetic moment in units of nuclear magnetons, and $f_{\mathrm{m}}$ the magnetic form factor. This result for spin magnetic moment can be extended in cases where there is orbital moment also. The form factor $f_{\mathrm{m}}$ arises from outer electron shells, so that there is a strong dependence on the scattering angle (see Fig. 28).

The formalisms of X-ray and neutron scattering are very similar, which makes many results directly comparable. However, the interactions with the atoms are fundamentally different, which make X-rays and neutron complementary probes. This has become more and more evident with the use of synchrotron radiation, as will be seen in the chapters where applications are discussed.

## II. BASIC FORMULAS FOR DIFFRACTION

Scattering from an atom can be generalized to concern any electron or nuclear density distribution $\rho(\mathbf{r})$. It is typical that $\rho(\mathbf{r})$ has several hierarchical levels. In a crystal the unit cell is repeated in three dimensions, whereas many organic materials are built of chainlike molecules, which form fibers, and these in turn structures of bundles, etc. Any object has a finite size, and in the following the effect of size in diffraction is considered. The results are based on the use of the *convolution theorem* of Fourier transform $\mathcal{F}$ (Guinier, 1963). Functions $g(\mathbf{r})$ and $h(\mathbf{r})$ are defined in real space and their Fourier transforms $G(\mathbf{K})$ and $H(\mathbf{K})$ in wavevector or reciprocal space. Then

$$F[g(\mathbf{r})h(\mathbf{r})] = G(\mathbf{K}) * H(\mathbf{K}) \tag{15}$$

where $*$ is the convolution operator.

Consider a statistically homogeneous object of volume $V$. When its shape is given by function $\tau(\mathbf{r})$, which is 1 inside the object and 0 outside, the electron (or nuclear) density is

$$\rho(\mathbf{r}) = \rho_\infty(\mathbf{r})\tau(\mathbf{r}) \tag{16}$$

i.e., the object is "cut" by $\tau(\mathbf{r})$ from the infinite, statistically homogeneous object with density $\rho_\infty(\mathbf{r})$. The scattered amplitude is

$$A(\mathbf{K}) = A_\infty(\mathbf{K}) * T(\mathbf{K}) \tag{17}$$

where T($\mathbf{K}$) is the transform of $\tau(\mathbf{r})$, and $A_\infty(\mathbf{K})$ that of $\rho_\infty(\mathbf{r})$.

The intensity is the amplitude multiplied by its complex conjugate, and it turns out to be the Fourier transform of the *autocorrelation (Patterson) function* $P(\mathbf{r})$ of the density of scatterers,

$$P(\mathbf{r}) = \int \rho_\infty(\mathbf{u})\rho_\infty(\mathbf{u}+\mathbf{r})\,\tau(\mathbf{u})\,\tau(\mathbf{u}+\mathbf{r})\,d^3\mathbf{u} \tag{18}$$

Here the product $\tau(\mathbf{u})\tau(\mathbf{u}+\mathbf{r})$ is unity when both $\mathbf{u}$ and $\mathbf{u}+\mathbf{r}$ are inside the object, and otherwise it is zero. This defines the function $V(\mathbf{r})$, which is the volume common to the object and its "ghost" at distance $\mathbf{r}$,

$$V(\mathbf{r}) = \int \tau(\mathbf{u})\,\tau(\mathbf{u}+\mathbf{r})\,d^3\mathbf{u} \tag{19}$$

The intensity can then be written as

$$\begin{aligned} I(\mathbf{K}) &= \int P_\infty(\mathbf{r}) V(\mathbf{r}) \exp[i\mathbf{K}\cdot\mathbf{r}]\,d^3\mathbf{r} \\ &= \mathcal{F}\,[P_\infty(\mathbf{r})] * |T(\mathbf{K})|^2 \end{aligned} \tag{20}$$

Here $|T(\mathbf{K})|^2$ is the Fourier transform of $V(\mathbf{r})$, and $P_\infty(\mathbf{r})$ corresponds to the average over unit volume. It is instructive to divide $P_\infty(\mathbf{r})$ in parts that correspond to diffraction from the average structure and from the fluctuations of the structure. The so-called *interference function* is

$$\begin{aligned} I(\mathbf{K}) = |T(\mathbf{K})|^2 / V V_c + \{1 + (1/V V_c) \\ \times \int V(\mathbf{r})[P(\mathbf{r}) - 1] \exp(i\mathbf{K}\cdot\mathbf{r})\,d^3\mathbf{r}\}. \end{aligned} \tag{21}$$

The first term is non zero only at small values of $\mathbf{K}$. It is proportional to the Fourier transform of $V(\mathbf{r})$, so it depends on the size and shape of the object and not on its internal structure. This is the small-angle scattering term (SAXS or SANS), and it is described in detail in another article (Kratky and Laggner, 1987).

The term in brackets depends on the distribution of scattering objects of volume $V_c$ (e.g., a group of atoms or nuclei) in homogeneous matter. If this distribution is statistically uniform, $P(\mathbf{r}) = 1$, and the interference function is equal to unity outside the SAXS (or SANS) region in reciprocal space. The variations of $I(\mathbf{K})$ about this average value show the fluctuations of the density of scatterers. This is the most complete result that a diffraction experiment can provide about the structure of a statistically homogeneous object. The effect of the finite size of the object is convolution of $I(\mathbf{K}) - 1$ by the Fourier transform of $V(\mathbf{r})$.

## III. KINEMATICAL THEORY OF DIFFRACTION BY CRYSTALS

The electron density distribution $\rho(\mathbf{r})$ of a group of atoms at the positions $\mathbf{r}_n$ may be taken as the sum of distributions of the individual atoms. In a solid the division of electron density is not obvious, and chemical bonding modifies the simple superposition of the atomic distributions. This is an important field of X-ray diffraction studies, but for the present purposes the effects of chemical bonding are ignored. The scattering amplitude is

$$\begin{aligned} G(\mathbf{K}) &= \int \sum_n \rho_n(\mathbf{r}) \exp\,[i\mathbf{K}\cdot(\mathbf{r}+\mathbf{r}_n)]\,d^3\mathbf{r} \\ &\cong \sum_n f_n(\mathbf{K}) \exp[i\mathbf{K}\cdot\mathbf{r}_n] \end{aligned} \tag{22}$$

If there are $M$ identical groups separated by a vector $\mathbf{a}$, the scattering amplitude of this row can be calculated in the same way as for diffraction of light in an optical grating. The amplitude involves a factor $(\sin M\mathbf{K}\cdot\mathbf{a}/2)/(\sin \mathbf{K}\cdot\mathbf{a}/2)$, which has maxima when $\mathbf{K}\cdot\mathbf{a} = 2\pi h$, where $h$ is an integer. The widths of the maxima are $\approx 1/Ma$. When there are similar periods of vectors $\mathbf{b}$ and $\mathbf{c}$, which are not coplanar with $\mathbf{a}$, the *Laue equations* for diffraction conditions by a three-dimensional crystal are obtained,

$$\mathbf{K}\cdot\mathbf{a} = 2\pi h; \quad \mathbf{K}\cdot\mathbf{b} = 2\pi k; \quad \mathbf{K}\cdot\mathbf{c} = 2\pi l \tag{23}$$

Here the Miller indices $(hkl)$ are integers (positive, negative, or zero). The geometrical meaning of the Laue equations is that they represent equidistant planes in $\mathbf{K}$-space (reciprocal space), which intersect at reciprocal lattice points $(hkl)$. These points (relps) define the values of $\mathbf{K}$ that satisfy the diffraction conditions.

The vectors $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$ span the *unit cell* shown in Fig. 2, the basic building block of the crystal. The atomic positions are given by fractional coordinates, $\mathbf{r}_j = x_j\mathbf{a} + y_j\mathbf{b} + z_j\mathbf{c}$. The structure factor is the scattering amplitude $G(\mathbf{K})$ at relp $(hkl)$,

$$\begin{aligned} F(hkl) &= \sum_j f_j(K) \exp[i\mathbf{K}_{hkl}\cdot\mathbf{r}_j] \\ &= V_c \int\int\int \sum_j \rho_j(xyz) \exp[2\pi i(hx_j \\ &\quad + ky_j + lz_j]\,dx\,dy\,dz \end{aligned} \tag{24}$$

The three-dimensional array of relps can be expressed using the *reciprocal lattice vectors*, $\mathbf{K}_{hkl} = 2\pi(h\mathbf{a}^* + k\mathbf{b}^* + l\mathbf{c}^*)$. When these are inserted in the Laue

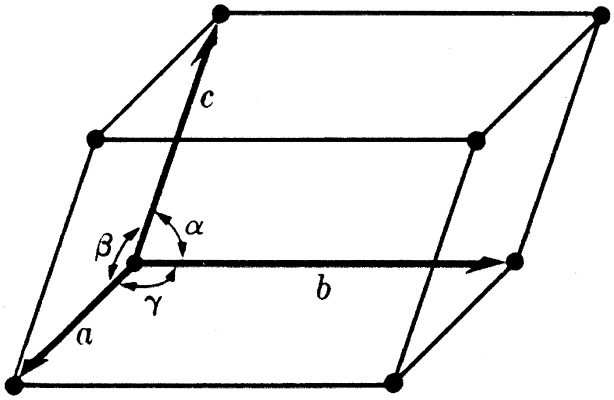



**FIGURE 2** Crystal unit cell.

equations, we obtain

$$\mathbf{a}^* = \mathbf{b} \times \mathbf{c}/V_c, \quad \mathbf{b}^* = \mathbf{c} \times \mathbf{a}/V_c; \quad \mathbf{c}^* = \mathbf{a} \times \mathbf{b}/V_c \tag{25}$$

where $V_c = \mathbf{a} \cdot \mathbf{b} \times \mathbf{c}$ is the volume of the unit cell.

The significance of these results is the following. Diffraction maxima occur at discrete values of the scattering vector $\mathbf{K} = \mathbf{k} - \mathbf{k}_o$. It is easy to show that $\mathbf{K}_{hkl}$ is the normal of the reflecting planes, and the Bragg equation is obtained,

$$2d \sin\theta = \lambda \quad \text{or} \quad K = 2\pi/d \tag{26}$$

where $d$ is the spacing of the lattice planes. The scattering amplitude at the diffraction maximum is called the structure factor $F(hkl)$, and it is in general a complex quantity. The structure factors are the Fourier coefficients of $\rho(xyz)$, so that

$$\rho(xyz) = V_c^* \sum_{hkl} F(hkl) \exp[-i\mathbf{K}_{hkl} \cdot \mathbf{r}] \tag{27}$$

The integral values $(hkl)$ span the reciprocal lattice, and the volume of the reciprocal unit cell is $V_c^* = 1/V_c$, which shows that the density of relp's and the number of reflections increases proportionally to the volume $V_c$ of the unit cell.

*Ewald's construction* illustrates the diffraction condition and reciprocal lattice (see Fig. 3). The vector $-\mathbf{k}_o/2\pi$ drawn from the origin of the reciprocal lattice defines the center of Ewald's sphere. Any relp lying on the sphere fulfills the diffraction condition, and an intensity maximum is observed in the direction $\mathbf{k}$. The diffraction pattern is recorded by rotating the crystal and the associated reciprocal lattice to make the relps to intersect Ewald's sphere, and the detector is placed in the direction of the diffracted ray (it may be a stationary two-dimensional detector). In reality, the relps are not points but small domains due to the finite width of $|T(\mathbf{K})|^2$ and mosaicity of the crystal. Also, the Ewald sphere is a bit "fuzzy" because of variations in $\mathbf{k}_0$. Therefore, the relevant quantity is the integrated reflection, which is obtained when the relp traverses the Ewald sphere during the scan. The other possibility of recording the integrated reflections is to use radiation with a continuous wavelength distribution. The center of the Ewald sphere covers a range along $\mathbf{k}_o/2\pi$, and relp's between the spheres of radii from $1/\lambda_{max}$ to $1/\lambda_{min}$ are intersected. This is the Laue method of X-ray and neutron diffraction. In either method there are several geometrical factors involved, but the main result is that the integrated reflection is proportional to the square of the structure factor, $|F(hkl)|^2$.

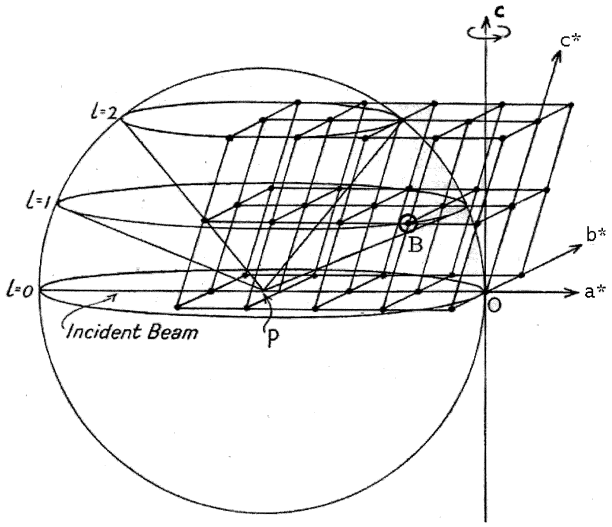



**FIGURE 3** The Ewald sphere and scanning of reflections. The center of the sphere is P, and the origin of reciprocal space O. Scanning of reflections *hk*0, *hk*1, and *hk*2 is shown. For instance, the diffraction condition is fulfilled for reflection −1, −1, 1 (point B).

Determination of crystal structure involves two basic steps: calculation of the unit cell vectors from the diffraction pattern, i.e., inversion from $\mathbf{a}^*, \mathbf{b}^*, \mathbf{c}^*$ to $\mathbf{a}, \mathbf{b}, \mathbf{c}$, and calculation of the electron or nuclear density from the structure factors by Fourier inversion [Eq. (26)]. However, the observed quantity is the intensity of the reflection, so that the information of the phase of $F(hkl)$ is lost. This is the famous *phase problem* of diffraction, and several ingenious schemes have been put forward for experimental and theoretical solutions. These are discussed in detail in another article (K. Ann Kerr, 1987). The basic assumption in the preceding discussion is that the scattering amplitudes of individual unit cells add up. This is the so-called kinematical approximation of diffraction, where interactions between the incident beam and diffracted beam are ignored. In many cases this is a good starting point for crystal structure determination, but usually corrections to the kinematical approximation are needed.

The crystal is not a static, perfectly periodic three-dimensional structure, but the atoms are displaced from their ideal positions by $\boldsymbol{\delta}_j$, and the structure factor becomes

$$F(\mathbf{K}) = \sum_j f_j(\mathbf{K}) \exp[i\mathbf{K} \cdot (\mathbf{r}_j + \boldsymbol{\delta}_j)] \tag{28}$$

The intensity from a crystal of volume $V$ with $N$ unit cells is, in electron units,

$$I(\mathbf{K}) = \sum_m^N \left[ \sum_n^N F_n F_{n+m}^* \exp(i\mathbf{K} \cdot \mathbf{r}_m) \right] = (1/V) \sum_m V(\mathbf{r}_m) y_m(\mathbf{K}) \exp(i\mathbf{K} \cdot \mathbf{r}_m) \tag{29}$$

Here $\mathbf{r}_m$ is the separation of unit cells located at $\mathbf{r}_n$ and $\mathbf{r}_{n+m}$. As before, the effect of the size of the crystal is

described by $V(\mathbf{r}_m)$, which allows the summation be extended to infinity. The use of $\mathbf{K}$ instead of $\mathbf{K}_{hkl}$ is due to the fact that in a distorted crystal there is scattered intensity between the relps. The average square of the structure factors is divided in two parts,

$$y_m(\mathbf{K}) = \left\langle F_n F^*_{n+m} \right\rangle = |\langle F_n(\mathbf{K})\rangle|^2 + \Phi_m(\mathbf{K}) \qquad (30)$$

The first part gives diffracted intensity from the average structure, the second from fluctuations of the structure. These may be due to static disorder, and the atoms are always in thermal motion. The motions are correlated, so that the thermal diffuse scattering (TDS) is not uniform in reciprocal space. In fact, long-wavelength in-phase vibrations (acoustic phonons) make the TDS peak under the Bragg reflections, and their contribution must be subtracted. If the average scattering and diffuse scattering can be resolved in components, a very complete picture of atomic positions and their displacements is obtained.

The preceding separation of the average structure is valid for X-rays, which travel at the speed of light, and in diffraction the instantaneous structure is "seen" by the X-rays. In the course of the experiment the average of the configurations is seen. This is not necessarily the case with neutrons, which may be slower than the elastic waves (phonons) in the crystal. In such a case the TDS is modified, and for instance, the peak under the Bragg reflections is not observed. A detailed discussion is found in *International Tables for Crystallography*, Vol. C, Ch. 7.4 (1999).

## IV. PERFECT CRYSTAL: DYNAMICAL THEORY OF DIFFRACTION

A complete description of diffraction from a crystal requires that the interaction between the incident and diffracted beams is taken into account. It may happen that there are several relps on the Ewald sphere at the same time, so that more than one diffracted beam is excited. In most diffraction measurements the goal is to record the intensities of reflections in the two-beam case, where for a given wavelength only one diffracted beam is excited at the time. However, multiple diffraction provides important information about the phases of the structure factors, and recently this has been used for an experimental solution of the phase problem of diffraction.

The interaction of the incident and diffracted beams can be treated on many levels of sophistication. All these treatments are called dynamical theories of diffraction. A good account of the approaches introduced by P. P. Ewald and M. von Laue is given by James (1962), but here we follow a simplified formulation given by Warren (1969), based on the work of C. G. Darwin. All the various forms of *dynamical diffraction theory* were formulated for X-ray diffraction, but the following discussion applies to neutron diffraction as well with appropriate changes of the meanings of the symbols.

The Darwin treatment begins with description of reflection from a layer of atoms. The resultant amplitudes of the reflected and transmitted beams are calculated using the formalism of Fresnel diffraction. The effects of electron binding, which lead to a complex scattering amplitude and absorption, are ignored, but these can be included without changing the essential results of the calculation. It is found that there is a $\pi/2$ phase shift in the reflected beam, and this shift plays a very important role in crystal diffraction, because a beam that has been reflected twice has suffered a phase shift of $\pi$ and is out-of-phase with the incident beam. The transmitted beam is the combination of the incident beam and the forward reflected beam, and this introduces a small phase shift and makes the real part of the refractive index slightly smaller than unity, as already discussed in Section I.A.

There are two distinctly different cases in perfect crystal diffraction. The geometry where the reflected beam exits at the same surface as the incident beam enters is called the reflection or Bragg case, and the geometry where the beam exits at the opposite surface is called the transmission or Laue case. The following discussion concerns the Bragg case, but the Laue case will be discussed briefly.

In Darwin's theory, diffraction in a crystal is treated by considering the propagation of the reflected and transmitted beams in a structure of parallel equidistant planes of atoms. The resultant reflected beam is a coherent sum of the beams reflected by one atomic layer and transmitted through the layers above. The beam in the direction of the incident beam is a sum of the transmitted beam and the beams reflected back to the forward direction. The phase shift of $\pi$ in two reflections decreases the amplitude of the transmitted beam. Calculation of the amplitudes leads to difference equations for the transmission and reflection coefficients of successive atomic layers. These are solved with the assumption that the changes in one layer are small, so that first-order approximations can be used. At the surface of the crystal the ratio of the reflected amplitude to the incident amplitude is

$$R_0/T_0 = ip/[i\varepsilon \pm (p^2 - \varepsilon^2)^{1/2}] \qquad (31)$$

where $p = \delta/\sin\theta\cos\theta$ and $\varepsilon = \theta - \theta_0$. Here $1-\delta$ is the real part of the refractive index, and $\theta$ the observed Bragg angle. By multiplying $R_0/T_0$ by its complex conjugate and reorganizing the terms several interesting features are observed:

- When $\varepsilon/p$ is between $-1$ and $+1$, the incident beam is totally reflected, i.e., $I/I_0 = 1$

- The center of the totally reflecting regime is shifted to $\theta_0$ because of refraction
- The width of the regime is $2p = 2r_e N\lambda^2 F(hkl)K_{\mathrm{pol,d}}/\pi \sin 2\theta$
- The integrated reflection is $E_\mathrm{d} = (8/3)p$.

The polarization factor $K_{\mathrm{pol,d}}$ is 1 when the electric vector of the incident beam is perpendicular to the plane of diffraction ($\sigma$-polarization). In the plane of diffraction ($\pi$-polarization) the polarization factor is $|\cos 2\theta|$, so that for an unpolarized incident beam $K_{\mathrm{pol,d}} = (1 + |\cos 2\theta|)/2$.

It is interesting to compare the preceding results with those in *kinematical diffraction*. In the same notation, the integrated reflection from a thick crystal in the symmetrical Bragg case is

$$E_\mathrm{k} = (\pi^2 \sin 2\theta/2\mu_0\lambda)(p/K_{\mathrm{pol,d}})^2 K_{\mathrm{pol,k}} = Q/2\mu_0 \tag{32}$$

where $Q$ is the integrated kinematical reflectivity per unit path length of the beam. There is a fundamental difference between kinematical and dynamical diffraction. In the latter case the penetration of radiation is limited to a thin surface layer of the crystal because of the strong interaction between the incident and reflected beams, whereas in the case of kinematical diffraction the penetration depth is limited by absorption. The range of total reflection is very narrow, typically in the arcsec range, so that the integrated reflection is much smaller than in diffraction from an imperfect crystal under kinematical conditions. For instance, if the 111 reflection from a perfect silicon crystal is measured with 1 Å X-rays, the integrated reflection is $2.8 \times 10^{-5}$, while the kinematical value is $42 \times 10^{-5}$. The kinematical and dynamical values converge when the interaction between the incident and reflected beams is weak, or when the interaction volume is very small. These conditions are discussed in the next section. Absorption in a perfect crystal can be treated by introducing a complex index of refraction, or complex atomic scattering factors. The principal result is that the total reflection is no more complete, and the reflectivity curve becomes asymmetric. An example is shown in Fig. 4 for the Bragg case.

When the crystal thickness is not infinite in the scale of beam attenuation, the boundary conditions become important. Exact solutions in closed form exist only for a parallel-sided crystal slab. The reflectivity curves show oscillations, which are essentially different in Bragg and Laue cases. When the thickness of the crystal is increased the integrated reflection approaches an asymptotic value in the Bragg case, whereas in the Laue case it oscillates around the value that is one-half of the Bragg case asymptote. The reason is that the transmitted beam exchanges

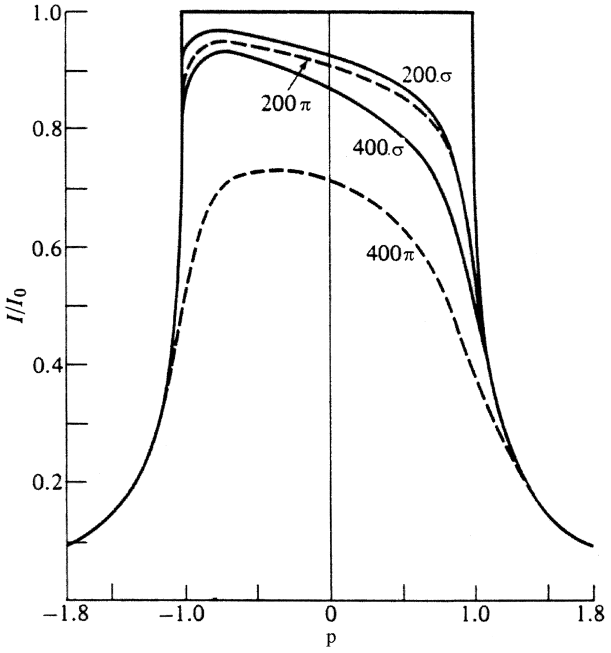



**FIGURE 4** Reflectivity curves for perfect crystal, when absorption is negligible (thick line), and with absorption (thin and broken lines). Approximate values for 200 and 400 reflections of NaCl with 1.54 Å radiation are used. [From Warren, B. E. (1969). *X-Ray Diffraction*, Addison-Wesley, Reading, MA.]

energy with the reflected beam, and in the Laue case the balance at the exit surface depends on the thickness of the crystal. In the Bragg case there is only one exiting beam; in the Laue case there are two, and their total power is that of the Bragg reflection, when the effects of absorption are negligible. The integrated reflections as functions of a normalized thickness $A$ are shown in Fig. 5.

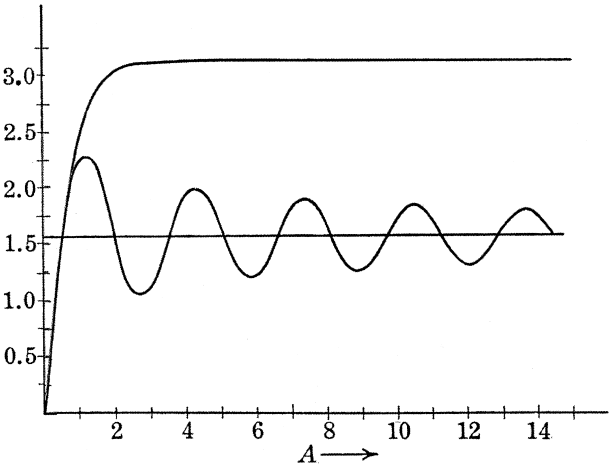



**FIGURE 5** Integrated reflectivity of a nonabsorbing perfect crystal plate in symmetrical Bragg and Laue diffraction. The thickness $D$ is given in units $A = D/\Lambda(\gamma_0\gamma_h)^{1/2}$. The asymptotic values of $\pi$ (Bragg case) and $\pi/2$ (Laue case) correspond to the Ewald solution of perfect crystal reflectivity.

## V. REAL CRYSTAL AND EXTINCTION, POWDER DIFFRACTION

Most single crystals that are used in studies of crystal structure are closer to the kinematical than the dynamical limit of diffraction. It is seen from Fig. 5 that at small values of $A$ the integrated reflection increases linearly with the thickness. This is the range of *kinematical diffraction*, where the reflected beam is weak and its power increases linearly with the diffracting volume. However, deviations from the conditions of kinematical diffraction are large enough to warrant correction. The effect is called *extinction*, and many different methods have been developed to correct for it.

Most extinction theories and correction methods are based on the concept of mosaic crystal. If the mosaic blocks are sufficiently small and their orientations vary appreciably, the blocks diffract independently, and the interaction between the incident and reflected beams inside the blocks is negligible. In such a case the crystal is called ideally imperfect, and the conditions of kinematical diffraction prevail. Although the concept of mosaic crystal is oversimplified, since imperfection in a crystal may be the result of dislocations and inhomogeneous strains, it has persisted. Darwin introduced two kinds of extinction, primary and secondary, and these concepts have been used since then.

*Primary extinction* is present when the integrated intensity from each of the mosaic blocks is less than predicted by kinematical theory. Before reaching an interior block the beam may have been diffracted by several other blocks, so that the beam is attenuated more than by ordinary absorption. This effect is called *secondary extinction*, and it becomes negligible only when the disorientation of the blocks becomes large, or the reflection is weak. Conceptually, primary extinction may be said to arise from coherent or amplitude coupling of beams, while secondary extinction is due to incoherent or intensity coupling of beams. With these concepts, explicit reference to the mosaic crystal model is avoided.

The effects of primary and secondary extinction on the integrated reflection are usually given by the expression

$$E_{\text{obs}}(hkl) = y_{\text{p}} y_{\text{s}} E_{\text{kin}}(hkl) \tag{33}$$

where $y_{\text{p}}$ and $y_{\text{s}}$ are the primary and secondary extinction coefficients, respectively. These are average values, which refer to the integral over the reflectivity curve. It would be more appropriate to use extinction coefficients that vary along the reflectivity curve, so that no assumptions of the distributions of the sizes and orientations of the mosaic blocks would be needed.

The size $D$ of a mosaic block is best measured in units of the extinction distance $\Lambda$,

$$t = D/\Lambda = D/[V_{\text{c}}/r_{\text{e}}\lambda K_{\text{pol,d}} F(hkl)] = A(\gamma_0\gamma_{\text{h}})^{1/2} \tag{34}$$

Here $\gamma_0$ and $\gamma_{\text{h}}$ are the direction cosines of the incident and reflected beams, respectively. It is seen that $t$ can be made small and the kinematical limit approached, if the X-ray wavelength or $K_{\text{pol,d}}$ is reduced. Model calculations indicate that

$$y_{\text{p}}(t) \approx \exp[-Ct^2] \to 1 - Ct^2 \quad \text{(at small } t\text{)} \tag{35}$$

where $C$ is a constant, which depends on the shape of the mosaic block. This can be used for extrapolation to zero primary extinction by changing $\lambda$ and/or $K_{\text{pol,d}}$ (Suortti, 1982). Recently, synchrotron radiation has provided new possibilities also in this area. In particular, wavelengths in the 0.1 Å regime can be used. These are an order of magnitude smaller than the ones used in traditional crystallography, so that $1 - y_{\text{p}}(t)$ can be reduced drastically.

*Secondary extinction* may be substantial even when primary extinction is negligible. For instance, some crystals have a layer-like structure, where the thin layers are highly parallel, but diffract incoherently. The interaction between the incident ($P_0$) and diffracted ($P_{\text{h}}$) beams is described by energy transfer equations (Zachariasen, 1945)

$$\begin{aligned} (\partial P_0/\partial s_0) &= -\mu_{\text{e}} P_0 + \sigma P_{\text{h}} \\ (\partial P_{\text{h}}/\partial s_{\text{h}}) &= -\mu_{\text{e}} P_{\text{h}} + \sigma P_0 \end{aligned} \tag{36}$$

Here $s_0$ and $s_{\text{h}}$ are the coordinates in the directions of the incident and diffracted beams, respectively, and $\mu_{\text{e}} = \mu_0 + \sigma$ the total attenuation coefficient, due to the linear absorption coefficient $\mu_0$ and the scattering coefficient $\sigma(s_0, s_{\text{h}}, \theta)$. If the transmitted and reflected beams are recorded as functions of $\theta$, and the geometrical factors are known, the transfer equations can be solved, if $\sigma$ depends weakly on $(s_0, s_{\text{h}})$. The extinction-corrected integrated reflection is $\int \sigma(\theta)\, d\theta$. This method has been used in a few cases, but with highly collimated, monochromatic SR experimental corrections for secondary extinction could be done in routine way.

Usually, an angular distribution $W(\Delta\theta)$ of $\sigma(\theta)$ is assumed without measuring the actual reflectivity curve, and the width of $W$ is a parameter of the model. For a symmetrical Bragg reflection from a thick sample,

$$E_{\text{obs}} = Q/2(\mu_0 + gQ) \tag{37}$$

where $gQ$ is the mean secondary extinction coefficient. For a Gaussian distribution of $W$, $g = 1/2\sqrt{\pi}\eta$, where $\eta$ is the standard deviation. In crystal structure determination a very large number of reflections is recorded, and the individual reflectivity curves are not examined. Extinction corrections are included in structure refinement, where the

model parameters are fitted by least-squares methods. The most widely used model for extinction corrections is that by Becker and Coppens (1974).

Diffraction from *powder samples* can be used to approach the kinematical limit. An ideal powder sample has a randomly oriented distribution of small crystallites. In terms of Ewald's construction, the reciprocal lattice takes all orientations with respect to $\mathbf{k_0}$, so that all reflections with $K \leq 2k_0$ take place simultaneously. The intensity of a given reflection is distributed uniformly on the cone inscribed by $\mathbf{k}$ about $\mathbf{k_0}$. The diffraction pattern or a part of it can be projected on a two-dimensional detector. The Bragg reflections form concentric rings, and the diffraction pattern is projected on $K$ by azimuthal integration. Secondary extinction is the same in all reflections, corresponding to the average scattering contribution to total attenuation. Primary extinction is usually small because of the small grain size of the powder particles, but broadening and overlap of the reflections make it difficult to separate the background from the diffraction pattern. Again, SR has improved the quality of powder diffraction data by making instrumental broadening of reflections nearly negligible. Powder diffraction is now used for structure determination, which is important because sufficiently large single crystals of many interesting materials are not available. However, the principal use of powder diffraction is in materials science, in studies of grain size, strain, and texture in metals, alloys, and ceramics. These quantities are extracted from the shapes and shifts of the reflection profiles, and from nonuniform distribution of intensity on diffraction cones. Accounts of developments in powder diffraction are found in books edited by Young (1993), and Snyder, Fiala, and Bunge (1999).

## VI. SYNCHROTRON RADIATION SOURCES AND PROPERTIES OF RADIATION

### A. Sources

Synchrotron radiation (SR) is emitted when light charged particles (electrons or positrons) moving with relativistic velocity undergo radial acceleration. Small synchrotrons have developed to large (circumference about 1 km) storage rings, which are dedicated to production of synchrotron radiation. Periodic magnetic structures called wigglers and undulators are inserted in the straight sections of the storage rings to enhance the flux of the SR sources. The latest advance in the field is development of special storage rings and linear accelerators, where free electron lasers provide tunable coherent radiation of high brilliance. The sources and properties of SR have been discussed by Brefeld and Gürtler (1991).

Third-generation synchrotron radiation laboratories are actually accelerator complexes, which include the source of particles (electron gun), a preaccelerator (microtron or linac), a booster synchrotron, where the final particle energy is achieved, and the storage ring. The layout of the European Synchrotron Radiation Facility (ESRF) is shown in Fig. 6. The storage ring is actually a polygon, where the particles travel in a vacuum tube, and dipole magnets bend the particle trajectory at the corners to make a closed orbit. The particles have formed bunches in the preaccelerators, and when they are injected to the storage ring, their energy is typically several GeV; the length of the bunch is of the order of 1 cm, and its diameter is smaller than 0.1 mm. The fill of the storage ring can be varied from single bunch to about 1000 bunches, so that the time difference between bunches is from a few microseconds to nanoseconds, and the bunch length corresponds to a few tens of picoseconds. The velocity is very close the velocity of light, because the electron or positron mass is about $10^4$ times the rest mass. When changing direction in the magnetic field the particles radiate along the tangent of the orbit, as will be seen in the following, and the loss of energy is compensated for in radio-frequency cavities. When scattering from residual gas molecules in the vacuum tube is small the lifetime of the particle beam is about 1 day.

The closed orbit of the particles can be described in phase space. The $1\sigma$-contours of position and angle are ellipses in horizontal and vertical directions, and the area/$\pi$ of each one is called the horizontal or vertical *emittance*, $\varepsilon_x$ or $\varepsilon_z$, respectively. A steady state is achieved in a few milliseconds after injection, and $\varepsilon_x$ and $\varepsilon_z$ are characteristic constants of the storage ring. The shape of the emittance ellipse changes along the ring depending on the focusing magnets, but the area is constant. The ratio $\varepsilon_z/\varepsilon_x$ is called coupling, and in modern storage rings it is a few percent. The horizontal emittance of the ESRF is $3 \times 10^{-9}$ m rad, which corresponds to a typical beam size $\sigma_x = 300\,\mu$m and divergence $\sigma_x' = 10\ \mu$rad. In the vertical direction, both values are reduced by a factor of 10.

### B. Properties of SR

The radiation pattern of a nonrelativistic electron orbiting in a magnetic field has the well-known dipole radiation distribution $I \propto \sin^2 \varphi$, where $\varphi$ is the angle between the observation direction and the direction of radial acceleration in the rest frame of the electron. For a relativistic electron, this radiation distribution has to be transformed by Lorentz transformation into the laboratory system. As a result, radiation is observed only within a narrow cone in the propagation direction of the electron (Fig. 7). The *opening angle* of this cone is approximately

$$\Delta\psi = 1/\gamma = E_0/E \tag{38}$$

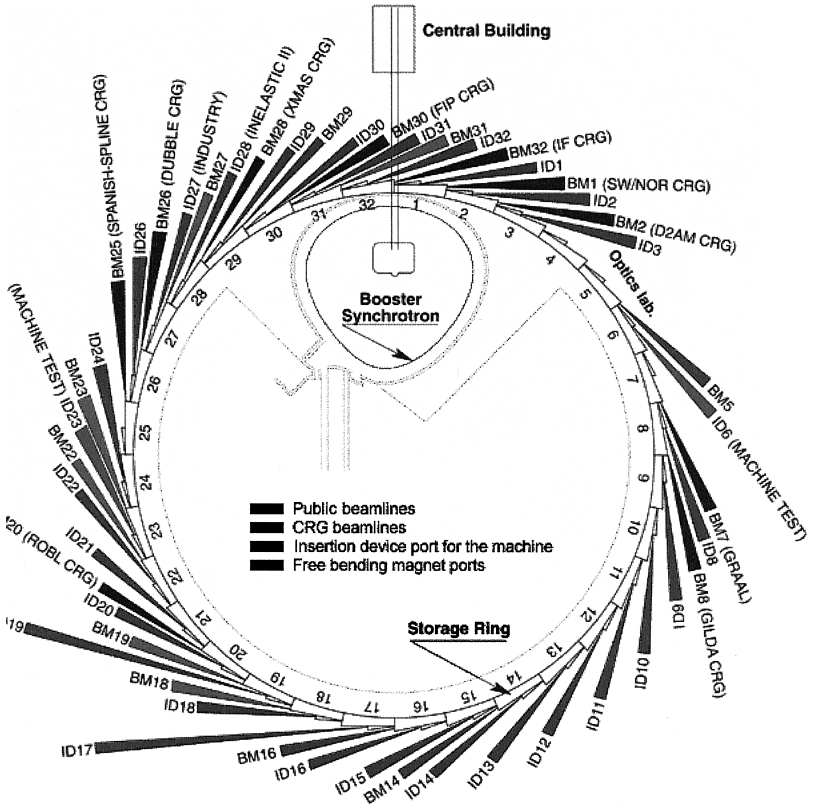



**FIGURE 6** Layout of the European Synchrotron Radiation Facility (ESRF) in Grenoble, France. The circumference of the storage ring is 845 m, and presently 40 beamlines are in operation. [*ESRF Highlights* 1999; reproduced with permission.]

where $E$ is the electron energy and $E_0$ is the electron rest energy, 511 keV. The frequency spectrum observed by a stationary observer is understood by considering the situation where the narrow radiation cone sweeps past in a short time approximately equal to $R/\gamma^3 c$, where $R$ is the radius of curvature of the orbit and $c$ the velocity of light. For the values relevant for the ESRF bending magnet, $\gamma = 1.174 \times 10^4$, $R = 23$ m, this single electron flash lasts $4.7 \times 10^{-20}$ s. This short pulse, Fourier transformed to frequency, contains a spectrum of harmonics up to $2.1 \times 10^{19}$ Hz, which corresponds to 0.14 Å wavelength or 90 keV photon energy. In a real storage ring with many emitting electrons, a continuous spectrum covering the range from infrared to hard X-ray regime is observed.

The spectrum of synchrotron radiation can be calculated precisely, and SR is actually used to calibrate instruments utilized in astrophysics. The spectral brilliance is usually given in units of photons per second per mm$^2$ source area per mrad$^2$ source divergence and per 0.1% bandwidth, $I = I(x, z, \theta, \psi, E, t)$. Integration over the source area yields the spectral brightness (intensity), and integration over all angles yields the spectral flux. Universal curves of spectral brightness and flux can be calculated for a bending magnet source, and those are illustrated in Fig. 8. The wavelength scale is given in units of the critical or characteristic wavelength $\lambda_c$, which is in practical units of the bending radius $R$, magnetic field $B$, and the electron energy $E$,

$$\lambda_c[\text{Å}] = 5.59R[\text{m}]/E^3[\text{GeV}^3] = 18.6/B[\text{T}]E^2[\text{GeV}^2] \tag{39}$$

The concrete meaning of $\lambda_c$ is that it divides the emitted power in equal halves. The universal curves can be used for calculating the brightness and flux of radiation from a given bending magnet.

SR from a bending magnet is linearly horizontally polarized when observed in the orbit plane. Out of the plane, the polarization is elliptical and can be decomposed into its horizontal and vertical components. These are shown in Fig. 7, and it is seen that the more intense horizontal component is closely approximated by a Gaussian of variance $\sigma_\psi$, which is related the full-width-half-maximum by $\text{FWHM}_\psi = 2.35\,\sigma_\psi$.

The sources of radiation in modern storage rings are mostly *wigglers* and *undulators*. These are periodic

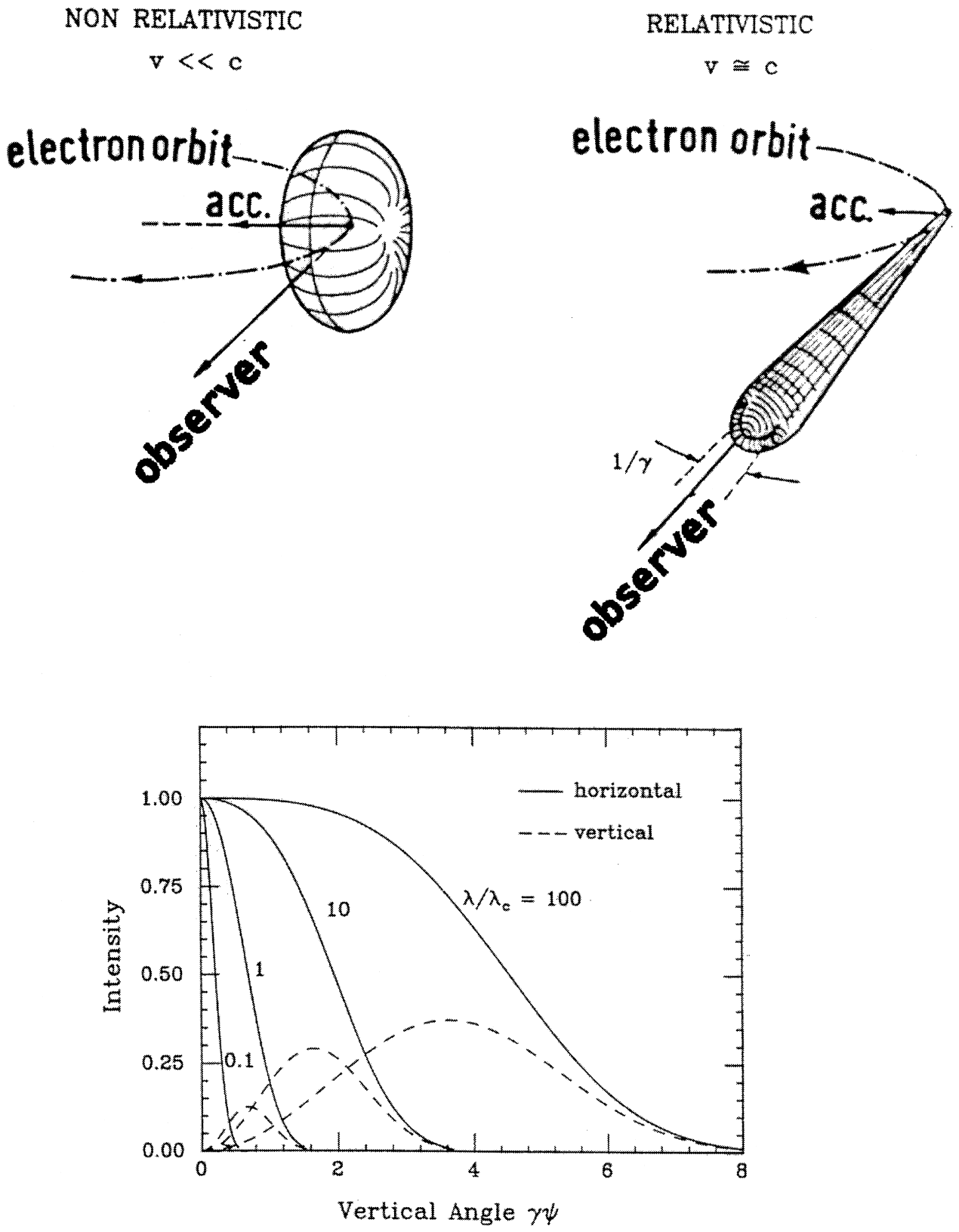



**FIGURE 7** Cone of synchrotron radiation from a relativistic electron, and vertical intensity distributions of the parallel (electric vector in the orbit plane) and perpendicular components.

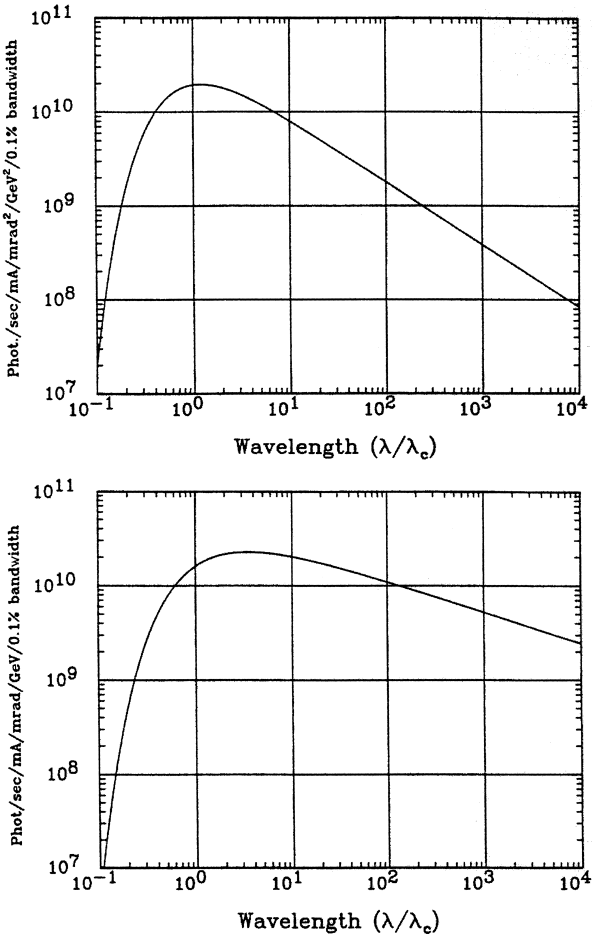



**FIGURE 8** Spectral brightness and flux of SR from a bending magnet source in universal units.

magnetic structures, where the electrons travel oscillating about the center line. In the simplest case the trajectory is sinusoidal, and it can always be described by a few Fourier components. In most cases, the magnetic field is vertical, so that the electron trajectory lies in the horizontal plane. Wigglers and undulators are similar structures, and they consist usually of permanent magnet blocks above and below the vacuum chamber. The magnetic field can be changed by opening or closing the gap between the upper and lower jaws. The wiggler or undulator is characterized by a parameter giving the ratio of the maximum angular deflection $\delta$ of the electron beam to the opening angle of the radiation cone, $1/\gamma$,

$$K = \gamma\delta = 0.934\,\lambda_0\,[\text{cm}]B_0[\text{T}] \qquad (40)$$

where $\lambda_0$ is the period length and $B_0$ the peak magnetic field. When $K \gg 1$ the radiation cone sweeps over a wide fan $2\delta$, typically a few mrad, and the device is called wiggler. The intensities from different source points add up incoherently, and the brightness and flux are approximately equal to that from $2N$ bending magnets, if there are $N$ magnet periods. When $K$ is of the order of 1, wave fronts from different periods interfere coherently, producing sharp peaks in the emitted spectrum. When an electron moves through an undulator of period $\lambda_u$ it undergoes transverse harmonic oscillation in its rest frame, and it emits at one frequency. In the laboratory frame, the wavelength of the radiation is

$$\lambda_1 = \left(\lambda_u/2\gamma^2\right)\left(1 + K^2/2 + \gamma^2\theta^2\right) \qquad (41)$$

where $\theta$ is the angle between the axis and direction of observation. When $K$ increases the displacements become larger, and the electron oscillates also in the longitudinal direction with double frequency. The oscillations are no longer harmonic, and shorter wavelengths $\lambda_i = \lambda_1/i$ appear. The odd harmonics have their maximum intensity on axis, while the on-axis intensity of even harmonics is zero. The intensity of the central beam and that of the angle-integrated spectrum are shown in Fig. 9.

The central brightness of an undulator beam is

$$I_i(\theta = 0) = 1.744 \times 10^{11}\, N^2 E^2[\text{GeV}^2] I[\text{mA}] F_i(\text{K}) \quad (42)$$

where $F_i(\text{K})$ is a function with a maximum value of about 0.5. For instance, with $K = 1.8$ and $N = 150$, an ESRF undulator delivers $1.4 \times 10^{16}$ photons/s/mm$^2$/0.1% BW in the third harmonic at 30 m from the source. The total power radiated by a wiggler or undulator is

$$P_T[\text{kW}] = 0.633 E^2[\text{GeV}^2] B_0^2[\text{T}^2] L[\text{m}] I[\text{A}] \qquad (43)$$

so that the foregoing undulator radiates 7.3 kW. The total wiggler power can be even higher, and therefore all beamline components exposed to the beam must be efficiently cooled. These numbers can be compared with high-power X-ray tubes, which may radiate 100 W to a solid angle of $2\pi$, while the undulator radiates to a solid angle of $10^{-8}$ rad$^2$.

## VII. X-RAY OPTICS FOR SYNCHROTRON RADIATION BEAMLINES

X-ray optics are based on certain reflecting and diffracting elements, which are used to select an energy band from the SR beam and to focus it on the sample. The transport of the SR beam from the source to the sample and from the sample to the detector can be best described by phase space analysis, where each optical element is a window in position–angle–wavelength space. The beamline should be seen as an integral system, where the optical components and the detector should be matched with the resolution of the sample and requirements of the experiment. In practice, beamlines are complicated structures, where

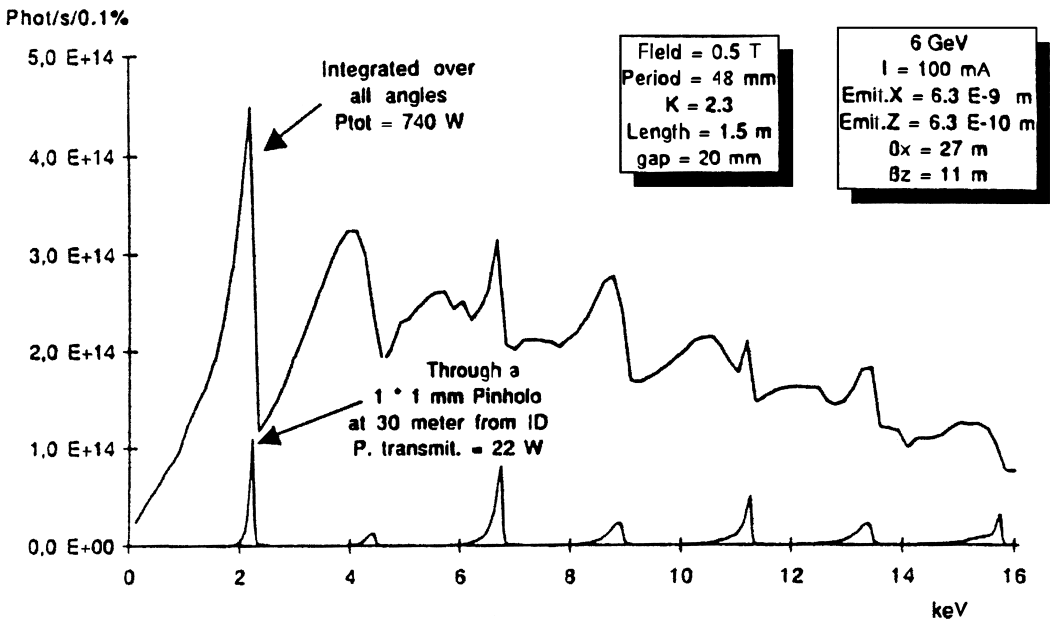



**FIGURE 9** Spectral flux from an undulator. The numerical values correspond to the early operation parameters of the ESRF.

compromises between optimal performance and technical feasibility must be made.

Two phenomena are utilized in X-ray optics: diffraction by single crystal or by synthetic multilayer structure, and total external reflection from a mirror. Focusing is achieved by curved crystals, multilayers, or mirrors.

### A. Mirrors

It was already mentioned that the refractive index is a complex quantity, where the real part $1-\delta$ is slightly smaller than unity. Therefore, total reflection takes place at the interface of vacuum and solid mirror, and if absorption is small, there is a well-defined critical angle,

$$\theta_c\,[\text{mrad}] = \sqrt{2\delta} = 2.3\lambda\,(\rho Z/A)^{1/2} \qquad (44)$$

where $\lambda$ is the X-ray wavelength in Å, $\rho$ the mirror density in units of g/cm$^3$, $Z$ the atomic number, and $A$ the atomic mass. Absorption rounds off the sharp edge at the critical angle. Because of the small opening angle of SR totally reflecting mirrors provide efficient solutions even when $\theta_c$ is only a few mrad in a typical case. Ideal point-to-point focusing is achieved with an elliptical mirror, when the source is at one of the focii. In practice, mirrors are ground to cylindrical shape in the sagittal direction and bent in the meridional direction. In addition to focusing, an X-ray mirror acts as a low bandpass filter, as the critical angle is inversely proportional to the X-ray energy. In many experiments, the high-energy harmonics from the monochromator cannot be allowed, and they are rejected by a mirror set to reflect the fundamental energy.

Mirror technology has advanced enormously in recent years. The rms surface roughness of 1 m long mirrors is on the level of atomic size, and the figure errors are less than 1 $\mu$rad. The usual mirror material is Si or SiC, and the mirror can be coated by a thin metal layer to provide the desired critical angle. In order to benefit from the high quality of the mirrors their shape must be maintained even under the high heat load of the SR beam. The mirror is water-cooled, and its shape is monitored by an optical sensor. The sensor has feedback to piezoelectric actuators, which correct for the effects of thermal deformations of the mirror.

### B. Multilayers

Multilayers are synthetic periodical structures, where alternating layers of light and heavy elements are deposited on a substrate. Typically, the layer thickness for each crystal element is 5 to 10 atomic layers, so that the period in the direction of the surface normal is 20 to 50 Å. Because of the large period, the Bragg angles are small, as seen in Fig. 10. The figure shows also the regime of total reflection at small incident angles. The relative width of the Bragg reflection from a multilayer is $1/N$ in the energy scale, where $N$ is the number of periods (typically between 100 and 1000), so that multilayers can be used as wide bandpass monochromators and focusing elements. Actually, the X-ray optical properties of multilayers are

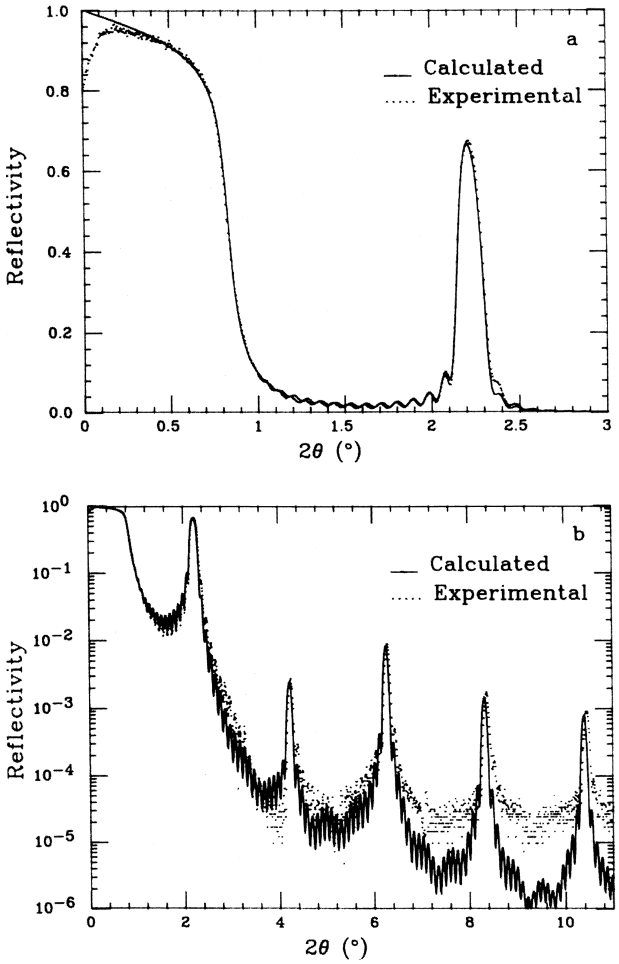



**FIGURE 10** Reflectivity of a W/C synthetic multilayer structure for 8 keV X-rays. The regime of total external reflection is seen at small angles, and the first multilayer Bragg reflection at 2.2°. The higher order reflections are strongly suppressed.

somewhere between those of mirrors and crystals, which make them useful in many applications.

## C. Perfect Crystals

It was already seen that perfect crystals are totally reflecting in a narrow range about the Bragg angle, when absorption is small. Perfect crystals, Si in particular, are used as X-ray optical elements that separate a narrow energy band from the polychromatic SR beam. The width of the relative energy or wavelength band is constant for a given reflection,

$$\delta E/E = r_e\,(2d)^2 K_{\mathrm{pol,d}} F(hkl)/\pi V_c \qquad (45)$$

which is easily derived from Eq. (31). The width of the relative energy band varies from $10^{-4}$ of the low order reflections to $10^{-7}$ or $10^{-8}$ for the high-order reflections used in back-reflection geometry. The angular width of the reflection is proportional to the wavelength, and it varies from a few arcseconds to the submicroradian range. The other contribution to the width of the energy band comes from the variations in the incident angle. From the Bragg law,

$$\Delta E/E = -\cot\theta\,\Delta\theta \qquad (46)$$

The standard monochromator construction is the so-called nondispersive or antiparallel setting of two identical crystals, as shown in Fig. 11. It is evident that the wavelength band reflected by the first crystal is reflected by the second, and the propagation direction is conserved. There is an offset of the beam, and various geometrical and mechanical solutions have been introduced to keep the offset constant when the energy is changed by rotating the crystals. The first crystal is cooled, because the heat load of the SR beam causes distortions and large changes of intensity and energy. Perhaps the most advanced solution is to cool the crystal to about 100 K by liquid nitrogen, because at that temperature the thermal expansion coefficient of Si goes to zero, and the thermal conductivity has a maximum. The second crystal may be bent sagittally for horizontal focusing, and different constructions have been put forward to maintain the cylindrical shape under dynamical bending.

Bent perfect crystals are used also for meridional horizontal focusing, particularly at high SR energies, where the geometrical aberrations of standard constructions become large due to small Bragg angles. Such an arrangement is shown in Fig. 12. The beam is first monochromatized by a cylindrically bent Laue-type crystal, and then focused in the horizontal plane by a Bragg-type crystal. Vertical focusing is obtained by a multilayer, which is placed between the monochromator and sample. Focus sizes of a few micrometers have been achieved, which allows probing local strains in bulk samples, for instance.

## D. Other X-Ray Optical Elements

All X-ray optical elements of SR beamlines are based either on refraction or diffraction. An interesting combination of the two phenomena is used in so-called Bragg–Fresnel lenses. Focusing by Fresnel zone plates is well-known in optics, and when the zone plates are made of perfect single crystals they can act as narrow-band monochromators at the same time. Photographs of linear and circular Bragg–Fresnel lenses are shown in Fig. 13. The special feature of the structure is that the X-rays reflected from the bottom of the grooves interfere constructively at the same place as those reflected from the elevated part of the lens. This is achieved by an appropriate depth of the groove, which is typically a few micrometers. Circular Bragg–Fresnel lenses form a point focus, but they can

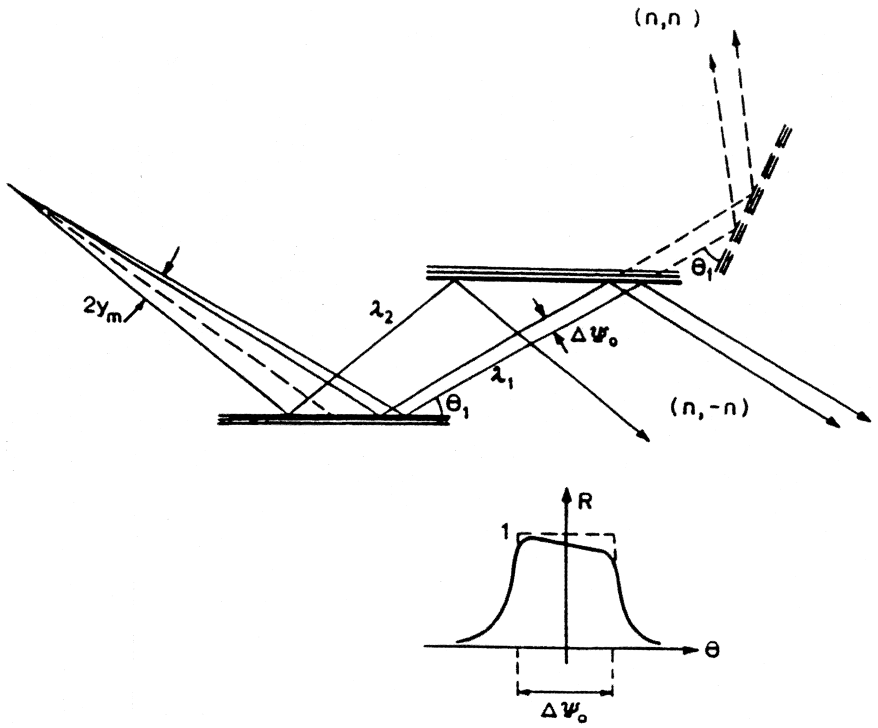



**FIGURE 11** Geometry of double-crystal monochromator in nondispersive ($n, -n$) (solid lines) and dispersive ($n, n$) (broken lines) settings.

be used only near normal incidence of the beam, which limits the available X-ray energies. Linear lenses form a line focus, but they can be bent cylindrically to focus also in the other direction. Focus sizes in the micrometer range have been achieved. The limitation of the Bragg–Fresnel lenses is their small size, which limits the useful aperture of the SR beam.

The fact that the refractive index of X-rays is smaller than unity can be utilized for construction of X-ray lenses. For a long time, this was considered impossible, because

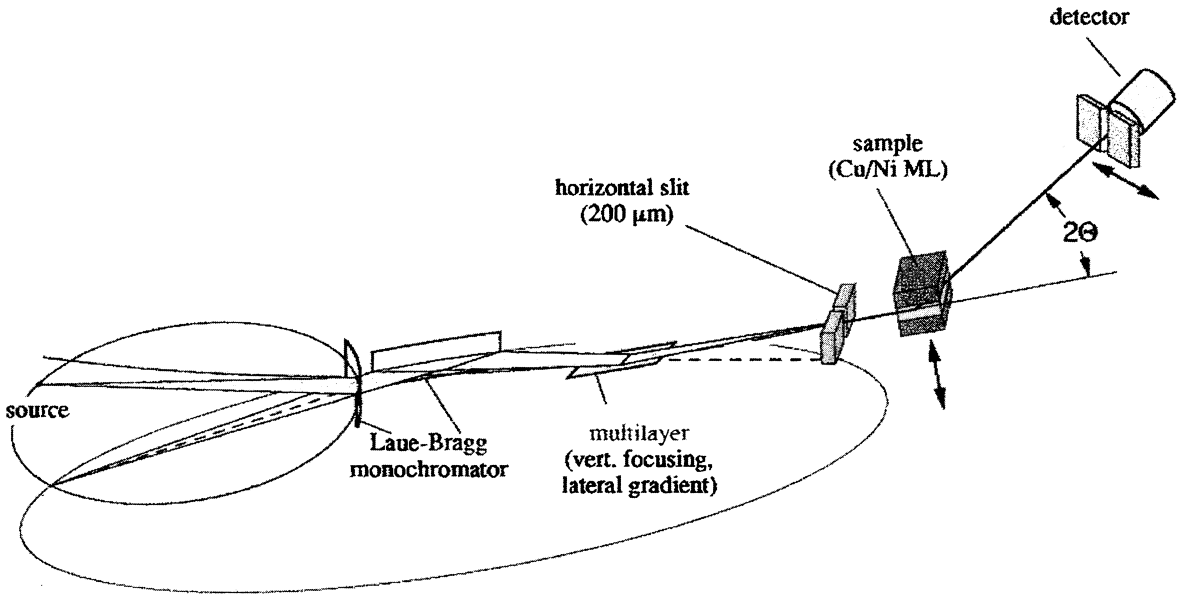



**FIGURE 12** Laue–Bragg monochromator. Horizontal focusing is achieved by bent crystals in the geometry where the source points and focii are on the Rowland circles. Vertical focusing is realized by a bent multilayer. [*ESRF Highlights* 1996/1997, reproduced with permission.]

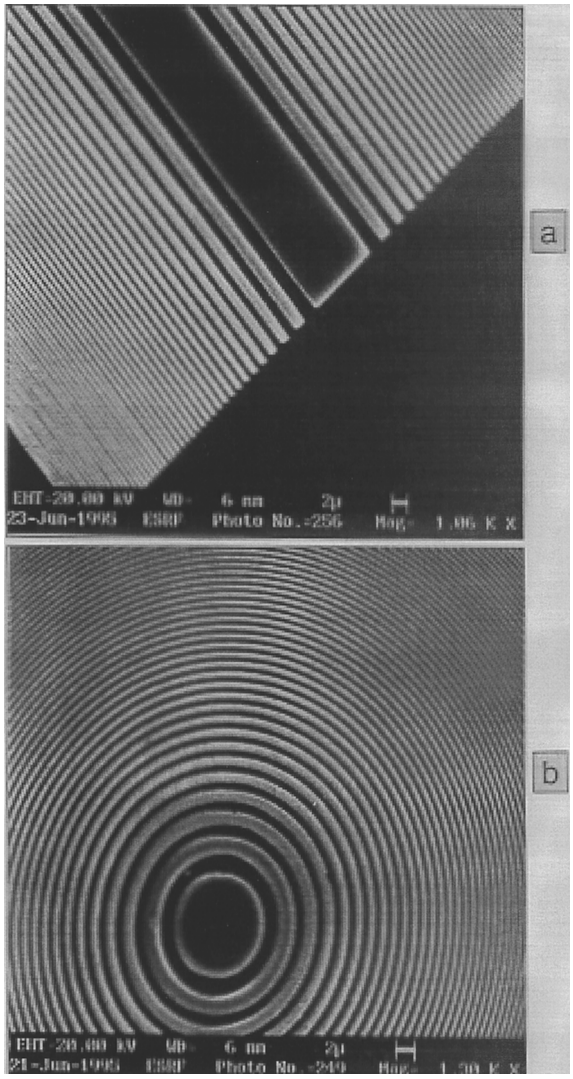



**FIGURE 13** Linear and circular Bragg–Fresnel lenses fabricated on Si wafers. [*ESRF Highlights* 1994/1995; reproduced with permission.]

typically $\delta$ is of the order of $10^{-5}$. However, a spherical cavity inside a solid acts as a condensing lens, and if a long row of holes is milled in a weakly absorbing material such as Be, a beam traveling along the row is focused. Focusing in the other direction is obtained by a crossed row of holes. Again, focus sizes in the micrometer range are achieved.

Still another method of microfocusing is that of using tapered capillaries. The SR beam enters the capillary at the wide end, and after successive total reflections from the inner walls of the capillary, exits from the other end. Submicron beam sizes have been achieved by this method. The methods described above have their advantages and disadvantages. With the exception of the bent crystal scheme, the useful aperture of the focusing element is small, typically less than 1 mm. There are losses due to scattering and absorption, and it always has to be remembered that the phase space density of photons cannot be increased. The price paid for high spatial resolution is the corresponding deterioration of the angular resolution, which limits the use of microfocusing in diffraction studies.

## E. Beamline

It was already mentioned that the power of the SR beam from an undulator or wiggler is typically several kW. Therefore, the optical elements and beamline instruments must be placed in "hutches," which provide radiation shielding. For instance, the spectrum of the radiation from the superconducting wavelength shifter at the ESRF extends beyond 1 MeV, and 60 mm of lead is needed for shielding. The beamline components are under vacuum to prevent contamination and production of ozone by the SR beam. An example of a beamline is shown in Fig. 14. The beam enters the first optics hutch through the shield wall. It is limited by fixed apertures, and usually the "soft" low-energy part of the spectrum is removed by absorbers. The beam position is monitored, and the size of the beam is limited by primary slits to match the optical elements downstream. An adaptive mirror focuses the beam to infinity and removes high energies, a narrow energy band is separated by a two-crystal monochromator, and the beam is focused on the sample by the second mirror. At high energies, where the critical angle of the mirrors becomes too small, the mirrors can be moved away, and horizontal focusing is achieved by the sagittally bent second crystal of the monochromator. All the components of the beamline are controlled remotely, and there are several feedback loops, which are used to maximize and stabilize the intensity of the exit beam. Usually the sample is enclosed in a chamber, where it is under vacuum or high pressure, it can be heated or cooled, and electric or magnetic fields can be applied. In diffraction studies, usually a four-circle or even six-circle diffractometer is used. Two angular motions are needed to orient the crystal for fulfillment of the diffraction conditions, and two motions are needed for rotation of the crystal during the scan and for tracking the diffracted beam by the detector. There may be an analyzer crystal in front of the detector to remove parasitic scattering. However, more and more two-dimensional detectors are used, which eliminates the detector movement and increases the speed of data collection. On the other hand, most such detectors operate in the integrating mode without any energy resolution, which increases the background in the observed diffraction pattern. A modern diffractometer is shown in Fig. 15.

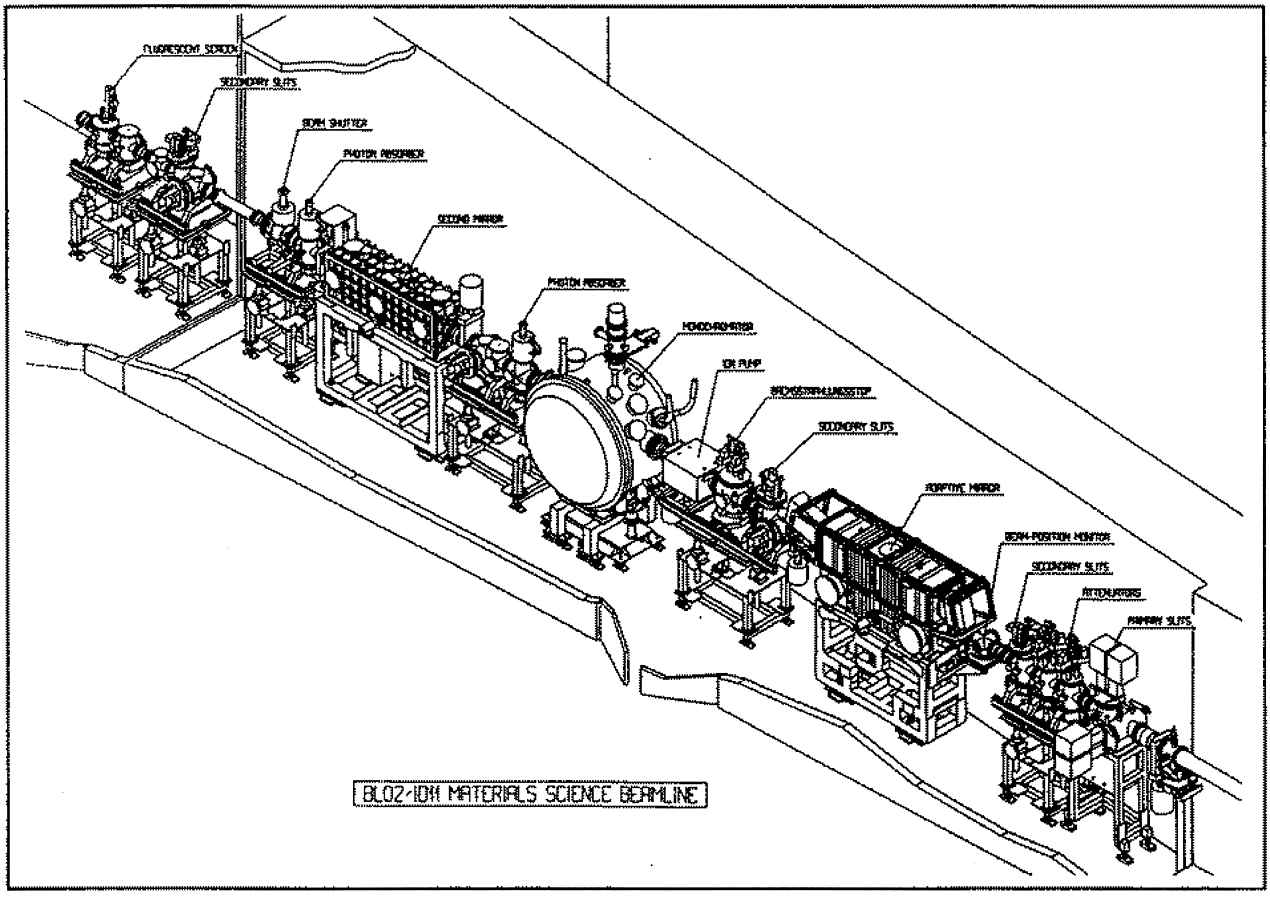



**FIGURE 14** X-ray optical components of a wiggler beamline at the ESRF.

## VIII. X-RAY DIFFRACTION METHODS AND APPLICATIONS OF SYNCHROTRON RADIATION

The conventional X-ray diffraction methods are covered in another article (K. Ann Kerr, 1987), and the present one concentrates on new developments, which originate mostly from SR laboratories. The methods are numerous, and only short descriptions of the main areas of research are possible.

### A. Structure Determination, Protein Crystallography

Recent reviews of protein crystallography based on the use of SR have been given by Bartunik (1991) and Helliwell (1992). The first steps of structure determination—growing the crystal, determination of the unit cell and the orientation matrix—have stayed the same as before, but there are some important advances. Very small crystals can be used, since $10^{10}$ to $10^{11}$ photons/s can be focused on a sample of 10 $\mu$m diameter. Peak search and orientation of the crystal is greatly facilitated by area detectors, most notably by image intensifiers coupled to fast readout CCDs. In general, area detectors with associated software have made collection and analysis of diffraction data a routine step in structure determination. On-line data analysis runs automatically as diffraction patterns are collected and keeps up with the data acquisition, so that the experimenter knows the quality of the data, completeness etc. A graphical user interface interacts with different processes (beamline control, data acquisition, and data analysis) running on different computers at different times.

Diffraction experiments with monochromatic radiation use mostly the rotating crystal method to scan reciprocal space. A diffractometer, such as the one shown in Fig. 15, is used for crystal orientation and scanning. The data rates are enormous, because the area detector has $10^6$ to $10^7$ pixels, and the dynamic range is 14 bits. It is estimated that the rate of uncorrected data may be up to 150 GB/day. Evidently, mass storage and structured data base are needed to handle the data. The essential step is reduction of the data to integrated intensities of the reflections, indexing the reflections, and making corrections for absorption, beam

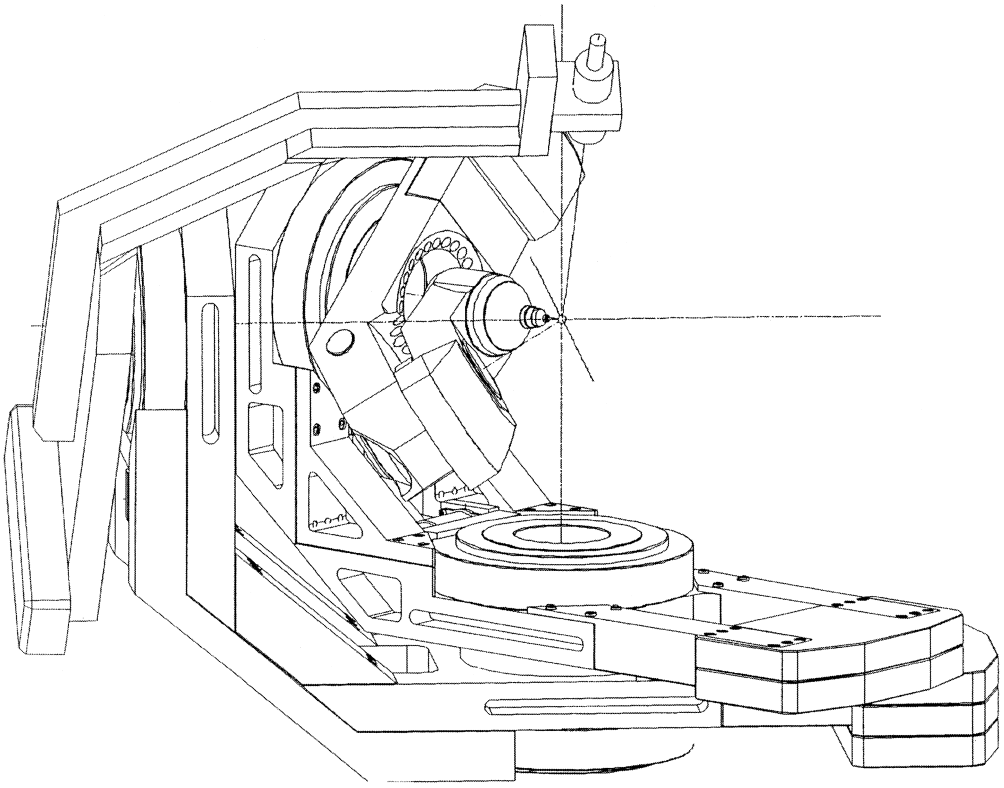


**FIGURE 15** Six-axis Kappa diffractometer at the Materials Science beamline of the ESRF. All axes are mechanically independent, and they allow free choice of the scattering plane. The detector arm can be equipped with a crystal analyzer or an area detector.

polarization, geometrical factors, and probably for extinction. The resulting structure factors without information about their phase angle give the starting point for crystal structure analysis.

Synchrotron radiation provides new possibilities for experimental determination of the phase angles of the structure factors. The use of multiple diffraction was mentioned earlier, but the methods based on anomalous dispersion are more important at the moment. The atomic scattering factor changes strongly near the absorption edges [cf. Eq. (4)]. The principle of phase determination from structure factors where the scattering factor of one type of atoms is changed is shown in Fig. 16. The changes $\Delta f'$ and $i\,\Delta f''$ can be calculated from theory, but the free-atom values may be modified in the crystal. The imaginary part, which is due to photoelectric absorption, can be determined experimentally by measuring the fluorescence signal as a function of energy, and the real part can be calculated from $f''$ by Kramers–Kronig relations. Unambiguous phase angles are obtained from measurements with two or three X-ray energies, where the changes in $f'$ and $f''$ are large. Most native proteins do not have heavy atoms with absorption edges in the range of energies suitable for diffraction studies, i.e., above 5 keV. Preparation of heavy-atom derivatives is a standard problem in protein crystallography, and there is much empirical knowledge. A quite general technique is to introduce Se during biosynthesis of proteins, and the K-edge of Se is at the convenient energy of 12.66 keV.

The use of continuous radiation and stationary crystal, i.e., the Laue method, has had a renaissance with the advent of synchrotron radiation. The intensity of SR as a function of energy can be calculated precisely, so that the integrated intensities of the reflections recorded with different energies can be converted to structure factors on a common scale. A diffraction pattern is shown in Fig. 17. It is remarkable that this pattern was taken in a single shot of 150 picoseconds using an image intensifier coupled to a CCD camera, i.e., the pattern is due to diffraction of X-rays from a single electron bunch of the storage ring. On one hand, this means that it is possible to determine structures of proteins which disintegrate rapidly due to radiation damage. It is also possible to make pump-and-probe experiments, where a conformational change of a protein is triggered by a laser pulse, and by varying the time difference to the X-ray pulse the evolution of a change in structure can be followed. The changes of the structure factors are observed, and the changes of the

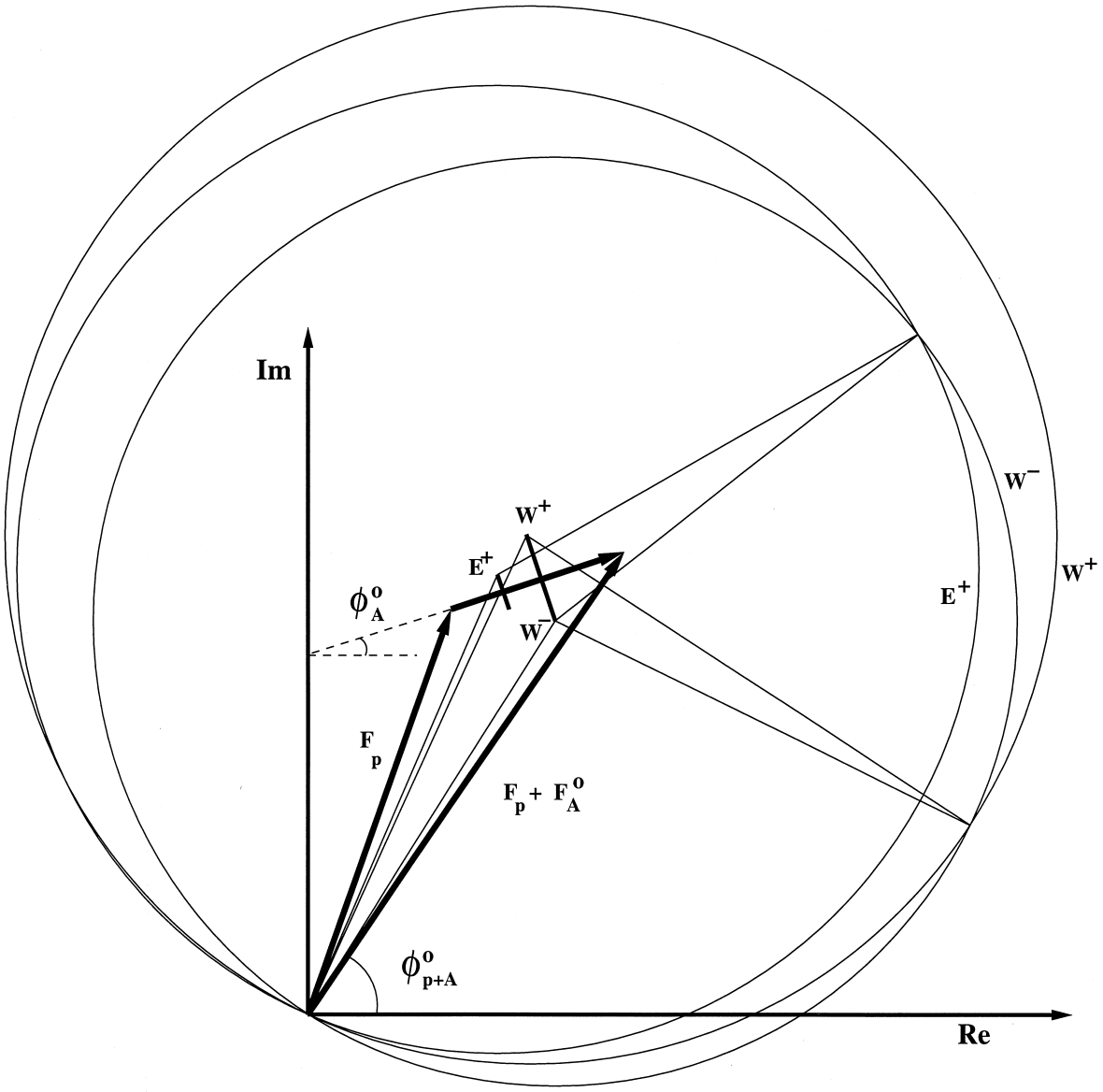



**FIGURE 16** Determination of the phase angle of the structure factor by using the multiple-wavelength anomalous dispersion (MAD) method. $F_A$ is the structure factor contribution of the anomalous scatterer, and $F_P$ that from the rest of the unit cell. The phase angle $\phi$ is solved from the intersection points of the circles of radius $|F|$ (Harker diagram).

electron density are calculated from Eq. (27). One recent example is the study of binding and release of CO at the heme site of myoglobin. The movements of CO have been studied extensively by infrared spectroscopy and molecular dynamics calculations, but time-resolved diffraction of SR provides a detailed picture of the conformational changes of the protein. The movement of the CO molecule to the docking site, where it stays ca 350 ns before leaving the pocket, can be seen. Then the molecule diffuses about in the outer protein for a fraction of millisecond and comes back to the Fe by random collisions.

Radiation damage is a limiting factor in protein crystallography. Most of the radiation falling on the crystal is not used, either as wavelengths that are not reflected in the Laue method, or during the time spent between the reflections in monochromatic diffraction. Radiation damage is not a linear function of dose, but also scales with the total duration of the experiment. Therefore, efficient data collection strategies are essential. It turns out that the damage caused by high-energy radiation is substantially less than damage at the energies normally used in protein crystallography.

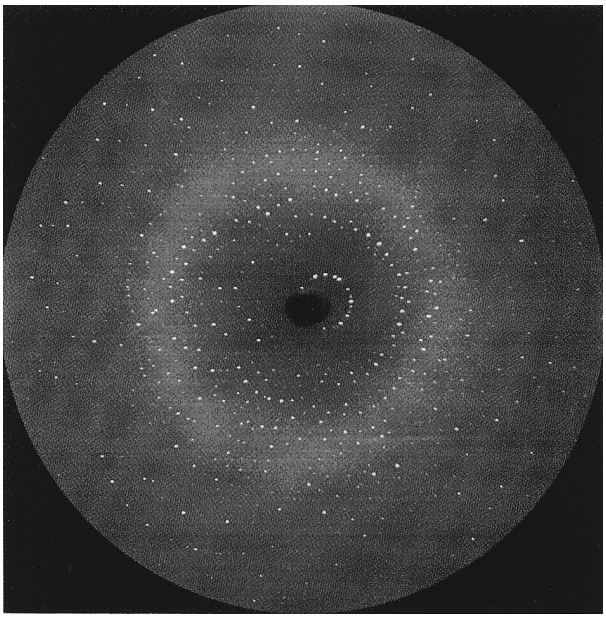


**FIGURE 17** Laue diffraction pattern from a protein crystal taken in a single shot of 150 ps from an ESRF undulator. The bandwidth is 7–11 keV, and the image was recorded by a CCD camera coupled to an image intensifier. [*ESRF Highlights* 1994/1995; reproduced with permission.]

## B. Fiber Diffraction

It was already mentioned that many materials and tissues found in nature have hierarchical structures. This means that there are many different length scales, and the diffraction pattern extends from ultrasmall angles to the wide-angle regime. Typically, the structures are built of long, chainlike molecules, which form fibrils, and these in turn are packed in ordered bundles. The interatomic distances in the molecule are seen in the wide-angle regime of diffraction, and the effects of the size and shape of the bundles in the ultrasmall-angle regime. Between these extremes, ordering is seen as diffraction maxima. Collagen is a good example of these hierarchical structures. There are many variants of collagen, and study of collagen structures and functions is a wide field of research by itself. Diffraction studies with conventional X-ray sources were limited to well-ordered collagen structures, such as tendon. The high intensity and collimation of SR has made possible high-resolution studies in the whole angular range. An example is given in the following.

Human dermis under the epidermis is mostly composed of fibroblasts in a matrix of collagen with small amounts of proteoglycans and elastin. The molecular structure of collagen is a triple helix, which is about 3000 Å long. The molecules form twisted fibrils, where adjacent molecules are parallel and shifted by 650 Å. In the dermis, these fibrils are also parallel forming bundles of 20 to 100 fibrils, and the bundles are more or less randomly oriented. The exact dimensions of these structures can be seen in the diffraction pattern. Only the small-angle and ultrasmall-angle part of the pattern is shown in Fig. 18. The diffraction maxima due to the 652 Å periodicity in the direction of the fiber axis are easily identified. One maximum arises from the 1150 Å spacing in the hexagonal packing of the fibrils in a bundle. In addition, there are oscillations that correspond to the side maxima of the Fourier transform of the cylindrical shape of the fibrils. The diameter of the fibrils is estimated to be about 1100 Å. The size of the bundles is too large to be evaluated reliably from the intensity at ultrasmall angles. This interpretation of the collagen structure in human dermis is summarized in Fig. 18.

## C. Surface Structures

The present understanding of the atomic structure of surfaces is to a large extent based on studies with SR. Because of the high intensity, a sufficient signal can be obtained from an atomic layer, and total reflection at grazing incidence largely eliminates the signal from bulk. The standard geometry of surface diffraction studies is shown in Fig. 19. The SR beam is incident at a small angle $\alpha$, and diffraction is observed as a function of the lateral angle $2\theta$ along a line perpendicular to the surface.

Surface diffraction takes place in a few top atomic layers of the surface, so that propagation of X-rays on the surface must be studied more closely. The critical angle $\alpha_c$, which is given by Eq. (44), corresponds to a wave vector transfer $K_c = 2k \sin\alpha_c \approx 2k\alpha_c$, which is independent of wavelength. Snell's law can be written at small angles as $\alpha^2 = \alpha'^2 + \alpha_c^2$. This implies that $\alpha'$ must be imaginary for $\alpha < \alpha_c$, which physically means that the transmitted wave at grazing angle $\alpha'$ is transformed into an evanescent wave propagating along the surface with an exponentially damped intensity profile. The penetration depth approaches $1/K_c$ at the limit $\alpha \ll \alpha_c$. The evanescent wave can be considered as an incident plane wave, which will be diffracted by the in-plane structures.

Surface structures can be classified as termination of the three-dimensional crystal at the surface, reconstruction of the topmost atomic layers of the surface, terraces of atomic layers on inclined surfaces, ordering of adsorbed mono- or submonolayers on crystal surface, and structure of films on surfaces. The termination of the crystal gives rise to so-called truncation rods, as the two-dimensional surface structure is Fourier-transformed to one-dimensional structure in reciprocal space. The relps of the bulk crystal lie on these rods, but the intensity is not zero between the relps. Formally, this is most easily seen by using the convolution theorem. The three-dimensional $\delta$-function of the crystal

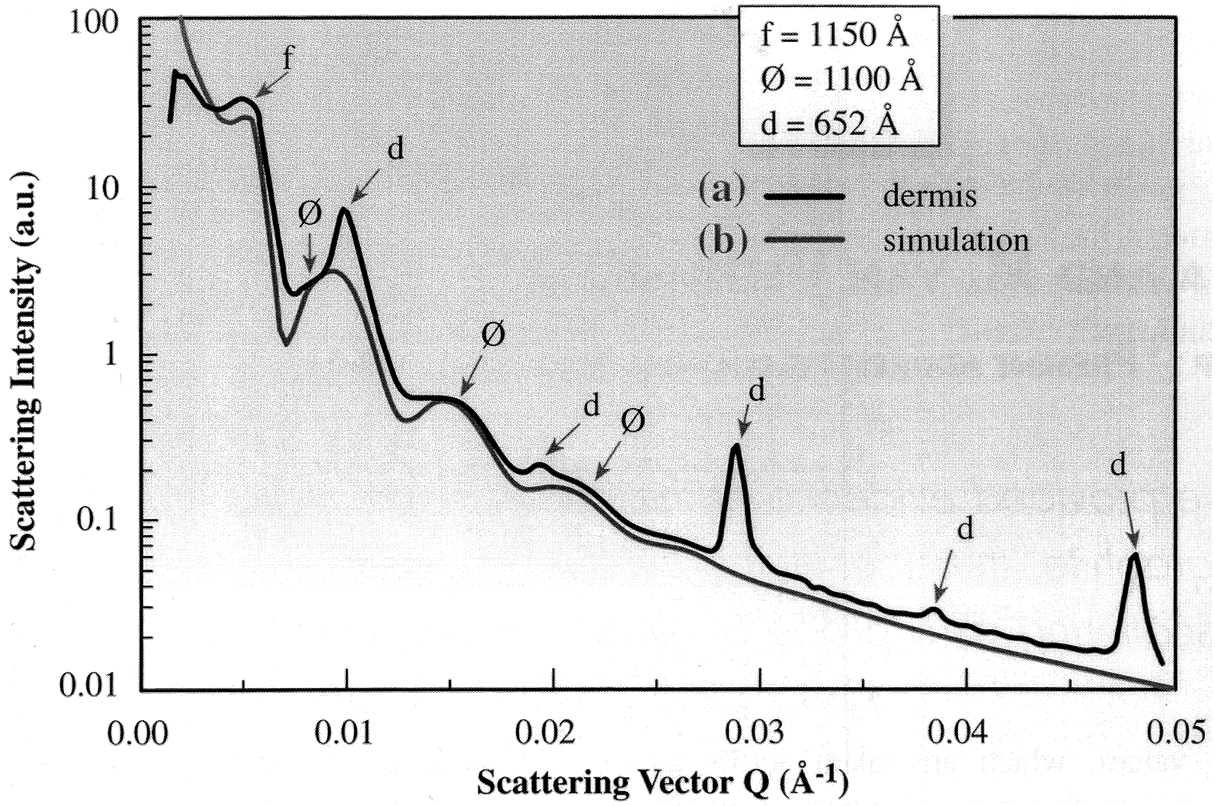



**FIGURE 18** Small-angle diffraction pattern from human dermis. The maxima in the diffraction pattern arise from the axial periodicity of the collagen, the diameter of the fibrils and their hexagonal packing. [*ESRF Newsletter* No. 29, 1997; reproduced with permission.]

lattice is multiplied by a step function. The reciprocal lattice is then convoluted by the Fourier transform of the step function, which is of form $1/(i\mathbf{K}\cdot\mathbf{c})$, if the lattice vectors **a** and **b** are in the surface plane. The atomic layers on the surface reorganize themselves spontaneously, forming new bonds with other atoms on the surface and in the layers underneath. This produces two-dimensional structures, which do not have the same periodicity as the terminated crystal, and in reciprocal space fractional order rods are observed. Similar structures develop when atoms are adsorbed on a surface. An example is shown in Fig. 20. Oxygen atoms on the (110) surface of Cu create a rectangular unit cell, the so-called c6×2 structure. Studies of the surface structures are important from the technological point of view, but also for understanding the mechanisms of phase transitions, for instance. A review has been given by Robinson and Tweet (1992).

Solid films of organic molecules can be grown on a gas–water interface; these monomolecular layers are called Langmuir layers. In these films the hydrophilic part of the long molecule is embedded in the water subphase, whereas the hydrophobic part, often aliphatic chains, points towards the gas. Studies of the structures and ordering of Langmuir films are important for many reasons. The films are model systems for studies of two-dimensional ordering processes, including superstructure and domain formation. Intermolecular interactions can be changed in wide limits by varying the surface charge density. Langmuir layers are also important models in membrane biophysics for the study of lipid–lipid and lipid–protein interactions. Films can be "peeled" from the water surface to form mono- or multilayers, called Langmuir–Blodgett films, which have unique optical and electronic properties. Information of the lateral structure of the Langmuir films is based mostly on surface diffraction studies by SR. Because of the relatively large thickness of the film, the contribution of the substrate is rather small. A review of these studies has been given by Als-Nielsen and Möhwald (1991).

### D. Microdiffraction

It was already mentioned that the intensity of SR at an ESRF crystallography beamline is $10^{10}$ to $10^{11}$ ph/s in a spot of 10 $\mu$m diameter. The beam size can be further reduced by capillary optics, and at the microfocus beamline of the ESRF the flux in a 2 $\mu m^2$ spot is $4 \times 10^{10}$ ph/s.

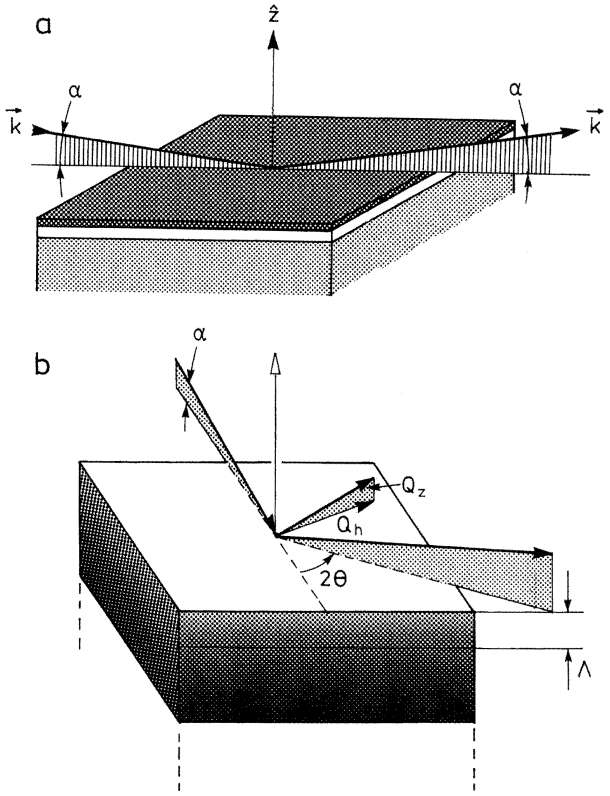



**FIGURE 19** Geometry of surface diffraction. Specular reflection (top) probes the density profile across the surface layer. When the angle of incidence is less than the critical angle the evanescent wave is diffracted by the in-plane structure of the surface layer (bottom).

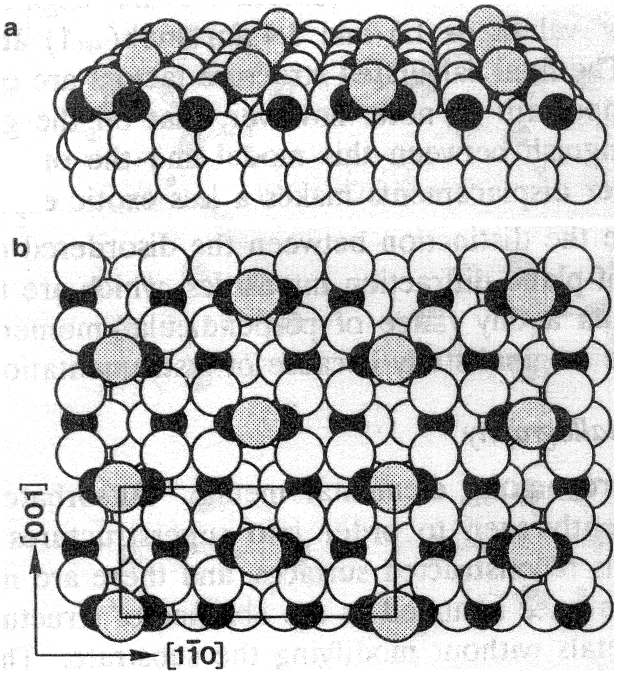



**FIGURE 20** Oxygen atoms (black) on the Cu(110) surface (white). There are two different positions for the O atoms, and their density is 2/3 of the Cu atoms in a nonreconstructed layer. The structure is called c6 × 2, and the unit cell is indicated by the rectangle. [From Robinson, I. K., and Tweet, D. J. (1992). *Rep. Prog. Phys.* **55,** 599–651.]

The divergences of the beam increase in proportion, so that such a beam is not suitable for studies where high angular resolution is needed. Some length scales are summarized in Fig. 21. An example of fiber diffraction was given earlier, but there the sample was relatively large. A microfocused beam allows diffraction measurements on a single fiber, and recent examples include structural studies of spider silk and cellulose fibrils in a single wood cell wall (*ESRF Highlights*, 1999).

Perhaps the most important applications are those where microdiffraction is combined with the use of other microprobes. One such combination is simultaneous mapping of X-ray fluorescence and diffraction patterns. Fluorescence from the sample is recorded by a solid-state detector, such as Li-drifted Si, and the diffraction pattern is recorded by a two-dimensional CCD detector. The probing beam is of the order of 1 $\mu$m, and the sample is scanned in even smaller steps. In general, with the resolution of 1 $\mu$m diffraction has become one of the imaging methods. In its most advanced form, tomographic imaging provides three-dimensional morphology and elemental

## MICROPROBES

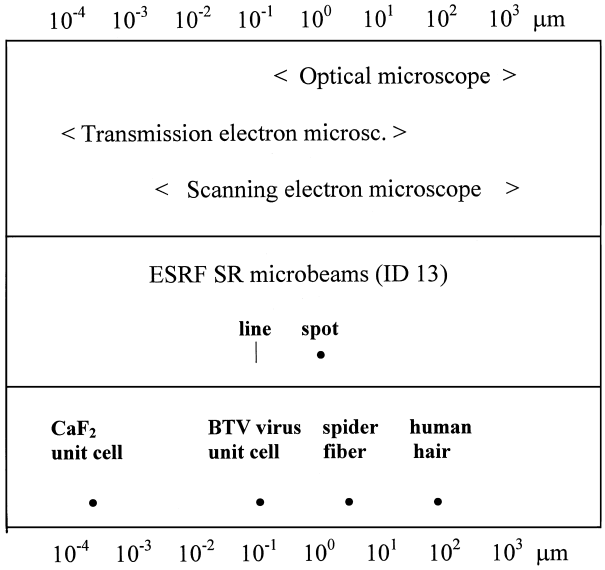



**FIGURE 21** Comparison of the spatial resolution of microdiffraction with the resolution of microscopes. Sizes of different objects are indicated in the bottom panel.

composition of an object, and diffraction adds another dimension by revealing the structures on the molecular and atomic level.

### E. Powder Diffraction

One of the most important applications of X-ray diffraction in materials science is powder diffraction. This method, which is widely used in nondestructive testing of materials, has experienced a renaissance as a tool in structure research. The high intensity, collimation, and narrow energy band of a monochromatic SR beam have virtually eliminated the instrumental broadening of powder reflections. It is possible to study small samples in special environments, to record the full diffraction cones by two-dimensional detectors for analysis of the texture of the sample, or to insert a crystal analyzer between the sample and detector to eliminate all incoherent scattering. Many interesting materials are not available as sufficiently large crystals for single-crystal diffraction, and powder diffraction may be the only possibility in structural studies of materials with coexisting phases. The density of reflections becomes very large when the unit cell dimensions approach 15 or 20 Å. Several pattern decomposition and indexing programs are available, and direct methods can be used for solving the structure. In the final phase, the atomic positions and even thermal parameters can be refined by the *Rietveld method*, where the calculated diffraction pattern is fitted to the observed pattern by least-squares methods (see Section X.B.)

### F. Other Diffraction Studies by SR

The preceding areas of diffraction studies by SR represent only a part of a very broad field. Small-angle scattering (SAXS) is covered by another article; topography has expanded from studies of crystal defects to phase contrast imaging where the coherence of the SR beam is utilized; structural changes due to magnetic ordering are seen by high-energy diffraction with submicroradian resolution. Magnetic structures themselves can be studied by SR diffraction, as a very large enhancement of the signal is seen near the absorption edges. A new field is diffraction where the energy of the incident beam is tuned across an absorption edge of the compound. This method provides at least in principle more complete information than EXAFS spectroscopy. Diffraction of coherent SR reveals correlations in the sample, which are averaged out when incoherent radiation is used. When coherent radiation is scattered from a disordered system, it gives rise to a random but reproducible "speckle" pattern, which is related to the exact spatial arrangement of the disorder. When this evolves with time the speckle pattern changes, and observation of intensity fluctuations at a single point of the pattern is a direct measure of the dynamics. The short wavelength of X-rays provides atomic-scale resolution, so that critical fluctuations near an order–disorder transition can be studied, for instance. Still another important field is studies of structures under high pressure. In diamond anvil cells small amounts of solids can be subjected to pressure that is close to that at the center of the earth. Understanding the structure of solids under such conditions is important in geophysics or even in astrophysics, and high-pressure studies also address the basic questions of interactions between atoms. On the methods side, the continuous and precisely known spectrum of SR makes quantitative diffraction measurements with energy-dispersive techniques possible. In this method, scattering is observed at a fixed angle, and the spectrum is analyzed by a solid-state detector, i.e., the diffraction pattern is recorded in the energy scale using the Bragg law [Eq. (26)].

In summary, the use of synchrotron radiation has made possible many new fields of research where X-ray diffraction is used. Perhaps the greatest impact is due to the use of the coherence properties of SR in imaging and studies of correlations, and on the other hand, due to combination of various methods in the submicron-length scales.

## IX. NEUTRON SOURCES AND NEUTRON OPTICS

### A. Nuclear Reactors and Spallation Sources

The neutrons for scattering experiments are always produced in large facilities, because there is no equivalent to the small X-ray laboratory unit. The size and operation practices of the research reactors and *spallation* sources are similar to those of synchrotron radiation facilities. The source is surrounded by a large number of experimental stations, which extract different distributions of X-rays or neutrons, and have different instruments for the experiments. The most common sources of neutrons are nuclear reactors, which are based on the continuous, self-sustained fission reaction. In the research reactors the power density is maximized in the volume that "leaks" the neutrons out. The fuel rods are made of highly enriched $^{235}U$. The distribution of neutron spectrum is centered about 1 MeV; most of them are moderated in the cooling liquid, $D_2O$ or $H_2O$, and are absorbed in fuel to propagate the reaction. As large a fraction as possible is allowed to leak out as fast neutrons into the surrounding moderator and to slow down to thermal equilibrium with this moderator. The

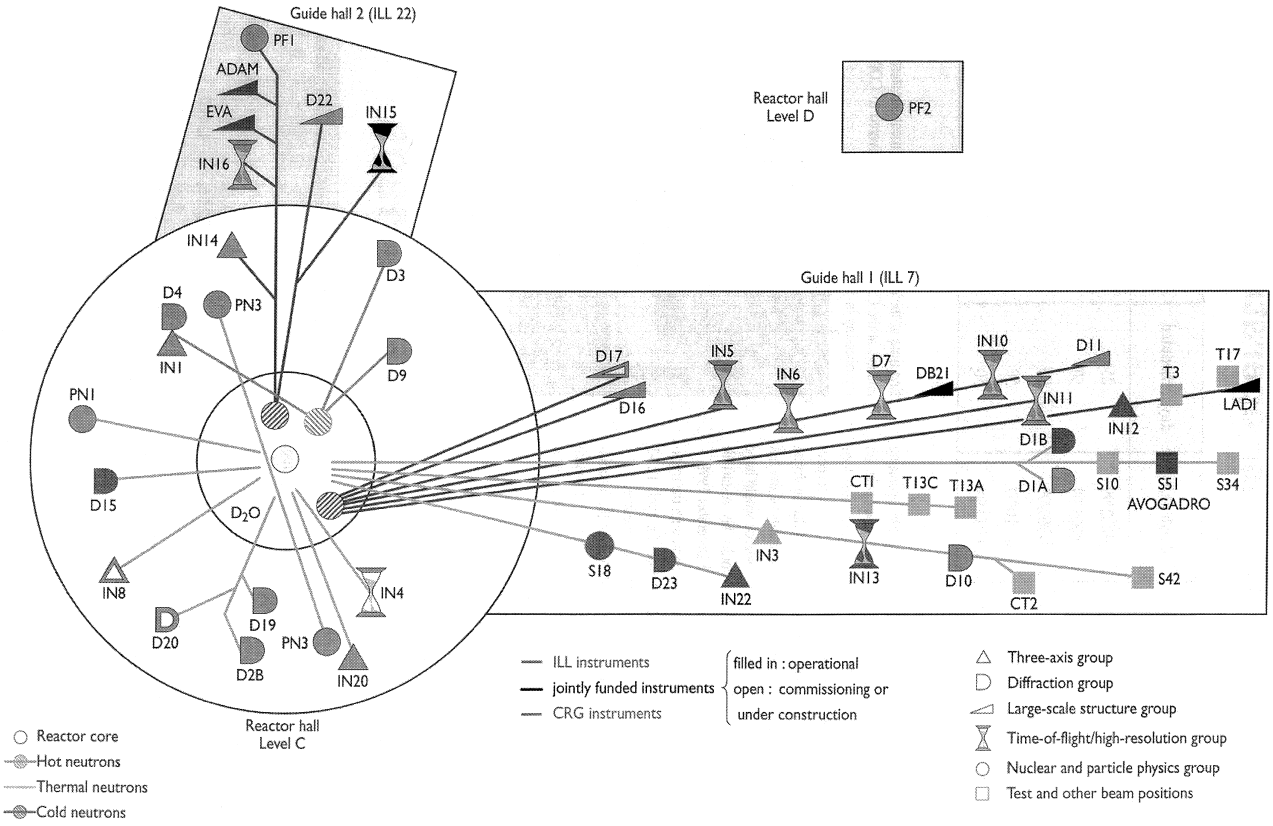



**FIGURE 22** Layout of the experimental facilities of the Institut Laue-Langevin (ILL) in Grenoble, France. The diffraction experiments are indicated by letter "D," and these include single-crystal and powder diffractometers (e.g., D1) and small-angle scattering instruments (e.g., D11). In addition, there are instruments for inelastic scattering ("IN") and other experiments. Also, the spectral range of neutrons (hot, thermal, cold) is indicated. [*ILL Annual Report* 1999; reproduced with permission.]

mean energy of the Maxwellian distribution at 300 K is 38 meV, which corresponds to 1.8 Å wavelength [cf. Eqs. (1) and (2)].

Neutron beams are extracted through holes that penetrate the moderator. The layout of the experimental stations around the High-Flux reactor at the Institut Laue-Langevin in Grenoble, France, is shown in Fig. 22. To shift the spectrum in energy, a cold source (liquid deuterium at 25 K) and a hot source (graphite at 2400 K) have been inserted in the $D_2O$ moderator. These extend the wavelength range between 20 Å and 0.2 Å.

The second method of producing neutrons is with charged particles (usually protons) striking target nuclei. Short bursts ($<1$ $\mu$s) of about 1 GeV protons produce high-energy "evaporation" neutrons, which have energies extending close to that of the incident proton beam. The neutrons are moderated as rapidly as possible in hydrogenous materials such as polyethylene to provide a short pulse of "slow" ($E_n < 10$ eV) neutrons. The frequency of the pulses is between 10 and 100 Hz.

There are fundamental differences between experiments carried out at steady reactor sources and those made at pulsed sources. This is illustrated in Fig. 23. A crystal monochromator is used to separate a narrow energy band from the distribution of thermal neutrons, and this is used for diffraction experiments much like the beam from an X-ray tube. On the other hand, the beam from a pulsed source has a wide energy distribution, and for an efficient use of the beam time-of-flight (TOF) method is used. The pulse length is 1 to 50 $\mu$s, depending on energy, and their separation is 10 to 100 ms. The average intensity is low, but very high within each pulse. With the TOF method, all neutrons scattered to a given angle are exploited. The diffraction pattern is analyzed as a function of the neutron wavelength, which is related to the flight time $\tau$ over distance $L$,

$$\lambda\,(\text{Å}) = 0.3966\tau(\mu\text{s})/L\,(\text{cm}) \qquad (47)$$

Actually, the TOF method is rather similar to the energy-dispersive diffraction method with X-rays, the difference

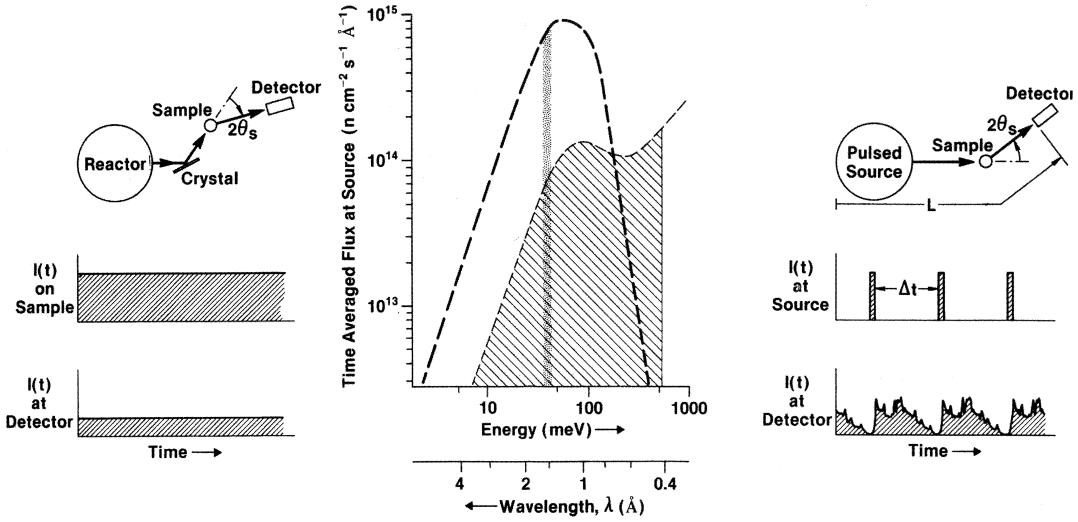



**FIGURE 23** Comparison of steady-state (reactor) and pulsed (e.g., spallation) neutron sources. An energy band is separated from the Maxwellian distribution of thermal neutrons by a crystal monochromator, while a wide spectral band from the pulsed source is used for time-of-flight experiments. [*International Tables for Crystallography*, (1999). Vol. C, reproduced with permission.]

being that the energy analysis of the neutron is done by their speed.

## B. Neutron Optics

Many of the optical elements used in neutron diffraction are similar to those used to focus and monochromatize X-rays or SR. The main differences come from the nature of the neutron source, which has large dimensions, and from the fact that the neutron velocity is typically five orders of magnitude smaller than the velocity of X-rays. Absorption of neutrons is weak, allowing construction of many devices that could not be used with X-rays.

Mirrors and multilayers are widely used in neutron optics. The refractive index is less than 1, so that total external reflection takes place below the critical angle $\theta_c$. This angle is small and proportional to the neutron wavelength [cf. Eq. (12)]. Of the elements, $^{58}$Ni has the highest critical angle, $0.12^0$ at 1 Å wavelength. Guide tubes with totally reflecting inner walls can be used to transport neutrons long distances, as seen in Fig. 22. If the guide tube is curved, it acts also as a low-pass energy filter. Focusing mirrors can be built, but these become very long, because the critical angle is smaller and the beam size larger than in the case of SR. Another solution is the use of capillary waveguides, which can be bent for focusing. These are manufactured of glass, and different focusing geometries are possible. For instance, an almost parallel beam is produced by a parabolic bundle of capillaries, or if this is used in the reverse direction, a wide beam from the end of a large-area waveguide can be focused to a small spot. It must be remembered, however, that the phase space density of neutrons cannot be increased, i.e., there is a trade-off between beam dimension and divergence.

Multilayers were already discussed in Section VII.B. They can be used as broad-band monochromators and focusing elements. The so-called "supermirror" is a multilayer structure where the layer thickness changes, superimposing the reflectivity curves over a range that extends from the regime of totally reflecting mirror to the limit determined by the smallest periodicity. In this range the reflectivity may exceed 90%, and the critical angle is three or four times larger than that of $^{58}$Ni.

Single crystals are used as monochromators in the same way as with X-rays. However, the narrow reflectivity curve of a perfect crystal is not matched with the low-brilliance neutron source, and perfect crystals are used only in some high-resolution applications, such as neutron interferometry or backscattering instruments. The widths of mosaic crystal reflectivity curves match the neutron beam divergences much better, but they increase the beam divergence and the width of the wavelength band. A better choice is a crystal where the reflecting Bragg planes are highly parallel, but the lattice spacing varies. Such crystals are called gradient crystals, and they are usually Si–Ge crystals where the lattice spacing increases with increasing Ge content. Another possibility is to introduce a thermal gradient in the direction of the scattering vector. The difference between a perfect, mosaic, and gradient crystal is shown in Fig. 24 as a reciprocal lattice diagram. Monochromators are also built from stacks of thin crystal

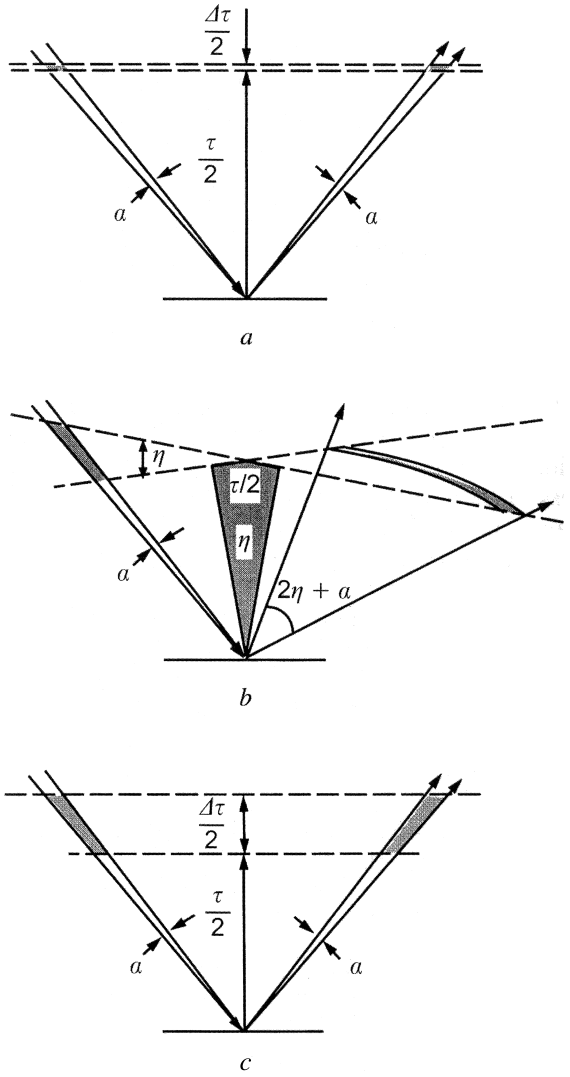



**FIGURE 24** Reciprocal space representation of different neutron monochromators. The divergence of the incident beam is $\alpha$, and $\Delta\tau$ the lattice constant width (relative energy band-pass). In the case of a perfect crystal the (small) reciprocal space element (shaded area) is retained upon reflection, in the case of a mosaic crystal with mosaicity $\eta$ the element it spread over an angle $2\eta + \alpha$, while in the gradient crystal the (large) element is retained, because the lattice spacing changes, but there is no mosaicity. [*International Tables for Crystallography*, (1999). Vol. C, reproduced with permission.]

wafers with slightly changing orientations. A controlled degree of mosaicity is introduced in one plane without deteriorating resolution in the perpendicular plane. Bent perfect crystals can be used in the same way as with X-rays and SR, but without the limitations arising from photoelectric absorption. Large gains of intensity and versatile focusing conditions are achieved in monochromator and analyzer constructions.

Neutron polarizers are important optical components, and many different polarizers exist. The polarizing efficiency is defined as

$$P = (N_+ - N_-)/(N_+ + N_-) \tag{48}$$

where $N_+$ and $N_-$ are numbers of neutrons with spin parallel (+) or antiparallel (−) to the guide field in the outgoing beam. A single crystal of a ferromagnetic compound can be used as a polarizer-monochromator. Atomic moments are saturated along magnetic field **B**, which is perpendicular to the scattering vector. The cross section of a Bragg reflection is

$$(d\sigma/d\Omega) \propto F_N^2 + 2F_N F_M(\mathbf{P}\cdot\boldsymbol{\mu}) + F_M^2 \tag{49}$$

where $F_N$ and $F_M$ are the nuclear and magnetic structure factors, respectively. **P** gives the polarization of the incident beam with respect to **B**, and $\boldsymbol{\mu}$ is the unit vector in the direction of atomic magnetic moments. For the neutrons with + spins, $\mathbf{P}\cdot\boldsymbol{\mu} = 1$, whereas for the neutrons with − spins, $\mathbf{P}\cdot\boldsymbol{\mu} = -1$. The polarizing efficiency is

$$P = \pm 2F_N F_M/\left(F_N^2 + F_M^2\right) \tag{50}$$

which can be either positive or negative and has the maximum value of unity for $|F_N| = |F_M|$. Usually ferromagnetic alloys are needed for a good structure factor matching and other properties required for polarizers. Other constructions include polarizing mirrors, because the refractive index of ferromagnetic material depends on magnetization, so neutrons with + spins have a critical angle different from that for neutrons with − spin.

There are also polarizing filters, and filters based on Zeeman splitting of neutron energies, which is large enough to be separated by a reflection from perfect Si crystal.

The finite speed of the neutron (a few km/s) is utilized in beam choppers and velocity selectors used in TOF measurements. The Doppler effect can be used for small variations of energy in backscattering instruments by moving the crystal along the scattering vector. The Doppler shift is also important in the operation of the rotating crystal TOF spectrometer, where the beam is either focused or defocused in time, depending on the sense of rotation.

A review of neutron sources and optics is found in Vol. C of *International Tables for Crystallography* (Chapter 4.4.) (1999).

## X. NEUTRON DIFFRACTION, METHODS, AND APPLICATIONS

The discussion in the preceding section indicated that the diffraction methods used in neutron diffraction are similar to those used in X-ray and SR diffraction, or the methods are analogous, such as the energy-dispersive

X-ray diffraction and the time-of-flight method in neutron diffraction. Therefore, the similarities of methods and applications are not discussed in the following, but emphasis is given to differences and special features of neutron diffraction.

### A. Neutron Crystallography

Properties of neutrons and X-rays for diffraction studies have been compared by Bacon (1975). Certain things in that comparison have changed with the use of SR, such as magnetic scattering of X-rays and free choice of wavelength in a wide range, but there are some basic differences, which make neutrons and X-rays complementary probes. A large fraction of atoms in biological structures is hydrogen, and the hydrogen contribution to the electron density is difficult to distinguish in X-ray diffraction, even with the best methods using SR. On the other hand, the neutron scattering amplitude of hydrogen is comparable with those of other nuclei, and the positions and thermal motion of hydrogen atoms in the unit cell can be determined with almost the same accuracy as the positions of heavier atoms. The scattering amplitude of normal hydrogen $^1$H is negative because of the large spin incoherence, whereas the scattering amplitude of deuterium $^2$H (or D) is positive, and incoherent scattering is small. The incoherent background is large when the hydrogen content of the sample is high, and particularly in powder diffraction it is advisable to use deuterated compounds, if possible. In general, comparison of diffraction data of normal and deuterated samples highlights the positions of the hydrogen atoms.

The structure of water and ice is a network of hydrogen bonds between the atoms, and depending on temperature and pressure, many different phases are formed. The phase diagram in Fig. 25 shows the richness of structures that arise from changes in the bonding network. This network collapses at sufficiently high pressures, and for instance, the common crystalline hexagonal ice $I_h$ is compressed into a high-density amorphous state (HDA), which transforms upon heating into low-density amorphous phase (LDA), and finally into the cubic phase $I_c$ . When water is confined to microporous solids, the boundary conditions have strong influence on the structure. The liquid phase prevails tens of degrees below $0^0$C, and the structure is the same as in the solid LDA phase (*ILL Annual Report* 1999). Structures and driving mechanisms of transitions in water and ice are of fundamental importance, and neutron diffraction has provided most of the experimental evidence.

Another field where the nonsystematic variation of the neutron scattering amplitude is exploited is studies of alloys. In many cases the alloys contain atoms that have

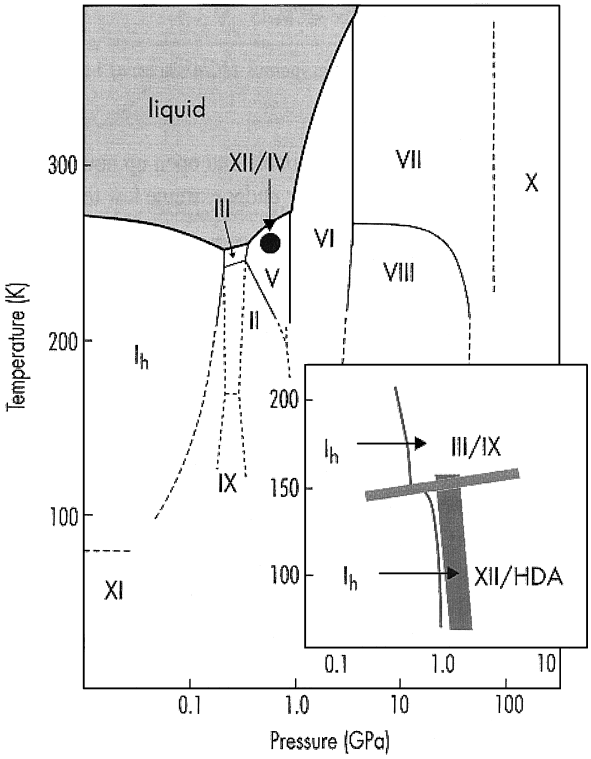



**FIGURE 25** Phase diagram of water. A new phase, ice XII, has been found at two isolated places of the pressure/temperature diagram, at (0.55 GPa, 260 K) and at (>1 Gpa, <150 K). In the latter domain, formation of ice XII competes with formation of the high-density amorphous ice (HDA). [*ILL Annual Report* 1999; reproduced with permission.]

small differences in atomic numbers. Sufficient contrast may be obtained when anomalous dispersion of X-rays is used, but this is not feasible in the case of light elements. For instance, the coherent scattering cross section for Mg is more than twice as large as that for Al. Isotopes of the same elements may have very different scattering amplitudes, as already seen in the case of hydrogen; this fact can be used to tailor the scattering amplitude, if isotopes can be enriched.

The amplitude of nuclear scattering is independent of the scattering vector, and the intensity falls off only because of the thermal motion of the atoms. The effect on the intensity is given by the Debye–Waller factor,

$$\mathrm{DW} = \exp[-2B(\sin\theta/\lambda)^2], \qquad (51)$$

where the thermal motion parameter is related to the mean-square displacement of the reflecting Bragg plane in the direction of the scattering vector by $B = 8\pi^2 \langle u^2 \rangle$. In the case of X-rays, the high-order reflections are suppressed because the atomic scattering amplitudes decrease with increasing scattering vector, and separation of the effects of thermal motion becomes difficult. Combination of

X-ray and neutron diffraction data in the so-called XN studies has been important in accurate analysis of charge density distributions of crystals.

## B. Powder Diffraction and Rietveld Method

Refinement of the crystal structure by the Rietveld method was already mentioned in connection of powder diffraction with SR. Actually, the method was developed for analysis of neutron diffraction data. Within the kinematical approximation of diffraction, the integrated intensities of powder reflections can be calculated from the positions and thermal motion amplitudes of the atoms. This distribution of $\delta$-functions in the scattering angle scale is convoluted by the so-called profile function. The model of the structure is refined by least-squares minimization of the residual,

$$M = \sum w_i [y_i(\text{obs}) - y_i(\text{calc})]^2 \qquad (52)$$

Here $y_i$ (obs) is the intensity measured at a point $i$ in the diffraction pattern, $w_i$ is its weight, and $y_i$ (calc) is the calculated intensity. The summation runs over all data points. The profile function is convolution of the instrument function by the function describing the diffraction domain in reciprocal space. Customarily, Gaussian shape of the profile function is assumed, although it is known that the tails of the reflections extend further out than those of the Gaussian. Other shapes that are used to describe the profile function include Lorentzian, Voigtian (convolution of Gaussian and Lorentzian), or Pearson VII functions. The main problem in Rietveld analysis is modeling the background, which is not constant or even smoothly varying because of disorder and thermal diffuse scattering. However, it is possible to reduce the parameters of the model to those describing the structure, thermal motion, grain size, strain, and texture of the powder sample, when the instrument function is calculated from the known diffraction geometry, and a proper model for background is included in the refinement (Honkimäki and Suortti, in Snyder *et al.*, 1999). An example of powder diffraction pattern and the fit of the model by the Rietveld method is shown in Fig. 26.

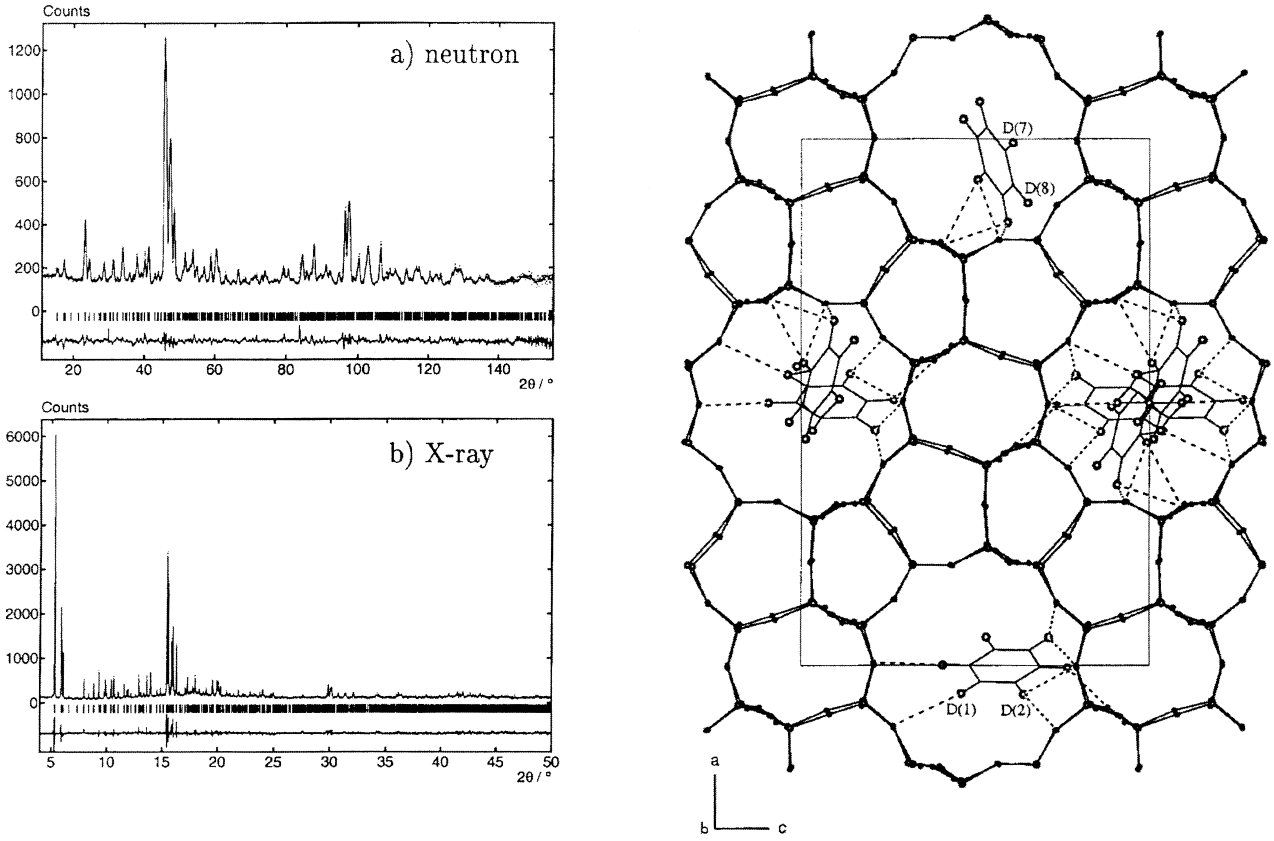



**FIGURE 26** Neutron and SR powder diffraction patterns of zeolite ZSM-5 with 3.4 benzene molecules absorbed per unit cell. The neutron wavelength is 3 Å, and the SR wavelength 1 Å, i.e., the Bragg angles in the neutron diffraction pattern are about 3 times those of the SR diffraction pattern. Note the superb resolution of the SR data, and the relatively high intensities of the neutron reflections at large scattering angles. The benzene positions in the channels of ZSM-5 are shown to right. [Courtesy of A. Fitch, ESRF.]

### C. Magnetic Structures

In spite of all the advances of magnetic SR diffraction, neutron diffraction remains the primary source of information on magnetization of solids. In the same way as in X-ray diffraction, there are two distinct areas of research: determination of magnetic structures, and studies of magnetization densities of known structures. Magnetic scattering arises from the interaction between the neutron magnetic moment and the magnetic moment of the atom. Magnetic scattering is superimposed on scattering of neutrons from the nuclei, which makes the situation much more complicated than that of X-ray scattering from electron charge distribution. The unpaired electrons of atoms give rise to a net magnetic moment. In the first transition series these are the $3d$ and $4s$ electrons, and in the rare-earths the $4f$ electrons. As to the interaction with the neighboring atoms, which results in magnetic ordering, these groups of atoms are very different.

In paramagnetic materials the directions of the atomic magnetic moments vary randomly, and the magnetic scattering is distributed in the background of the diffraction pattern. The scattering cross section is [generalized from Eq. (14)]

$$d\sigma_{\mathrm{pm}}/d\Omega = (2/3)J(J+1)(g\gamma/2)^2 r_{\mathrm{e}}^2 f^2 \qquad (53)$$

where $g$ is the Landé splitting factor, $J$ the total angular momentum, $\gamma$ the neutron magnetic moment ($-1.91$ nuclear magnetons), $r_{\mathrm{e}}$ the classical electron radius, and $f$ the magnetic scattering factor. The magnetic scattering factor is defined analogously to the X-ray atomic scattering factor,

$$f(\mathbf{K}) = \int \mathbf{M}(\mathbf{r}) \exp(i\mathbf{K}\cdot\mathbf{r})\, d^3\mathbf{r} \qquad (54)$$

where $\mathbf{M}(\mathbf{r})$ is the normalized magnetization density. If other contributions to the background (incoherent nuclear scattering, multiple scattering, and TDS) can be subtracted, the magnetic scattering factor can be determined.

In a ferromagnet all the moments in a single domain are aligned parallel. A simple antiferromagnetic structure is shown in Fig. 27. In an antiferromagnet the atoms may be considered to lie on two sublattices whose spins are oppositely aligned. In the picture of local magnetic moments, exchange interaction couples the spins of neighboring atoms. The potential energy is $V_{ij} = -2J_{ij}\, S_i\, S_j$, where $S_i$ and $S_j$ are the spin quantum numbers of the two atoms, and $J_{ij}$ is the exchange integral. For a ferromagnet this integral is positive, leading to parallel orientation of spins, while for an antiferromagnet $J_{ij}$ is negative. The exchange integral is due to overlap of atomic orbitals, and the value and sign of the integral depend on the distance between the atoms. The magnetic unit cell of a ferromagnet is the same as the chemical unit cell, so that magnetic reflections coincide with the nuclear reflections. In an antiferromagnet the magnetic unit cell is usually larger than the chemical unit cell, so that the magnetic structure gives rise to superlattice reflections between the nuclear reflections. The dependence of the integrated intensity on the structure factors is given by Eq. (49). It is seen that when the neutron beam is unpolarized the middle term averages to zero. This simple argument applies only to ferromagnets and antiferromagnets, where all the magnetic moments are either parallel or antiparallel to a single direction. The detailed treatment of polarization is much more complicated, particularly for noncollinear arrangements of spins such as are found in helimagnetic compounds. In general, the neutron scattering cross section consists of five separate terms: coherent nuclear scattering, incoherent isotope and nuclear spin scattering, nuclear–magnetic interference term, purely magnetic scattering, and polarization dependent term in complicated structures (Bacon, 1975).

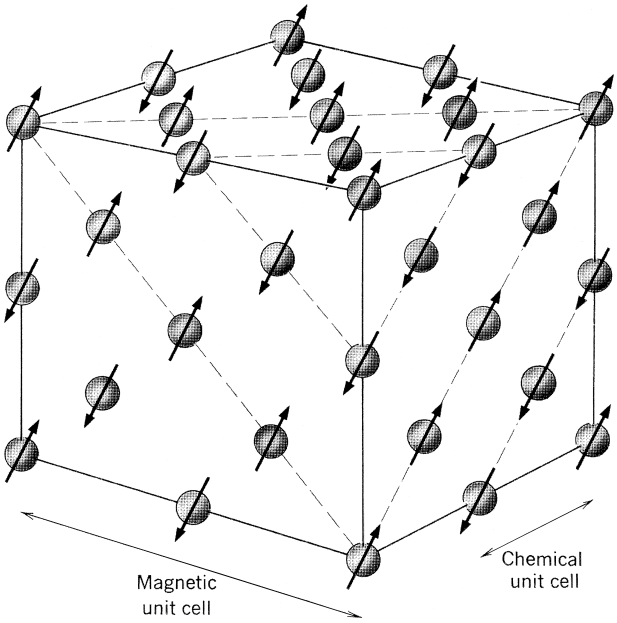



**FIGURE 27** Magnetic structure of antiferromagnetic MnO. Magnetic moments are parallel to each other in an (*hhh*) plane of Mn atoms, and the direction of the moments is reversed in the next plane. The magnetic unit cell is twice the chemical unit cell in all three directions.

When the purely magnetic scattering contribution to the total diffraction pattern can be separated and the structure is known, the magnetic form factors of atoms can be calculated. This is important for two reasons: The form factors are needed for solving magnetic structures from intensity measurements, and secondly, electron distributions can be calculated by Fourier transform. The uncompensated spin and orbital magnetic moments are due to the

outer orbitals of the atoms, so that they are very sensitive to the neighboring atoms. Even early measurements on Fe demonstrated strong directional differences in the magnetic form factor. This was seen directly in reflections that have the same absolute value of **K**, but correspond to different directions in the crystal. The unpaired electron density can be calculated from Eq. (27) using the experimental magnetic structure factors. The results were interpreted in terms of the relative occupancy of $e_g$ and $t_{2g}$ orbitals, which have different directional dependencies. In recent years, the methods for calculation of electron wave functions in solids have improved essentially, so that electron densities of magnetic subbands can be calculated separately. Accurate measurements of magnetic structure factors are stringent tests for the accuracy of these calculations. The anisotropy of the electron distribution is particularly important in the transition element metals and their compounds, because the 3d electrons are exposed to the crystal field. This leads to quenching of the orbital magnetization.

The determination of form factors of magnetic ions in ferromagnetic and paramagnetic materials is based on the measurement of the flipping ratio in polarized-neutron diffraction experiments. Production of polarized neutron beams was described in Section IX.B. When the polarization vector of the incident neutron beam is parallel to the field direction at the sample, the intensity of a reflection is proportional to $(F_N + F_M)^2$, whereas when the polarization vector and the field are antiparallel, the intensity is proportional to $(F_N - F_M)^2$. The direction of neutron polarization is reversed by a radio-frequency coil, which is placed between the monochromator-polarizer and the sample, i.e. the neutron spin is "flipped". When the polarization efficiency is known [cf. Eq. (48)] the nuclear and magnetic structure factors can be separated. For most antiferromagnetic structures the classical flipping-ratio measurement is not applicable, because the neutron scattering cross-section is not polarization dependent in a simple way. However, it is possible to extract the form factor of a magnetic ion in an antiferromagnetic compound by making polarization analysis of the beam scattered by the sample. The result for a particularly difficult case of the $Cr^{3+}$ ion in $Cr_2O_3$ is shown in Fig. 28. The periodicity of the magnetic and nuclear structures is the same, so that the magnetic and nuclear scattering are superimposed. The low value of the chromium moment suggests that there is significant covalent transfer to the oxygen ligands, and the directional differences are evident in the high-angle reflections.

The magnetization densities in rare-earth compounds are often quite complex. The 4f electrons responsible for the net magnetic moment are rather deep in the atom, so that the orbital moment of a rare-earth ion is not quenched

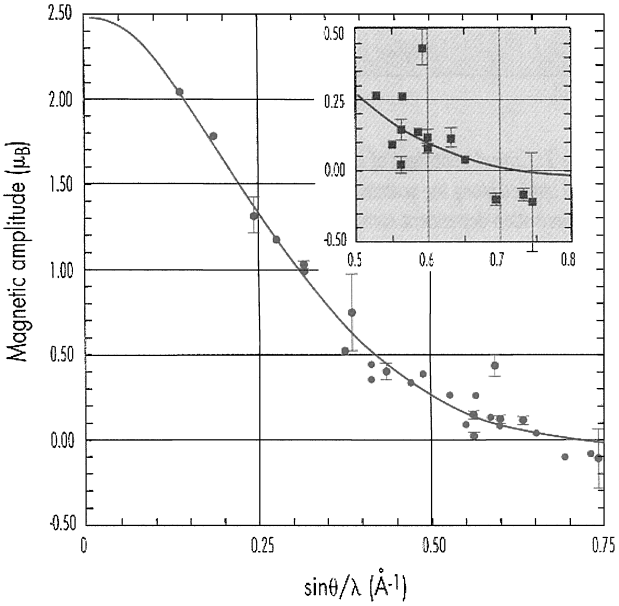



**FIGURE 28** Experimental magnetic form factors derived from the *hkl* Bragg reflections of $Cr_2O_3$. The smooth curve is the spin-only free ion form factor for $Cr^{3+}$ normalized to 2.5 Bohr magnetons [*ILL Annual Report* 1999; reproduced with permission.]

by the crystal field, and the Hund rules apply. The magnetic structure of the rare-earth metals is helical in many cases. Magnetic ordering gives rise to weak satellite reflections, which are studied as function of temperature. Because of resonant enhancement, these satellites can be measured also by magnetic scattering of SR, which provides complementary information. In the case of holmium, for instance, magnetic moments order helically below the Néel transition temperature of 132 K (see Fig. 29). Neutron diffraction measurements yield the total magnetic moment, but the orbital ($L$) and spin ($S$) contributions could be

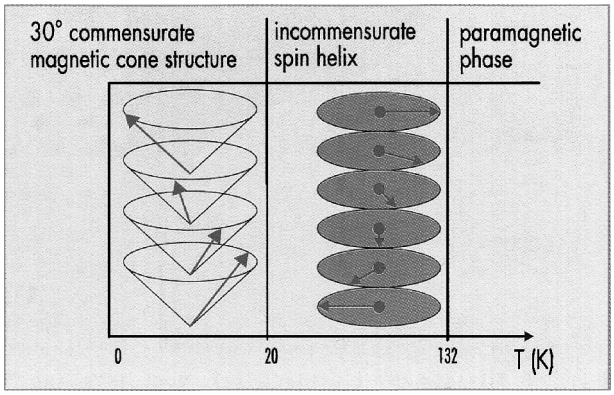



**FIGURE 29** Magnetic ordering of Ho in the direction of the *c*-axis as a function of the temperature. [*ILL Annual Report* 1999; reproduced with permission.]

separated by polarization analysis of the diffracted SR beam. The order is ferromagnetic in the basal plane, i.e., in a monolayer of atoms. Then from one layer to the next, the orientation turns a certain angle, setting up a magnetic spiral perpendicular to the planes. However, the periodicity of the helix does not coincide with the crystal planes, i.e., the spin helix is incommensurate. At the transition temperature the turn angle is 50° decreasing continuously to 30° at 20 K. At this temperature the magnetic moments tilt upward out of the basal plane, locking into a commensurate helical magnetic-cone order. Recently magnetic ordering in a thin epitaxial Ho(00.1) film could be studied by observing the $(00.0 + \tau^*)$ magnetic satellite reflection. It turns out that the Néel temperature is lowered to 105 K and there is no lock-in transition to commensurate magnetic structure. These results are technologically important, as magnetic nanostructures are integrated in new electronic devices.

## XI. FUTURE DEVELOPMENTS

Some 20 years ago, X-ray and neutron diffraction could be considered mature fields of research, where well-established methods were used for structural studies in physics, chemistry, and biology. However, the increasing use of synchrotron radiation was changing this picture, and in the past decades the synchrotron laboratories have been motors of new developments. Research in many traditional fields has been revolutionized by the new tools, and totally new areas of research have been introduced. The examples given above emphasized this aspect, and it may be useful to reiterate the most important advances:

- Microdiffraction: Samples of micron size can be studied, even crystals that were considered as small powder grains; diffraction under very high pressure is possible
- Anomalous scattering: Experiments can be performed with wavelengths close to any absorption edge, and this links diffraction to spectroscopic methods
- Time-resolved diffraction: A Laue pattern can be recorded with radiation from a single bunch of electrons, and evolution of structures under external stimuli can be followed by pump-and-probe experiments
- Surface structures: The high intensity and collimation of SR allows probing the structures of atomic layers, interfaces, and thin films
- Coherence: The SR beam from an undulator is partially coherent, which makes possible phase-contrast imaging and studies of spatial and temporal correlations
- Magnetic diffraction: The weak interaction of the electromagnetic field of the photons with the magnetization of solids is enhanced near the absorption edges, and this has made magnetic diffraction of SR complementary to neutron diffraction
- Diffraction at very high energies: Bulk properties of solids are probed in the same way as in neutron diffraction; extinction-free measurements of structure factors have become possible.

In many areas there has been important synergy between synchrotron radiation and neutron diffraction studies, and this has also stimulated neutron research, in areas such as structures of thin films and fibers, and magnetism-driven structural changes. It seems evident that there are many new, exciting developments over the horizon, and there the new sources of neutrons and synchrotron radiation are expected to play leading roles.

Safety concerns have made building and operation of nuclear research reactors more and restricted, and in recent years rather few reactors have been built while some of the existing ones have been closed down. This is a great concern to scientists, because neutrons are very versatile probes of condensed matter, and their complementarity with SR has opened up many new areas of research. There is a clear "shortage of neutrons," and to meet the growing demand new projects of spallation sources have been put forward. A project called European Spallation Source (ESS) has been initiated and, at the time of writing, is in study phase. The corresponding project in the United States is more advanced. The Spallation Neutron Source (SNS) will be located at Oak Ridge National Laboratory, Tennessee, and funding for the design and first construction phases has been appropriated. The proton beam energy is 1.0 GeV, and there are $2 \times 10^{14}$ protons in each 0.7-$\mu$s long pulse. The repetition rate is 60 Hz, so that the proton beam power on the target is 2 MW. The target is liquid mercury, which is not damaged by radiation, and because of the high $Z$ the yield of spallation neutrons is high. The intensity of a neutron pulse is two orders of magnitude larger than the flux from steady-state reactors, and this high flux is efficiently exploited in TOF experiments. On the other hand, there are many experiments where the much higher average flux of reactors is needed.

The third generation SR sources, such as the ESRF (Grenoble, France), APS (Chicago), and SPring-8 (Hyogo, Japan), are approaching the ultimate performance of circular accelerators, but there is quest to ever higher brilliance and coherence of the source, i.e., to an X-ray laser. These fourth-generation sources are being studied in several places, and the most advanced projects are those at DESY (Hamburg, Germany) and at SLAC (Stanford, California). In both cases, linear electron accelerators are used. A test facility has been built at DESY, and the existing linear accelerator will be used at Stanford. It has been predicted and experimentally demonstrated that in

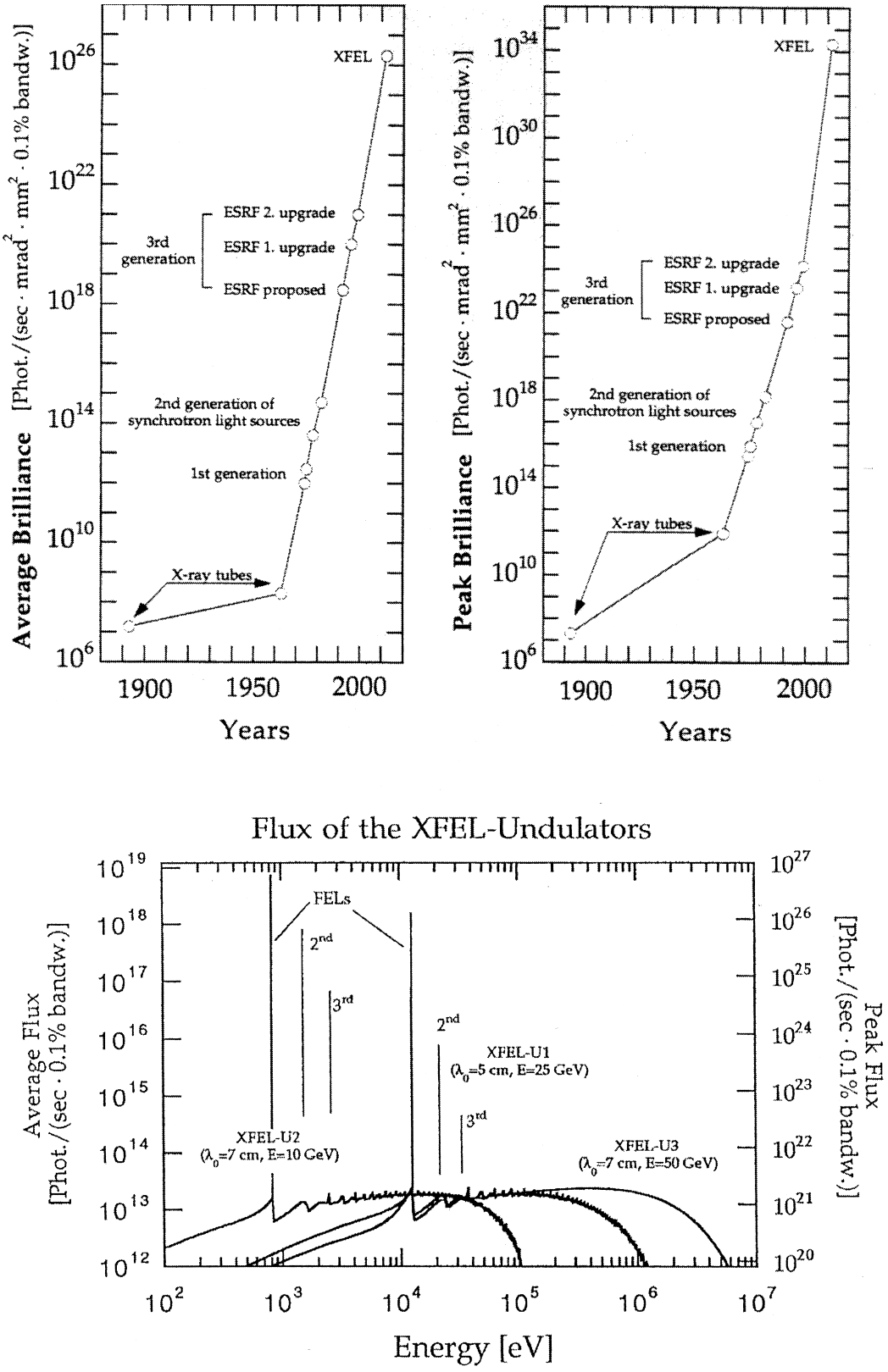



**FIGURE 30** Average and peak flux from undulators planned for the TESLA facility of DESY in Hamburg. The development of average and peak brilliance of different X-ray and SR sources since the discovery of X-rays in 1895.

a very long undulator the electron and photon beams, which travel together, exchange energy. The undulator wavelength is given by Eq. (42). The electric vector of the photon field is parallel to the transverse velocity of the electron, so that energy is transmitted from the radiation field to the electron or vice versa. Depending on their precise longitudinal position some electrons are accelerated while others are decelerated, leading to formation of equidistant microbunches. The length of the microbunch is smaller than the emitted wavelength, so that electrons in a microbunch emit coherently as a point charge. The radiated power is proportional to the square of the number of electrons in the microbunch. Furthermore, the interbunch distance is related to the wavelength leading to a coherent superposition of radiation from all microbunches. The process is called self-amplification of spontaneous emission (SASE). The necessary condition is a very short electron bunch, on the order of 10 $\mu$m, and this is achieved by special electron guns and pulse compression. Radiation from a free electron laser (FEL) is coherent, and the peak brilliance is many orders of magnitude higher than that of the third-generation storage rings. On the other hand, the duty cycle of a FEL is low, but still very large gains of the average brilliance are expected, as seen in Fig. 30.

The high brilliance, time structure, and coherence of the FEL radiation gives prospects for many new types of studies, as discussed by Doniach (1996). One interesting possibility is determination of the structure of individual macromolecules. Simulations indicate that a diffraction pattern from a single molecule may be recorded using intense ($10^6$ X-ray photons/Å$^2$) femtosecond pulses. The structure of the molecule can be reconstructed from a large number of patterns from randomly oriented molecules using Fourier transforms. Another possibility is Fourier holography, where the reference beam is obtained by diffraction from a thin film of crystallites. It is evident that the use of femtosecond pulses improves the resolution of time-resolved protein crystallography by several orders of magnitude.

It may be conjectured that most of the applications of FEL radiation are still over the horizon. The synchrotron radiation diffraction methods and their fields of application may be revolutionized in the same way as the traditional X-ray diffraction was by the use of SR. It is also likely that new links to neutron scattering develop, leading to increasing synergy of the two methods. The first FEL sources in the X-ray regime should become operative in the first decade of this millenium.

## SEE ALSO THE FOLLOWING ARTICLES

X-RAY PHOTOELECTRON SPECTROSCOPY • X-RAY SMALL ANGLE SCATTERING

## BIBLIOGRAPHY


Als-Nielsen, J., and Möhwald, H. (1999). *In* "Handbook on Synchrotron Radiation," Vol. 4 (Ebashi, S., Koch, M., and Rubenstein, E., eds.), Elsevier, Amsterdam.

Bacon, G. E. (1975). "Neutron Diffraction," Clarendon Press, Oxford.

Bartunik, H. D. (1999). *In* "Handbook on Synchrotron Radiation," Elsevier, Amsterdam.

Becker, P. J., and Coppens, P. (1974). *Acta Crystallogr.* **A30,** 129–147.

Brefeld, W., and Gürtler, P. (1999). *In* "Handbook on Synchrotron Radiation," Vol. 4 (Ebashi, S., Koch, M., and Rubenstein, E., eds.), Elsevier, Amsterdam.

Brunel, M., and de Bergevin, F. (1999). *In* "Handbook on Synchrotron Radiation," Vol. 3 (Brown, G., and Moncton, D., eds.), Elsevier, Amsterdam.

Doniach, S. (1996). *J. Synchrotron Rad.* **3,** 260–267.

European Synchrotron Radiation Facility (1994–1999). "Highlights."

Ewald, P. P., ed. (1962). "Fifty Years of X-Ray Diffraction," Oosthoek, Utrecht.

Guinier, A. (1963). "X-Ray Diffraction," Freeman, San Francisco.

Helliwell, J. R. (1992). "Macromolecular Crystallography with Synchrotron Radiation," Cambridge University Press, Cambridge, U.K.

Institute Laue-Langevin (1998, 1999). "Annual Reports."

James, R. W. (1962). "The Optical Principles of the Diffraction of X-Rays," G. Bell and Sons, London.

Jarlskog, G., Mjörnmark, U., and Sjöstrand, T., eds. (1999). "Workshop on Future Linear Electron–Positron Colliders," Lund.

Kerr, K. Ann (1987). "Crystallography." *In* "Encyclopedia of Physical Science and Technology," Vol. 4, pp. 45–80, Academic Press, San Diego.

Kratky, O., and Laggner, P. (1987). "X-ray small-angle scattering." *In* "Encyclopedia of Physical Science and Technology," Vol. 14, pp. 693–742, Academic Press, San Digeo.

Lovesey, S. W., and Collins, S. P. (1996), "X-Ray Scattering and Absorption by Magnetic Crystals," Oxford Science Publ., Oxford.

Robinson, I. K., and Tweet, D. J. (1992). *Rep. Prog. Phys.* **55,** 599–651.

Snyder, R. L., Fiala, J., and Bunge, H. J., eds. (1999). "Defect and Microstructure Analysis by Diffraction," Oxford University Press, Oxford.

Suortti, P. (1982). *Acta Crystallogr.* **A38,** 642–647.

Warren, B. E. (1969). "X-Ray Diffraction," Addison-Wesley, Reading, MA.

Wilson, A. J. C., and Prince, E., eds. (1999). "International Tables for Crystallography," Vol. C, Kluwer, Dordrecht.

Young, R. A., ed. (1993). "The Rietveld Method," Oxford University Press, Oxford.

Zachariasen, W. H. (1945). "Theory of X-Ray Diffraction in Crystals," John Wiley & Sons, New York.