

Accelerator Physics and Engineering

Frank T. Cole, deceased
Maury Tigner
Cornell University

Alexander W. Chao
Stanford Linear Accelerator Center

I. Introduction
II. History
III. Applications of Accelerators
IV. Types of Accelerators
V. Physical Principles of Beam Motion
VI. Accelerators of the Future

GLOSSARY

Betatron Circular induction accelerator for electrons. The magnetic guide field rises during acceleration to keep particles on a circle of constant radius.

Circular accelerator Cyclic accelerator in which particles are bent by magnetic fields around closed paths, passing many times through the same accelerating system.

Colliding beams System in which the fixed target is replaced by a second beam of accelerated particles moving in the opposite direction. The collisions of moving particles produce very high-energy phenomena.

Cyclic accelerator Particle accelerator in which each particle passes many times through a small potential drop to be accelerated to high energy.

Cyclotron Circular accelerator in which protons or heavy ions spiral outward from the center as they are accelerated by a radio-frequency voltage.

Electron volt (eV) Unit used to describe the energy of particles in an accelerator.

Emittance Area in the transverse phase space (in units of millimeter–milliradians) occupied by the distribution of particles in a beam.

Fixed target Target fixed in location that the beam strikes after acceleration to produce physical changes of interest.

Focusing system System for confining divergent particles in a beam close to the ideal orbit during the course of acceleration and storage.

High-voltage accelerator Particle accelerator in which each particle passes once through a high potential. Examples are Cockcroft–Walton accelerators, Van de Graaff generators, and Marx generators.

Induction accelerator Accelerator in which particles are accelerated by electric fields that are generated from a changing magnetic field by Faraday's law of induction.

Linear accelerator Cyclic accelerator in which acceleration takes place along a straight line as particles pass sequentially through repeated accelerating units in synchronism with an electromagnetic wave.

Microtron Circular accelerator in which electrons move in circles that are all tangent at one point where a radio-frequency voltage accelerates them.

Radio-frequency system System where particles are accelerated in cyclic accelerators. Accelerating field is provided by an electromagnetic wave at a microwave frequency. The oscillating field must synchronize with the time of arrival of the particles to be accelerated.

Storage ring Circular accelerator with magnetic field fixed in time so that one or more beams of particles can circulate continuously for a long period of time.

Strong-focusing system System of alternating focusing and defocusing lenses which produce a strong net focusing effect.

Synchrotron Circular accelerator in which particles are kept in a circle of constant radius by a magnetic guide field that rises in time as they are accelerated by a radio-frequency system.

Wake field Electromagnetic microwave fields excited in the vacuum chamber by a passing beam.

A PARTICLE BEAM is an ensemble of particles that move together in close proximity. A beam is characterized by a few basic parameters such as the particle species and the average energy of the particles. The beam of particles must be well collimated so that all particles stay in close proximity throughout its motion. Deviation of the motion of individual particles from the average motion of the beam must be kept sufficiently small.

The species of particles in the beam is most commonly electrically charged. Most common examples of particle species are electrons and protons. Less common particles include charged particles such as muons, pions, or neutral particles such as some specific atoms or molecules. The physics of accelerators also significantly overlaps with the physics of light optics and lasers, as light can be treated as a beam of photon particles.

A *particle accelerator* is a device that manipulates the motion of charged particle beams. The most common manipulation is to increase the energy of the particles, thus the term "accelerators." In modern times, however, other variations which do not accelerate or for which acceleration plays only a minor role have been introduced. All these devices are also customarily considered as particle accelerators.

There are also devices that manipulate neutral particles. The physics of these devices sometimes is considered as part of accelerator physics because they can be described by very similar physical principles. Atomic beam devices and particle traps are examples of this category.

I. INTRODUCTION

A. Parameters Characterizing the Beam

Other than the particle *species*, the final *energy* of the accelerated particles is the most important parameter of a particle accelerator. The particles have electric charge equal to the electron charge e or a multiple of it, and they are accelerated by potentials measured in volts (V). Therefore, a natural unit of energy is the electron volt (eV), the energy acquired by one electron charge in passing through a potential difference of 1 V.

The electron volt is a very small unit of energy (1 eV $= 1.6 \times 10^{-19}$ Joule), more directly applicable to energy levels in atoms than to accelerators. There are therefore multiples of the electron volt that are used to describe accelerators.

$$1\ \text{keV} = 10^3\ \text{eV}$$

$$1\ \text{MeV} = 10^6\ \text{eV}$$

$$1\ \text{GeV} = 10^9\ \text{eV}$$

$$1\ \text{TeV} = 10^{12}\ \text{eV}$$

The energies of particle accelerators now being operated range from a few hundred keV to 1 TeV. The sizes of particle accelerators range from table-top devices to devices stretching over several miles. One of the largest, the Fermi National Accelerator Laboratory near Chicago, is shown in an aerial view in Fig. 1.

A second important parameter used to characterize an accelerator is its *intensity*, usually the number of particles moving together in a bunch, or, in the case of a continuous flow of particles, the number of particles accelerated per second.

Other parameters characterizing a particle beam include parameters that specify the degree of collimation of the beam. In particular, the spread of particle energies

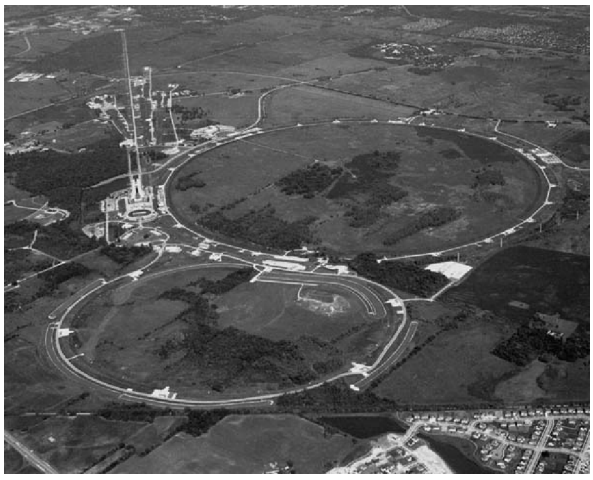


FIGURE 1 Aerial view of the world's highest energy accelerator, the Fermi National Accelerator Laboratory. The accelerator is in an underground tunnel. The accelerated beam is extracted for use in experiments in the areas stretching toward the top left of the photograph. [Courtesy of Fermilab National Accelerator Laboratory; Batavia, IL.]

around its average value is one such parameter. This *energy spread*, or momentum spread, which is related, is denoted by $\Delta E/E$, and typically ranges from 10^{-2} to 10^{-4}. Two more parameters, called *transverse emittances*, specify the degree the beam is bunched into a tight bundle throughout its motion. The emittances are denoted by ϵ_x and ϵ_y, and are in units of millimeter–milliradians. A tightly bundled beam will require small values of the emittances. These three parameters, $\Delta E/E$, ϵ_x, and ϵ_y, will have to meet the requirements of the accelerator application in hand.

Still other parameters may characterize special beam properties. One example is the beam *polarization*, which characterizes the degree of alignment of all the spins of the particles in the beam. A high degree of polarization is a very useful tool in analyzing some of the high-energy physics experiments.

B. Accelerator Physics Research

Accelerator physics is a branch of physics that studies the dynamics of the beams in accelerators. Sometimes it is also called *beam physics*, although, strictly speaking, accelerator physics studies only the part of beam physics encountered in accelerators, while beam physics may also contain the study of neutral particle devices, particle traps, and cosmic ray mechanisms.

Particle motion in an accelerator has a close analogy with light optics. One branch of accelerator physics, called *beam optics*, studies the motion of particles in the channel of accelerator elements. The accelerator elements are arranged in an optimal manner to guide and focus the beam along. For some high-performance accelerators, such as in an electron microscope or in a storage ring, it has been necessary to consider very clever element arrangements in such a way that nonlinear optical aberrations are compensated or minimized. In a storage ring, the beam is stored for typically much longer than 10^{10} revolutions, or many more times than the earth has circulated about the sun. Particle motion will have to be stable for this long a storage time. This branch of accelerator physics therefore requires a highly sophisticated knowledge of nonlinear dynamics and chaos physics.

As the beam intensity is increased, the self electromagnetic fields gets stronger. The beam interacts with its surroundings to create a perturbing *wake field*, which in turn may cause the beam to become unstable. A large number of various types of *collective beam instabilities* occur due to the high beam intensity. Understanding and analysis of this branch of beam physics is closely related to plasma physics.

Some high-performance accelerators require collimated beams with very small energy spreads and emittances. Still another branch of accelerator physics addresses this issue by innovations of several types of *beam cooling* techniques used to reduce the energy spread and emittances.

Accelerator physics is both a fundamental research and an applied research. The above-mentioned aspects are examples of fundamental research in accelerator physics. Applied research in accelerator physics concentrates on the development of accelerator technology, which constitutes a research area in its own right because of the complication and depth involved. As accelerator applications put forward increasingly demanding requirements on the beam, research in accelerator technology becomes increasingly specialized and sophisticated.

C. Accelerator Technology

Technology provides the means to manipulate the beams in accelerators. Accelerator research therefore also covers areas of the physics and engineering of accelerator technology. Notable examples include the technologies of high-power microwave devices, room-temperature iron magnets, superconducting magnets, large-scale ultra-high vacuum, intense particle sources, computer control and networking, fast electronics, high-power switches, and materials developments.

Microwave (also referred to as *radio-frequency wave*) technology is the main way to accelerate particles in accelerators. The microwave involved typically has a frequency ranging from a few hundred mega-Hertz (MHz) to a few tens of giga-Hertz (GHz), where Hertz is a frequency unit of one cycle per second. Developing high-power

microwave sources is one of the major technology research activities.

Superconducting magnets provide high magnetic fields for the purpose of guiding and focusing particle motion in accelerators. The higher the magnetic field can reach, the more compact the accelerator can be made. In addition, superconducting magnets also save operating power compared with room-temperature iron and copper counterparts. Developing high-field superconducting magnets is one important technology issue in accelerator physics.

Superconductivity also benefits radio-frequency devices. In particular, by replacing the room-temperature copper radio-frequency cavities by *superconducting cavities*, one can reach high accelerating fields and save operating power. The development of superconducting radio-frequency cavities has made substantial progress in the past two decades.

There are also accelerator applications using high-temperature superconductors. Research in this direction has also made progress in recent years.

II. HISTORY

During the nineteenth century, physicists experimented with *Crookes tubes*, evacuated glass systems containing internal electrodes. A current of electrons will flow when a voltage is applied between these electrodes. J. J. Thomson used a Crookes tube in his discovery of the electron in 1890. Röntgen discovered X-rays using a Crookes tube in 1896. The X-ray tube was later made into a practical device for use in medicine by Coolidge.

Rutherford spurred the development of particle accelerators for use in nuclear physics research in a famous lecture in 1920. He pointed out the need for higher energy particles for the further understanding of the atomic nucleus. During the next decade, both high-voltage and cyclic accelerators were invented. Many different methods of producing high voltage were demonstrated, but there was always great difficulty in avoiding sparks at high voltage. Cockcroft and Walton, in Rutherford's Cavendish Laboratory, developed a successful accelerating tube. They used an existing voltage-multiplying circuit and built a 300-keV proton accelerator, which they used in 1932 to do the first nuclear physics experiment with accelerated particles.

Ising was the first to conceive (in 1925) a cyclic accelerator, a *drift-tube linear accelerator*. Wideröe expanded the idea and built a working accelerator in 1928, and accelerated mercury ions.

Perhaps the most important consequence of Wideröe's work was to stimulate Lawrence to conceive the *cyclotron*. Lawrence and Livingston built the first operating cyclotron (at 1.2 MeV) in 1932. A succession of cyclotrons was built in Lawrence's laboratory through the 1930s and reached 5 to 10 MeV. Their original cyclotron is shown in Fig. 2.

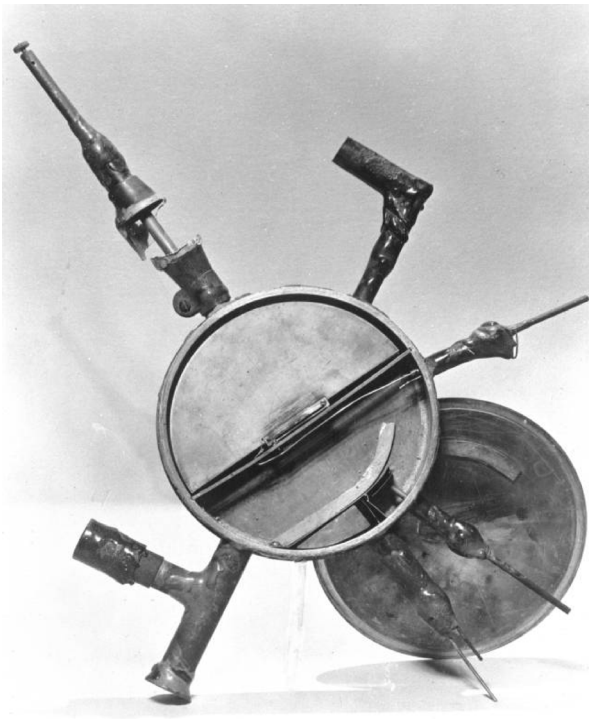


FIGURE 2 The vacuum chamber of the first Lawrence cyclotron. [Courtesy of LBNL.]

Lawrence's cyclotrons and *electrostatic generators*, which had been conceived and demonstrated by Van de Graaff in 1931, were used for nuclear physics research throughout the 1930s. Both were limited to energies of 15 MeV or less, and reaching energies beyond this limit was a major topic of research in the 1930s. Thomas proposed the *azimuthally varying-field (AVF) cyclotron* in 1938, but his work was not understood until much later. Kerst, with the help from an orbit dynamist, Serber, built the first successful *betatron* in 1941 and built a second 20-MeV machine before World War II intervened. A 100-MeV betatron was built in 1945, and a 300-MeV betatron a few years later. This model of betatron was built in large quantities for use in X-ray testing of large castings, particularly for the armor of military tanks.

The next great step in energy began in 1944 and 1945 when Veksler in the U.S.S.R. and McMillan in the U.S. independently conceived the *principle of phase stability*, permitting frequency modulation of the accelerating voltage in a cyclotron to overcome effects of relativity and making the *synchrotron* possible. In 1946, the first synchrocyclotron was operated and a number of 300-MeV electron synchrotrons came into operation in the next few

years. The research done with these accelerators was important in studying the properties of the pion.

World War II radar work had stimulated the development of radio-frequency power sources and these were used in linear accelerators. *Traveling-wave* electron linear accelerators, at frequencies of approximately 3 GHz, were extensively developed by Hansen, Ginzton, Panofsky, and their collaborators. This effort led over the years to the 50-GeV Stanford Linear Accelerator of today. Alvarez extended the concept of the drift-tube accelerator for heavier particles and built the first of many drift-tube accelerators used in physics and chemistry research and as injectors for synchrotrons.

Work also began in the late 1940s to build proton synchrotrons. The first proton synchrotron, the 3-GeV cosmotron at Brookhaven, NY, came into operation in 1952, and the 6-GeV Bevatron at Berkeley, CA, came into operation in 1954. An interesting precursor, a 1-GeV proton synchrotron, conceived in 1943 independently of the principle of phase stability, was built in Birmingham, U.K. However, the project was too ambitious for the resources available at the time, and the machine did not come into operation until July 1953. The proton synchrotrons were used for research in heavier mesons, the "strange" particles that had been observed in cosmic-ray experiments, and in antiprotons.

Speculative discussions aimed toward increasing the highest accelerator energy led to the conception of the *strong-focusing principle* by Courant, Livingston, and Snyder in 1952. It was later found that Christofilos had developed the principle independently in 1950. Strong or alternating-gradient focusing keeps the oscillations of particles about the ideal orbit small and makes possible compact and economical magnets.

The discovery of strong focusing led to an explosion of ideas. In 1953, Kitigaki and White independently conceived the *separated function* strong-focusing synchrotron, which made possible higher guide fields and more economical designs. In 1958, Collins conceived the *long straight section*, which made possible economical configurations with space for injection, acceleration, extraction, and detection equipment. As well as making it possible to go to a higher energy with proton synchrotrons, the strong-focusing principle gave impetus to new thinking in many other directions. Linear accelerators were greatly improved in performance by the addition of strong focusing along the orbit, first conceived by Blewett. The AVF principle was rediscovered by a number of people, among them Kolomensky, Ohkawa, Snyder, and Symon. It was extended to *spiral-sector focusing* by Kerst in 1954. Kerst then proposed that successively accelerated beams could be "stacked" in circulating orbits in a *fixed-field alternating gradient* (FFAG) accelerator, a variant of the AVF configuration. The intense stacked beam can then be used in colliding-beam experiments. The concept of *colliding beams* had been known for many years (it was patented in 1943 by Wideröe), but *beam stacking* is essential to achieve useful rates of collisions. Shortly afterwards, a number of people (Newton, Lichtenberg, Ross, and, independently, ONeill) proposed the concept of a *storage ring* separate from the accelerating device. The storage ring is a better colliding-beam system than the FFAG accelerator because it is less costly and because it provides more free space for detectors.

Experimental confirmation of these ideas did not lag far behind. The first strong-focusing electron synchrotron was a 1-GeV synchrotron operated by Wilson and his collaborators in 1955 at Cornell University, followed by several other electron synchrotrons. The FFAG principle and beam stacking were demonstrated by Kerst and his collaborators in the 1950s. Traveling-wave electron linear accelerators reached 1 GeV in this same era, and a series of important experiments on electron-proton scattering was done that elucidated the structure of the proton. The first electron storage rings were built and operated in the early 1960s at Stanford, CA; Frascati, Italy; Novosibirsk, then-U.S.S.R.; and at Cambridge, MA. Two large proton synchrotrons of 28- and 33-GeV energy were built by CERN, a new international laboratory in Europe, and by the Brookhaven Laboratory. These became the foundation of major advances in high-energy physics, with the discovery of many new particles and the beginning of a conceptual ordering among them and understanding of them. The electron synchrotrons and the Stanford Linear Accelerator, which reached 20 GeV in 1966, added to this understanding.

In the late 1960s, the first major proton storage ring, the ISR, was built at CERN. It stored and collided two proton beams of 28 GeV each. Colliding these two beams is equivalent to a fixed-target accelerator of over 1500 GeV energy.

The second generation of strong-focusing synchrotrons also began to be built in the late 1960s. These incorporated the more efficient separated-function magnet system and long straight sections. A proton synchrotron that reached 400 GeV was completed at the new Fermilab in Illinois in 1972. A similar synchrotron was later built at CERN. These workhorses incorporated new beam-sharing methods and each could provide beams simultaneously to several targets and a dozen major experiments. An important feature that has made this multiple use possible is the very high degree of precision in beam handling and manipulation. The data from these and ISR experiments led to the development of quantum chromodynamics and electroweak theory, large advances in our understanding of the basic building blocks of nature.

Important experimental evidence also came from the second generation of electron storage rings, now always with positrons as the second beam, the first of which reached 3 GeV in each beam at Stanford in 1972. Electron-positron storage rings have now reached 100 GeV per beam. As in the case of proton synchrotrons, electron–positron storage rings have been developed to a high art. Radio-frequency systems to replace the energy lost in synchrotron radiation are a major factor in the design and cost of these rings, but at the same time the synchrotron radiation also provides a beam cooling mechanism that makes the beam very small in size, making precise manipulations possible and increasing the colliding-beam interaction rate. In fact, synchrotron radiation has become a valuable experimental tool in its own right for use in atomic physics and materials science research, and a number of single-beam electron (or positron) storage rings have been built as dedicated synchrotron radiation facilities, providing in particular X-ray beams many orders of magnitude stronger than can be obtained from a laboratory X-ray tube.

The two large proton synchrotrons were developed further in quite different directions. At Fermilab, superconducting magnets underwent long, arduous development, and a superconducting magnet ring was built and installed in the tunnel of the 400-GeV accelerator. It reached 800 GeV in 1983. The CERN synchrotron was converted to a proton-antiproton storage ring by the addition of a small ring to accumulate antiprotons, making use of the new technique of *stochastic beam cooling* invented by van der Meer. Colliding-beam experiments have been carried out there, culminating in the discovery of the W and Z particles in 1983.

The spectacular successes of these accelerators and storage rings have led to a number of new initiatives. A large electron–positron (LEP) ring to initially reach energies of 50 GeV, and later upgraded to reach over 100 GeV, has been built at CERN. An electron–proton ring (HERA) has been built in Germany. A single-pass, colliding-beam system SLC has been built at Stanford. Here, the two 50-GeV beams collide only once in a linear system, not a circulating configuration. A useable event rate is achieved by very small beam sizes (thus increasing the density). Construction is in progress at CERN on a proton–proton collider, a large hadron collider, with 7 TeV in each beam.

In parallel with the above efforts based on more traditional approaches, research work has been carried out for many years on new methods of particle acceleration, making use of plasmas and lasers, with the goal of achieving substantially higher accelerating fields.

The historical development of the energy of particle accelerators is plotted in Fig. 3, the famous *Livingston chart*. One can see from the chart that the development

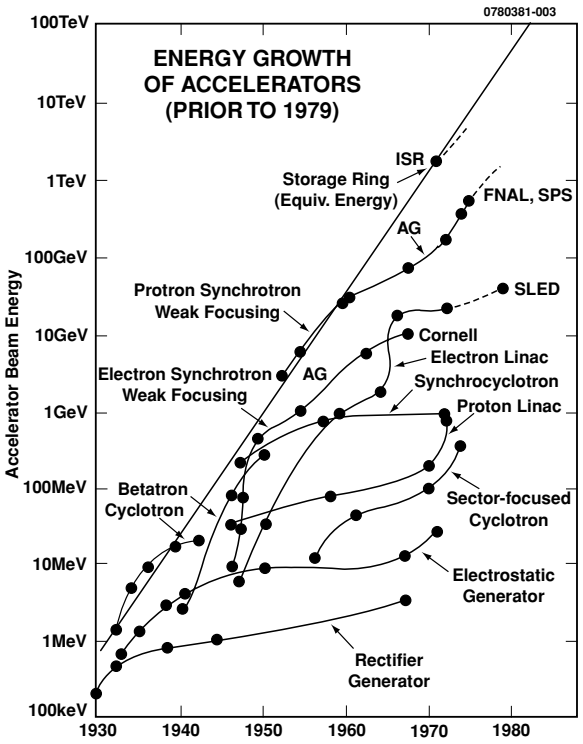


FIGURE 3 Livingston chart.

of each new type of accelerator gave an energy increase. Each accelerator type is eventually replaced by another as it reaches its limit for producing higher energy beams. It is evident from the chart that the field of accelerator research has been active and productive over the last several decades.

III. APPLICATIONS OF ACCELERATORS

Many types of accelerators are being used today. Examples range from daily appliances such as television sets and microwave ovens to medium-sized accelerators for medical and industrial uses to gigantic devices such as those used in high-energy physics and nuclear physics research.

A. Household Appliances

Some household appliances are miniature accelerators. Most notable examples are vacuum tubes, television sets, and microwave ovens.

The recent electronic revolution has been based on semiconductors. However, the first electronic revolution was made possible by the invention of the *vacuum tubes*. A vacuum tube is a miniature accelerator consisting basically of a electric heater, an electron-emitter called the *cathode*,

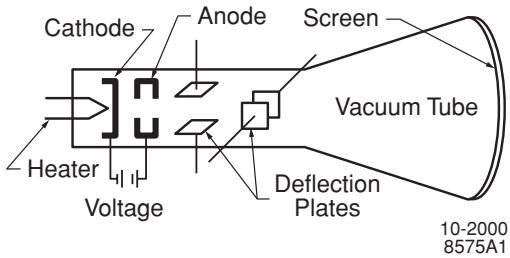


FIGURE 4 Schematic of a cathode ray tube. Actual devices, especially at the anode, are typically more complex with additional consideration of collimation and providing accurate focusing of the beam image on the screen.

and an electron-collector called the *anode* which is maintained at a positive potential relative to the cathode. Such a vacuum tube can be useful in rectifying oscillating currents into a dc current, but by adding an additional control grid between the cathode and the anode, the vacuum tube becomes an amplifier, a wide application of which triggered the first electronic revolution. Most functions of the vacuum tubes have been replaced by semiconductor devices. However, important applications remain in areas where high beam power is involved.

A *cathode ray tube* consists of a heater, a cathode, and an anode, just like a vacuum tube, but the anode here serves only to provide the accelerating voltage for the electrons and not as the electron collector. Electrons are made to pass the anode through the passage hole, and strike a fluorecent screen downstream to produce an image on the screen. The direction of motion of the electrons is controlled by a set of deflecting plates. Cathode ray tubes are the basic device for scientific instruments such as the oscilloscope and the streak camera, but most commonly they are used in the television set. Figure 4 illustrates the schematic of a cathode ray tube.

The miniature accelerator used in a microwave oven is a *magnetron*, which consists of a cylindrical vacuum tube surrounded by a magnetic solenoid. A cylindrical vacuum tube has a cathode at the center of the cylinder and the anode at the outer cylindrical surface of the cylinder. By imposing on the tube with an oscillating solenoidal magnetic field, the cathode electrons will move in an oscillatory pattern, which in turn generates microwaves. High-power microwaves are used in radars, but more common household use is in the microwave oven. In a microwave oven, the microwave from the magnetron is directed into the oven from the anode.

B. High-Energy and Nuclear Physics

Particle accelerators are essential tools of high-energy, or elementary-particle, physics, as it is also called. In a high-energy physics experiment, particles with very high energies are directed to bombard a target. In the interaction between a projectile particle and the target particle it strikes, new kinds of particles can be produced that provide clues to the nature of matter. Systematic study of these interactions requires controlled, copious production using accelerators. A similar requirement holds for the research of nuclear physics. Compared with accelerators of the household, these accelerators for high-energy and nuclear physics research are much larger in size as well as in complexity.

When an accelerated particle strikes a stationary particle in a target in a *fixed target* experiment, a large fraction of the energy so laboriously put into the particle goes to move all the products forward in the direction of motion of the beam, because momentum is conserved in all collisions. Thus, a 100-GeV proton accelerator has available only approximately 27 GeV for making new particles. When the proton energy is increased to 1000 GeV, the available energy increases to only 43 GeV.

A way to make all the energy useful is to utilize a second accelerated beam as a target. If the two beams are moving in opposite directions, the total momentum of the system is zero and none of either beam's energy need be used in moving products downstream. This *colliding beams* method is a more economical method of creating new, higher energy interactions and new particles, as depicted schematically in Fig. 5.

The difficulty of a colliding beam configuration is that the collision rate in colliding beams is much lower than in a fixed target because the particle density in a beam is much

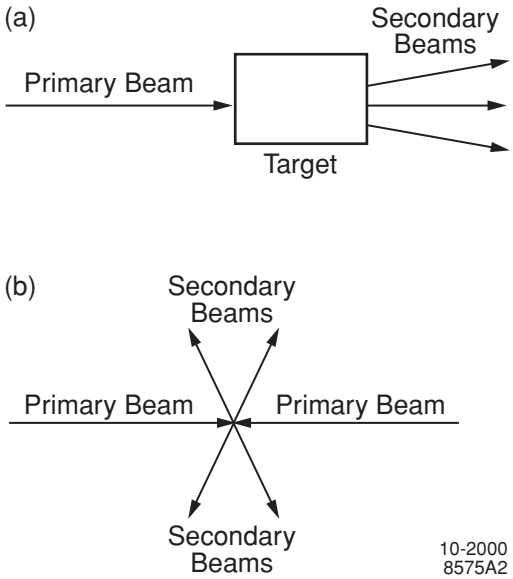


FIGURE 5 Schematic diagram of fixed-target and colliding-beam experimental methods.

less than in a solid target; therefore, there are many fewer particles to interact with in a beam. This difficulty can be partially overcome by storing two circulating beams of particles in a *storage ring*. There they pass many times through one another to increase the effective collision rate.

In all these experiments, higher energies are needed to probe to smaller distances within the atomic nucleus. As a result, there has been a constant drive toward higher energy.

The highest energies now used for physics experiments are approaching 1 TeV in fixed-target research and two beams each of 1-TeV energy in colliding-beam research. At this time, a collider, the large hadron collider (LHC), with beams up 7 TeV each is being constructed at CERN.

Reaching higher and higher beam energies is not the only frontier in high-energy physics accelerators. One modern class of storage-ring colliders aims for extremely high rate of events, even at moderate energies. These colliders, called *factories*, demand a deeper understanding of the accelerator physics involved. Examples include the ϕ-meson factory at Frascati, and the B-meson factories at Tsukuba, Japan, and Stanford, CA. Both B-factories started operation in 1999.

A large nuclear physics accelerator, the relativistic heavy ion collider (RHIC), is newly commissioned at BNL. By colliding two beams of heavy ions, such as gold ions, at 100 GeV/nucleon per beam, the RHIC is intended to create a new type of matter, the quark–gluon plasma, by the violent collision. Figure 6 shows a glimpse of the RHIC as seen in its tunnel.

The development of particle accelerators for high-energy and nuclear physics research is continuing actively. The above-mentioned LHC, the electron–positron linear collider research in several laboratories, and the exploratory design research of a muon collider are examples of current efforts being carried out around the world.



FIGURE 6 Inside the tunnel of the relativistic heavy ion collider (RHIC) during installation work. [Courtesy of BNL.)

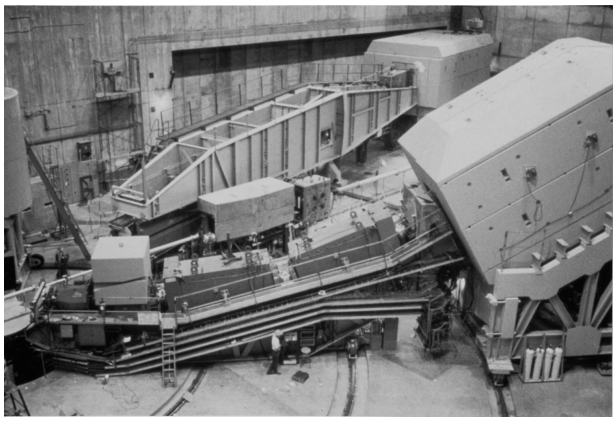


FIGURE 7 Spectrometer at End Station A of SLAC. [Courtesy of Stanford Linear Accelerator Center.]

C. Spectrometer

Spectrometers are devices for the precise measurement of energies or masses of the particles in a beam. These devices are used, for example, in the secondary beams to sort out the products of high-energy or nuclear reactions. Charged particles produced in these reactions are guided through a spectrometer to analyzing stations. Although no acceleration is performed on these particles, accelerator physics is needed to manipulate them. The working principle is similar to that of the prisms in light optics. In order to be able to detect rare events, of particular concern are issues of high resolution, large angular and energy acceptance, and being able to handle a wide variety of beams and targets. For short-lived products, it is also necessary to make the spectrometer path as short as possible. A spectrometer, used in the fixed-target experiment at the Stanford Linear Accelerator Center, is shown in Fig. 7.

D. Medical Accelerators

Accelerators have been extensively used in medicine. For example, a common X-ray machine is a particle accelerator (typically a 5- to 30-MeV electron linear accelerator). In it, electrons are accelerated and made to strike a heavy-metal target to produce X-rays.

Recently, charged particle beams from accelerators have been used to treat cancers directly instead of being used to produce X-rays first. The main limitation of the use of X-rays to treat cancers is that they deposit most of their energy where they originally enter the body and, in order not to damage healthy tissue, the overall dose has to be rather limited. On the other hand, particle beams offer

the advantage of depositing most of the energy in a rather narrow region just before they are stopped in the tissue.

Particle accelerators are also used in medicine to produce radioactive isotopes, which are then used to trace the movement of chemicals through the human system as an aid in diagnosis.

E. Electron Microscope

Electron microscopes are small accelerators of high optical precision and mechanical stability. The wave nature of the electrons allows them to be used in a microscope to replace optical light. The very short wave length of the elctrons makes it possible to achieve resolutions not achievable with the optical microscopes. With close attention paid to its beam optics, including compensation of their high-order aberrations, modern electron microscopes now achieve unprecedented resolutions in the Angstrom range, permitting views of single atoms.

F. Synchrotron Radiation and Free Electron Lasers

Another scientific application of accelerators that has wide use is in the production of *synchrotron radiation* in the form of intense ultraviolet light or X-rays. This synchrotron radiation is produced by an electron (or positron) beam stored in a storage ring. The radiation has the property of high intensity and excellent collimation. The X-ray intensity from a synchrotron radiation facility, for example, is typically several orders of magnitude higher than common X-ray sources. Availability of this intense source has opened up many new areas of research, including atomic and molecular physics, biology, chemistry, surface and material sciences, and micromachining.

A potentially even more intense light source is provided by *free electron lasers* (FELs). Infrared and ultraviolet FELs are being developed all over the world. X-ray FELs based on electron linear accelerators are being designed at Stanford and Deutsches Elecktronen Synchrotron in Germany. These high-performance accelerators are becoming the radiation sources of the future.

G. Microwave Sources

Accelerators are used to generate *microwaves*, which are electromagnetic radiation whose wavelength is in the range between one millimeter to many meters. Microwaves are useful for radar and for long-distance communication purposes. They are also used as the acceleration mechanism in all cyclic accelerators. As mentioned earlier, magnetrons are a microwave-generating accelerator. There are many variation of accelerators that generate microwaves useful under a variety of circumstances. One notable example of such accelerators is the *klystron*. Compared with the magnetron, a klystron has a linear architecture instead of a cylindrical one, and it typically is capable of generating microwave power in the multiple tens of megawatt range in a pulsed operation mode.

H. Industry, Material Science

The use of accelerators in industry has some similarity to their use in medicine. The particle energies are usually low, hundreds of keV to 10 MeV in most cases. A major industrial use is in diagnosis and testing. Pressure vessels, boilers, and other large metal castings are routinely X-rayed to search for internal flaws and cracks. Particle energies of 20 MeV or more are often used for greater penetration of thick castings. Such accelerator devices are also used to detect contraband at air- and seaports.

Particle accelerators are also used in materials treatment. Precise concentrations of impurity ions are implanted in metal surfaces for solid-state electronics manufacture. Particle beams are used to etch microchips in the production of integrated circuits.

Many manufactured objects are sterilized by accelerators. Such sterilization is the preferred method for bandages and surgical instruments because it damages them less than heat sterilization. The accelerator energies used for sterilization are low enough that no radioactivity is induced in the object being sterilized.

Materials are also changed chemically by accelerator radiation. A notable application is in the polymerization of plastics. Transparent shrink wrapping is treated by accelerators to produce the desired shrinkability with the application of heat. Cables are radiated to increase their durability.

Food preservation by accelerator radiation is also being carried out, mostly on a trial basis at this time, but with some large-scale application by the military services.

A recent accelerator application is the destruction of harmful bacteria in sewage by accelerator beams so the treated sewage can then be used as fertilizer.

Some military uses of accelerators have been suggested but none has as yet been put into actual practice.

I. Chemistry

In addition to industrial chemical applications, accelerators are also used for research in chemistry. Typically this involves low-energy (keV or lower) cold beams of neutral or charged molecules of specific species. Interactions between the molecules in the beam with a gas target or another beam gives information of the interaction potential,

the dynamics, and the rotation or excitation electronic energy level structures of the molecules.

J. Neutron Source, Fusion Driver

An intense beam of protons or heavy ions, led to bombard a target, can serve as an intense source of neutrons, which are useful for material research of industrial or military applications. Such a device is an alternative or complementary to research reactors. An accelerator that is capable of producing such intense proton beams is technically very demanding. A project called Spallation Neutron Source (SNS), at Oakridge National Laboratory, TN, with the design goal of 1 MW beam power, is under construction.

Research and development is being carried out to test the applicability of accelerators to confined fusion as energy source. In the systems envisaged, beams of high-energy particles would be used to bombard a small deuterium–tritium pellet. The inertia of the beams implodes the pellet, and in the process, the pellet is heated to the point at which deuterium and tritium in the pellet would fuse. A very large accelerator system would be needed to produce fusion energy economically. Ideas have also been proposed to use such a system to dispose of nuclear waste while producing energy.

IV. TYPES OF ACCELERATORS

Accelerators can be divided into two classes, those in which acceleration is carried out by use of a high dc voltage, and those in which acceleration is carried out by a lower but oscillating voltage, which are called *cyclic accelerators*. Cyclic accelerators are further divided into linear accelerators and circular accelerators.

A. High-voltage Accelerators

In a *high-voltage accelerator*, a terminal or electrode is charged to high dc voltage and particles are accelerated from it to ground potential. If the terminal is charged to a voltage V, a singly charged ion will gain energy eV in the accelerator. The maximum possible voltage is limited by sparking to ground. It is also limited by less dramatic corona discharge. With ample space to ground and scrupulous attention to detail in design, terminals have been built that hold 25 MV.

The simplest way to produce high voltage is with an ac electrical step-up transformer system. X-ray machines produce voltages up to 1 MV by this method. The beam is only accelerated on one-half the ac cycle and varies in energy throughout the pulse.

In order to avoid scattering of beam particles by residual gas, there must be an accelerating tube that is evacuated to low pressure. The voltage drop must be distributed somewhat uniformly along this tube to avoid sparking. For voltages above approximately 1 MV, the accelerating tube is almost always insulated outside the vacuum by a pressurized gas of high dielectric strength, often sulfur hexafluoride.

To produce a dc voltage, electric charge is brought to the terminal. Charge may be brought electronically by voltage-multiplying circuits (e.g., the Cockcroft–Walton set and the Marx generator), or mechanically by a moving belt (e.g., the Van de Graaff accelerator).

1. Voltage Multiplying Devices

The first successful high-voltage accelerator, the *Cockcroft–Walton accelerator*, made use of a voltage-doubling circuit, the Greinacher circuit, shown in Fig. 8. Two rectifiers act on opposite sides of the ac sine wave to charge a capacitance to twice the voltage. The principle can be extended to many stages. Cockcroft and Walton accelerated protons to 300 keV and in 1932 demonstrated the first nuclear reaction with artificially accelerated particles. Modern Cockcroft–Walton generators are available commercially with voltages up to approximately 1 MV. Special pressurized systems have been built to 3 MV. Cockcroft–Walton generators are used often as the first stage of higher energy accelerator systems because they produce beams with very good energy regulation. This application however has nowadays been replaced by the radio-frequency quadrupoles.

The *Marx generator* is similar in principle. The rectifier system is external to the capacitor stack. In essence, the capacitors are charged in parallel, then discharged in series through spark gaps. Marx generators were originally used in the 1920s to produce surges of high voltage

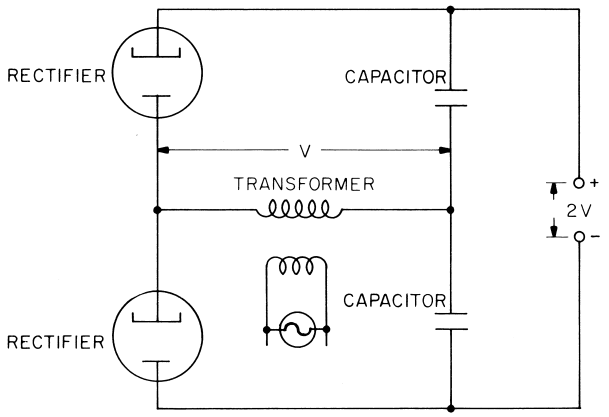


FIGURE 8 Schematic diagram of a voltage-multiplying circuit.

to test electrical generating and transmitting equipment. They are used today to produce very intense (1 to 10 kA), short (10 to 50 nsec) pulses of 1 to 10 MeV particles. The particle energy is poorly regulated during the pulse.

2. Charge Carrying Devices

The most prominent of the accelerators in which charge is carried mechanically to the terminal is the electrostatic generator or *Van de Graaff accelerator*. In this device, depicted schematically in Fig. 9, clharge is carried by a moving belt to the high-voltage terminal. In most modern accelerators, the belt is made of a series of insulated metal links, looking a little like a bicycle chain. Van de Graaff accelerators played an important role in nuclear physics development during the 1930s and are still in use in that field. They are now frequently used in industrial applications. Because of their inherent outstanding regulation, these machines can be used as sensitive isotope separator for geological age determinations.

The Van de Graaff can be expanded to the tandem Van de Graaff, in which negative ions are accelerated to high voltage, then stripped of electrons so that they become positive ions and are accelerated back to ground, thus receiving twice the kinetic energy corresponding to the terminal potential. Cockroft–Walton and Marx generators can reach approximately 1 MV. Van de Graaff generators have reached 25 MV and tandems have accelerated particles to 50 MeV. This tandem Van de Graaff produces a higher energy, but with lower intensity because the process of stripping cannot be made perfectly efficient. The dc beam current in a single electrostatic generator can be as large as 10 to 20 μA, but in a tandem is of the order of 1 μA. An electrostatic generator installation, a *tandem pelletron*, is shown in Fig. 10.

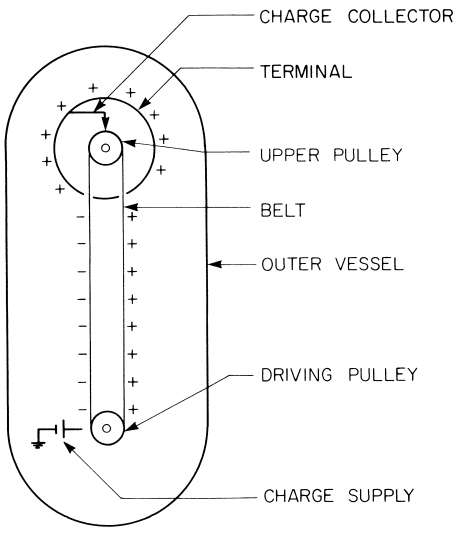
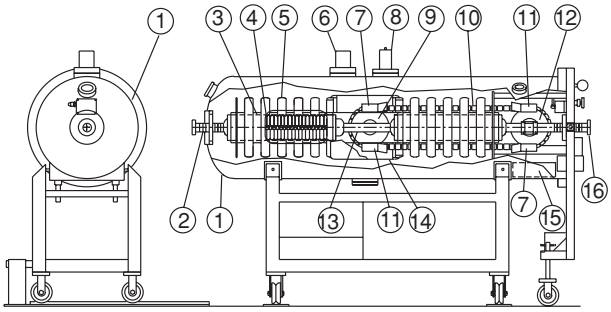


FIGURE 9 Schematic diagram of an electrostatic generator, or Van de Graaff accelerator.



1. Pelletron Tank
2. Beam Input Drift Tube
3. Insulating Plastic Support
4. Accelerating Tube-Cutaway view
5. Potential Distribution Ring
6. Generating Voltmeter
7. Discharging Inductor
8. Capacitor Pickup
9. Chain Idler Wheel
10. Pellet Charging Chain
11. Charging Inductor
12. Chain Drive Wheel
13. Gas Stripper Assembly
14. High Voltage Terminal
15. Dessicant Tray
16. Beam Output Drift Tube

FIGURE 10 Schematics of a 1-MV tandem pelletron electrostatic accelerator built by National Electrostatics Corporation. Voltage is charged by pellet chain. (Figure courtesy of World Scientific Pub.) [Courtesy of World Scientific, Singapore.]

B. Linear Accelerators

There are two kinds of cyclic accelerators: linear accelerators (linacs), in which each particle passes once through a sequence of accelerating structures, and circular accelerators, in which each particle traverses a closed path (not necessarily exactly a circle) and passes repeatedly through the same accelerating structure. In a cyclic accelerator, the accelerating forces must vary with time, in contrast with the dc forces in high-voltage accelerators.

In a linac, the particles being accelerated follow paths that are approximately straight. These particles are accelerated in the desired direction by the action of electric fields. In the transverse directions, the particles are confined or focused into a beam by the action of lenses employing static electric or magnetic fields, or in some cases by time-varying harmonic fields, as in the *radio-frequency quadrupole* (RFQ) focusing system.

The high-voltage dc accelerators discussed earlier, the most elementary form of linac, are limited by the maximum electric potential that can be supported by an array of conductors. In practical cases this is a few million volts. On the other hand, if the accelerating electric field is time varying, continuous acceleration can be achieved, and there is no physical limit to the maximum particle kinetic energy that can be achieved. In the time-varying field linac, a substantial fraction of the accelerating field

energy is contained in its accelerating wave. Particles traveling with the crest of the wave gain energy from it much as a surfer gains energy from a traveling water wave near a beach.

Preaccelerators for linacs are often high-voltage dc or pulsed-dc accelerators, operating at energies from a few hundred keV to a few MeV. Protons and heavier ions have rest energies (masses) of one to many GeV so that they emerge from the preaccelerator with velocities only a small fraction of the velocity of light. Electrons have a mass of 511 keV so that even a preaccelerator of only 80 keV boosts their velocity to one half the velocity of light. The arrangements needed for efficient generation of accelerating waves depend markedly on the desired wave velocity, and the designs of linacs for proton or heavier ions and for electrons differ markedly.

1. Proton and Heavy-Ion Linacs

The traveling-wave system does not work for particle speeds very much less than that of light because the wave cannot be efficiently slowed down enough to match the particle speed. If the wave velocity is greater than the particle velocity, the wave passes by each particle. As it passes, a particle will experience decelerating forces during the negative-field portion of the sinusoidal variation of the wave. To accelerate these low-energy particles, instead of a traveling wave a *standing wave* is used, and conducting drift tubes are placed around the particle trajectories in the regions of negative fields in order to shield particles from these fields. These drift tubes are shown schematically in Fig. 11. Acceleration then takes place in the gaps between *drift tubes* and the particles drift through the tubes unaffected by negative fields. In most standing-wave linacs, the drift tubes contain focusing devices to confine the transverse motion. Drift-tube linacs are used for accelerating protons or heavy ions. They are capable of producing high intensities of accelerated beam.

The presence of drift tubes strongly affects the choice of frequency in a standing-wave linac. The accelerating frequency cannot be so high that the drift tubes are too small to contain focusing elements. Proton linacs usually utilize frequencies of approximately 200 MHz. Heavy-ion linacs, which inject at even lower speed, often have frequencies in the 60 to 80-MHz range.

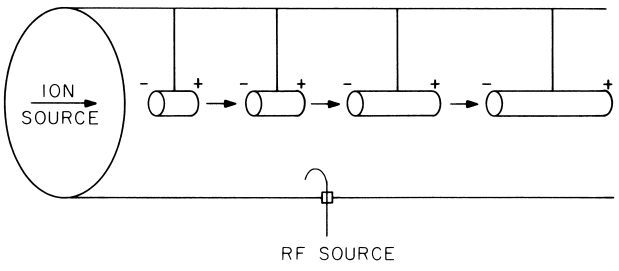


FIGURE 11 Schematic diagram of a standing-wave drift-tube linear accelerator.

The linac cavities that form the outer envelope of the accelerator and contain the electromagnetic field are built in sections for convenience in manufacture. A 200-MeV proton linac is approximately 500 ft long. The pulse length is short enough that the average radio-frequency power is only a few kilowatts. The separate amplifiers for each cavity are synchronized by a master oscillator. Peak currents of 200 mA of 200-MeV protons are achieved in injectors for proton synchrotrons. Linacs with longer pulse length are built for applications in which very high intensity is desired. For higher energy, a standing-wave linac can be used to inject into a traveling-wave linac when the particle speed is comparable to that of light. The largest proton linac is the 800-MeV accelerator at Los Alamos National Laboratory.

The copper lining and drift tubes of a standing-wave linac can be replaced by superconducting metals such as a niobium or lead, and the entire system cooled down to liquid helium temperature (2 to 4 K). Several superconducting heavy-ion linacs (e.g., ATLAS at Argonne National Laboratory) are is in operation. The use of superconductivity is economically justifiable for long-pulse or continuous wave (CW) linacs.

Particles can also be accelerated by the electric fields induced by time-varying magnetic fields according to Faraday's Law. In an *induction linac*, magnetic fields are pulsed sequentially in synchronism with particle motion through the accelerator. Induction linacs are particularly useful for producing short (10 to 50-nsec) pulses of very high peak intensity (1000 A) at energies of 10 to 50 MeV.

2. Electron Linac

In this kind of linac, the wave velocity is made constant at light velocity over almost the entire length. An efficient conductor arrangement that supports the needed longitudinal electromagnetic accelerating wave is shown in Fig. 12. This waveguide is a cylindrical pipe periodically loaded with diaphragms spaced between one-fourth and one-half of the free-space wavelength of the driving field. The wave velocity is controlled by the diameter of the pipe, about equal to the wavelength, while the rate at which power flows down the waveguide is controlled by the size of the hole in the diaphragm. The operating wavelength of such linacs is set by the simultaneous need for efficient acceleration and for efficient generation of the microwave power carried by the accelerating wave. Operating wave lengths between 30 and 3 cm have been used, with 10 cm being the most common today.

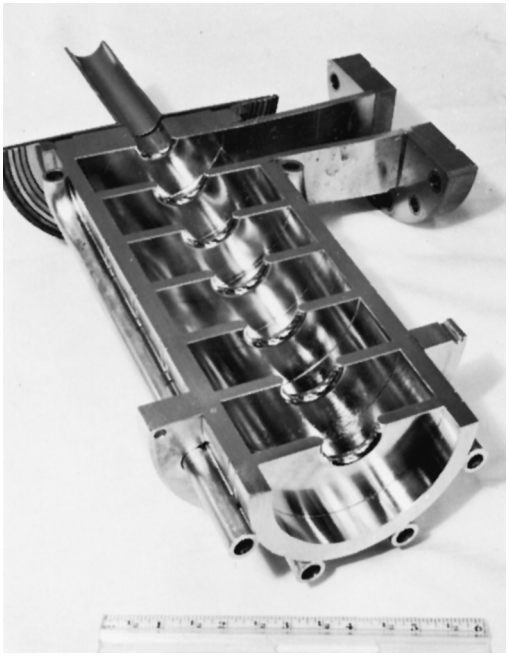


FIGURE 12 Cutaway view of a traveling-wave wave guide. The radio-frequency wavelength used is approximately 10 cm. The beam enters from the upper left and passes through the small hole. Radio-frequency power is brought in at the upper right, and the wave travels with the beam. [Courtesy of Stanford Linear Accelerator Center.]

Rather tight tolerances must be maintained in construction of the waveguides. At 10-cm operating wavelength, tolerances of about 0.02 mm must be held. At shorter wavelengths, the tolerances are correspondingly tighter. Maintenance of the correct wave velocity also requires regulation of the wave-guide temperature to a fraction of a degree Centigrade. In spite of these technical difficulties, very high frequency accelerating structures (e.g., at W-band of 90 GHz) are being explored as a way to produce high-gradient acceleration for the future linacs for high-energy physics research.

With proper synchronization by a master oscillator, a large number of units such as those of Fig. 12 can be strung together end to end to produce as high a beam energy as needed. Copper accelerating wave guides available today can dependably maintain effective accelerating fields of 20 MV per meter of length for a power expenditure of 3 to 5.4 MW per meter. Thus, the output energy of the 3000-m Stanford Linear Accelerator will be about 50 GeV with each of its 240 power amplifiers pulsing at 50 MW. A view of the SLAC linac is shown in Fig. 13.

While magnetron tubes are sometimes used as power sources for linacs of a few MeV output energy, *klystron amplifiers* are the usual choice at microwave frequencies. Today, accelerator klystrons capable of higher than 50 MW peak power with pulse lengths of a few microseconds are being made. Tubes capable of up to 1 GW for a fraction of a microsecond are being contemplated.

As with proton or heavy-ion linacs, it is possible to replace the normal copper conductor with a superconductor, operating at liquid helium temperatures. The amount of microwave power needed to establish the accelerating fields is then reduced by a factor of 10^5 to 10^6. This benefit is somewhat offset by the need of powerful refrigerators. The technology has matured substantially over the past decade or two, and has come into use more readily now for the new generation of proton and electron linacs.

C. Circular Accelerators

Like linear accelerators, circular accelerators utilize time-varying fields to accelerate particles. In addition, there must be magnetic fields to bend particles around a closed path that returns them to the accelerating structure. The configuration of the accelerating structures and magnetic fields can take many different forms in different circular accelerators.

1. Cyclotron

The earliest circular accelerator was the cyclotron. In the cyclotron, particles are injected at the center of the cyclotron and spiral outward as they are accelerated. A uniform, time-independent magnetic field provides the

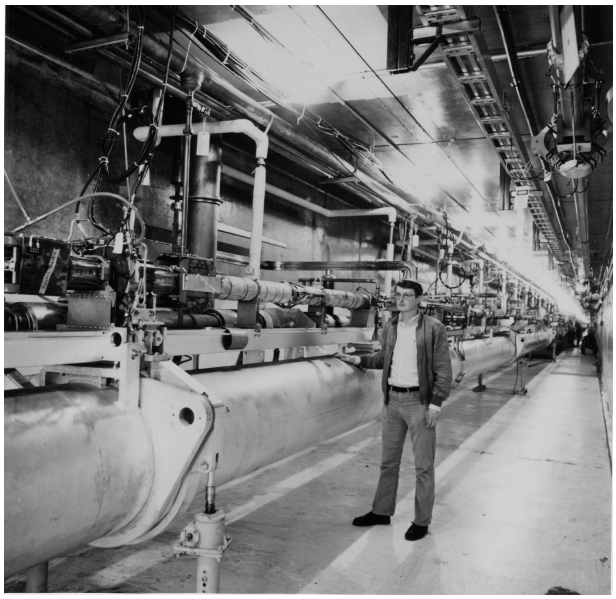


FIGURE 13 Tunnel of the Stanford linear accelerator. The accelerator structure is the white horizontal tube at the man's eye level. It sits on a strong girder system for alignment and stability. Radio-frequency power is fed to the waveguide from klystron amplifiers in separate buildings through the rectangular tubes. [Courtesy of Stanford Linear Accelerator Center.]

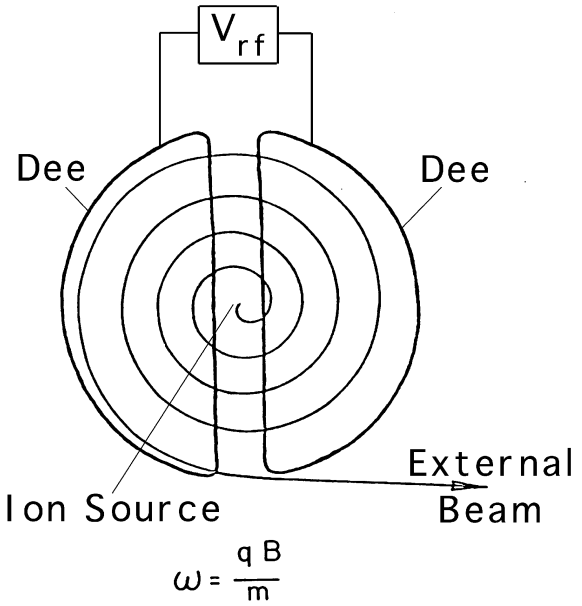


FIGURE 14 Schematic diagram of cyclotron orbits. (Figure courtesy of World Scientific Pub.)

bending to make these spirals. The drift tubes in a cyclotron are in the form of hollow boxes called "dees" after their shape. Acceleration takes place as a particle crosses the gap between the two dees. Cyclotrons have been built to operate up to 800 to 900 MeV.

Figure 14 is a sketch of cyclotron orbits. A particle of higher energy moves faster along the orbit but has a longer distance to go between gap crossings. The frequency of gap crossings, and therefore accelerations, is constant in the cyclotron, and it produces a steady stream of bunches of accelerated particles. Quantitatively, as will be explained in Eq. (6), $mv = e\rho B$, but $v = \omega\rho$, so that $m\omega = eB$, and ω is independent of the radius ρ and the energy.

According to the special theory of relativity, when a particle's velocity approaches that of light, the mass of a particle increases as its energy is increased. As a consequence, for the same increase in energy, the velocity of a particle of higher energy does not increase as much. The gap-crossing frequency therefore decreases, and particles fall out of step. This effect makes a noticeable difference at 15 MeV for protons, which limits the peak energy of the Lawrence cyclotron. Several systems to circumvent this problem have been invented. One of these is the *synchrocyclotron*, in which the frequency of the accelerating voltage is reduced to keep in step with a group of bunches of particles as they are accelerated. The synchrocyclotron produces bursts of a series of bunches of accelerated particles.

Synchrocyclotrons have been built to produce protons of 750-MeV energy. They have been largely superseded by *azimuthally varying-field* (AVF) cyclotrons, which are also called *sector-focused cyclotrons*. Here the magnetic field varies periodically around the azimuth of the cyclotron in such a way as to keep the frequency of gap crossings constant. Like cyclotrons, AVF cyclotrons produce a steady stream of bunches. The intensity achievable is much larger than in a synchrocyclotron, and many are being used for scientific research.

2. Betatron

As with linear accelerators, it is possible to accelerate particles by induction in a circular configuration. Circular induction accelerators are called *betatrons*. They are used to accelerate electrons. Many betatrons are used to provide 20 to 30 MeV electrons for medical or industrial work. The energy limit is due to the effect of synchrotron radiation energy loss. Charged particles radiate electromagnetic waves when their trajectories are bent. In accelerators, this is called *synchrotron radiation*. For light particles such as electrons, the energy radiated away is large enough that it limits the maximum achievable energy of the betatron to 300 MeV.

3. Microtron

The relativistic effects that limit cyclotrons set in at much lower energies for electrons, and a different configuration, the *microtron*, is more appropriate. In the simplest microtron, electron orbits are a series of circles tangent at the position of the accelerating radio-frequency cavity. The periods of revolution on each of these circles differ by an integral number of periods of the accelerating voltage and the electron bunches therefore stay in phase with the accelerating voltage. These simple orbits are sketched in Fig. 15. Microtrons can also be built in the shape of racetracks and have been used to accelerate electrons to energies of several GeV.

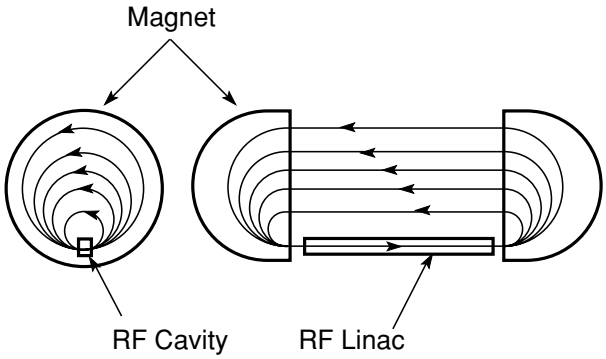


FIGURE 15 A classical microtron and a racetrack microtron. [Courtesy of World Scientific, Singapore.]

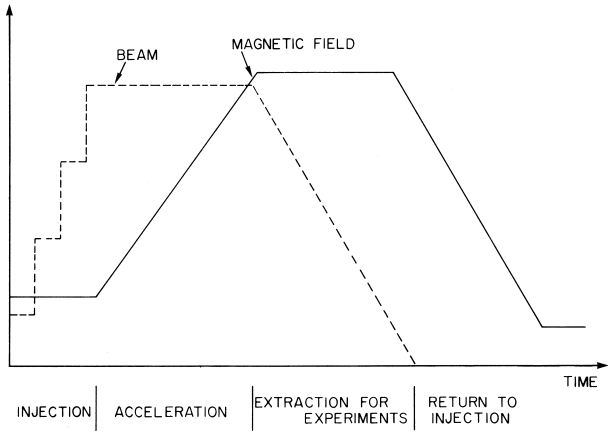


FIGURE 16 Operation cycle of a synchrotron.

4. Synchrotron

Energies greater than 1 GeV require a different configuration, the *synchrotron*. The magnetic guide fields in all the circular accelerators discussed previously are constant in time. In the synchrotron, the magnetic field is increased in time as the particle energy is increased. A graph of a synchrotron cycle is shown in Fig. 16. The radius of the particle orbit is held constant. Synchrotron magnetic fields need to extend over a relatively small aperture rather than over the entire area of the circle, as in constant-field configurations. The synchrotron is therefore a far more economical design for energies in the GeV range. The largest synchrotron, 6 km in circumference, accelerates protons to an energy close to 1 TeV (Fig. 1).

Electron synchrotrons, although sharing the principle and magnetic-field configuration with proton synchrotrons, have a separate feature, that is, the synchrotron radiation energy loss suffered by the electron must be replaced if the electrons are not to spiral inward and strike the walls. The energy loss per revolution ΔT by an electron of kinetic energy T following a circle of radius ρ is

$$\Delta T = 88.5T^4/\rho \tag{1}$$

where ΔT is in keV if T is in GeV and ρ is in meters. Thus, the energy loss increases very rapidly as T is increased. The accelerating system must make up this energy loss, as well as provide voltage for acceleration. Radio-frequency systems to carry out these functions become so large that electron synchrotrons become uneconomical compared with linear accelerators at energies beyond about 100 GeV.

5. Storage Ring

Storage rings are very similar in general configuration to synchrotrons. The magnetic guide fields are constant in time and a beam of particles circulates continuously. In some designs, two storage rings are intertwined with one another and beams of particles circulate in opposite directions, colliding at the intersection points, one example of which is shown in Fig. 17. In other designs, beams of particles and antiparticles (electrons and positrons or protons and antiprotons) circulate in opposite directions on the same path in the same magnetic field, often with small auxiliary fields to keep them from colliding except at the designated collision points. Like electron synchrotrons, electron storage rings are limited in energy by synchrotron radiation. Present-day technology utilizing superconducting radio-frequency cavities has allowed the electron–positron collider LEP to reach beyond 100 GeV per beam. Beyond that point, all designs of electron–positron colliders are of the type of linear colliders. Electron or positron storage rings are also widely used as synchrotron radiation facilities.

As another approach to circumvent the limit of synchrotron radiation energy loss on electron-positron colliders, it has been suggested that similar high-energy physics goals can be reached by colliding positive and negative muons in a *muon collider*. Synchrotron radiation would be much reduced because the muons are about 200 times more massive than the electrons. The disadvantage is that muons decay in a relatively short time, and experiments must be done and collision data taken before they decay.

6. New Concepts

Although it is possible in principle to build a synchrotron for an almost arbitrarily high energy, at least for protons,



FIGURE 17 One of the intersecting points in the intersecting storage ring at CERN. The large blocks are quadrupole magnets. The intersection of the two vacuum chambers can be clearly seen. [Courtesy of CERN.]

it eventually becomes impossible economically. There are accelerator scientists looking beyond to new ways of accelerating particles, perhaps using the intense electromagnetic fields available in lasers or in plasmas for acceleration and guidance or focusing of particles to very high energies. This work is still in the research phase, with many interesting ideas being proposed and tested.

V. PHYSICAL PRINCIPLES OF BEAM MOTION

A. Linear Accelerators

A charged particle moving in an electric field experiences a force in the direction of the field (or the opposite direction if it is a negative charge) and is accelerated in that direction. Thus, its velocity and kinetic energy will increase. The force $F = qE$, where q is the particle charge and E the electric field. If q is measured in Coulombs and E in volts per meter, then F is given in Newtons. This is the basic mechanism of operation of a particle accelerator. The electric field can be either constant in time, in which case particles are accelerated continuously and come out in a steady stream, or varying in time, in which case particles are accelerated only when the electric field is in the right direction and come out in bunches.

In a traveling-wave linac, the accelerating force will be proportional to the strength of the electric field along the axis of the accelerator (i.e., $F = qE$). The traveling accelerating field will vary with time t and with distance z along the accelerator as:

$$E = E_0 \cos\left[\omega\left(t - \frac{z}{v_w}\right) + \phi_i\right] \tag{2}$$

where ω is the frequency of the wave, v_w is the phase velocity of the wave, and ϕ_i is a constant that measures the initial value of the wave strength at the beginning of acceleration.

If the particles being accelerated have a velocity v_p along the accelerator axis, their position is given by $z = v_p t$. As a consequence, if the linac is designed so that $v_w = v_p$, the particles are accelerated by a constant force in the z direction:

$$F = qE_0 \cos\phi_i \tag{3}$$

As a consequence of the wave nature of the acceleration, only particles near the wave crests (i.e., $\phi_i = 0, 2\pi, 4\pi, \ldots$) receive useful acceleration. Thus, the beams from linacs employing time-varying fields are bunched, the separation of the bunches being $\lambda v_w/c$, where λ is the free-space wavelength of the accelerating field. As the particles are accelerated, their velocity increases according to the relation,

$$\frac{v_p}{c} = \left\{1 - \left[\frac{1}{1 + \left(T/M_0c^2\right)}\right]^2\right\}^{1/2} \tag{4}$$

where T is the kinetic energy of the particle being accelerated, M_0c^2 is its rest energy, and c is the speed of light. For kinetic energies much less than the rest energy, $T \ll M_0c^2$, this expression simplifies to $v_p/c \sim (2T/M_0c^2)^{1/2}$, the classical relation between velocity and kinetic energy. For $T \gg M_0c^2$, $v_p/c \sim 1$.

B. Circular Accelerators

Compared with linear accelerators, a circular accelerator has, in addition to the accelerating electric field, a magnetic field for bending particles to follow a circular path. A charged particle in a static magnetic field experiences a force in the direction perpendicular to the plane formed by two vectors, the magnetic-field vector and the particle's velocity vector. A particle moving in the direction of the magnetic field experiences no magnetic force. Because the force on a particle in a static magnetic field is perpendicular to its velocity, no work is done on the particle and its energy does not change. The energy increase in a circular accelerator comes from the accelerating electric field, not from the magnets. The force on a particle moving at velocity v perpendicular to a magnetic field B is

$$F = qvB \tag{5}$$

If B is measured in Tesla (IT $= 10^4$ Gauss) and v in meters per second, the force F is in Newtons. When the magnetic field provides the centripetal force that bends the particle in a circle of radius ρ, then

$$mv^2/\rho = qvB \tag{6}$$

and there is a relation between the momentum p and the product of radius and field,

$$p = mv = q\rho B \tag{7}$$

Thus, as the momentum of a particle is increased during acceleration, either the radius of curvature must increase, as in a cyclotron, or the magnetic field must increase, as in a synchrotron. This relation $p = q\rho B$ is very basic for circular accelerators. It holds for slow and fast particles, including effects of special relativity. For practical calculations, if B is in Tesla, ρ in meters, and p in MeV/c (where c is the speed of light), then

$$p = 300B\rho \tag{8}$$

There is a sense in which magnetic fields can increase the energy of particles. The induction linac and betatron

discussed above make use of a magnetic field changing in time to induce an electric field that can be used to accelerate particles. In circular accelerators, the paths of particles during the course of acceleration are very long, sometimes many thousands of kilometers. A particle injected at an angle with respect to the ideal path will stray farther and farther during this long distance and will leave the confines of the magnetic field and be lost unless some means is provided to focus the particles back toward the ideal path. This focusing is accomplished by building in carefully designed focusing magnets, or *quadrupoles*. Focusing is one of the most important considerations of accelerator design.

We emphasize here the difference between a particle accelerator and a nuclear reactor. In the accelerator, the particles are focused and form a beam, all moving in the same direction with the same energy, whereas in a reactor, the particles are heated and move with the random directions and wide energy distribution of a hot gas. In addition, the reactor accelerates particles by nuclear forces, whereas the accelerator makes use of electromagnetic forces. These electromagnetic forces cease when the electric power supplied to the accelerator is interrupted. Thus, the radiation from an accelerator stops, except for small residual effects, when the accelerator is turned off. The situation is quite different from that of a reactor.

The motion of particles in an accelerator is one of the most important aspects of accelerator science. The importance of *beam dynamics* arises in part from the fact that in many kinds of accelerators particles travel very large distances in the course of acceleration or storage (sometimes many millions of miles or kilometers). Stability against small perturbations and errors is vital if the particles are to stay in the accelerator for such distances.

In addition, in most accelerators it is desirable to accelerate as high an intensity of particles as possible. At high intensity, the mutual electromagnetic forces among the charged particles (e.g., *space charge* force), or the forces induced by the wake fields as a result of beam–environment interaction, can be important, thus disturbing the stability that each particle would have by itself.

To produce a stable beam in an accelerator, particles will have to execute only small oscillations about the ideal path during acceleration. These oscillations will occur in both of the two directions transversed to the ideal path and in the longitudinal direction along that path. Although there can be situations in which the transverse and longitudinal motions are coupled, in most accelerator configurations the coupling is weak, and the two kinds of motion can be discussed separately, as we shall do.

C. Transverse Motion

Transverse oscillations are called *betatron oscillations* because Kerst and Serber gave the first clear discussion of them in connection with Kerst's betatron. The equations of motion of transverse oscillations had, in fact, already been given by Walton and in a form more useful for cyclotrons by Thomas.

The overall objective is to make the transverse motion dynamically stable, so that particles injected in the vicinity of the ideal orbit will remain in that vicinity. In some short accelerating systems, this can be achieved by focusing the beam at the particle source, but in longer accelerators, restoring forces along the orbit are required. These restoring forces are supplied by external electric and magnetic fields.

1. Weak Focusing

Let us consider the magnetic field that bends the particle around a closed path in a circular accelerator. Here r, is the radial direction in the plane of the closed-orbit path, the median plane, and z is the dimension perpendicular to the orbit plane. The distance along the orbit is s. If the vertical magnetic field B_z varies as a function of r (that is, if it has a gradient), then there is a radial field B_r, at positions off the median plane, as follows from the Maxwel1 equation $\nabla \times \mathbf{B} = 0$. A particle moving in the s direction experiences a vertical force whenever it is away from the median plane. If the vertical field decreases with radius ($\partial B_z/\partial r < 0$), the force deflects the particle back toward $z = 0$ for z either positive or negative. On the other hand, the force of the particle is always in the direction away from the median plane if $\partial B_z/\partial r > 0$.

Thus, if the guide field decreases with radius, motion off the median plane is stable in the sense that a particle starting off the median plane will not move to ever-larger z. This vertical focusing was found experimentally in the earliest cyclotrons and understood qualitatively at that time. It was made quantitative by Kerst.

In a decreasing field, the radial force on a particle decreases with radius. However, the centripetal force mv^2/r needed to keep the particle of mass m and speed v in a circle radius r decreases as $1/r$. Thus, if the field decreases less rapidly than $1/r$, a particle at larger radius feels a larger force focusing it back toward the ideal orbit and the particle has horizontal or radial focusing. Kerst expressed these results in terms of the relative derivative, or *field index*,

$$n = -\frac{r}{B}\frac{\partial B_z}{\partial r} \tag{9}$$

and the condition for focusing in both transverse directions is

$$0 < n < 1 \tag{10}$$

In this *weak focusing* arrangement, horizontal and vertical focusings are complementary to each other in the sense

that each decreases as the other increases, and a balance must be struck in design between vertical and horizontal aperture. For example, the fields in cyclotrons decrease very little with radius, in order to keep as close as possible to isochronous motion, and the vertical focusing is very weak, if at all.

2. Strong Focusing

Many synchrotrons were built with weak focusing and operated well. But the amplitudes of oscillations about the ideal orbit are larger in a larger accelerator (approximately proportional to the radius), reaching 10 to 20 cm in the cosmotron (3 GeV) and betatron (6 GeV) proton accelerators. It would be extremely costly to make use of weak focusing in a much higher energy accelerator, when oscillation amplitudes could be as large as several meters.

Strong focusing overcomes this difficulty by using an alternating series of gradients, thus *alternating-gradient* focusing, to focus both horizontally and vertically. An alternating series of gradients focuses a particle in a manner similar to an alternating series of optical lenses. As we can see in Fig. 18, a ray is farther from the axis in the converging (focusing) lenses than in the diverging (defocusing) lenses and so is bent more sharply, so that the net result is focusing. A gradient that is focusing for horizontal motion is defocusing for vertical motion, but the alternation produces focusing in both. The complementarity that limits weak focusing is avoided, and the net focusing can be much stronger.

The gradients vary periodically around the circumference of the accelerator, with a fixed number of periods per revolution. Oscillation amplitudes in a large synchrotron are a few centimeters or less, and the vacuum chambers and magnet apertures are correspondingly small.

3. Betatron Oscillation

A particle that is not injected on the ideal orbit will execute betatron oscillations about that orbit. These oscillations are characterized by ν, the number of complete oscillations per revolution around the accelerator. There

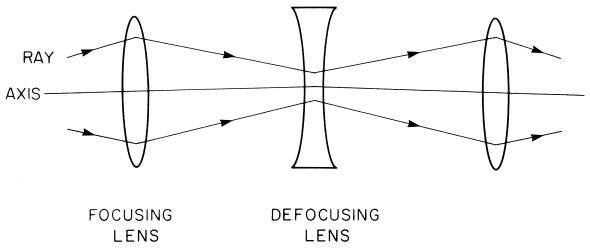


FIGURE 18 Focusing of rays in a series of alternating lenses.

are separate values for horizontal (ν_r) and vertical (ν_z) oscillations. In a weak-focusing accelerator, the oscillations are sinusoidal, and both ν values are less than unity. In a strong-focusing accelerator, these oscillations are sinusoidal on the average, with periodic excursions around the average sine wave, and ν_r and ν_z are usually considerably larger than unity. Thus, in a weak-focusing accelerator, the oscillations have the form:

$$\begin{aligned} r &= r_0 + A_r \cos(\nu_r s/R + \theta_r) \\ z &= A_z \cos(\nu_z s/R + \theta_z) \end{aligned} \tag{11}$$

where $2\pi R$ is the accelerator circumference, and the amplitudes A_r and A_z and the phases θ_r and θ_z are determined by the initial conditions at injection.

In a strong-focusing accelerator, the oscillations have the form

$$\begin{aligned} r &= r_0 + A_r \sqrt{\beta_r(s)} \cos[\phi_r(s) + \theta_r] \\ z &= A_z \sqrt{\beta_z(s)} \cos[\phi_z(s) + \theta_z] \\ \phi_r(s) &= \int_0^s \frac{ds}{\beta_r(s)} \\ \phi_z(s) &= \int_0^s \frac{ds}{\beta_z(s)} \end{aligned} \tag{12}$$

The periodic functions $\beta_r(s)$ and $\beta_z(s)$ are the *betatron amplitude functions*. The amplitude varies periodically with $\sqrt{\beta(s)}$. The phase advances as $1/\beta(s)$, so that β is the instantaneous wavelength of the oscillation. The amplitude functions and the ν values are related because the total phase advance per revolution is

$$\phi(2\pi R) = 2\pi\nu = \int_0^{2\pi R} \frac{ds}{\beta(s)} \tag{13}$$

for either r or z.

A group of particles is injected with angles and positions distributed around the ideal orbit. It is instructive to plot the motion of the group of particles in a space whose axes are position and angle at a given point s, as in Fig. 19. This is called a *phase space*. As the group moves along the accelerator and s varies, the envelope containing the group will vary in shape, but its area will remain constant. If instead of the angle, the product of angle and total momentum is taken as a phase-space coordinate, the phase-space area remains constant even during acceleration. This is an example of a general dynamical rule called *Liouville's theorem*. We shall return to this theorem in the discussions of beam stacking and cooling later.

The horizontal and vertical motions are independent in an ideal accelerator and each has its own separate two-dimensional phase space. In a real accelerator, nonlinear restoring forces, magnetic-field imperfections, or magnet misalignments can introduce horizontal–vertical

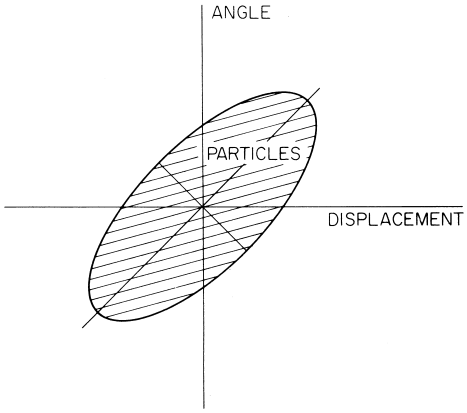


FIGURE 19 Transverse phase space. The shaded elliptical area represents a group of particles. This ellipse will oscillate as the group moves along the accelerator, but its area will remain constant.

coupling, and the two motions can affect each other. The phase spaces are then not independent, but the four-dimensional volume of the combined phase spaces occupied by particles will still remain constant.

Magnetic-field errors and magnet misalignments can also distort the beam path by inducing forced oscillations in the beam. The central orbit then moves in a periodic forced oscillation, the *closed orbit*, and all particles oscillate about this closed orbit. The occupied region of phase space then moves in this forced oscillation. If the ν value is close to an integer, the closed-orbit oscillation becomes very large, and the beam can rapidly leave the accelerator. This *integral resonance* can be kept under control by careful construction and alignment of the magnets and by careful control of the ν value to avoid integers.

In a strong-focusing accelerator, errors in magnetic-field gradients can make the oscillations about the closed orbit unstable if the ν value is close to a half-integer. In this case, the occupied region of phase space becomes elongated, even though area is still preserved, and particles reach very large oscillation amplitudes. *Half-integral resonances* are not as serious as integral resonances, but care must be taken in construction and alignment to avoid them, too.

Even in an ideal accelerator, not all particles will have exactly the same momentum. Each particle's momentum will also vary relative to the ideal momentum during longitudinal oscillations. A particle whose momentum is different from the central momentum will undergo forced transverse oscillations about the closed orbit. These will appear in phase space as overlapping groups of particles. This phenomenon is called *dispersion* in analogy to light optics.

To be focused toward one another, two particles on different orbits must encounter different magnetic fields, as in a gradient magnet or quadrupole, or go through different lengths of field. Different lengths can be achieved by building magnets whose edges are not perpendicular to particle orbit. This edge focusing is used in AVF cyclotrons. If an edge is slanted so that the path length increases with radius, the edge is horizontally defocusing and vertically focusing.

In a radial sector AVF cyclotron, both upstream and downstream magnet edges are vertically focusing. Radial focusing is provided by the increase of guide field with radius. This system is called *Thomas focusing*. In a spiral-sector AVF cyclotron, one edge is vertically focusing and the other is vertically defocusing, giving alternating-edge focusing, analogous to alternating-gradient focusing.

4. Transverse Motion in Linacs

There is no centripetal force in a linac, so there is no analog of weak focusing. Before strong focusing was developed, many proton linacs had wire grids installed in the drift-tube bore opening to change the variation of the electric field with longitudinal distance and radius to provide some focusing. But the grids intercepted many beam particles and were unsuitable for high-intensity beams because they were heated and melted by the beam.

After strong focusing was developed, quadrupole magnets were built into the drift-tube interiors, and linacs became high-intensity accelerators. More recently, methods of shaping the radio-frequency field to produce quadrupole focusing, the *radio-frequency quadrupoles* (RFQs), have been developed and provide even higher intensities and smaller beam losses.

Beams extracted from an accelerator and secondary beams produced in a target can also be steered and focused by sequences of bending magnets and strong-focusing lenses. These beam transport lines are used to bring beams to the point of use in an optimally focused configuration.

5. Synchrotron Lattice

The *lattice* of a synchrotron refers to the periodic arrangement of bending and focusing magnets around the circumference. The betatron amplitude functions, the ν values, and the dispersion all depend on the lattice. Two developments make it flexible to vary these functions and to achieve optimal desired orbit properties.

Separated-function magnet. Particles are bent around the accelerator by *dipole* fields that are independent of radius. They are focused to stay close to the central orbit by *quadrupole* fields that vary linearly with distance from the center. In the original conception of strong focusing, these two functions of bending and focusing were

combined in one gradient magnet, in a *combined-function* lattice. A *separated-function* lattice carries out these two functions in separate magnets. This lattice is more efficient because the bending field is the same throughout the magnet and is not limited by the maximum field attainable at the high-field side of the magnet. The bending field in a separated-function conventional iron dipole can easily be 1.8 to 2 T, while in a gradient magnet it is difficult to achieve more than 1.3 T without significant field distortion. The difference is even more striking in superconducting magnets, where it is more difficult to design a gradient magnet. The focusing is also more efficient in a separated-function lattice, because focusing magnets are concentrated at locations where the amplitude function β is large and defocusing magnets are concentrated at locations where β is small. Combined-function and separated-function magnets are shown in cross section in Fig. 20. Separated-function lattices are now almost always used in synchrotrons and storage rings.

Long straight section. The usual lattice has straight sections (i.e. field-free spaces between magnets) whose maximum length is of the order of the magnet length. For the introduction of necessary auxiliary apparatus, such as accelerating cavities, injection equipment, and particle detectors for experiments in storage ring colliders, the long straight section has proven most useful. If the normal bending arcs of the accelerator are simply interrupted by field-free regions of size necessary to accomodate necessary auxiliary equipment, the natural divergence of the beam will result in excessive aperture requirements. This divergence can be avoided by use of a few separate focusing quadrupoles to maintain the focusing properties of the lattice without significantly encumbering the needed space. Periodic arrays, or sublattices, of concentrated quadrupoles permit design of straight, almost field-free regions of arbitrary length.

6. Beam Diagnostics

It is possible to "fly blind" and operate a particle accelerator without measurements of the beam. Indeed, early accelerators operated this way, with the only indication of a beam coming from the final accelerated beam striking a target and producing X-rays or radioactivity. However, an accelerator can be operated much more easily and at much higher intensity if the beam position and size are known during operation.

The first methods of beam measurement were movable probes to stop the beam at an adjustable radius in a cyclotron. In early synchrotrons, probes were replaced by fluorescent screens which were observed through transparent windows. These rudimentary devices are still sometimes used in the early stages of searching for circulating beams, although nowadays the energy is high enough that fluorescent screens are viewed remotely using television cameras.

The center of mass of a circulating beam can be measured continuously by detecting the electric fields of the beam bunch with *pickup electrodes*, or the magnetic fields with pickup coils, in each case surrounding the beam. By using these methods, it is possible to measure the transverse position at given locations, as well as the phase of the beam bunches relative to the accelerating voltage. The transverse position as a function of azimuth is just the closed orbit discussed above. The closed-orbit information can be used to set currents in *correction magnets* to reduce its distortion and to analyze magnet misalignments. The closed-orbit and beam-phase information are used together as input to the radio-frequency feedback systems for beam orbit control.

Another important practical aspect of particle bending and focusing is scattering by the residual gas in the accelerator. Beams suffer significant scattering and diffusion in even a few feet of air at normal pressure. Accordingly, the beam must pass through an evacuated space. Usually a vacuum-tight metal or ceramic tube surrounds the beam and is evacuated by pumps. In most accelerators, a residual pressure of 10^{-8} atm is acceptable. In storage rings, where the effective path length may be several billion kilometers, a substantially better vacuum (10^{-11} to 10^{-12} atm) is required.

It is also possible to measure the beam shape making use of ionization of a dilute gas by the beam or by analysis of the *frequency spectrum* of the beam electric field, with knowledge of the dispersion function at the location of the pickup. In places where each beam particle passes only once, grids of *wire scanners* can be used, or a single wire can be moved rapidly through the beam. Two measurements arriving at the acceleration system at different locations can be combined to give the distribution in phase space.

D. Longitudinal Motion

In dc high-voltage accelerators or in induction accelerators (betatrons), particles that are accelerated at different times experience the same acceleraring field. But, in accelerators that utilize radio-frequency fields to accelerate particles (linacs, microtrons, cyclotrons, and synchrotrons), particles arriving at different times will experience different acceleraring fields and will consequently have different motions.

1. Longitudinal Stability and Acceleration

How the longitudinal motion evolves will depend on how the frequency of revolution depends on particle energy,

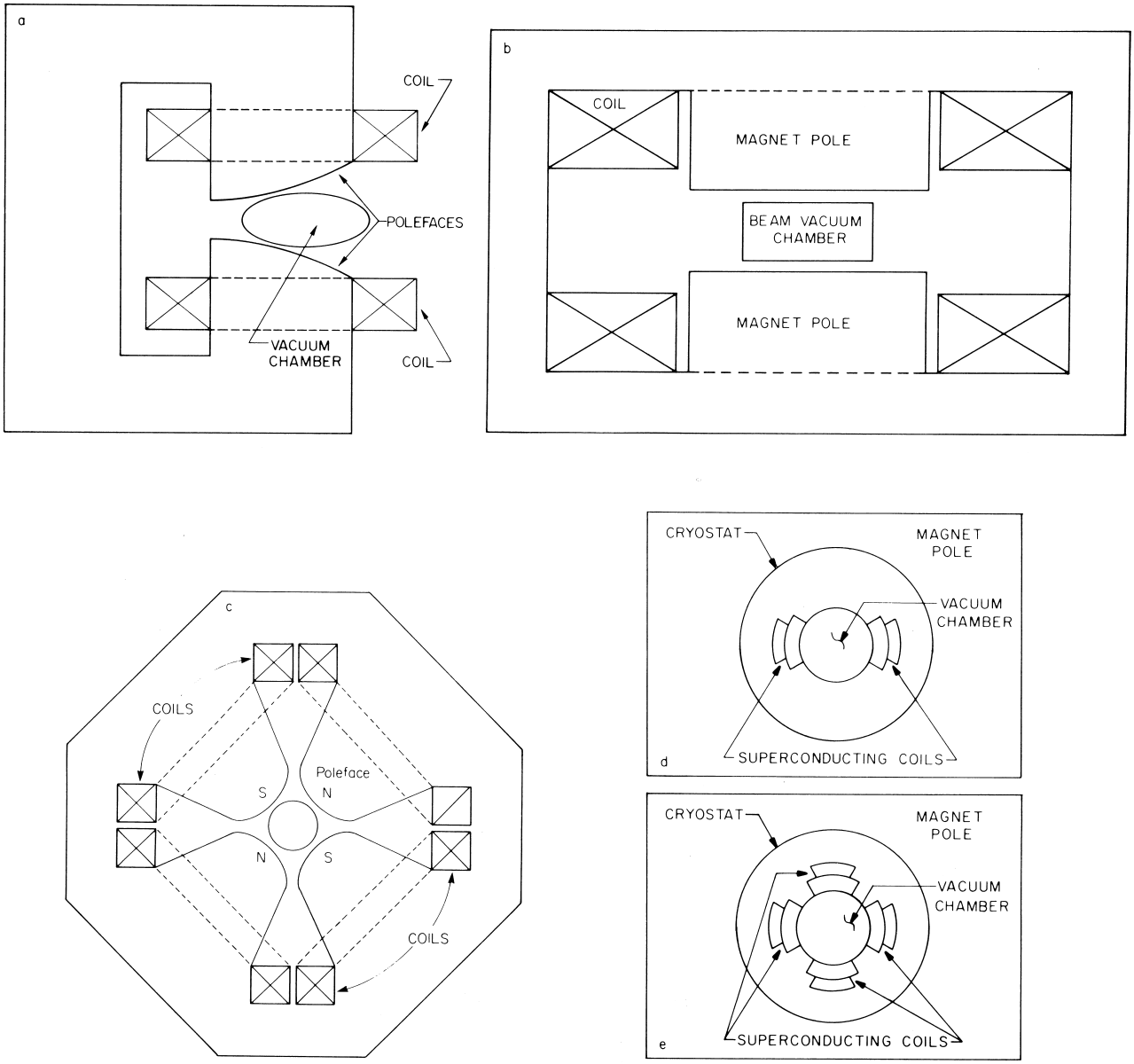


FIGURE 20 Alternating-gradient synchrotron magnet cross sections. (a) combined-function magnet, (b) dipole magnet, (c) quadrupole magnet, (d) superconducting dipole magnet, and (e) superconducting quadrupole magnet.

as illustrated in Fig. 21. The ordinate is the accelerating voltage and the abscissa is time. Thus, on this plot the accelerating voltage is a sine wave. Consider now two particles that cross the accelerating gap at the same time on one revolution. The voltage is rising as they cross. Particle 1 is in step with the accelerating voltage and crosses the gap the second time at the same phase. Particle 2 has higher energy than particle 1 at the first crossing. Let us consider two alternatives:

1. Revolution frequencies increases with energy ($df/dE > 0$). The revolution period is smaller for particle 2, and it will arrive earlier, lower on the sine wave, and will gain less energy than particle 1. The energy difference between the two particles will be decreased after the second pass. Similarly, a particle of lower energy would arrive later, higher on the wave, and would gain more energy, again decreasing the energy difference. Thus, the energy difference across the entire group of particles will not increase and they will be accelerated together. The accelerating longitudinal motion is stable.

2. Revolution frequency decreases with energy ($df/dE < 0$). Now, particle 2 arrives later and gains more energy than particle 1. It continues to gain more energy at

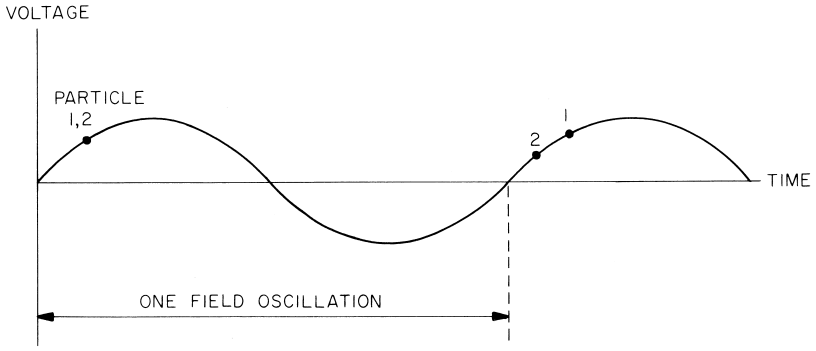


FIGURE 21 Particles riding on a voltage wave during acceleration.

every pass and the energy difference increases continuously. Thus, the longitudinal motion is unstable, and the group of particles will break up and not be usefully accelerated.

Luckily, there is a saving grace in the second alternative. There is another side to the voltage wave, where the voltage is still in the correct direction to accelerate, but is falling. Now the arguments of less or more voltage are just reversed and the acceleration longitudinal motion is stable in an accelerator with $df/dE < 0$.

What affects df/dE? There are two factors: a particle with higher energy goes faster, which always increases its frequency, but in some accelerators, it goes on a different orbit between gap crossings, which in almost all cases decreases its frequency of revolution. These two factors are in opposite directions; how they balance depends on the kind of accelerator.

In linacs, all particles travel the same path and there is no difference in path length, only a difference in speed, thus df/dE is always positive. On the other hand, in weak-focusing synchrotrons, orbits corresponding to different energies are relatively widely separated because the guide-field variation with radius is small. Here the path-length difference always overbalances the speed difference and df/dE is always negative. The same is true in a microtron. In strong-focusing synchrotrons, the speed difference is larger than the path-length difference at low energy in proton accelerators, so $df/dE > 0$, but the path-length difference is constant and the speed difference decreases as particle speeds approach the speed of light, so that $df/dE < 0$ at high energy. There is thus a *transition energy* at which df/dE is zero. At this energy, the radio-frequency accelerating voltage must be turned off, then turned back on within a few milliseconds at a different phase relative to the beam particles. Acceleration then continues on the back side of the wave. This has not been difficult in practice in proton synchrotrons. It is not necessary at all in strong-focusing electron synchrotrons because electrons move at close to the speed of light at much lower energy and the transition energy in an electron synchrotron is therefore typically lower than the injection energy, and transition is never crossed.

In this discussion, cyclotrons are an anomaly. In principle, speed differences and path-length differences just balance in cyclotrons, and they are always exactly at transition energy. In practice, there are small effects that give enough marginal stability that particles are accelerated.

2. Phase Oscillation

The above discussion of longitudinal stability can be made quantitative. The results can be described graphically in a plot like that of Fig. 22, giving angular momentum of the particle (almost the same thing as energy) against phase of a particle relative to the radio-frequency accelerating voltage. This phase can vary between 0 and 2π (360°). If the accelerating voltage is turned off and a beam of particles is simply coasting around the ring, this beam is represented by a band stretching horizontally across the entire range of 2π in phase and stretching vertically across a range in angular momentum (and energy) that corresponds to the energy spread of the beam.

When there is an accelerating voltage, there is a region of closed curves representing stable oscillations. These curves surround an equilibrium phase, which appears on the plot as a point at the center. A particle that starts at this phase and angular momentum will remain there as the whole plot rises vertically during acceleration. Particles that start within the stable region will move on a closed curve around the equilibrium phase, oscillating in momentum and phase. These oscillations are called *phase oscillations* or *synchrotron oscillations*. In almost all cases, the frequency of these oscillations is very much smaller than the frequency of revolution, so that many revolutions are needed to complete one circuit around the diagram. Particles of different energy have different orbits in a circular

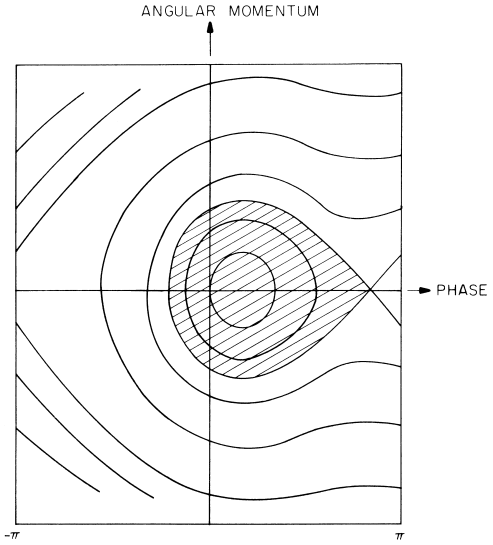


FIGURE 22 Accelerating bucket; particles in the shaded area will move on closed curves as the bucket moves up in energy during acceleration, while particles outside the bucket will slip in phase and be lost.

accelerator and there is a radial oscillation as the energy of a particle oscillates.

The stable region is called a *bucket*. The edge of a bucket is called the *separatrix*. Particles starting beyond the separatrix will slip in phase relative to the accelerating voltage and will not be accelerated continuously. During acceleration, the particles form a bunch within the bucket. Even if they fill the bucket to the separatrix, they do not occupy the full 2π in phase; a bucket that accelerates occupies less than 2π, because the voltage is decelerating for one-half the range in phase.

It is also possible to have a stationary bucket, where the equilibrium particle crosses the accelerating gap at the moment the voltage is zero. In a stationary bucket, particles can occupy the entire 2π in phase. It is also possible to accelerate a bunch in a stationary bucket by starting it near the bottom and continuing through one-half a phase oscillation, where it is near the top. This is how acceleration takes place in an electron linac.

In circular accelerators, the accelerating frequency is an integral multiple of the revolution frequency. Use of a higher frequency will make it possible to use smaller accelerating cavities and radio-frequency amplifiers. The integer multiple h is the *harmonic number*. There are now h buckets stretched across the range of 2π in phase. Each of these buckets has the properties discussed above.

The operation of an accelerating system in a synchrotron can be improved considerably by a beam-feedback system. Pickup electrodes are used to measure the phase and radius of the beam bunches. This information is fed back electronically to correct the phase and voltage of the accelerating system to keep the beam centered in the vacuum chamber and the beam and radio-frequency accelerating system in phase with each other. External signals can be used at particular times in the cycle to move the beam either laterally or longitudinally for purposes of extraction, targeting, or stacking.

In Fig. 22, the area occupied by a beam of particles remains constant during acceleration. This combination constitutes a *longitudinal phase space*, analogous to the phase spaces discussed in connection with transverse motion.

Even though phase-space area is preserved, in many cases the bunch will filament into many small threads that wind around the bucket. The empty areas between filaments are carried along with bunches in acceleration and the effective area of the bunch can increase. It is possible to avoid this decrease of density by turning accelerating voltages on and off very slowly or by overfilling buckets at the start so they stay full through acceleration, while some particles are thrown away.

3. Beam Stacking

A batch of accelerated particles can be left to circulate in a storage ring. Then another batch can be injected, accelerated, and put next to it. This *stacking* process can be repeated many times. If care is taken to avoid filamentation, the total phase-space area occupied is the sum of the areas of the individual batches. The particle density in physical space can be increased greatly by beam stacking and this makes colliding proton–antiproton beams feasible.

Synchrotron radiation in electron or positron storage rings provides a natural means for stacking, because the emission of radiation damps motion and increases the density in phase space and the particle density in physical space.

E. Multi-Particle Effects

In almost every use of accelerators, higher intensity is desirable. However, higher intensities bring with them new phenomena arising from the electromagnetic forces either directly between particles or through their interaction with the vacuum chamber environment. These forces can affect the focusing of particles and can also introduce new kinds of instabilities.

1. Effects on Focusing

The electrostatic repulsion between particles in a beam decreases the restoring forces that focus the beam. This decreases the transverse oscillation frequency of the focusing. As more particles are added, the frequency continues to decrease until it reaches a resonance. Then additional

particles will be driven to large amplitude by the resonance and lost. The beam is *space-charge limited*.

Two charges moving on parallel paths repel each other electrostatically, but they also form two parallel currents, which are attracted to each other by magnetic forces. The magnetic forces reduce the electrostatic repulsion and increase the space-charge limit. The repulsive force is always greater, but the magnetic force increases as the particles are speeded up, so space-charge forces become small at high energies. Both the electrostatic and magnetic forces are modified at high energy by the presence of the conducting vacuum chamber surrounding the beam walls.

In a colliding-beam arrangement, particles in one beam are perturbed by the electric and magnetic fields of the oncoming beam. In this case, the electric and magnetic forces add instead of cancelling each other, and this *beam–beam interaction* does not go away at high energies. Beam–beam perturbation is one of the main limitations on the performance of storage ring colliders as well as linear colliders.

2. Collective Instabilities

The space-charge and the beam–beam effects are examples of the phenomena that arise from the electromagnetic fields of the beam. Another source of electromagnetic self-fields of the beam comes from the interaction of the beam charge and current with its metallic vacuum-chamber surroundings. As the beam passes by a discountinuity on the vacuum chamber, for example, it leaves behind a *wake field*. An intense beam will generate a strong wake field, which can drive the beam either longitudinally or transversely into instability. These phenomena are called *collective instabilities* because all the beam particles feel the same forces and move collectively together in some transverse or longitudinal pattern. Because it is collective, the motion can be detected by a beam pickup and an opposing electromagnetic force applied to electrodes inside the chamber. The damaging effects of collective instabilities can be overcome to some extent by such feedback systems. In some cases, however, the frequencies involved are too high or the instability growth too rapid to make feedback practical, and the instabilities must be controlled by design changes.

F. Beam Cooling

We have emphasized in our discussion the constancy of phase-space area, both longitudinal and transverse. There are methods to reduce the phase-space area and thus to increase the density of particle beams, which is advantageous for applications such as colliding beams. These methods are called *beam cooling*. In a sense, one has to defeat Liouville's theorem in the corresponding phase spaces of the beam.

1. Synchrotron Radiation

Synchrotron radiation is usually significant only for the lightest charged particles, electrons, and positrons, although at the multi-TeV energies now being discussed, it is also important for protons.

As they are bent around the curved path by centripetal acceleration in a circular accelerator or storage ring, electrons and positrons emit a narrow cone of radiation, tangent to their instantaneous orbits. The radiated energy is mostly in the ultraviolet and X-ray regions. The energy radiated increases rapidly as the particle energy is increased (see Eq. (1)). Most of the acceleration voltage in a multi-GeV electron synchrotron is needed to make up the energy radiated away.

This synchrotron radiation decreases both longitudinal and transverse particle oscillations about the equilibrium orbit. Longitudinal oscillations are reduced because particles with energy exceeding the equilibrium value will radiate more than equilibrium particles and thereby be damped toward the equilibrium energy. Particles with energy less than the equilibrium value radiate less than equilibrium particles and are restored toward the equilibrium energy by the accelerating system. Transverse oscillations are damped because the synchrotron radiation is emitted along the direction of motion, reducing both longitudinal and transverse momentum components. The accelerating system restores only the longitudinal momentum component. Neither longitudinal nor transverse oscillations are reduced all the way to zero because the sudden random emission of the photons that make up the synchrotron radiation excites small longitudinal and transverse oscillations. The equilibrium beam size, typically about a millimeter, results from the balance between the average *radiation damping* effect and this stochastic *quantum excitation* effect.

2. Electron Cooling

A gas of protons (or heavy ions) and a gas of electrons will interact with each other by Rutherford scattering. If the proton gas has more thermal energy, it will give this energy to the electrons through the scattering, thus decreasing the phase-space area of the proton gas while increasing that of the electron gas.

Electron cooling can reduce both the longitudinal and transverse energy spreads of the proton beam. Electron cooling is done in practice by arranging an electron beam to move at the same speed as the proton beam through a straight section of a ring. At the same speed, the electron beam has much less momentum and is easily bent in and out of the proton beam at the ends of the straight section. The proton beam is repeatedly cooled by multiple traversals of the straight section as it circulates around, while the cold electrons are resupplied at each passage.

Electron cooling is more effective at lower energy, although higher energy cooling has been discussed. Its efficiency can be improved by the addition of an external longitudinal magnetic field.

3. Stochastic Cooling

Decrease of the amplitude of a collective oscillation by means of electronic feedback was discussed earlier. The root mean square amplitude of a beam can also be reduced by a feedback system using the technnique of *stochastic cooling*. The basic plan of stochastic cooling is a beam pickup that measures the average position of the beam, an amplifier system, and a kicker that transmits the amplified signal to the beam.

Consider a ring full of N circulating particles and break the circulating beam up into N parcels so that each parcel around the ring contains only one particle. If the electronic system has enough frequency bandwidth to respond to this small a parcel, it can give a signal to the kicker to correct each individual particle. Thus, particles give up their phase-space area to the electronic system.

If, as in practice, the electronic system has smaller bandwidth, each parcel has more than one particle. Each of the other particles in a parcel gives electronic noise interfering with a given particle's correction signal and the system corrects more slowly. Eventually, as more particles are added, the noise masks the signal completely and cooling ceases.

Stochastic cooling has been used as the basis for a spectacularly successful effort to cool antiproton beams and to collide them with proton beams. Important new results in high-energy physics have been achieved with this system.

4. Laser Cooling

Intense lasers can be used for cooling an electron beam. The beam is made to collide more or less head on with the laser beam at its focus, inducing a large energy loss of the elctrons. Reaccelerating the beam in the forward direction refurbishes the beam energy, but the transverse divergence of the beam is reduced. This cooling can be applied to an electron linac, or in a storage ring.

Lasers can also be used to cool a neutral atomic or ionic gas in a trap. By choosing the laser frequency to be slightly higher than the energy difference between two atomic energy levels, the laser cools the stored gas due to the Doppler shift of the escaping atoms.

5. Ionization Cooling

Cooling can also be provided by passing a beam through a dense material, causing the beam particles to lose energy, and then accelerating the beam afterwards in the forward direction. In this application, the beam loses energy by multiple scattering in the material. In order for the beam not to be lost in the material, the interaction between the beam and the material must not be too strong. This limits the applicability of ionization cooling to muon beams which might be used for neutrino sources or muon colliders.

VI. ACCELERATORS OF THE FUTURE

As discussed earlier in the history section, remarkable improvements in accelerator capabilities have been achieved over the past several decades. In applications to science, medicine, and industry, these improvements have increased maximum energy capabilities by nine orders of magnitude in 60 years, increased beam current capacity by an even larger factor, increased the variety of atomic particles that can be accelerated, and lowered the cost of particle acceleration dramatically. These improvements have been brought about by a combination of means: New acceleration methods have been devised and technological improvements to existing methods have continually emerged.

Developments in accelerators for basic scientific research have emphasized energy increase at lowered cost per unit beam energy. Improvements in accelerators for medicine have focused on increasing the variety of particles that can be accelerated, the precision of control over the beams, and the compactness and cost effectiveness of the devices. Industrial applications have continuously broadened through a reduction in accelerator costs, increases in beam current and variety of accelerated particle, and through miniaturization and increased portability.

Even though accelerators have become very sophisticated, the rate of improvement has remained steady. While this progress has occurred across the entire range of accelerator applications, the most dramatic developments have usually occurred in accelerators to be used for basic science. It is the expected developments in this area that are emphasized here. We may classify these developments as improvements to existing accelerator types and as new and improved acceleration methods.

Developments of existing methods come through deepening understanding of the physics of the method and through application of new and improved materials and techniques both in the design and in the manufacture of the accelerator. We can foresee significant improvements to both circular accelerators and microwave linacs.

A. Circular Accelerators

Basic high-energy physics research with colliding proton beams of 100 TeV or more per beam can probably

be carried out with circular accelerators based on the synchrotron principle. Continuing improvement in our understanding of nonlinear dynamics and of collective effects in dense beams will permit use of smaller and therefore more economical magnet and beam-channel cross sections while accelerating denser beams. Superconducting materials now under study in the laboratory may make possible accelerator magnets that operate at 15 Tesla or higher. With such magnets, an accelerator of 100 TeV with a 190-km circumference would be possible. Expected improvements in cryogenic technology and electronic controls will make it possible to operate such an accelerator with only a few people and at power demands about the same as those of present research accelerator complexes.

Circular accelerators producing synchrotron radiation for basic physical, biological, and chemical research will also be markedly improved through basic understanding of the physics of these accelerators and of the mechanisms of coherent radiation production. The ongoing generation of such accelerators is typically about 1 km in circumference and is outfitted with special devices called *wigglers* and *undulators* for enhancing the emitted radiation for particular research purposes. The wavelength of the radiation will be adjustable and concentrated in a narrow band and its intensity will be increased, permitting a broader range of use. The next generation will likely concentrate on the production of coherent X-ray radiation with long undulators and electron beams from high-performance linacs.

B. Microwave Linacs

Because of the intense synchrotron radiation emitted by electrons confined in circular accelerators, high-energy physics research with electrons at energies above 100 GeV will have to be carried out with colliding beams produced by linacs. We can foresee the possibility of extending current microwave linac technology up to perhaps a few TeV per beam. Such a collider might employ two 10-km linacs. Required for economic viability will be an improved understanding of beam collective effects, permitting acceleration of denser beams, as well as accelerating structures capable of better energy transfer efficiency to the beam while supporting effective accelerating fields of more than 100 MV per meter. Such gradients have been produced in the laboratory. An additional ingredient must be microwave power sources capable of several hundred megawatts peak pulse power with high efficiency. A microwave generator based on a relativistic electron beam formed by photoemission from a pulsed laser-irradiated photocathode gives promise of meeting these goals. Other generators under development may also turn out to be useful. Although being developed at a later stage, the superconducting microwave linac is playing an increasingly significant role in research applications requiring long pulses or continuous beams. One such accelerator, CEBAF, is already in operation at the Jefferson Laboratory. Increasingly higher accelerating gradients using superconducting cavities are a research and development priority.

As mentioned earlier, one approach of achieving a high acceleration gradient has been to develop the technology of high-frequency microwaves. Research in microwave structures and sources (e.g., in the W-band) has been carried out. Further work along this line will include the use of a laser to replace the microwave source, and a crystal to replace the microwave structure.

C. Novel Acceleration Methods

The Livingston chart, Fig. 3, indicates an exponential growth of effective beam energy over the past several decades. It has been and remains a challenge to maintain this exponentiation. The recent trend, however, is that the currently frontline technologies are becoming too expensive and novel concepts are needed.

The key to future accelerators beyond the multi-TeV energy range is the efficient production of high acceleration gradients. Acceleration mechanisms can be distinguished by how the accelerating electromagnetic wave is created and how its velocity is controlled. In principle, these necessary conditions for acceleration can be arranged in free space far from any material body, in a space near a specially designed array of conductors or dielectrics, or in some material medium.

In free space, only plane electromagnetic waves exist. In these waves, electric and magnetic fields are transverse to the wave velocity. A single-frequency plane wave is incapable of continuous acceleration of a charged particle. If two waves of different frequencies are used, however, continuous acceleration can be achieved. One wave serves to give the particle a slightly sinuous orbit so that its velocity has a component parallel to the accelerating field of the second wave. This is the *inverse free-electron laser*. Because the orbit is slightly curved, it also is limited in its maximum energy capability by synchrotron radiation. Although no such accelerator has yet been operated, it is estimated that electron energies of up to 300 GeV might be achieved. As currently conceived, it would not be a useful accelerator for protons or heavier particles.

Various arrangements of conductors have been devised to support longitudinal electromagnetic waves at wave velocities approaching that of light. The classical microwave linac is one such example. By operating such a device at shorter wavelengths, it is believed that higher accelerating fields can be supported. At wavelengths of 1 cm or less, a free-electron laser might be used as the source of the driving electromagnetic wave and is believed to have the

potential for high conversion efficiency. The low-energy, high-current beam of the free-electron laser is made to accelerate a low-current, high-energy beam. This idea has been called the *two-beam accelerator*. A free-electron laser of almost 100 MW peak power at 1 cm has been operated. In this application, the free-electron laser serves as a replacement of the more conventional klystron.

The maximum accelerating field capability of all accelerators using conductor arrays to control the wave type and velocity is limited by damage to the conducting material. A possible way to avoid this fundamental limit is to use a different array for each pulse so that damage is not relevant. The possibility of forming a suitable periodic array of liquid droplets made conducting by creating a plasma on their surfaces has been suggested. To take advantage of the enormous peak power available from lasers as an electromagnetic wave source, the droplets would be of microscopic size. The liquid drops could also be replaced by disposable optical grating. No such accelerator has been constructed. Accelerating fields of several hundred MeV per meter might be possible.

The accelerators discussed above are driven by harmonic power sources. A wide frequency band, conductor-controlled accelerator, the *wake-field transformer accelerator*, has been proposed. Similar in some respects to the two-beam accelerator, it employs as an energy source a high-current, low-energy beam consisting of a series of rings propagating along their axis of symmetry at essentially the velocity of light. The energy carried by these rings is deposited as a pulse of electromagnetic energy near the periphery of a conducting cylinder. This pulse of energy propagates toward the center of the cylinder being compressed thereby. The resulting high fields at the center are then used to accelerate the low-current, high-energy beam. Acceleration fields of as much as 200 MV/m are expected. Initial experiments on wake-field acceleration have been successful.

In a medium such as a gas, plasma, or charged-particle beam, electromagnetic waves propagate more slowly than in free space. One of these can be used to achieve continuous acceleration in a variety of ways. One of the most interesting is the *plasma beat-wave accelerator*. In this device, two superposed laser beams of slightly different frequency travel with and just ahead of the particle beam to be accelerated. If the difference of the two laser frequencies is just equal to the plasma frequency, the laser pulses will resonantly drive the plasma into oscillation. The electric fields resulting from the charge separation in the plasma wave are calculated to be enormous, as much as several gigavolts per meter. Such a high accelerating gradient is possible because plasmas do not have breakdown limits. The existence of beat waves has been established experimentally, as has their ability to accelerate particles.

The two beating lasers can be replaced by a single sharp laser pulse to generate a similar effect. In this case, the laser pulse drives the plasma wave as a shock response instead of a resonant driving. One can also replace the laser pulse by a short pulse of electron beam and accomplish a similar result.

Improvements in accelerator development will result from all of these studies. Which, if any, of these approaches will provide the front-line accelerators of the future remains to be seen.

SEE ALSO THE FOLLOWING ARTICLES

CHARGED-PARTICLE OPTICS • CP (CHARGE CONJUGATION PARITY) VIOLATION • HEAVY IONS (HIGH ENERGY PHYSICS) • LASERS, FREE ELECTRON • X-RAY, SYNCHROTRON AND NEUTRON DIFFRACTION

BIBLIOGRAPHY

Chao, A. W., and Tigner, M. (1998) "Handbook of Accelerator Physics and Engineering," World Scientific, Singapore.

Lawson, J. D., and Tigner, M. (1984). "The physics of particle accelerators," *Ann. Rev. Nuclear Part. Sci.* **34,** 99.

Livingston, M. S., and Blewett, J. F. (1962). "Particle Accelerators," McGraw–Hill, New York.

Scharf, W. (1986). "Particle Accelerators and Their Uses," Harwood Academic, New York.

Wilson, R. R. (1980). *Sci. Am.* **242,** 26.



Atomic and Molecular Collisions

Robert E. Johnson
University of Virginia

Joel M. Bowman
Emory University

I. Introduction
II. Impact Parameter Cross Sections
III. Elastic Scattering
IV. Wave Mechanics of Scattering
V. Interaction Potentials
VI. Inelastic Collisions
VII. Reactive Collisions

GLOSSARY

Born approximation First-order estimate of the collision cross sections.

Born–Oppenheimer approximation Separation of the electron and nuclear motion. The latter is often treated classically.

Charge exchange Process of transferring an electron from one of the colliding particles to the other; usually from a neutral to an ion.

Cross section Probability of an interaction between two colliding particles expressed as an area.

Differential cross sections Cross sections that are functions of one of the collision results (e.g., angle scattered; energy transfer). Summing over all possible results gives the cross section.

Elastic collision Collision involving a deflection of the colliding particles but no change in their internal state.

Impact parameter Defines the closeness of a collision as that component of the distance between the two colliding particles that is perpendicular to their relative velocity.

Inelastic collision Collision resulting in a change in the internal states of the colliding particles.

Interaction potentials Net change in potential energy of the two colliding particles. A potential exists for each set of initial states of the particles.

Scattering length Effective range of the scattering interaction for a very low energy collisions.

Semiclassical method Calculation of wave-mechanical effects using classical quantities.

Resonance A long-lived, metastable, state formed in a collision.

ATOMIC AND MOLECULAR COLLISIONS involves the study of the effects produced by the motion of atomic and molecular particles when they approach each other.

The effects produced are generally described in terms of the amount of energy transferred between the colliding particles. This may be simply an elastic energy transfer corresponding to a deflection or an inelastic energy transfer producing changes in the internal states of the particles. The study of this energy transfer is used to determine the details of the forces of interaction between atomic and molecular particles. In addition, knowledge of these effects is used to describe phenomena in which collisions play an important role, such as the behavior of gases and plasmas and the modification of solids by ion bombardment.

I. INTRODUCTION

A. Overview

The need to understand the behavior of colliding atoms and molecules is self-evident as we live in a world constructed from atomic building blocks. It is a dynamic construction of moving particles governed by a few fundamental forces. The interaction between a pair of moving atoms or molecules is thought of as a collision, and the effects produced by these collisions are a primary concern of this chapter. In addition, ever since Rutherford's discovery of the nucleus, collisions between atoms have provided a means of determining atomic structure and the forces of interaction. Therefore the field of atomic and molecular collisions has been sustained both by investigations into the nature of the interactions and by application of the results to help understand our atomic and molecular environment.

Collision events are correctly described via quantum mechanics (wave mechanics), but it has become customary when discussing such events to employ classical notions. In some cases this simplifies the understanding of the physics, in which case wave-mechanical effects, such as interference and diffraction, can be incorporated as corrections. Such an approach is referred to as a semiclassical method, of which there are a variety. Basically they all have the same justification: that is, they are employed when the quantum-mechanical wavelength associated with the collision is small compared to the dimensions of the system. This is the same basis for using geometric optics to approximate the passage of light through a medium. When discussing collisions, the incident "radiation" is a beam of particles and the medium is the field of a "target" atom or molecule. The wavelength is then given by $\lambda = h/p$ where h is Planck's constant and p is the momentum of the particles. Comparing λ to an atomic radius (e.g., a_0, the Bohr radius), we establish a rough criteria for the usefulness of semiclassical methods: collision energies much greater than a Rydberg (27.2 eV) for incident electrons and much greater than hundredths of an electron-volt for incident ions, atoms, or molecules. For the heavy particles, therefore, this criterion is satisfied for most energies of interest. Care must be taken, however, in making such a statement. Diffraction regions (i.e., scattering at small angles) are *always* dominated by wave-mechanical effects, as are regions in which transitions take place. In many cases such regions determine the nature of the collision process. Therefore, the above, very useful criterion must be applied cautiously.

B. Cross Section Defined

The effect of atomic particles on each other is generally described via an interaction cross-section. The conceptually simplest cross section, the total collision cross section, is obtained from an experiment like that in Fig. 1. A beam of particles is incident on a target containing atoms, and the change in intensity of the beam is monitored. If the target is "thin" so that an incident particle is only likely to make a single collision, then the change in intensity, ΔI, for a small change in thickness, Δx, is written

$$\Delta I = -\sigma n_B \Delta x \tag{1}$$

where n is the density of target atoms and I is the measured intensity (particles/cm^2/sec). In Eq. (1) the proportionality constant σ is the cross section, indicating, roughly, the range of the interaction between the colliding particles. Integrating Eq. (1), the intensity of unscattered particles versus thickness x can be found for thick samples:

$$I = I_0 e^{-n_B \sigma x} \tag{2}$$

Differentiating I in Eq. (2), the quantity $dI/I_0 = (n_B\sigma)e^{-n_B\sigma x}\,dx$ is the Poisson probability of the first collision occurring between x and $x + dx$, and $(n_B\sigma)^{-1}$ is called the mean free path between collisions.

One of the first results found from such measurements is that the total cross section varies slowly with velocity

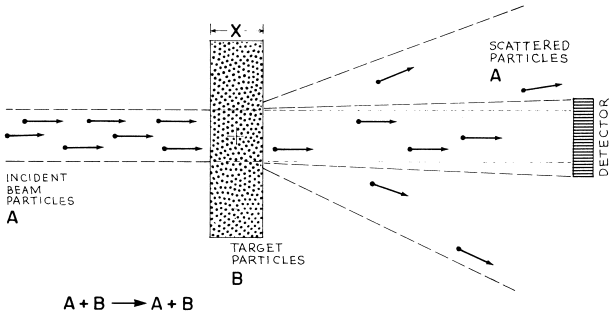


FIGURE 1 Beam experiment to obtain the scattering cross section of A by particles B. [From Johnson, R. E. (1982). *Introduction to Atomic and Molecular Collisions*. Plenum Press, New York.]

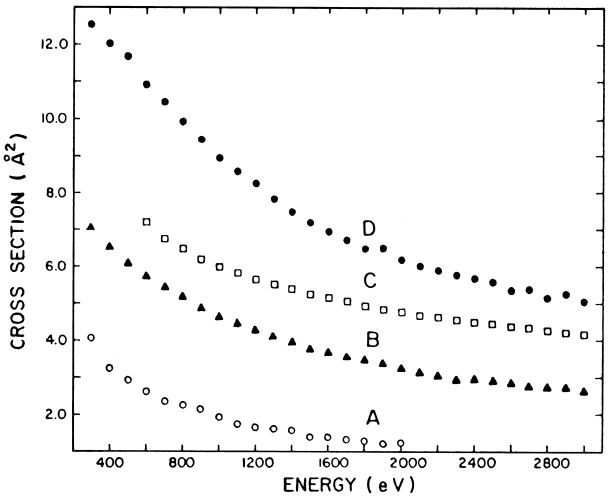


FIGURE 2 He + He. Detector apertures: A 0.57°, B 0.26°, C 0.11°, D 0.056°. [From W. J. Savola, Jr., F. J. Eriksson, and E. Pollack (1973). *Phys. Rev.* **A7**, 932.]

as shown in Fig. 2. That is, atoms have diffuse boundaries and the effective range of the interaction region changes with velocity. This dependence is wave mechanical in nature, as the cross section is determined by the amount of scattering at small angles, which is the diffraction region mentioned above. Total cross sections that are calculated strictly classically from *realistic* potentials are always infinite. Hence, we see in this simplest of examples the need for cross-section approximations that incorporate wave-mechanical effects.

A second set of experiments for studying atomic interactions requires detection of those particles that are scattered rather than those not scattered. For a beam of atoms A incident on a target containing atoms B, the angular differential cross section

$$d\sigma = (d\sigma/d\Omega)_{\mathrm{A}}\, d\Omega$$

relates the number of incident particles per unit time scattered into a region of solid angle $d\Omega$ to the incident flux. Similarly,

$$d\sigma = (d\sigma/d\Omega)_{\mathrm{B}}\, d\Omega$$

relates the number of target particles per unit time ejected into a region of solid angle to the flux of incident particles. Of course, collecting all particles A scattered into all solid angles is equivalent to detecting all scattered particles. Hence, integrating $d\sigma$ over all angles yields the total scattering cross section σ described above. In the experiment described, one might also discriminate between any internal changes that have occurred (e.g., changes in mass, charge, energy, etc.) The cross sections determined from such an experiment, therefore, would indicate both the nature of the interaction between A and B and the likelihood of an internal change (transition) in either A or B. These two aspects of the cross section will be brought out more clearly in the subsequent discussion. When no internal changes occur the collisions are called elastic; when such changes occur we refer to them as inelastic.

In this chapter we first elaborate on the nature of the cross section using a semiclassical description referred to as the impact parameter method. We then discuss the forces between atomic particles and subsequently use those forces to calculate cross sections and transition probabilities. Such calculations are divided into those methods useful for transitions in fast collisions and those useful in slow collisions. For incident ions and atoms, collisions are fast or slow depending on whether the ratio τ_c/τ_0 is less than or greater than 1. Here τ_c is the collision time ($\sim d/v$, where d is a characteristic dimension associated with A and B, e.g., $d \approx a_0$ for outer shell electrons, and v is the relative speed) and τ_0 is the characteristic period of the system of particles. For ionization of outer shell electrons, the characteristic period is $\tau_0 \approx 10^{-16}$–10^{-17} sec, whereas for molecular processes, $\tau_0 \approx 10^{-14}$ sec for vibrational motion and $\tau_0 \approx 10^{-13}$ for rotational motion. Setting $\tau_c \approx \tau_0$, such characteristic times translate into incident-particle energies of the order 1–100 keV/amu, 0.1 eV/amu, and 0.001 eV/amu, where amu is the atomic mass unit. When $\tau_c/\tau_0 \approx 1$, the collision time is about the same size as the characteristic period so that the transition probabilities and cross sections are large. At much larger or much smaller collision times, the cross sections decrease.

As the interaction between even the simplest atoms and molecules can be quite complex, many of the details of the following discussion are treated qualitatively. The purpose of the presentation following is to let the reader have a "feel" for the complexities and yet acquire the ability to understand the nature of the approximate expressions and formulas used by many workers in the field of atomic and molecular collisions. The discussion starts using the classical impact parameter concept to formulate cross sections of various types so the reader has a clear idea of the definitions of the quantities calculated and used later.

II. IMPACT PARAMETER CROSS SECTIONS

A. Formulation

The trajectory of an incident particle A interacting repulsively with an initially *stationary* target particle B is shown in Fig. 3. The quantity b in that figure indicates the closeness of approach and is referred to as the impact parameter. It is the perpendicular distance between the incident velocity vector v and the position R of particle A measured

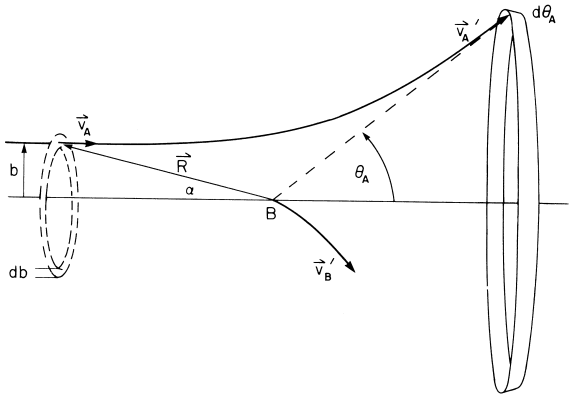


FIGURE 3 Scattering of A by B: b is impact parameter, **R** shows the position of A with respect to B, $\mathbf{v_A}$ is the initial velocity, $\mathbf{v'_A}$ and $\mathbf{v'_B}$ are the final velocities, and θ_A is the scattering angle for A: $\mathbf{v_A} = \mathbf{v}$ here.

from B. The impact parameter also indicates the angular momentum of the colliding particles, (i.e., $|L| = |M_A \mathbf{R} \times \mathbf{v}| = M_A v b$). The impact parameter (or angular momentum) and the relative velocity v (or energy) are sufficient to characterize collisions between spherically symmetric particles. As the impact parameter between two colliding particles determines the likelihood of a deflection, we assign a collision probability $P_c(b, v)$ for each impact parameter. If the incident and/or target particle has a spin or is a molecule, then initial orientations with respect to the collision axis have to be specified, for example, $P_c(b, \phi, \omega, v)$ where ϕ is the azimuthal angle of approach and ω is an orientation angle. However, if there are no preferential aligning fields (e.g., outside fields or target B in a crystalline lattice), then the incident and target particles are presumed to be randomly oriented and $P_c(b, v)$ is an average over collisions involving all possible orientations.

Based on the experimental definition in Eq. (1), the interaction cross section between A and B can now be written

$$\sigma(v) = 2\pi \int_0^\infty P_c(b, v) b \, db \tag{3}$$

That is, the particles passing through the ring of area $2\pi b \, db$ about a single target atom, (e.g., Fig. 3) will be scattered from the beam with a probability P_c and, hence, contribute to the observed change in intensity of the beam, ΔI in Eq. (1). In classical mechanics, processes are deterministic and hence, P_c is either 0 or 1. For a finite-range interaction (e.g., collision between two spheres of radius r_A and r_B), $P_c = 0$ for $b > r_A + r_B$ and the cross section is finite, $\sigma = \pi(r_A + r_B)^2$. For interactions between atomic or molecular particles the forces are infinite in range (i.e., $P_c = 1$ for all b) yielding a classical cross section which is infinite. Quantum mechanics, on the other hand, gives finite cross sections for most realistic interactions, like the measured values in Fig. 2. That is, $P_c(b, v) \xrightarrow{v \to \infty} 0$ in a quantum-mechanical calculation for interactions that decay rapidly at large separation. This deficiency in the classical estimate of the cross section is not of practical importance in many applications. That is, one generally wants to know whether a particle experiences a deflection or loses an amount of energy larger than some prescribed minimum size and *not* the likelihood of being deflected even at infinitesimally small angles (i.e., extremely large b).

From Fig. 3 it is seen that if the forces between the particles are independent of their orientations (i.e., no azimuthal dependence), then those particles passing through the ring of area $2\pi b \, db$ will be scattered into the angular region described by $2\pi \sin\theta_A \, d\theta_A$. Therefore

$$d\sigma = (d\sigma/d\Omega)_A 2\pi \sin\theta_A \, d\theta_A = 2\pi b \, db$$

so that the differential cross section discussed earlier is

$$\left(\frac{d\sigma}{d\Omega}\right)_A \equiv \sigma(\theta_A) = \left|\frac{b \, db}{\sin\theta_A \, d\theta_A}\right| \tag{4a}$$

where the simplified notation for azimuthally symmetric cross section, $\sigma(\theta_A)$, is often used instead of $(d\sigma/d\Omega)_A$. Similarly, the cross section for scattering of target particles is

$$\left(\frac{d\sigma}{d\Omega}\right)_B \equiv \sigma(\theta_B) = \left|\frac{b \, db}{\sin\theta_B \, d\theta_B}\right| \tag{4b}$$

For an elastic binary collision of the type we have been discussing, the energy and momentum of each particle after the collision can be related to the incident energy and the scattering angle using the conservation of energy and momentum. The scattering angle can be expressed either in the laboratory or in the center of mass (CM) system. In a practical problem the former is more useful, but in describing the relationship between the atomic interactions and the resulting deflections the latter is much more convenient. In Table I the relationships between laboratory and CM quantities are summarized for the case we have been discussing, a moving particle A incident on a stationary particle B. We write the CM deflection angle versus b as $\chi(b)$ (the deflection function) and, therefore, by analogy with the above discussion, the classical differential cross section in the CM system is

$$\sigma(\chi) = \left|\frac{b \, db}{\sin\chi \, d\chi}\right| \tag{5}$$

This can be transformed (see Table I) to give the laboratory scattering cross sections in Eqs. (4) for the incident or target particle.

TABLE I Relationship between Laboratory and CM Variables

Velocity of CM	$\mathbf{V_c} = (M_A \mathbf{v_A} + M_B \mathbf{v_B})/(M_A + M_B) \equiv \dot{\mathbf{R}}_c$
Relative velocity in CM	$\mathbf{v} = (\mathbf{v_A} - \mathbf{v_B}) \equiv \dot{\mathbf{R}}$

Total laboratory quantities	**CM quantities**
$M = M_A + M_B$	$m = M_A M_B/(M_A + M_B)$
$E_A + E_B = \frac{1}{2} M V_c^2 + E$	$E = \frac{1}{2} m v^2$
$\mathcal{P} = M\mathbf{V}_c$	$\mathbf{P} = 0$
$\mathcal{L} = M\mathbf{R}_c \times \mathbf{V}_c + \mathbf{L}$	$\mathbf{L} = m\mathbf{R} \times \mathbf{v}$

Transformations (for v_B initially zero and elastic collisions):

$$\theta_B = \tfrac{1}{2}(\pi - \chi)$$

$$\tan\theta_A = \mu \sin\chi/(1 + \mu\cos\chi); \qquad \mu = M_B/M_A$$

$$T = \gamma E_A \sin^2(\chi/2); \qquad \gamma = 4M_B M_A/(M_A + M_B)^2$$

$$(d\sigma/d\Omega)_A = \sigma(\chi)\left|\frac{d\cos\chi}{d\cos\theta_A}\right| = \sigma(\chi)\frac{(\mu^2 + 2\mu\cos\chi + 1)^{3/2}}{\mu^2|\mu + \cos\chi|}$$

$$(d\sigma/d\Omega)_B = \sigma(\chi)\left|\frac{d\cos\chi}{d\cos\theta_B}\right| = \sigma(\chi)|4\sin(\chi/2)|$$

B. Energy Loss Cross Sections

For elastic collisions the energy transfer to B is simply related to χ when B is initially stopped (see Table I) and, therefore, it is often useful to consider the elastic energy transfer cross section instead of $\sigma(\chi)$. Calling T the energy transfer to particle B, one writes

$$\frac{d\sigma}{dT} = \frac{4\pi}{\gamma E_A}\sigma(\chi) \tag{6}$$

where γ is given in Table I and γE_A is the maximum energy transfer. The reader will find both $\sigma(\chi)$ and $d\sigma/dT$ used in the literature. In addition to transferring kinetic energy T to particle B, there is also a probability that one or both particles will experience a change in internal energy (a transition), which we label $P_{0\to f}(b, v)$. The subscripts indicate the initial (0) and final (f) states of the colliding particles, and those collisions for which $P_{0\to f} \neq 0$ are inelastic. As in Eq. (3), an *inelastic* cross section can be written

$$\sigma_{0\to f}(v) = 2\pi \int_0^\infty P_{0\to f}(b, v) b\, db \tag{7}$$

C. Reaction Cross Sections

The classical expression for the *reaction* cross section is

$$\sigma_{r\to p}(v) = 2\pi \int_0^\infty P_{r\to p}(b, v) b\, db, \tag{8}$$

where $P_{r\to p}(b, v)$ is the reaction probability for the reactants in their initial quantum state r to collide and form products in their quantum state p. A relatively simple and well-studied example of such a process in molecular scattering is the reaction $F + H_2 \to HF + H$ in which the product HF is formed vibrationally excited.

The reaction cross section is related to reaction rate coefficient (also referred to as the rate constant) for a bimolecular reaction that occurs in the gas phase at a temperature T through the expression

$$k(T) = \langle v\sigma_{r\to p}(v)\rangle, \tag{9}$$

where the brackets imply a thermal average over the distribution of relative speeds, v, and internal quantum states of the reactant and a sum over the internal quantum state of the products. For the simplest model of a chemical reaction with a barrier, the "hard-sphere-line-of-centers" model, the translational energy dependence of the reaction cross section (summed over final states) is given by

$$\sigma_r(E_t) = \begin{cases} 0 & E < E_0 \\ \pi d^2(1 - E_0/E_t) & E \geq E_0 \end{cases} \tag{10}$$

where E_0 is the *threshold energy* and d is the range of chemical interaction, the corresponding rate constant is given by

$$k(T) = \langle v\rangle \pi d^2 \exp(-E_0/k_B T), \tag{11}$$

where $\langle v\rangle$ is the average relative speed and k_B is the Boltzmann constant.

III. ELASTIC SCATTERING

A. Classical Deflection Function

The determination of $\chi(b)$, the CM deflection function, begins with angular momentum conservation. In the CM system the two colliding particles follow trajectories that are equivalent and the collision can be described as the scattering of a particle of reduced mass m by a stationary center of force. Using the coordinates given in Fig. 3, we write the angular momentum as

$$L = mvb = mR^2\dot{\alpha} \tag{12}$$

Rearranging Eq. (12) and integrating over time, we obtain

$$\chi(b) = \pi - \int_{-\infty}^{\infty} \dot{\alpha}\, dt = \pi - vb \int_{-\infty}^{\infty} \frac{dt}{R^2}$$

Finally, assuming an interaction potential $V(R)$ that depends only on the separation R, energy conservation is expressed as

$$\frac{1}{2}m\dot{R}^2 + \frac{L^2}{2mR^2} + V(R) = E$$

Employing Eq. (12), this expression can be written as a radial velocity,

$$\dot{R} = \pm v\left[1 - \frac{b^2}{R^2} - \frac{V}{E}\right]^{1/2} \tag{13}$$

At the start of the collision the radial velocity is $(-v)$, that is, the atoms approach each other at a speed v. As R decreases, $\dot{R}$ approaches zero. R then begins to increase, as the particles recede, until it becomes $(+v)$ at large separations. Using Eq. (13) the deflection function above can be written

$$\chi(b) = \pi - 2b \int_{R_0}^{\infty} \frac{dR}{R^2}\left[1 - \frac{b^2}{R^2} - \frac{V}{E}\right]^{-1/2} \tag{14}$$

where R_0 is the distance of closest approach at which $\dot{R} = 0$. This expression can be integrated analytically for a few potentials of the form $V(R) = C_n/R^n$ but otherwise is treated by a simple numerical procedure given in Table II.

B. Impulse Approximation

For fast incident ions [ratio V/E small for all R in Eq. (14) the deflections are small and it is useful to replace the above expressions for $\chi(b)$ by an impulse approximation. That is,

$$\chi(b) = \frac{(\Delta p)_\perp}{p_0} \approx \frac{\int_{-\infty}^{\infty} F_\perp\, dt}{p_0} = -\frac{d}{db}\left[\frac{1}{2E}\int_{-\infty}^{\infty} V\, dZ\right] \tag{15}$$

TABLE II Elastic Collision Expressions

$$\chi(b) = \pi - 2b\int_{R_0}^{\infty}\frac{dR}{R^2}\left(1 - \frac{b^2}{R^2} - \frac{V}{E}\right)^{-1/2}$$
$$\approx -\left(\frac{1}{2E}\right)\frac{d}{db}\int_{-\infty}^{\infty} V(R)\, dZ$$

$$\eta^{sc}(b) = \frac{p_0}{\hbar}\left[\int_{R_0}^{\infty}\left(1 - \frac{b^2}{R^2} - \frac{V}{E}\right)^{1/2} dR - \int_{b}^{\infty}\left(1 - \frac{b^2}{R^2}\right)^{1/2} dR\right] \approx \frac{-1}{2\hbar v}\int_{-\infty}^{\infty} V(R)\, dZ$$

$$L\chi(b) = (2\hbar b)\frac{\partial\eta^{sc}(b)}{\partial b}, \qquad L = p_0 b = mvb$$

$$\sigma(\chi) = \frac{\gamma E_A}{4\pi}\frac{d\sigma}{dT}\ (B \text{ initially stopped}), \qquad S_n = \frac{\gamma E_A}{2}\sigma_d$$

$$\rho = \chi \sin\chi\sigma(\chi), \qquad \tau = \chi E,\ E = \frac{M_B}{M_A + M_B}E_A$$

Quadratures (m = number of integration points)

$$g(x) \equiv \frac{b}{R_0}\left[\frac{1 - x^2}{1 - (b/R_0)^2x^2 - \dfrac{V(R_0/x)}{E}}\right]^{1/2}$$

$$\chi(b) \approx \pi\left[1 - \frac{1}{m}\sum_{i=1}^{m} g(x_i)\right],$$

$$x_i = \cos[(2i-1)\pi/4m]$$

$$\eta^{sc}(b) \approx \frac{2L}{\hbar}\left\{\frac{\pi}{(2m+1)}\sum_{j=1}^{m}\sin^2\left(\frac{j\pi}{2m+1}\right)\left[g(x_i)^{-1} - 1\right]\right\},$$

$$x_j = \cos\left[\frac{j\pi}{2m+1}\right]$$

Impluse estimates ($V/E \ll 1$)

for $V = C_n/R^n$, $\qquad \chi(b) \approx a_n V(b)/E$,

$$a_n = \pi^{1/2}\Gamma\left(\frac{n+1}{2}\right)\Big/\Gamma\left(\frac{n}{2}\right) \xrightarrow{n\to\infty} (\pi n/2)^{1/2}$$

$$\eta^{sc}(b) \approx \frac{-L\chi(b)}{2\hbar(n-1)}$$

[$\Gamma(x+1) = x\Gamma(x)$, $\Gamma(1) = 1$, $\Gamma\left(\frac{1}{2}\right) = \pi^{1/2}$; gamma function tabulated]

$$\rho \approx n^{-1}(a_n C_n/\tau)^{2/n}$$

for $V = \left(Z_A Z_B e^2\right)\exp(-\beta R)/R$

$$\chi(b) \approx \frac{\left(Z_A Z_B e^2\right)\beta}{E}K_1(\beta b) \xrightarrow{b\to\infty} \left(\frac{\pi\beta b}{2}\right)^{1/2}\frac{V(b)}{E}$$

$$\eta^{sc}(b) \approx -\frac{\left(Z_A Z_B e^2\right)}{\hbar v}K_0(\beta b) \xrightarrow{b\to\infty} \frac{-L\chi(b)}{2\hbar(\beta b)}$$

(K_n are the modified Bessel functions; tabulated)

Massey–Mohr [Eq. (30)]

$$\sigma \approx 2\pi(\bar{b})^2\left(1 + \frac{1}{2n-4}\right) \qquad (n > 2)$$

$$\bar{b} \approx \left[\frac{2a_n C_n}{(n-1)\hbar v}\right]^{1/(n-1)}$$

Continues

Continued

Born approximation

for $V = \left(Z_A Z_B e^2\right) \exp(-\beta R)/R$

$$\frac{d\sigma}{dT} \approx \pi \mathscr{A}^2 \Big/ (\gamma E_A)(T + T_0)^2$$

$$\mathscr{A} = \left(\frac{2M_A}{M_A - M_B} Z_A Z_B e^2\right)$$

$$T_0 = \frac{(\hbar\beta)^2}{2M_B}$$

$$\sigma \approx \pi \mathscr{A}^2 \Big/ T_0(\gamma E_A + T_0)$$

$$S_n \approx \left[\ln\left(\frac{\gamma E_A}{T_0} + 1\right) - \frac{\gamma E_A}{(\gamma E_A + T_0)}\right]$$

In this equation we assumed a straight line trajectory, i.e., $R^2 \approx b^2 + Z^2$, with $Z = vt$. For the general class of power-law potentials ($V = C_n/R^n$), χ (b) calculated from Eq. (15) is given in Table II. It is seen, that the deflection is determined primarily by the nature of the potential *near the distance of closest approach* of the colliding particles ($R_o \approx b$). Therefore, it is *not* necessary to know the potential accurately at all R to obtain a reasonable estimate of the deflection function for collisions between ions and atoms as long as $V/E \ll 1$. Since V is of the order of electron-volts, this criterion is satisfied for a large number of problems involving incident ions, atoms, or molecules.

Before discussing the determination of interaction potentials for specific collision pairs, we examine the nature of differential cross section for two forms of the potential, a purely repulsive and a long-range attractive plus short-range repulsive potential. The latter potential is characteristic of the interaction of an ion with an atom or molecule in its ground state. In Fig. 4 are shown the deflection functions and ρ versus τ plots. Based on Table II it is clear that the deflection function should approximately follow the form of the potential. Therefore, for the repulsive potential shown, χ is always positive, attaining a maximum value of π for head-on collisions [$b = 0$, Eq. (14)]. For the other potential, χ is negative at large b, eventually becomes positive, and again reaches π at $b = 0$. However, χ goes through a minimum, χ_r, at the impact parameter labeled b_r. At this minimum, $d\chi/db = 0$, and, therefore, the classical calculation of cross section in Eq. (5) becomes *infinite*. This large *enhancement* in the scattering probability is similar to the effect that produces rainbows in the scattering of light from water droplets; hence, χ_r is called the rainbow angle. When $\chi_r < \pi$, then for angles greater than χ_r, only one impact parameter contributes to the cross section. However, for angles less than χ_r, three impact parameters contribute that have the same value of cos χ, as seen in Fig. 4.

Using a Lennard–Jones form for the longrange attractive plus short-range repulsive potential, $V = C_{2n}/R^{2n} - C_n/R^n$, and for the powerlaw results in Table II the rainbow angle (for $V \ll E$) is

$$\chi_r \approx \frac{a_n^2}{a_{2n}} \left(\frac{V_{\min}}{E}\right) \tag{16}$$

That is, the rainbow angle is determined by the depth of the potential minimum $V_{\min}$ (which is $V_{\min} = -C_n^2/4C_{2n}$ for the Lennard–Jones potential), and the shape of the potential via a_n and a_{2n}. Because the ratio a_n^2/a_{2n} changes slowly with n, measuring the rainbow angle can directly give an estimate of potential well depth.

As the collision energy decreases, χ_r can become much larger than π, implying that the two particles may orbit each other before separating. In fact, for very small velocities there is a particular impact parameter for which the particles can be trapped in orbit (i.e., $\ddot{R} = 0$ at $\dot{R} = 0$). Using only the attractive part of potential, along with Eq. (13) and the power-law results in Table II, the orbiting impact parameter is

$$b_0 \approx \left[\left(\frac{n}{2}\right)\frac{C_n}{E}\right]^{1/n} \Big/ \left(\frac{n-2}{n}\right)^{(n-2)/2n} \tag{17}$$

Therefore, for $b \leq b_0$, the interaction times can be very long and the particles approach each other closely, whereas for $b > b_0$ the particles simply scatter. This fact has been used extensively in estimating ion–molecule interaction cross sections.

Near the rainbow angle (or when orbiting occurs), a number of impact parameters contribute to the scattered flux at a given observation angle; hence, interference

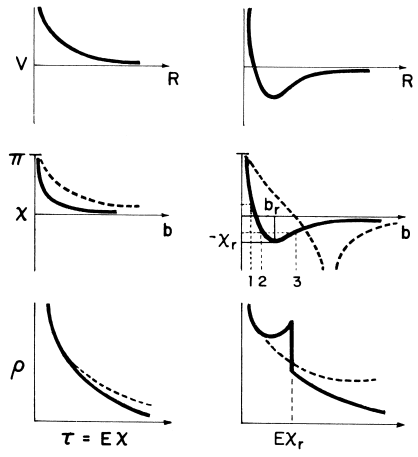


FIGURE 4 Values of χ versus b and ρ versus τ for a repulsive and an attractive potential, Rainbow angle and impact parameter are χ_r and b_r. Three impact parameter's contribution for $|\chi| < |\chi_r|$ are labeled. Solid line higher energy, dashed line low energy. [From R. E. Johnson (1982). "Introduction to Atomic and Molecular Collisions," p. 53. Plenum Press, New York.]

phenomena occur and classical cross-section estimates are not valid. However, if the angular resolution is not large the simple addition of the contribution to the scattered flux from each impact parameter gives an adequate representation of the cross section. This was used to give ρ versus τ in Fig. (4). That is,

$$\sigma_{(\chi)} \approx \sum_i \left| \frac{b\,db}{\sin\chi\; d\chi} \right|_{b=b_i} \tag{18}$$

where the b_i are all those impact parameters giving the same value of $\cos\chi$ [i.e., $\chi \to \pm(\chi + 2\pi m)$, m an integer]. As the interference phenomenon can give important additional information about the *details* of the interaction potentials, wave-mechanical calculations are useful. In addition, the effects at very small angles and at low energies can only be described via wave mechanics. In the following we briefly review that wave-mechanical description of the elastic scattering cross section that parallels the above discussion. We thereby obtain results for diffraction (e.g., total cross sections σ) and interference (e.g., rainbow) phenomena. The results have parallels in light scattering.

IV. WAVE MECHANICS OF SCATTERING

A. The Scattering Amplitude

In wave mechanics the scattering of the particle is replaced by scattering of a wave, but here again we can describe this scattering in the CM system, as the transformation between the laboratory and CM quantities is the same as in Table I. The beam of particles incident on a target is replaced by a plane wave, $\exp(i\mathbf{K}\cdot\mathbf{R} - i\omega t)$, where $\hbar\mathbf{K}$ replaces the momentum and $\hbar\omega = \hbar^2K^2/2m$ is the energy per particle in the beam. K^{-1} is often written as $\lambda\!\!\!^{-}$ where $\lambda\!\!\!^{-} = \lambda/2\pi$, with λ the wavelength. Describing the beam of particles as a plane wave implies that the particles are nonlocalized, that is, it is equally probable to find a particle at any point in the beam. When this wave is scattered, as in light scattering, the outgoing wave at very large distances from the scattering center has the form

$$\left[\exp(i\mathbf{K}\cdot\mathbf{R}) + f(\chi)\frac{e^{iKR}}{R}\right]e^{-i\omega t} \tag{19}$$

where we have assumed, as in the above discussions, spherical scattering centers (i.e., no azimuthal dependence). The magnitude of the scattered wave in an angular region about χ is given by $|f(\chi)|^2/R^2$ in Eq. (19), where $f(\chi)$ is referred to as the scattering amplitude and R is the distance from the scattering center. The plane wave in Eq. (19) is the unscattered portion of the wave, which is assigned unit amplitude. This implies that the scattering potential is only a small disturbance. The differential cross section (probability of scattering into a unit solid angle about χ) is simply

$$\sigma(\chi) = |f(\chi)|^2 \tag{20}$$

Therefore, the scattering problem reduces to solving the Schroedinger wave equation subject to the boundary condition at large R, which is expressed by the form of the wave function in Eq. (19).

To draw analogies with the classical calculation of cross section, it is customary to express the plane wave in terms of multipole moments,

$$\exp(i\mathbf{K}\cdot\mathbf{R}) = \sum_{l=0}^{\infty} i^l(2l+1)P_l(\cos\chi)j_l(KR)$$

The $j_l(KR)$ are the spherical Bessel functions, which have the asymptotic form, as $R \to \infty$

$$j_l(KR) \to \frac{\sin(KR - l\pi/2)}{KR}$$

The P_l are the Legendre polynomials, where l labels the various moments ($l=0$, spherical; $l=1$, dipole; $l=2$, quadrupole; etc.). In the present problem, however, l also is the angular momentum index [$L^2 = l(l+1)\hbar^2$]. Therefore, l replaces the impact parameter b, which we used in the classical description

$$b \to \sqrt{l(l+1)}\,\hbar/mv \approx \left(l+\tfrac{1}{2}\right)/K = \left(l+\tfrac{1}{2}\right)\lambda\!\!\!^{-} \tag{21}$$

Upon intersecting a scattering center, each spherical wave j_l experiences a phase shift (η_l), becoming $\sin(KR - l\pi/2 + \eta_l)/KR$ in the asymptotic region. Therefore, one can use these expressions to write

$$f(\chi) = \frac{1}{2Ki}\sum_{l=0}^{\infty}(2l+1)\left[\exp(2i\eta_l) - 1\right]P_l(\cos\chi) \tag{22}$$

Substituting Eq. (22) into Eq. (20) indicates that the *determination of the scattering cross section* reduces to the *determination of the phase shifts*. Before calculating η_l we use the form for $f(\chi)$ in Eq. (22).

The total cross section, obtained by integrating over the angle in Eq. (20), becomes

$$\begin{aligned}\sigma &= 2\pi\int_{-1}^{1}|f(\chi)|^2\,d(\cos\chi)\\ &= \frac{8\pi}{K^2}\sum_{l=0}^{\infty}\left(l+\tfrac{1}{2}\right)\sin^2\eta_l\end{aligned} \tag{23}$$

If the phase shifts are zero, no scattering has occurred and the cross section is zero. Further, for large-momentum collisions (small wavelengths) the sum in l above can be approximated by an integral in b using Eq. (21). The expression for σ can then be written in the same form as the classical cross section in Eq. (3) if we define $P_c(b) \approx 4\sin^2\eta(b)$.

(Here the dependence of the phase shift η on l is written as a dependence on b.) This expression for P_c clearly demonstrates that at large impact parameters (or l) when $\eta(b)$ goes to zero the wave-mechanical scattering probability $P_c(b)$ goes to zero.

Comparing Eqs. (22) and (23), it is also seen that

$$\sigma = \frac{4\pi}{K}\mathrm{Im}\, f(0) \tag{24}$$

This result is referred to as the optical theorem, based on analogy with light scattering, and it shows directly what we stated much earlier. The cross section σ is determined by scattering at zero degrees, which for wave scattering is a diffraction region. Note also that P_c above has an average value of 2 at small b and not unity as it is classically. These facts both emphasize that the total scattering cross section σ is a nonclassical quantity. Note that, although the sum in Eq. (22) can also be replaced by an integral over b at small wavelengths, the expression for differential cross section so obtained is quite *different* from the classical expression in Eq. (18). That is, in wave mechanics the *scattering amplitudes* for each impact parameter (or l) that contribute to scattering at angle χ are added, whereas classically the *cross sections* that contribute at angle χ are added [e.g., Eq. (18)]. Hence, the wave mechanical cross section can exhibit an *oscillatory* behavior.

The behavior of the scattering amplitude and cross section for very low energy collisions is of current interest in the context of Bose-Einstein condensation. At very low collision energies only the $\ell = 0$ partial wave (the so-called *S-wave*) contributes to f and to σ. Further it can be shown that as K goes to zero

$$\sin\eta_0 \to \eta_0 = -Ka, \tag{25}$$

where a is called the scattering length. Thus, the cross section at very low collision energies is given by

$$\sigma = 4\pi a^2. \tag{26}$$

This is four times the cross sectional area of sphere of radius a and is in fact identical to the low energy limit of the quantum cross section of the scattering by a hard sphere of radius a. The sign of a also has significance; if a is negative the interaction at very low collision energies is attractive, if a is positive the interaction is repulsive.

Another important aspect of low energy collisions is the possibility of forming a scattering *resonance*. This is a phenomenon in which the scattering partners form a "sticky" collision complex. The signature of an ideal, narrow resonance in atom-atom scattering is a jump in the phase shift by π as a function of the collision energy. In this ideal case the resonance energy is identified as the energy where the phase shift has increased by $\pi/2$. The

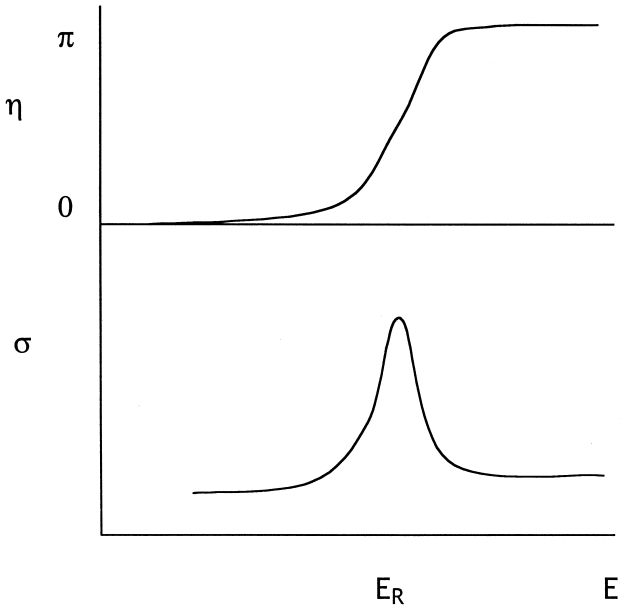


FIGURE 5 Schematic of the dependence of the phase shift and the cross section in the vicinity of the resonance energy E_R.

manifestation of this behavior of the phase shift is seen in both the integral and differential cross sections. In the former case the total cross section will exhibit a rapid change in value as a function of the collision energy in the vicinity of the resonance energy, E_R. The differential cross section at E_R may exhibit a strikingly different angular dependence than at other "off-resonance" energies. Often the differential cross section will show a near symmetric forward-backward symmetry at E_R. The situation for ideal resonances is illustrated schematically in Fig. 5.

B. Semiclassical Cross Section

The relationship between the classical and wave-mechanical cross sections becomes clearer if the sum in Eq. (22) is examined in more detail. At those impact parameters b_i producing classical scattering into the angular region χ, *constructive interference* occurs in the sum in Eq. (22). Therefore, the classical trajectories are like the light rays in geometrical optics. The angular differential cross section

$$\sigma(\chi) = |f(\chi)|^2 \approx \left|\sum_i \sigma[\chi(b_i)]^{1/2}\exp[i\alpha(b_i)]\right|^2 \tag{27}$$

replaces the classical expression in Eq. (18), where the $\sigma[\chi(b_i)]$ are the classical cross sections, $|bd\,b/\sin\chi d\chi|_{b=b_i}$, used in Eq. (18). Equation (27) directly exhibits the interference between contributions from

different impact parameters. In this expression the phase factor is $\alpha(b) = A/\hbar \pm \varepsilon$, where ε is a multiple of $\pi/4$. Here A is the difference in the classical action between an undeflected particle and a scattered particle, $A = 2\hbar\eta^{sc}(b) - L_\chi$ with

$$\hbar\eta^{sc}(b) = \int_{R_0}^{\infty} p(R)\,dR - \int_{b}^{\infty} p_0(R)\,dR$$

evaluated in Table II. The regions of constructive interference are those for which the *classical action is a minimum* ($\partial A/\partial b = 0$), so that

$$\chi = \left(\frac{2}{K}\right)\frac{\partial\eta^{sc}(b)}{\partial b} \tag{28}$$

On substituting the form for $\eta^{sc}(b)$ given above, Eq. (28) becomes equivalent to the expression for $\chi(b)$ in Eq. (14). Therefore, each region of constructive interference corresponds to a classical trajectory that contributes at angle χ. If only one impact parameter contributes in Eq. (27), the semiclassical result for $\sigma(\chi)$ is *identical* to the classical result. When more than one impact parameter contributes at a given scattering angle, $\sigma(\chi)$ is oscillatory. For rainbow scattering, which we discussed earlier, a schematic diagram of the differential cross section is shown in Fig. 6 indicating the difference between the wave-mechanical and classical behavior. That is, at large τ (close collisions) one impact parameter contributes and the semiclassical cross section follows the classical calculation. At small τ interferences occur and the semiclassical cross section oscillates about the classical result.

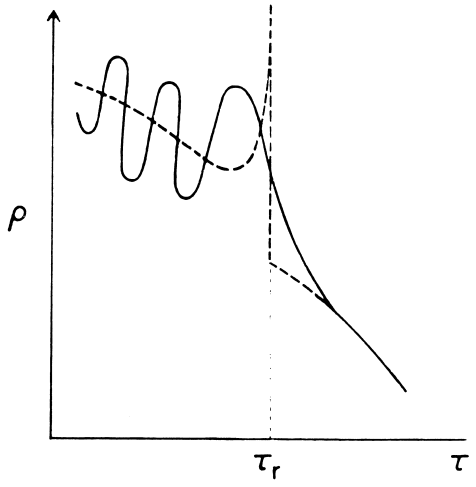


FIGURE 6 Schematic diagram of angular differential cross section shown as ρ versus τ plot. τ_r indicates rainbow value for $E\chi_r$. Dashed curved, classical cross section; oscillation occurs where more than one trajectory contributes. [From R. E. Johnson (1982). "Introduction to Atomic and Molecular Collisions," p. 88. Plenum Press, New York.]

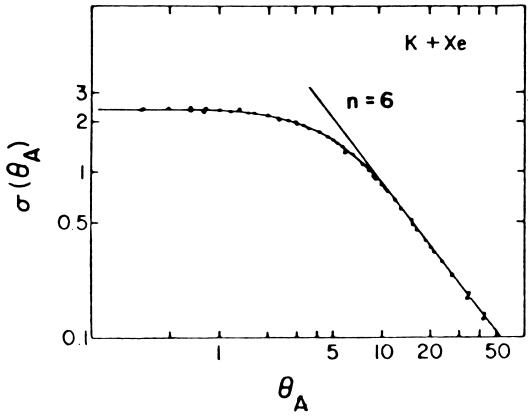


FIGURE 7 Elastic scattering cross section (arbitrary units) versus laboratory scattering angle. Line indicates classical calculation using vander Waals potential. [Data from Helbing and H. Pauly (1964). *Z. Physik* **179**, 16. Figure from R. E. Johnson (1982). "Introduction to Atomic and Molecular Collisions," p. 163. Plenum Press, New York.]

C. Other Approximations

Using the semiclassical expression for $\eta(b)$ given above and the impulse approximation, a simple estimate of $\eta^{sc}(b)$ that is applicable when (V/E) is small, is given in Table II. Using this in Eq. (22) *and* an approximation to $P_l(\cos\chi)$ valid at small angles, one obtains an estimate of $f(\chi)$ valid at small angles,

$$f(\chi) \simeq -\frac{m}{2\pi\hbar^2}\int d^3R\, V(R) e^{-i\Delta\mathrm{p}\cdot\mathrm{R}/\hbar} \tag{29}$$

This is referred to as the first Born approximation, and results for $\sigma(\chi)$ calculated using Eq. (29) are given in Table II.

Since $\eta(b) \overset{b\to\infty}{\longrightarrow} 0$ (see Table II) for potentials that decrease faster than $1/R$, the impulse approximation to $\eta(b)$ can be used to estimate the integrated cross section σ. That is, we replace $\sin^2\eta(b)$ by $\frac{1}{2}$ in Eq. (23) out to some large impact parameter $\bar{b}$ beyond which $\eta(b)$ is always small. Then $\sin^2\eta(b) \approx \eta^2(b)$ at larger b, and σ is written

$$\sigma \simeq 2\pi\overline{b^2} + 8\pi\int_{\bar{b}}^{\infty}\eta^2(b) b\,db \tag{30}$$

This is referred to as the Massey–Mohr approximation, and the result for power-law potentials is given in Table II. The expression given is finite for $n > 2$, unlike the classical result, which is never finite for a potential of infinite range. For a repulsive potential, the cross section decreases monotonically with increasing energy as shown earlier in Fig. 2. Using the optical theorem [Eq. (24)] in reverse, we see that, if σ is finite, then the differential cross section must be finite as $\chi \to 0$ (see Fig. 7), unlike the classical result. By contrast, a similar calculation of the diffusion cross section σ_d (or S_n) gives the *same form* as the

classical expression at small wavelengths. This comes about because scattering at $\chi = 0$ is excluded in the expression for σ_d. Therefore, classical expressions of the diffusion cross, and hence the nuclear stopping cross section, are accurate over a broad range of incident ion energies, allowing extensive use of classical estimates for ion penetration of gases and solids.

In the opposite extreme, when the wavelength is *large* compared to the dimensions of the scattering center ($K^{-1} \gg d$), then no classical analogies exist. However, only the lowest l values contribute in Eq. (22); hence, only one term is kept in the sum, so that

$$f(\chi) \simeq \frac{1}{K} \sin\eta_0 \exp(i\eta_0); \quad \sigma(\chi) = \frac{\sin^2\eta_0}{K^2} \tag{31}$$

(Note that if $\eta_0 \approx n\pi$ then higher terms are needed: this is referred to as the Ramsauer–Townsend effect.) The expression for the cross section in Eq. (31) represents the *same region of impact parameter* as the semiclassical result. However, at very low velocities (i.e., large χ, small K), it is seen from Eq. (21) that an increase of l from 0–1 can make b very large. When b is much greater than the size of the colliding particles, d, then all interactions are encompassed, which at low velocities may only require a single value of l. In this limit the differential cross section is *independent of angle for any potential*, a result obtained classically only for the collision of spheres. Because of its simplicity, an isotropic scattering cross section is often used in low-energy collisions, even when the largewavelength criterion does not rigorously apply. At large wavelengths, the integrated cross section in Eq. (23) is simply written as

$$\sigma \approx \frac{4\pi}{K^2} \sin^2\eta_0 \tag{32}$$

which is also equal to the diffusion cross section σ_d at low velocities.

In the above we have considered very general properties of cross sections. In wave mechanics, as in the scattering of light, the cross sections will exhibit oscillatory behavior and/or resonances for a variety of potential forms. For example, when the long-range potential is attractive, a forward glory is seen in the integrated cross section. In the following sections we will briefly review the nature of interaction potentials, but in doing so we will also consider inelastic effects such as excitations and ionizations.

V. INTERACTION POTENTIALS

A. Overview

The primary force determining the behavior of colliding atoms or molecules is the Coulomb interaction. This force acts between each of the constituent electrons and nuclei. Using Coulomb potentials the complete interaction potential for all of the particles can be immediately written down. However, each of the constituents moves relative to the center of mass of its parent molecule. As this motion is superimposed on the overall collisional motion, the description of a collision can be quite complex even for the simplest atoms. Rather than solve the wave equation for the complete, many-body system, one often introduces the concept of single-interaction potentials averaged over the relative motion of the constituent particles. In using this concept we again exploit the huge mass difference between electrons and nuclei. This mass difference allows us, with reasonable accuracy, to separate the motion of the electrons and nuclei, a procedure referred to as the Born–Oppenheimer separation.

The behavior of the electrons during a collision depends on the relative motion of the nuclei. Therefore, interaction potentials are generally calculated in two limits. If the collisions are fast relative to the internal motion of the electrons ($v \gg v_e$), then during the collision the electronic distribution is static, except for abrupt changes (transitions) that occur when the particles are at their closest approach. The transitions reflect the ability of the molecules to absorb (emit) energy when exposed to the time-varying field of the passing particle in the same way that these molecules absorb or emit photons. Before and after the transition the potentials are determined from the separated charge distributions.

In the opposite extreme (slow collisions, $v \ll \bar{v}_e$) the electrons adjust continuously and smoothly to the nuclear motion, returning to their initial state at the end of the collision. This collision process is called adiabatic as the electrons do not gain or lose energy. That is, even though the molecules may be deflected and change kinetic energy, their initial and final *electronic states* remain the same. The electronic distribution evolves from a distribution in which electrons are attached to separate centers at large R to a distribution in which the electrons are shared by the two centers at small R, a covalent distribution. Therefore, for every possible initial state there is a corresponding adiabatic potential, resulting in rather complex potential diagrams. Such potentials also determine the ability of the two particles to bind together to form a molecule.

As molecules in a collision are not moving infinitely slowly, the motion of the nuclei can induce transitions between the adiabatic states. For slow collisions these transitions occur at well-defined internuclear separations—for example, at those internuclear separations at which the atomic character of the wave function gives way to the molecular, covalent character. Because a large computational effort is required to obtain a set of potential curves, and an additional large computational effort is required to describe the collisions when transitions occur, simplifying

procedures are very attractive and approximate potentials are often constructed.

As the electrons in different shells have very different velocities, the above separation into fast and slow collisions allows us to treat the orbitals separately. For instance, when a collision is fast with respect to the outer-shell electrons, it may be adiabatic with respect to the inner-shell electrons. Therefore, the inner-shell electrons return to their initial state. Their effect is to only screen the nuclei and, hence, they play a passive role in the collision. Alternatively, when inner-shell excitations occur the outershell electrons can be considered to be static during the collision. As a point of reference, it is useful to remember that a nucleus with a speed equivalent to an electron in the ground state of a hydrogen atom has an energy of about 25 keV/amu. Further, the orbital speed of an electron in an atom can be scaled to the speed of an electron in the ground state of hydrogen using the effective nuclear charge.

The interaction potential between two atoms can be written as a sum of the nuclear repulsion and the electronic energy $\varepsilon_j(R)$,

$$V_j(R) = \frac{Z_A Z_B e^2}{R} + [\varepsilon_j(R) - \varepsilon_j] \tag{33}$$

In this expression j labels the electronic state. Z_A and Z_B are the nuclear charges, and ε_j is the total electronic energy of the colliding atoms, $\varepsilon_j(R)$, as $R \to \infty$. Each state j is associated with a pair of separated atomic states at large R. In the *electrostatic limit* ($v \gg \bar{v}_e$), the electronic energy $\varepsilon_j(R)$ is a sum of the electronic energies of the separated atoms, $(\varepsilon_{Aj} + \varepsilon_{Bj}) = \varepsilon_j$, *plus* the averaged interaction of the electrons on each atom with the electrons and nucleus of the other atom, V_j^e. The quantity V_j^e is written

$$\begin{aligned} V_j^e(R) = &-Z_B e^2 \int \frac{\rho_{A_j}(\mathbf{r}_A)}{|\mathbf{R} - \mathbf{r}_A|}\, d^3 r_A \\ &- Z_A e^2 \int \frac{\rho_{B_j}(\mathbf{r}_B)}{|\mathbf{R} - \mathbf{r}_B|}\, d^3 r_B \\ &+ e^2 \int \frac{\rho_{A_j}(\mathbf{r}_A)\, \rho_{B_j}(\mathbf{r}_B)}{|\mathbf{R} - \mathbf{r}_A + \mathbf{r}_B|}\, d^3 r_A\, d^3 r_B \end{aligned} \tag{34}$$

where the ρ_{Aj} and ρ_{Bj} are the electron densities on atoms A and B. V_j^e is the classical electrostatic interaction and is evaluated in many texts.

In the adiabatic approximation, $\varepsilon_j(R)$ in Eq. (33) is calculated from the full electronic wave equation at each R. An approximation that is particularly useful for light atoms is to estimate $\varepsilon_j(R)$ via the molecular-orbital method used by chemists. For heavy atoms, on the other hand, the Thomas–Fermi method is often employed to estimate $\varepsilon_j(R)$. In this model the electrons are treated as a gas subject to the Pauli principle. In the following, rather than calculate, $V_j(R)$ for all R, we will consider its behavior for various regions of R. Such an approach is reasonable as the deflection function is determined primarily by a narrow region of R about the distance of closest approach.

B. Short-Range Potentials

For close collisions ($R \ll \bar{r}_A, \bar{r}_B$ where $\bar{r}_A$ and $\bar{r}_B$ are the mean atomic radii), the nuclear repulsion dominates. The electrons essentially screen the nuclear repulsive interaction; hence, one often approximates $V_j(R)$ by

$$V_j(R) = \frac{Z_A Z_B e^2}{R} \Phi\left(\frac{R}{a_j}\right) \tag{35}$$

where a_j is a screening length and Φ the screening function. Considerable effort has been expended determining Φ and a_j for many-electron atoms and often Φ is written as $\exp(-R/a_j)$. At small distances of closest approach which are usually associated with fast collision, an electrostatic calculation [Eq. (34)] can be used to estimate Φ and a_j. In fact, the electrons in different shells have different screening lengths. Therefore, for a bare ion colliding with an atom, a potential of the form

$$V_j(R) \approx \frac{Z_A e^2}{R} \sum_i N_{B_i} e^{-R/a_i} \tag{36}$$

is used, where N_{B_i} is the number of electrons in the ith shell on B. A good approximation for light atoms is to assume that the ith shell has a screening constant determined by the ionization energy I_i, $a_i \approx a_0/Z_i'$ with the effective charge $Z_i' = (I_i/I_0)^{1/2}$, where I_0 is the ground-state, hydrogen atom ionization potential. For collisions between large atoms ($Z \gtrsim 10$), the Thomas–Fermi screening constant has been used, $a_{TF} = 0.8853 a_0 (Z_A^{2/3} + Z_B^{2/3})^{-1/2}$. Noting that $V_j(R)$ in Eq. (35) can be written in terms of a scaling energy ($Z_A Z_B e^2/a_j$) and scaled distances (R/a_j), a socalled "universal" potential has been constructed in the repulsive regime by Ziegler *et al.* based on scattering data. Expressions for cross section and stopping cross section have been given in Table II for the potential in Eq. (35) using the exponential screening function and "universal" scaling. Results obtained by Lindhard for the Thomas–Fermi screening function are also given. As the short-range collision region has been extensively studied, one should use experimentally determined parameters when available.

C. Long-Range Interactions

For slow collisions even small disturbances can lead to deflections; therefore the interaction potential at long range (large separations between colliding partners, $R \gg \bar{r}_A, \bar{r}_B$) is of interest. If the electronic distributions of each of the colliding particles are distorted very little by the presence

of the other, the interaction potential is written in powers of R,

$$V_j(R) = \sum C_n / R^n \qquad (37)$$

and, generally, only the largest term (lowest power of n) is kept in a calculation. Therefore, the power-law expression for cross section that we developed earlier can also be used for weak-interaction, long-range collisions. The lead terms in Eq. (37) can be obtained from the electrostatic interaction in Eq. (34) by expanding the denominators in powers of R. For ion–ion collisions the lead term in Eq. (37) is $n = 1$, $C_1 = \bar{Z}_A \bar{Z}_B e^2$, which is the coulomb interaction with $\bar{Z}_A = Z_A - N_A$ (the nuclear charge of A minus the number of electrons on A or the net charge). For ion collisions with a neutral molecule having an electric dipole moment μ_B (e.g., water molecule), the lead term is $n = 2$, $C_2 = \bar{Z}_A \mu_B e \cos\theta$ where $\cos\theta$ is the angle between the internuclear axis and the dipole moment. For neutral molecules that do not have dipole moments (e.g., O_2, N_2), the ion–quadrupole interaction dominates ($n = 3$). For two neutral molecules having dipole moments, the dipole–dipole interaction ($n = 3$) dominates, and so on.

For colliding atomic particles the electrostatic multipole moments of the charge distribution are zero and for some molecules (e.g., H_2) the multipole moments are small. However, the field of the other particle induces moments in the separated charge distribution. The lead term for ion neutrals is the ion-induced dipole interaction ($n = 4$),

$$C_4 = -\frac{\alpha_B}{2}(\bar{Z}_A e)^2 \qquad (37)$$

where α_B is the polarizability of the neutral particle B. For collisions between two neutrals there is not average field at B due to A or vice versa. However, instantaneous fluctuations in charge density on either atom produces short-lived fields that induce moments in the other. Therefore, for neutrals the lead term is the induced-dipole–induced-dipole interaction, which is the well-known van der Waals potential used to describe the behavior of realistic gases. This interaction ($n = 6$) is always attractive (i.e., C_6 negative) and the coefficients have been extensively evaluated from collision experiments.

D. Intermediate Range Potentials and Charge Exchange

When R is the order of the atomic radii ($\bar{r}_A$, $\bar{r}_B$), the electron clouds on the two interacting particles overlap. In this region the distortion of the charge clouds on each center becomes too large to treat as simply a perturbative polarization of the separated electronic distributions. This is an especially important region for determining transitions between colliding particles but is also important in molecular structure determinations. The emphasis here is on those aspects important in collisions.

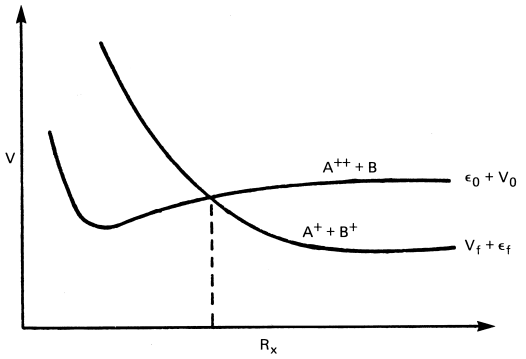


FIGURE 8 Interaction potentials versus internuclear separations; curve crossing for the $A^{2+} + B \rightarrow A^+ + B^+$ collisions.

The overlap of charge on the two centers allows for the possibility of charge-exchange collisions (e.g., $H^+ + O_2 \rightarrow H + O_2^+$). When charge exchange occurs the potentials before and after the collisions can be drastically different. For example, for $O^{2+} + S \rightarrow O^+ + S^+$, a process of interest in the Jovian magnetosphere plasma, the initial long-range interaction is attractive and is determined by the polarizability of S. On the other hand, the final interaction is clearly repulsive. Depending on the final states of O^+ and S^+, the initial and final potential curves may cross as shown in Fig. 8. Therefore, electron exchange between O^{2+} and S can occur with *no net change* in the total electronic energy *at* the crossing point. Of course, the change in state gradually results in a net change in the *final* electronic energy as the particles separate. Therefore, such crossings indicate transition regions and the likelihood of charge exchange is determined by the overlap of the charge distributions at the crossing point.

For this region of intermediate R the covalent (electron exchange) interaction can be examined by considering the $H^+ + H$ (i.e., H_2^+) system. In such a system, the exchange $H^+ + H \rightarrow H + H^+$ occurs *without* a net change in internal energy, unlike the case discussed above. The two identical states at large R ($H + H^+$ and $H^+ + H$) split at smaller R, due to electron sharing, forming both attractive and repulsive potentials, as shown in Fig. 9. If the wave functions placing the electron on centers A and B in identical states of the hydrogen atom are ϕ_A and ϕ_B, then the linear combination appropriate to the H_2^+ molecule are $\psi_{g,u} \cong (1/\sqrt{2})(\phi_A \pm \phi_B)$. The labels g and u refer to symmetric and antisymmetric (gerade and ungerade) states of H_2^+. In the scattering of protons by hydrogen, as the initial state is *either* ϕ_A *or* ϕ_B, half the collisions will be along repulsive potentials (u state) and half along an attractive potential (g state). The difference in energy between these states is referred to as the exchange energy

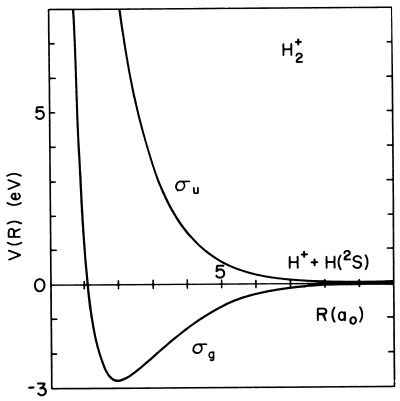


FIGURE 9 Interaction potential energy curves versus internuclear separation for ground state H_2^+: nuclear repulsive plus electronic (ε_g or ε_u).

and it determines the behavior of charge-exchange collisions. This exchange energy decays as the overlap of the wave functions on A and B ($\langle\phi_A/\phi_B\rangle \alpha e^{-R/\bar{a}}$), which is an exponential function of R. The exchange interaction is eventually dominated at very large R by the long-range, power-law potentials in Eq. (37), discussed earlier.

Molecular orbital potentials for H_2^+ can also be constructed for each excited state of H, or, more usefully, they can be constructed for one electron in the field of two identical nuclei *and* the other electrons. In Fig. 10 we give a diagram showing the general behavior of these one-electron, molecular orbital binding energies $\varepsilon_j(R)$. As this quantity does not include the nuclear repulsive term in Eq. (33), we can follow the behavior of these states down to $R = 0$. This diagram correlates the one-electron states of the separated centers ($R \to \infty$) with those of the united atom ($R \to 0$), indicating how the electronic binding energy changes with R. The states are labeled by their symmetry under inversion (g or u) and by the component of electronic angular momentum along the internuclear axis ($|m_l| = 0, 1, 2, \ldots$ as $\sigma, \pi, \delta, \ldots$). The correlation between states at large and small R is determined by the fact that the potential energy curves associated with states of the same symmetry [rotation $|m_l|$ and inversion (g, u)] do not cross (i.e., do not become degenerate in energy). Two sets of notation are used for these states in order to indicate which are the corresponding atomic states at either $R \to 0$ or $R \to \infty$. For example, the two lowest states we considered above for H_2^+ are labeled σ_g 1s and σ_u 1s at large R, indicating they originate from the ground state (e.g., ls state of H). At small R they are labeled $1s\sigma_g$ and $2p\sigma_u$, indicating they correlate with the 1s and 2p states of the united atom (e.g., He^+ for the H_2^+ molecule, Fig. 10).

To obtain interaction potentials for a particular pair of atoms the net binding energy of each electron, determined from the correlation diagram and the Pauli principle, is combined with the nuclear repulsive potential in Eq. (33). Therefore, two hydrogen atoms (H_2) interact via a singlet state $(\sigma_g 1s)^2\ ^1\Sigma_g^+$ which is *attractive* at intermediate R, and a triplet state $(\sigma_g 1s)(\sigma_u 1s)\ ^3\Sigma_u^+$. The latter is *repulsive* at intermediate and small R (i.e., ignoring the polarization at very large R). Therefore, for the collision between two H atoms $\frac{3}{4}$ of the collisions are repulsive and $\frac{1}{4}$ attractive. Two helium atoms (He_2) interact via a $(\sigma_g 1s)^2 (\sigma_u 1s)^2\ ^1\Sigma_g^+$ ground repulsive state. The lowest state for two oxygen atoms is $(\sigma_g 1s)^2\ (\sigma_u 1s)^2\ (\sigma_g 2p)^2\ (\pi_u 2p)^4\ (\pi_g 2p)^4\ (\sigma_g 2s)^2\ ^3\Sigma_g^+$. In the above the $\Sigma, \Pi, \Delta, \ldots$ represent the *net* angular momentum along the internuclear axis ($|\Sigma_i m_l| = 0, 1, 2, \ldots$, where i implies a sum over orbitals). These states are all double degenerate (i.e., the angular momentum vector can be pointed either way) except the Σ states, which are, therefore, also labeled according to their behavior under reflection ($+$ or $-$) about a plane containing the nuclei.

The molecular orbital diagrams are extremely useful for determining which transitions are likely to occur. Following the molecular orbital states at large R into small R will result in a number of curve crossings as seen in the potential diagram for certain states of He_2^{2+} in Fig. 11. In a full adiabatic calculation of the electronic energies, for which the electrons *totally* adjust to one another, crossings are avoided for states of the same total symmetry. Therefore

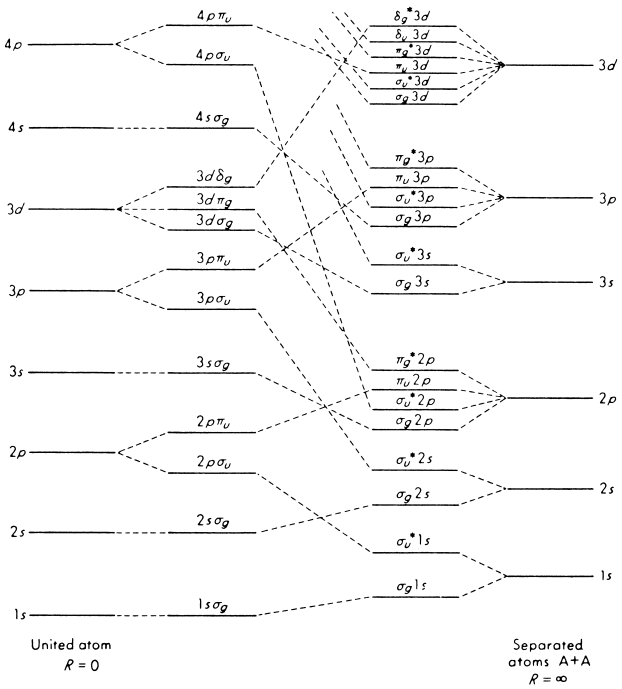


FIGURE 10 Correlation diagram for $\varepsilon_j(R)$ in Eq. (33) for one electron on two identical centers with effective charge Z'_A [From R. E. Johnson (1982). "Introduction to Atomic and Molecular Collision," p. 124. Plenum, New York.]

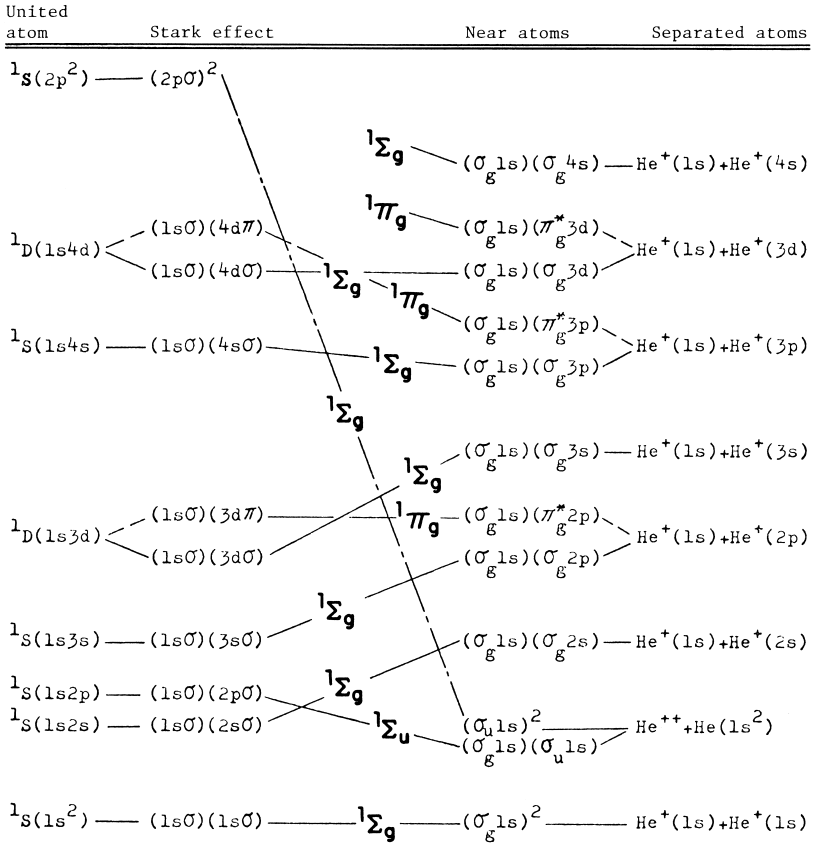


FIGURE 11 Correlation diagram for certain states of He_2^{2+}. $He^{2+}+$ He gives a $^1\Sigma_g$ that crosses many states and a $^1\Sigma_u$ that does not. [From R. E. Johnson (1982). "Introduction to Atomic and Molecular Collisions," p. 123. Plenum, New York.]

the crossings in the molecular orbital diagram indicate that value of R at which the character of the wave function is changing and hence indicate the transition regions. The $(\sigma_u 1s)^2\ ^1\Sigma_g^+$ state of He_2^{2+} is seen to make a series of crossings as both of the electrons are promoted to (2p) electrons in the united atom limit. Therefore, half of the collisions for He^{2+} on He will proceed along an attractive $^1\Sigma_u^+$ for which transitions *are unlikely* at low velocities and half along a strongly repulsive $^1\Sigma_g^+$ state for which transitions are *very likely* to occur if the distance of closest approach is small enough. As pointed out by Licthen and Fano, such electron promotion occurs in all colliding systems as there are two atomic orbital states at large R for every atomic orbital state at small R (see Figs. 9 and 10).

A molecular orbital correlation diagram can also be constructed for an electron in the field of two positive centers having different effective charges Z_A and Z_B. As the inversion symmetry is broken, the correlations are determined using $|m_l|$ only, thereby lessening the promotion effect. The importance of promotion, and hence transition, depends in part, therefore, on how similar the effective charges are on each center. For instance, for a proton interacting with an argon atom (ArH^+ potential), the ground-state potentials at intermediate R are well described by a single electron shared by an Ar^+ core and a proton. As these have binding energies $I_{Ar} = 15.2$ eV and $I_H = 13.6$ eV, the effective charges are very similar ($Z'_{Ar} = 1.06$, $Z'_H = 1$). Therefore, the lowest Σ states of ArH^+ are similar to those of H_2^+ at intermediate R. Stated another way, states of the same symmetry that are close in energy at large R are strongly coupled; therefore, they tend to "repel" (diverge) from each other. The covalent nature of these states is associated with an exchange interaction similar to that in H_2^+, which we write in a general form as

$$\Delta\varepsilon \approx A\exp(-R/\bar{a}) \tag{38}$$

with an effective radius, $\bar{a} = 2a_0/(Z'_A + Z'_B)$.

This confusion of potentials presents problems for the user. For example, although the $O^+ + O$ (O_2^+ system) has a lowest-lying attractive (bound) state $^2\Pi_g$, in a collision of a ground-state oxygen ion with a ground-state oxygen atom seven different potential curves evolve, each with a different multiplicity. Therefore, simplifying procedures are desirable. Although it is not strictly correct to use a

single average potential, the repulsive curves often dominate and a simple repulsive potential is often used to approximate the interaction at intermediate R. Based on the form for electron exchange interaction discussed above, this potential is generally represented as an exponential, $V \approx Ae^{-R/\bar{a}}$. Such a potential is called a Born-Mayer potential and is commonly used in particle penetration calculations. It can be thought of as an extension of the short-range, "universal" repulsive interaction discussed earlier.

E. Interraction Potentials for Molecular Collisions

The interaction between an atom and a molecule or between two molecules depends on the internal coordinates of the molecule and the relative orientations of the atom and molecule, in addition to the distance R (the distance between their centers of mass). These additional degrees of freedom relative to the atom-atom case make the calculation of molecular interaction potentials very computationally intensive. Nevertheless, there has been great progress in doing this recently. A particularly important class of atommolecule interactions goes under the heading of van der Waals systems. A particularly simple form for this interaction obtains for closed shell molecules and atoms. For an atom-diatomic molecule the interaction potential is written as

$$V(R, r, \gamma) = \sum_{\lambda=0} V_\lambda(R, r) P_\lambda(\cos\gamma),$$

where r is the diatomic internuclear distance, R is the distance between the atom and the center of mass of the diatomic, and is the orientation angle such that γ equal to zero corresponds to the one collinear arrangement of the three atoms and γ equal to π corresponds to the other collinear arrangement of the three atoms. P_λ are Legendre polynomials of order λ and the V_λ are functions of R and r.

In many instances over the energy range of interest or at moderate or long ranges, the dependence on r for these closed shell systems is minor and the V_λ are essentially functions of R only. In this case the interaction depends only on a single distance and the molecular orientation. The "stereochemistry" of the combined system is determined by V and to a useful level of approximation the interaction of many-body system, e.g., a cluster, is determined by the pairwise sum of the molecule-molecule or atom-molecule interactions.

VI. INELASTIC COLLISIONS

A. Overview

The calculation of the changes in internal motion of a molecular system of particles induced by a collision with another atom or molecule is a complicated many-body problem for which a number of approximate models have been developed. When the interacting particles are moving, as in a collision, then the interactions described in the preceding sections become dynamic. In addition to the overall deflections, calculated by the potentials described, the time-dependent fields produce changes in the internal state of the molecule. The motion of each of the constituent particles of the molecules can be characterized by a frequency ω and a mean radial extent from the center of mass, $\bar{r}$. The interaction with a passing particle of velocity v at a distance b from the center of a target atom or molecule is said to be nearly adiabatic if $\tau_c \gg \omega^{-1}$ where $\tau_c \approx b/v$. For nonadiabatic collisions, transitions become likely, as discussed in Section I.B.

In most instances, the approximations used involve only two states. Such models can be divided into two categories: strong interactions, for which the evolution of the electronic states during the collision is important (e.g., curve-crossing transitions) and weak interactions, for which the initial charge distributions can be considered static. The former case generally applies to incident ion velocities comparable to or smaller than the velocities of the constituent particles of the target and the latter to large velocities. If, in addition, $b \ll \bar{r}$, then the collision is a close collision and the incident particle can be thought of as interacting with each of the constituents of the target molecule separately. This is referred to as the binary-encounter limit. For $b \gg \bar{r}$, a distant collision, the target atom or molecule must be viewed as a whole. Classically the constituent particles are often treated as bound oscillators of frequency ω. These oscillators are then excited by the time-dependent field of the passing particle. (Of course, in wave mechanics the impact parameter is not a well-defined concept and the close and distant criterion is replaced by whether the momentum transfer to the constituent particles is large or small.) In the following we calculate, classically, the large-momentum transfer (close collision) and small-momentum transfer (distant collision) contributions to the energy-loss cross section. These results are synthesized in the Bethe–Born approximation to the cross section. Following this we consider models for strong interactions such as charge exchange.

B. Classical Oscillator

For distant-collisions, and low-momentum transfer collisions, we can treat the bound electrons and/or nuclei as classical oscillators that are excited by the time-varying field of the passing particle. The motion of an electron oscillator in a field $\mathscr{E}_j(t)$ is

$$m_e\ddot{\mathbf{r}}_j + \Gamma_j\dot{\mathbf{r}}_j + m_e\omega_j^2\mathbf{r}_j = -e\mathscr{E}_j(t) \qquad (39)$$

where ω_j is the binding frequency and Γ_j is a damping constant. Writing the energy transfer as

$$Q_j = \int_{-\infty}^{\infty} \dot{\mathbf{r}}_j \cdot [-e\mathscr{E}_j(t)]\, dt$$

then as $\Gamma_j \to 0$ it is straightforward to show that

$$Q_j \to \frac{\pi}{m_e} |e\mathscr{E}_j(\omega)|^2$$

where $\mathscr{E}_j(\omega)$ is the Fourier transform of $\mathscr{E}_j(t)$, that is,

$$\mathscr{E}_j(\omega) = \frac{1}{\sqrt{2\pi i}} \int_{-\infty}^{\infty} \mathscr{E}_j(t) \exp(-i\omega t)\, dt$$

Note that $|e\mathscr{E}_j(\omega)|$ has units of momentum, indicating the net impulse to the oscillator.

Describing $\mathscr{E}_j$ as the field associated with a screened coulomb potential, $V = (Z_A e^2/R)e^{-\beta R}$ and assuming straight line trajectories ($R^2 = b^2 + v^2t^2$), the energy transfer is

$$Q_j = \frac{2\left(Z_A e^2\right)^2}{m_e v^2}\left(\frac{1}{b^2}\right)\left[(\beta_j' b)^2 K_1^2(\beta_j' b) + \left(\frac{\omega_j b}{v}\right)^2 K_0^2(\beta_j' b)\right] \tag{40a}$$

where $\beta'^2 = \beta^2 + (\omega_j/v)^2$ and K_1 and K_0 are modified Bessel functions. This energy transfer behaves asymptotically as

$$Q_j \to \frac{2\left(Z_A e^2\right)^2}{m_e v^2}\left(\frac{1}{b^2}\right)\begin{cases} \left.\begin{matrix} 1 \quad \text{for} \quad b\beta \ll 1 \\ \dfrac{\pi}{2}(\beta b)\exp(-2\beta b) \\ \text{for} \quad b\beta \gg 1 \end{matrix}\right\} \dfrac{\omega_j b}{v} \ll 1 \\ \pi\left(\dfrac{\omega_j b}{v}\right)\exp\left(\dfrac{-2\omega_j b}{v}\right); \\ \dfrac{\omega_j b}{v} \gg 1,\ \dfrac{\omega_j}{v} \gg \beta \end{cases} \tag{40b}$$

It is important to note in these expressions that screening of the collision is a result *both* of the direct screening in the potential via $\beta[\beta \to a_j^{-1}$ in Eq. (35)] and that due to the motion of the electron via ω_j.

C. Bethe–Born

The first-order estimate of the inelastic cross section in quantum mechanics is given by the Born approximation. The Born scattering amplitude for a transition from an initial state (0) to a final state (f) is given by

$$f_{0\to f} \approx \frac{-m}{2\pi\hbar^2} \int e^{-i\mathbf{K}_f \cdot \mathbf{R}} V_{f0}(R) e^{i\mathbf{K}_0 \cdot \mathbf{R}}\, d^3R \tag{41a}$$

This leads to a cross section, as in Eq. (20), of the form

$$\sigma_{0\to f}(\chi) = \frac{K_f}{K_0} |f_{0\to f}|^2 \tag{41b}$$

where $\hbar K_f$ and $\hbar K_0$ are the final and initial momentum and $\cos\chi = \hat{K}_f \cdot \hat{K}_0$. In Eq. (41a) the exponentials represent incoming and outgoing plane waves and $V_{f0}(R)$ is the interaction potential averaged over the final and initial states. Equation (41) can be related to Eq. (29), the elastic scattering result. For elastic scattering, the final and initial electronic states are identical; therefore $\mathbf{K}_j$ and $\mathbf{K}_0$ are the same size but differ only in direction. In describing elastic scattering, the potential $V_{00}(R)$ is simply the electrostatic potential for the ground state that we called $V_0(R)$ in Eq. (33). In addition, $e^{i(\mathbf{K}_0 - \mathbf{K}_f)\cdot\mathbf{R}} = e^{-i\Delta\mathbf{p}\cdot\mathbf{R}/\hbar}$, yielding the result in Eq. (29).

For a fast collision of an incident ion with a neutral, Bethe approximated the cross section above as

$$d\sigma_{0\to f} \simeq \frac{2\pi\left(Z_A e^2\right)^2}{m_e v^2} Z_B \frac{dQ}{Q^2} |F_{0\to f}(Q)|^2 \tag{42}$$

where $Q = \Delta\mathbf{p}^2/2m_e$ and the quantity $F_{0\to f}(Q)$ is the interaction matrix element of the target atom. Note that if $|F_{0\to f}|^2 = 1$, this expression becomes identical to the Rutherford cross section above. Equation (42) differs from the BEA in that $\Delta\mathbf{p}$ is *not* the momentum transfer to a single electron but to the system as a whole. Hence, even for very small net change in momentum, electronic transitions, which require significant internal energy changes, *can* occur. The interaction matrix element is a weighting factor, which has the property that

$$\sum_f (\varepsilon_f - \varepsilon_0)|F_{0\to f}(Q)|^2 = Q$$

so that the *average* effect of all the electronic transitions is an energy change Q. $F_{0\to f}(Q)$ contains the screening effect of the moving electron, so that for large-momentum transfers, and hence large Q,

$$|F_{0\to f}(Q)|^2 \approx \begin{cases} 0 & \varepsilon_f - \varepsilon_0 \neq Q \\ 1 & \varepsilon_f - \varepsilon_0 = Q \end{cases}$$

This means that for large energy transfers, the momentum transfer is, with a *high probability*, equal to that transferred to a single electron raising it to an excited state. At small momentum transfers, hence low Q,

$$Z_B |F_{0\to f}(Q)|^2 \approx \left(\frac{Q}{\varepsilon_f - \varepsilon_0}\right) f_{0f}$$

This is the limit in which the excitations are predominantly dipole excitations and f_{0f} is the dipole oscillator strength, which is also used to determine the polarizability of an atom in the field an ion.

The quantities above can also be used to determine the leading terms at large v for total ionization cross section σ_I and the straggling cross section $S_e^{(2)}$. The lead term at large v is

$$\sigma_I \equiv \sum d\sigma_{0\to f} \approx \frac{4\pi\left(Z_A e^2\right)^2}{m_e v^2}\left(\frac{2m_e\overline{r_0^2}}{3\hbar^2}\right)\ln\left(\frac{2m_e^{v^2}}{I}\right) \tag{43}$$

where f implies all final states leading to an ionization and $\overline{r_0^2}$ is the mean-squared radius of the initial state.

D. Two-State Models: Charge Exchange

For collision velocities comparable to or smaller than the speed of the constituents of the target particles, the details of the interaction potentials described earlier can control the transition probability. When the coupling between neighboring states is strong and the existence of other states can be treated as a weak perturbation, a number of two-state models can be used for calculating these probabilities. The models are based on the impact parameter cross section in Eq. (7), and forms for the transition probabilities $P_{0\to f}$ are given in Table III. A requirement for strong coupling is that the energy difference between the states at any point in the collision is small compared to the uncertainty in the energy of the states during the collision. This uncertainty is estimated as $\Delta E \approx \hbar(\Delta R_x/v)$ where ΔR_x is the extent of the transition region. The general nature of these two-state inelastic cross sections was described earlier. At low velocities (ΔE above much less than the state spacing $\varepsilon_f - \varepsilon_0$), the levels are well defined and transitions are not likely. At high velocities the uncertainty ΔE becomes large so that the states effectively overlap. However, at high velocities, the time for a transition to occur becomes short so that transition probabilities again become small and the Born approximation described above should be used. When the spacing between states during the collision is comparable to ΔE, then the cross section is large (i.e., $\sim\pi a_0^2$ for transition involving outer-shell electrons).

Approximate models for the two-state impact parameter cross section in Eq. (7) can be written in the form

$$\sigma_{0\to f} \approx \bar{P}_{0\to f}\,\pi b_x^2 \tag{44}$$

where b_x is that impact parameter giving an onset for transitions and $\bar{P}_{0\to f}$ is an averaged transition probability. For symmetric resonant charge exchange (e.g., $H^+ + H \to H + H^+$) the probability of charge exchange is an oscillatory function at low v (Fig. 12 and Table III). The average probability of a transition (exchange) is roughly $\frac{1}{2}$ as the initial and final states are identical (i.e., $\varepsilon_0 - \varepsilon_f$). As $\Delta\varepsilon(R)$ from Eq. (38) divided by $\hbar$ is roughly the rate of electron

TABLE III Inelastic Collision Expressions

Impact parameter results

$$\left(\sigma_{0\to f} = 2\pi\int_0^\infty P_{0\to f}(b)\,db,\ \varepsilon_f - \varepsilon_0 \equiv \hbar\omega_{f0} \equiv Q_{0\to f}\right)$$

First-order (Born)

$$P_{0\to f}(b) \approx \left|\frac{1}{i\hbar}\int_{-\infty}^{\infty}\Delta\varepsilon_{0f}(R)\exp(i\omega_{f0}t)\,dt\right|^2$$

for $\Delta\varepsilon_{0f} = V_0 e^{-\beta R}$

$$P_{0\to f}(b) \approx \left(\frac{2V_0 b}{\hbar v}\right)^2\left(\frac{\beta}{\beta'}\right)^2 K_f^2(b\beta')$$

$$\beta'^2 = \beta^2 + \frac{\omega_{f0}^2}{v^2}$$

$$\sigma_{0\to f} \approx \frac{4\pi}{3}\left(\frac{2V_0}{\hbar\omega_{f0}\beta}\right)^2\frac{(\omega_{f0}/v\beta)^2}{[1+(\omega_{f0}/v\beta)^2]^3}$$

Landau-Zener–Stueckleberg

(curve crossing, $\varepsilon_f + V_f = \varepsilon_0 + V_0$ at $R = R_x$

$$P_{0\to f}(b) \approx 2\bar{P}_{0\to f}\sin^2\left[\frac{1}{\hbar}\int_0^{t_x}(\varepsilon_f - \varepsilon_0 + V_f - V_0)\,dt\right]$$

$$\bar{P}_{0\to f} = 2p_{0f}(1-p_{0f}), \quad p_{0f} = 1 - \exp(-\delta)$$

$$\delta = \left|\frac{2\pi}{\hbar}\Delta\varepsilon_{0f}^2\Big/\left(\frac{dR}{dt}\right)\frac{d}{dR}(V_f - V_0)\right|_{R=R_t}$$

$$\sigma_{0\to f} \approx \pi R_x^2\,\bar{P}_{0\to f}$$

Useful form: $\Delta\varepsilon_{0f} \approx V_0(R/\bar{a})\exp(-0.86R/\bar{a})$, $V_0 = (I_A I_B)^{1/2}$,

$$\bar{a} = a_0(Z'_A + Z'_B), \quad Z'_A \equiv \left(\frac{2a_0}{e^2}I_A\right)^{1/2}$$

Rosen-Zener (noncrossing)

$$P_{0\to f}(b) \approx \bar{P}_{0\to f}\sin^2\left[\frac{1}{\hbar}\int_{-\infty}^{\infty}\Delta\varepsilon_{0f}\,dt\right]$$

$$P_{0\to f} = \frac{1}{2}\left|\frac{\int_{-\infty}^{\infty}\Delta\varepsilon_{0f}\exp(i\omega_{f0}t)\,dt}{\int_{-\infty}^{\infty}\Delta\varepsilon_{0f}\,dt}\right|^2$$

Demkov (exponential coupling, $\Delta\varepsilon_{0f} = Ae^{-R/\bar{a}}$)

$|\Delta\varepsilon_{0f}|_{R=R_x} = \frac{1}{2}|\varepsilon_f - \varepsilon_0 + V_f - V_0|_{R=R_x}$, defines R_x

$$P_{0\to f}(b) \approx \bar{P}_{0\to f}\sin^2\left[\frac{1}{\hbar}\int_0^{t_x}(\varepsilon_f - \varepsilon_0 + V_f - V_0)\,dt\right]$$

$$\bar{P}_{0\to f}(b) \approx \frac{1}{2}\mathrm{sech}^2\left[\frac{\pi\bar{a}}{2\hbar}(\varepsilon_f - \varepsilon_0 + V_f - V_0)\Big/\left(\frac{dR}{dt}\right)\right]_{R=R_x}$$

$$\sigma_{0\to f} \approx \pi R_x^2\,\bar{P}_{0\to f}(b)$$

Resonant charge exchange

$v \lesssim \bar{v}_e$ (Firsov, Smirnov)

$$P_{ct}(b) = \sin^2\left[\frac{1}{2\hbar}\int_{-\infty}^{\infty}\Delta\varepsilon(R)\,dt\right]$$

$$\sigma_{ct} = \frac{1}{2}\pi b_x^2, \quad \left[\frac{1}{2\hbar}\int_{-\infty}^{\infty}\Delta\varepsilon\,dt\right]_{b_x} \approx \pi^{-1}$$

gives, $[(\pi b_x\bar{a}/2)^{1/2}\Delta\varepsilon(b_x) \approx 0.28v\hbar]$

$v \gg \bar{v}_e$ for $(H^+ + H)$

$$P_{ct} \to \frac{64\pi(b/a_0)^3}{(v/v_0)^7}\exp\left(-\frac{bv}{a_0 v_e}\right), \quad \sigma_{ct} \propto v^{-12}$$

Continues

Continued

Langevin

$\sigma_{0\to f} = \bar{P}_{0\to f}\,\pi b_0^2 \quad [b_0^2 \text{ in Eq. (20)}]$

For ion-molecule reaction:

$$b_0^2 = \frac{2}{v}\left(\frac{\alpha_B Z_A e^2}{m}\right)^{1/2}, \quad \alpha_B = \text{polarizability}$$

Born approximation

$$\sigma_{0\to f}(\chi) = \frac{K_f}{K_0}\left[\frac{m}{2\pi\hbar^2}\int d^3R\, V_{f0} e^{i(\mathbf{K}_0-\mathbf{K}_f)\cdot\mathbf{R}}\right]^2$$

using $V_{f0} = V_0 \exp(-\beta R)$

$$\sigma_{0\to f}(\chi) = \frac{K_f}{K_0}\left[\frac{4m\beta V_0}{\hbar^2}\right]^2 \frac{1}{[\beta^2 + |\mathbf{K}_0 - \mathbf{K}_f|^2]^4}$$

$$\sigma_{0\to f} = \frac{\pi}{3K_0^2}\left[\frac{4m\beta V_0}{\hbar^2}\right]^2 \times \left\{\frac{1}{[\beta^2 + (K_0 - K_f)^2]^3} - \frac{1}{[\beta^2 + (K_0 - K_f)^2]^3}\right\}$$

transfer, then at large R, when the time for sharing of an electron becomes longer than the collision time, the cross section goes to zero. Firov, therefore, estimated the size of the cross section by finding the largest value of b for which the number of transfers is small,

$$\frac{1}{2\hbar}\int_{-\infty}^{\infty} \Delta\varepsilon\, dt \approx \pi^{-1} \tag{45a}$$

This is similar to the Massey-Mohr procedure used for estimating the total elastic cross section in Eq. (30). As

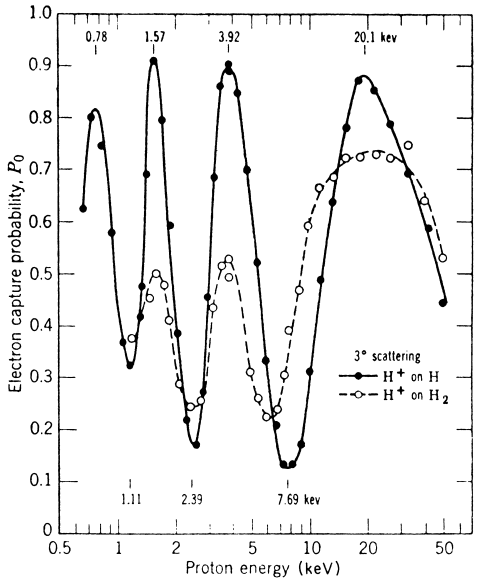


FIGURE 12 Charge-exchange probability at fixed scattering angle versus incident proton energy. Peaks nearly equally spaced in v^{-1}. [From G. J. Lockwood and E. Everhart (1962). *Phys. Rev.* **125**, 567.]

the exchange energy at large R decreases exponentially, Eq. (45a) reduces to the well-known result

$$b_x = b_{ct} \approx \bar{a}(B - \ln v) \tag{45b}$$

where B is a very slowly varying function of v and $\bar{a}$ is the mean radius of the atomic electron cloud. Using Eq. (45b) with Eq. (44), it is seen that the charge exchange cross section increases slowly with decreasing velocity at intermediate velocities.

At very high velocities, exchange of an electron from a stationary atom to a fast ion requires a significant change in momentum. Therefore, $\bar{P}_{0\to f}$ goes to zero rapidly ($\sim v^{-12}$ for $H^+ + H$). This begins to occur for $v > \bar{v}_e$. Therefore, at high velocities the ionization cross sections dominate over the charge transfer cross sections as seen in Fig. 13. At very low velocities the ion and neutral can orbit due to the long-range polarization potential. Orbiting occurs at an impact parameter determined in Eq. (17), which for the polarization interaction ($n = 4$) changes as $v^{-1/2}$. As this value of b increases more rapidly with decreasing v than b_{ct} in Eq. (45b) orbiting can dominate as $v \to 0$. A schematic diagram of the net cross section over many orders of magnitude of v is given in Fig. 13b.

It is seen in Fig. 13b that the cross section grows even at very low velocities when $\varepsilon_f = \varepsilon_0$. On the other hand, charge exchange between nonsymmetric systems (e.g., $S^- + O$ or $H^+ + O$) requires a small change in energy. Therefore, the cross section will exhibit a maximum as discussed above and indicated in Fig. 13. When ΔE is much greater than the state separation $|\varepsilon_f - \varepsilon_0|$, then the cross section behaves like the symmetric resonant collision [i.e., b_a given by Eq. (45) and $\bar{P}_{0\to f} \approx \frac{1}{2}$]. At velocities for which ΔE is comparable to or smaller than $|\varepsilon_f - \varepsilon_0|$, transitions generally occur in a narrow range of internuclear separations ΔR_x about a particular value R_x. In this case the transition probability is a rapidly varying function of v. Such transitions are divided into two classes, those for which a curve crossing exists (e.g., Fig. 8), and those for which the states do not cross. Therefore, knowledge of the potentials discussed earlier is required.

For the curve crossing case, b_x in Eq. (44) is set equal to the crossing point R_x. For impact parameters less than b_x, the colliding particles pass through the point R_x twice, as indicated in Fig. 14. At each passage we assign an average transition probability p_{0f} between the states. After the collision a *net* change in state will have occurred if a transition occurred on the first passage but not the second $[p_{0f}(1 - p_{0f})]$ and vice versa $[(1 - p_{0f})p_{0f}]$, yielding $\bar{P}_{0f} = 2p_{0f}(1 - p_{0f})$. Using this, Eq. (44) becomes

$$\sigma_{0f}^{LZS} \approx 2p_{0f}(1 - 2p_{0f})\pi R_x^2 \tag{46}$$

The fact that two different pathways lead to the same result implies interference occurs; hence, the differential cross

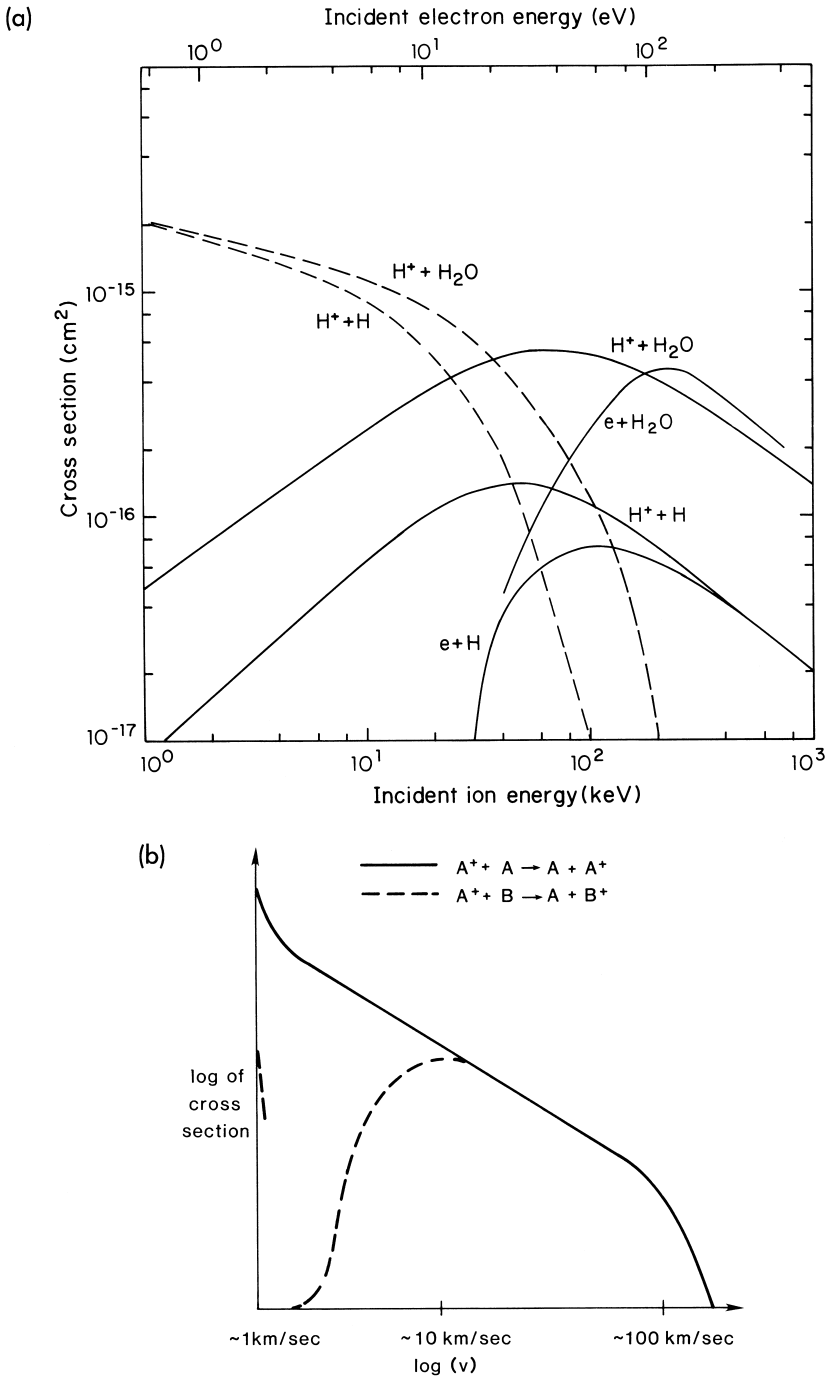


FIGURE 13 (a) Solid lines: ionization cross sections as indicated: dashed lines: charge exchange cross sections as indicated. [From R. E. Johnson (1990). "Physics and Chemistry in Space," Vol. 19, Springer-Verlag, Berlin]. (b) Schematic diagram of the resonant and nonresonant charge exchange cross sections: decrease at very high energy due to momentum required; cross section increases with decreasing velocity at very low energies due to orbiting, for nonresonant only if exothermic.

section is oscillatory. The Landau–Zener–Stueckleberg expression for $p_{0\mathrm{f}}$ is given in Table III. These expressions apply up to velocities where ΔE becomes greater than the energy splitting $|\varepsilon_\mathrm{f} - \varepsilon_0|$, at which point the cross section behaves like the resonant case of Eq. (45). Figure 15 shows the results for three collision processes [see Fig. 13a also].

Demkov and Rosen and Zener have considered the case of the noncrossing interaction potentials, for example,

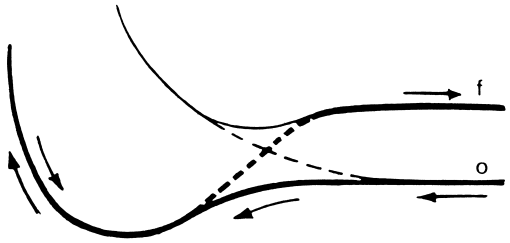


FIGURE 14 Effective potentials for the two trajectories leading to a transition. Dashed lines are approximate diabatic potential curves that cross as in Fig. 8. [From R. E. Johnson (1982). "Introduction to Atomic and Molecular Collisions," p. 139. Plenum Press, New York.]

$A^+ + B \rightarrow A + B^+$. The transition region R_x is defined as that point at which the exchange interaction between the initial and final states [e.g., Eq. (38)] $\Delta\varepsilon_{0f}(R)$ is approximately equal to half the spacing between states $|\varepsilon_f - \varepsilon_0 + V_f - V_0|$, where V_f and V_0 are the interaction potentials for these states [e.g., Eq. (33)]. Demkov developed a very simple and useful expression for $\bar{P}_{0\rightarrow f}$ given in Table III. The cross section so calculated would again join smoothly onto the symmetric-resonant-like result at higher velocities where the energy difference $|\varepsilon_f - \varepsilon_0|$ becomes unimportant.

At very low velocities, the cross sections for both the crossing and noncrossing cases may again increase *if the transitions are exothermic*. That is, even though the transition probability is small for any one pass through the transition region, if orbiting occurs [see Eq. (17)] then after many passes the cumulative effect can lead to a significant probability for a change in the electronic state as suggested in Fig. 13b. As the collision energy is low this can only occur when internal energy is *released* due to an exothermic change in the internal state. In such a case $b_x \approx b_0$ of Eq. (17) and $\bar{P}_{0\rightarrow f}$ in Eq. (44) is equal to the statistical probability of populating any of the exothermic states.

Although we have emphasized charge-exchange collisions and, further, only collisions involving ions and atoms, these procedures apply generally to molecular collisions and to all varieties of internal change in state including molecular reactions. For example, the above discussion of orbiting applies to low-energy ion-molecule reactions yielding the often-used Langevin cross section (see Table III). In all cases, including the processes discussed here, readers are urged to refer to the specific literature for accurate measured and calculated cross sections and use the approximations here as a guide or for rough estimates when more accurate results are not available.

E. Detailed Balance

We briefly consider a property of the inelastic collision cross sections that can be quite usefully exploited in some cases. The equations of motion, both quantum-mechanical and classical, are such that

$$|f_{0\rightarrow f}(\chi)|^2 = |f_{f\rightarrow 0}(\chi)|^2 \quad \text{or} \quad P_{0\rightarrow f}(b) = P_{f\rightarrow 0}(b) \quad (47a)$$

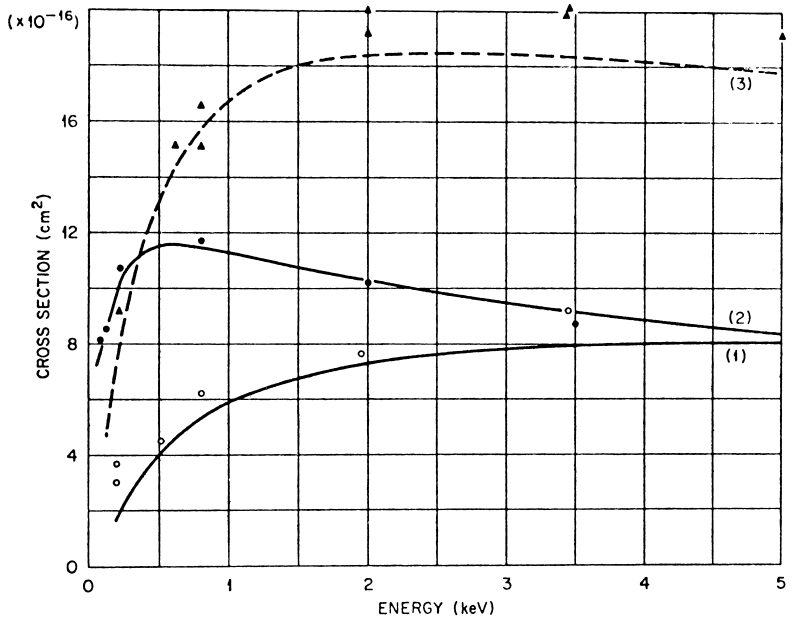


FIGURE 15 Single-electron capture by doubly charged ions. Curves, LZS calculation; points are data: (1, ○) $Ar^{+2} + Ne \rightarrow Ar^+ + Ne^+$; (2, ●) $N^{-2} + He \rightarrow N^- + He^+$; (3, △) $Ne^{+2} + Ne \rightarrow Ne^+ + Ne^-$. [From R. A. Mapleton (1972). "*Theory of Charge Exchange*," p. 212. Wiley, New York.]

This is due to the time-reversal symmetry of the collision process. Using Eq. (41b) with Eq. (47a), the differential cross sections for the forward and reverse reactions can be related. Further, integrating over angle, the integrated inelastic cross sections are related by

$$p_0^2\sigma_{0\to f}(p_0) = p_f^2\sigma_{f\to 0}(P_f) \tag{47b}$$

where p_0 and p_f are the initial and final momenta. In the semiclassical region $p_0 = p_f$, and therefore the forward and reverse reactions have roughly the same cross section. That is, in this region endothermic and exothermic processes behave similarly. At low velocities, near threshold for the endothermic process, the P_0 and P_f differ significantly, and therefore the forward and reverse reactions can differ markedly as pointed out when discussing orbiting. However, these processes are simply related by Eq. (47b). This relationship is a statement of the principle of detailed balance used when describing equilibrium in statistical mechanics. Based on the notions of statistical mechanics, Eq. (47b) can be extended to cases where there are a number of equivalent initial states ξ_0 and/or final states ξ_f (e.g., spin or angular momentum states),

$$p_0^2\xi_0\sigma_{0\to f}(p_0) = p_f^2\xi_f\sigma_{f\to 0}(p_f)$$

Such a relationship allows one to determine, for instance, deexcitation cross sections from data on excitation cross sections and provides a constraint when calculating cross sections by approximate methods.

VII. REACTIVE COLLISIONS

A. Overview

Reactive scattering here refers to a chemical reaction

$$A + B \to C + D$$

that takes place electronically adiabatically (that is on a single potential energy surface) or possibly electronically nonadiabatically, and where "A" and "B" are called the *reactants* and "C" and "D" are the *products* (these may be atoms or molecules). Reactive scattering is at the heart of chemical reaction in the gas phase and thus of all the types of scattering presented here this is the one that makes the closest contact to chemistry.

The methodology for describing quantum reactive scattering is now well developed in both time-independent and dependent approaches, and we briefly review these approaches below.

B. Time-Independent Approach

The time-independent approach to reactive scattering is based on the *coupled channel* method in which the time-independent wavefunction is expanded in terms of a basis of known internal wavefunctions $\phi_i'(\boldsymbol{Q})$ times unknown scattering wavefunctions $G_{i,j}(\boldsymbol{R})$ as

$$\psi_j(E) = \sum_i \phi_i(\boldsymbol{Q})G_{i,j}(R), \tag{48}$$

where the subscript j refers to an initial quantum state of the reactants and R is the scattering coordinate. As a result of this expansion of the wave function, the time-independent Schrödinger equation is recast as a matrix equation given by

$$(\boldsymbol{H}(\boldsymbol{R}) - \varepsilon)\,\boldsymbol{G}(R) = 0, \tag{49}$$

where $\boldsymbol{H}$ is the matrix representation of the Hamiltonian H in the basis of the internal functions used in (xx) and ε is a diagonal matrix of translational energies $E - E_i$, where E_i is the energy of the internal state i and E is the total energy. This equation is solved numerically, and from the solution the scattering matrix $\boldsymbol{S}$ is obtained, in terms of which state-to-state transition probabilities are given by the square of the absolute value of elements of $\boldsymbol{S}$.

For the simplest class of chemical reactions, i.e., an atom + diatomic reaction, the generalization of the expression for the total cross section given by Eq. (23) is

$$\sigma_{r\to p} = \frac{1}{K_r^2}\sum_{J=0}^{\infty}(2J+1)\left|\boldsymbol{S}_{r\to p}^J\right|^2, \tag{50}$$

where J is the total (nuclear) angular momentum of the three-atom system. In the methods currently in use the entire $\boldsymbol{S}$-matrix is obtained at each value of the total energy E. For some problems this is neither practical nor desirable and, in part, this motivated the development of computational approaches based on the time dependent description of scattering.

C. Time-Dependent Approach

The time-dependent approach to collision theory is an insightful as well as a computationally effective alternative to the time-independent theory. In this approach an initial wave packet, denoted $\Psi(0)$, is propagated in time according to the time-dependent Schrödinger equation

$$i\hbar\frac{\partial\Psi}{\partial t} = H\Psi \tag{51}$$

The common form for $\Psi(0)$ is

$$\Psi(0) = (\pi\alpha^2)^{-1/4}\exp\left[-(R - R_0)^2/2\alpha^2\right] \times \exp(-iK_0R)\phi_i(Q) \tag{52}$$

which is a Gaussian function in the scattering coordinate centered at R_0, which is chosen to be in the noninteracting region. In the conjugate momentum space the wave packet is Gaussian function centered at K_0. Clearly $\Psi(t)$ is not an

eigenfunction of H and the propagation in time must be done numerically. There are two widely used propagation methods that are used in atomic and molecular scattering. These are the Split-Operator method and the Chebychev method. In the Split-Operator method the packet is propagated for a time step Δt according to

$$\Psi(t+\Delta t) = \exp\left(-\frac{i}{\hbar}T\Delta t/2\right)\exp\left(-\frac{i}{\hbar}V\Delta t\right) \times \exp\left(-\frac{i}{\hbar}T\Delta t/2\right)\Psi(t) + \mathrm{O}(\Delta t^3), \quad (53)$$

where for simplicity we assume H is given by $T+V$, as it would be in atom-atom scattering. The utility of this expression is that it permits the action of the kinetic and potential operators on $\Psi(t)$ to be carried sequentially, and most efficiently is $\Psi(t)$ is represented on a grid. The operation involving $\exp(-\frac{i}{\hbar}T\Delta t/2)$ is facilitated by first Fourier transforming $\Psi(t)$ to momentum space where $\exp(-\frac{i}{\hbar}T\Delta t/2)$ becomes a simple scalar operator, then back transforming to coordinate space, where $\exp(-\frac{i}{\hbar}V\Delta t)$, is a scalar operator, transforming to momentum space, acting with $\exp(-\frac{i}{\hbar}T\Delta t/2)$, and finally back transforming to coordinate space. The forward and backward Fourier transforms can be done efficiently using "Fast Fourier Transform" methods.

The Chebychev propagation method is a direct expansion of the time evolution operator $\exp(-\frac{i}{\hbar}Ht)$ in-terms of Chebychev polynomials, P_n, whose argument is a scaled Hamiltonian operator, H_s. Thus, letting

$$H_s = (E_{\max} + E_{\min} - 2H)/(E_{\max} - E_{\min}), \quad (54)$$

where $E_{\min}$ and $E_{\max}$ are estimates of the minimum and maximum eigenvalues of H. Then, denoting ΔE as $E_{\max} - E_{\min}$,

$$\exp\left(-\frac{i}{\hbar}Ht\right) = \sum_n a_n(\Delta E t/2\hbar)P_n(-iH_S), \quad (55)$$

where the a_n are related to regular Bessel functions of order n. As in the Split-Operator method the wave packet is represented on a grid. However, in the Chebychev approach the action of the Chebychev polynomials in H_S on the wave packet is facilitated by a simple recursion relation.

In order to apply these time-dependent methods to scattering problems special techniques have been developed to eliminate reflections of the wave packet at the grid boundaries. These involve either the explicit use of negative imaginary potentials or damping functions, both techniques attenuate the wave packet near the grid boundaries and effectively eliminate reflections.

An example of a reactive system that has been extensively studied is the $D + H_2$ reaction to form $HD + H$. The exact quantum reaction probability for zero total angular momentum is plotted in Fig. 16 as a function of the total energy and for the indicated initial rotational state of H_2. As seen, the reaction probability depends strongly on the initial rotational state, j, with the maximum occurring for $j = 1$. This dependence on j is typical for reactions that occur mainly through a collinear arrangement of the atoms.

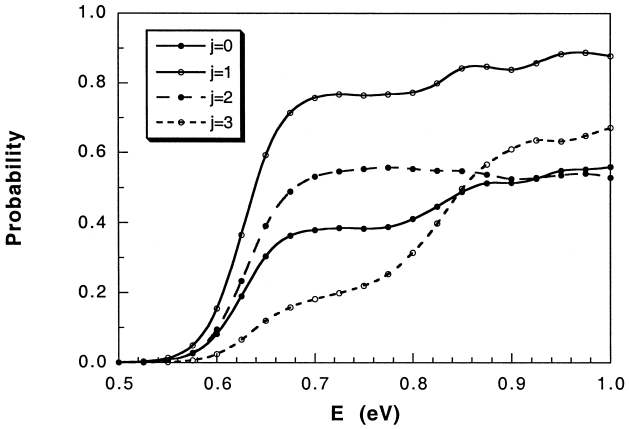


FIGURE 16 Exact quantum reaction probabilities for the $D + H_2 \rightarrow DH + H$ reaction as a function of the total energy E for zero total angular momentum and for initial rotational states of H_2, j, as indicated.

Resonances occur in molecular collisions, and their dependence on the molecular interaction potential make them an important object of experimental and theoretical investigation. They are also of importance for the kinetics of recombination and dissociation. One diagnostic for the presence of resonances is the behavior of the Smith collision lifetime matrix, $\boldsymbol{Q}$, which is defined by

$$\boldsymbol{Q} = i\hbar\left(S\frac{dS^\dagger}{dE}\right) \quad (56)$$

where $\boldsymbol{S}$ is the scattering matrix. In the vicinity of a narrow resonance the trace of $\boldsymbol{Q}$ shows an abrupt increases in value. Note that $\boldsymbol{Q}$ has units of time and thus its value at the resonance energy is a measure of the collision lifetime.

SEE ALSO THE FOLLOWING ARTICLES

ATOMIC PHYSICS • COLLISION-INDUCED SPECTROSCOPY • POTENTIAL ENERGY SURFACES • SCATTERING AND RECOILING SPECTROMETRY

BIBLIOGRAPHY

Bernstein, R. B. (1979). "Atom-Molecule Collision Theory," Plenum, New York.

Bates, D. R., and Bederson, B. (1965-1982). "Advances in Atomic and Molecular Physics," Vols. 1–18. Academic Press, New York.

Child, M. S. (1974). "Molecular Collision Theory," Academic Press, New York.

Hasted, J. B. (1972)."Physics of Atomic Collisions," 2nd ed. Am. Elsevier, New York.

Hirschfelder, J. O., Curtiss, F., and Bird, R. B. (1964). "Molecular Theory of Gases and Liquids," Wiley, New York.

Johnson, R. E. (1982). "Introduction to Atomic and Molecular Collisions," Plenum, New York.

Johnson, R. E. (1990). "Energetic Charged-Particle Interactions with Atmospheres and Surface." Vol. 19 of "Physics and Chemistry in Space" (L. Lanzerotti, ed.). Springer-Verlag, Berlin.

McDowell, M. R. C., and Coleman, J. P. (1970). "Introduction to the Theory of Ion-Atom Collisions," North-Holland, Amsterdam.

Massey, H. S. W. (1979). "Atomic and Molecular Collisions," Halsted Press, New York.

Massey, H. S. W., Burhop, E. H. S., and Gilbody, H. G. (1970–1974). "Electronic Ionic Impact Phenomena," 2nd ed., Vol. 1–5. Oxford Univ. Press, London and New York.

Mott, N. F., and Massey, H. S. W. (1965). "The Theory of Atomic Collisions," 3rd ed. Oxford Univ. Press, London and New York.

Ziegler, J. F., Biersack, J. P., and Littmark, U. (1985). "The Stopping and Ranges of Ions in Solids," Pergamon, New York. (Contains "universal" scattering and tables of collisions quantities.)



Atomic Physics

Francis M. Pipkin (deceased)

Department of Physics, Harvard University

Mark D. Lindsay

Department of Physics, University of Louisville

I. Historical Development
II. The Hydrogen Atom
III. The Helium Atom
IV. Rydberg and Exotic Atoms
V. Complex Atoms
VI. Interaction of Atoms with Radiation

GLOSSARY

Atomic number Positive charge on the nucleus of an atom in units of the electron charge, denoted by Z; the number of protons or number of electrons on a neutral atom.

Bohr magneton Natural unit in which to measure the magnetic moment of an electron or atom; denoted by $\mu_B = e\hbar/m$.

Classical (not quantized) Physical quantities can have a continuous range of values—any value allowed by basic laws of conservation of energy, momentum, and angular momentum.

Electric dipole radiation Principal form of electromagnetic radiation emitted by atoms. Requires a change in parity and $\Delta J = 0, \pm 1$ where J is the total angular momentum ($J = 0$ to $J = 0$ is not allowed.)

Electron volt Energy acquired by an electron when it is accelerated by a potential difference of 1 V, denoted by eV; equal to 1.602×10^{-19} J.

Fine structure Shifts in energy levels of atomic electrons due to the interaction of the electron's orbital angular momentum and spin.

Fine structure constant Measure of the strength of the interaction between a charged particle and the electromagnetic field. Denoted by $\alpha = e^2/4\pi\varepsilon_0 c\hbar = 1/137.036$ (unitless).

g-factor Ratio of the magnetic moment of an atom or an electron to the product of the largest component of the angular momentum and the Bohr magneton. For nuclei, the nuclear magneton is used instead of the Bohr magneton.

Ground state Lowest energy state allowed by the laws of quantum mechanics.

Hartree-Fock Method of calculating the energy levels of electrons in atoms taking into account the many-body interactions between electrons.

Hyperfine structure Shifts in energy levels of atomic electrons due to the interaction of the electron's spin and nuclear spin.

Ionization potential Smallest energy required to remove completely a bound electron from an atom.

Isotope Elements with the same atomic number Z but with different mass number A. They have the same number of protons in the nucleus but different numbers of neutrons.

jj coupling Angular momentum coupling scheme for atoms, in which for each electron the spin angular momentum is tightly coupled to its own orbital angular momentum. The total angular momentum is obtained by first adding for each electron its spin angular momentum and orbital angular momentum and then adding the angular momenta of the individual electrons.

LS coupling Angular momentum coupling scheme for electrons in an atom, in which the spin and orbital angular momentum of individual electrons are independently added together to form the total spin and total orbital angular momentum; these angular momenta are then added together to obtain the total angular momentum (also called *Russell Saunders coupling*).

Orbital Wave function describing the distribution of an electron in one quantum state of an atom (also called *quantum state* or *energy level*).

Parity Behavior of the wavefunction when the coordinate system is inverted and for each electron (x, y, z) is replaced by $(-x, -y, -z)$. The parity is even or odd depending on whether or not the wave function changes sign.

Photon The smallest quantized unit of electromagnetic radiation.

Quantized (not classical) Physical quantities can have only certain allowed discrete values and cannot have intermediate values between the allowed values.

Selection rules Allowed changes in quantum numbers when an atom undergoes a quantum transition from one state to another; for example, $\Delta J = 0, \pm 1$, but not $J = 0$ to $J = 0$.

Spontaneous emission Radiation of an electromagnetic photon by an atom when there are no photons present.

Stimulated emission Radiation of an electromagnetic photon by an atom triggered by the presence of a matching photon.

ATOMIC PHYSICS is that branch of physics that deals with the structure of electrons in atoms, with the structure of ions formed by addition or removal of electrons from neutral atoms, with the interaction of atoms and ions with one another, and with their interaction with the electromagnetic field and free electrons.

I. HISTORICAL DEVELOPMENT

The idea that all matter consists of indivisible particles called *atoms* can be traced to the third- to fifth-century B. C. Greek philosophers Leusippus, Democritus, and Epicurus, who speculated that everything was composed of indivisible particles of a small number of simpler substances in different combinations and proportions. The first-century B.C. poem entitled *De Rerum Natura* by the Latin poet Lucretius provides an elegant summary of early atomism. This idea was attractive and provided encouragement for the alchemists in the Middle Ages who were trying to convert cheaper metals such as iron and lead to the noble metals silver and gold by changing the relative proportions of a few basic constituents.

The modern theory of atoms was founded by the English chemist John Dalton at the beginning of the nineteenth century. Dalton realized that when the chemical elements were combined to form compounds, the masses with which they combined always occurred in well-defined proportions, which were always ratios of integers. For any given compound, the mass of each type of atom—each element—was proportional to an integer. For hydrogen, the integer, or mass unit, was 1. A good example is provided by the combination of nitrogen and oxygen to form the several oxides of nitrogen. Fourteen parts by mass of nitrogen combine with 8, 16, 24, 32, or 40 parts by mass of oxygen to form the five common oxides of nitrogen. Dalton interpreted these proportions in terms of the combination of different numbers of atoms, each of which has a fixed mass. He was thus able to understand the oxides of nitrogen in terms of the combination of nitrogen atoms (N) with a mass of 14 units with oxygen atoms (O) with a mass of 16 units in the proportions N_2O, NO, N_2O_3, NO_2, and N_2O_5, where the subscripts denote the number of atoms of each species in the smallest units of the compounds of nitrogen and oxygen. Dalton had no way of knowing the mass or size of one atom, but this gave him a way of knowing the ratio of the masses of the atoms. Through a study of the available chemical reactions, chemists were able to develop a table of combining masses for the elements. This was called the *periodic chart*. Thus the *mass* of atoms was found to be quantized.

Studies of gases by Amedeo Avogadro and others showed that common gases such as hydrogen, nitrogen, and oxygen consisted of molecules, each of which was composed of two atoms, and that equal volumes of gases at the same temperature and pressure contained the same number of molecules. Rare gases such as helium, neon, and argon, which are chemically inert and occur in the gas phase as single atoms, were not clearly identified until the last decade of the nineteenth century.

In the first few decades of the nineteenth century, when scientists began to study the behavior of solutions when an electric current was passed through them, they realized that in solutions atoms could become positively or negatively charged. These positively or negatively charged atoms were called *ions*. In 1833, Michael Faraday found that the passage of a fixed quantity of electricity through a solution containing a compound of hydrogen would always cause the appearance of the same amount of hydrogen gas at the negative terminal, irrespective of the kind of hydrogen compound that had been dissolved and the strength of the solution. More generally, it was found that the masses of the material given off or deposited at both the positive and negative terminals by the same quantity of electricity were proportional to the chemical combining masses. The electrical behavior of ions in solutions led to the speculation that all electrical currents were formed by the motion of charged particles. The *charge* of any particle was found to be quantized in integer multiples of the electron's charge. The measurements on solutions were used to determine the ratio of the charge to the mass (e/m) for the ions in solution. The largest ratio was obtained for the positive hydrogen ion.

In the middle decades of the nineteenth century, physicists developed kinetic theory. This theory describes the macroscopic behavior of gases in terms of the behavior of independent atoms and molecules that move according to Newton's laws of motion and occasionally collide elastically with one another. The temperature of the gas is related to the average kinetic energy of the constituents; the pressure is related to the number of constituents per second striking the walls of the container. For a typical gas at room temperature, the average velocity is 500 m/sec. The average distance a constituent travels before it strikes another constituent is called the *mean free path* and it depends upon the diameter of the constituents and the number of constituents per unit volume. Kinetic theory relates the *macroscopic* quantities of temperature, viscosity, rate of diffusion, and thermal conductivity of gases to the *microscopic* characteristics of diameter and average velocity of the constituents.

It was also discovered that when atoms were excited in a flame or an electrical discharge, they emitted light at discrete wavelengths or frequencies that were characteristic of the particular elements. The observed frequencies did not display the characteristic harmonic structure found in mechanical vibrations. All attempts to understand the pattern of observed frequencies in terms of vibrating mechanical structures failed. In 1885, a Swiss school teacher, Johann Jacob Balmer, discovered that the frequencies of the known spectrum lines of hydrogen could be represented by a simple empirical formula, which, with hindsight, can be written in the form:

$$f = Rc\left(\frac{1}{2^2} - \frac{1}{n^2}\right)$$

where f is the frequency, R an empirical constant, c the speed of light, and $n = 3, 4, 5, \ldots$. The constant R and the equation are now named in honor of J. R. Rydberg, a Swedish physicist who made major contributions to our understanding of the spectra of atoms. The theoretical basis of the Rydberg equation was not known at the time and was explained only 35 years later with the advent of Bohr quantum theory.

In the last third of the nineteenth century, physicists studying the conduction of gases in discharge tubes at low pressure identified cathode rays, which moved toward the positive terminal (anode), and canal rays, which moved toward the negative terminal (cathode). The two components were studied by using cathodes and anodes with holes in them to prepare beams of the rays. Initially it was not clear whether they were waves or particles.

In 1897, Joseph John Thomson, an English physicist, showed that the cathode rays were negative particles whose properties did not depend on the nature of the gas. They were called electrons—a name first suggested in 1891 by G. J. Stoney for the natural unit of electricity. Thomson found that the charge-to-mass ratio of the electron was roughly 1836 times larger than that found for the hydrogen ion in solutions.

In 1898, Wilhelm Wien showed that the canal rays have ratios of charge to mass that depend on the gas and are similar to those found for ions in solution. It was later shown that for discharges in hydrogen gas there are canal rays with the same charge-to-mass ratio as that found for the hydrogen ion in solutions. This particle, the hydrogen ion, is now called the *proton*. The study of canal rays led Thomson to the development of mass-spectrometric methods for measuring precisely the masses of the elements.

The discovery of the electron led to a model for the atom as a structure made up of positively charged protons and negatively charged electrons. The charge of the electron was subsequently measured by observing the influence of an electric field on the motion of droplets that were charged due to the removal or addition of one or more electrons. The most extensive set of measurements was made by Robert A. Millikan, using oil drops. Millikan's experiments also showed that the charge of the proton had exactly the same magnitude (but opposite sign) as the charge of the electron.

In 1900, Max Planck, a German physicist, found that he could predict the distribution of electromagnetic radiation emitted by a black body—a body that absorbs all the radiation incident upon it—by postulating that the electromagnetic radiation was emitted and absorbed in characteristic units called *quanta*, or *photons*, whose energy E is related

to the frequency f of the electromagnetic radiation by the equation:

$$E = hf$$

where h is the Planck constant. Intense beams of electromagnetic radiation are made up of many photons. Einstein later used the Planck hypothesis to explain the photoelectric effect, in which light caused metals such as sodium to emit electrons. Thus *electromagnetic radiation* (light) is quantized.

In 1911, Ernest Rutherford and his coworkers discovered through experiments in which alpha particles were scattered from thin gold foils that all the positive charge in an atom was concentrated in a very small region of space much smaller than the size of the atom determined from studies of gases, liquids, and solids. This led to the picture of the atom as a miniature solar system consisting of a heavy, positively charged nucleus about which a group of electrons moved. Like the solar system, most of the atom was empty space.

Niels Bohr, a Danish physicist, subsequently combined the insight provided by the Rutherford experiment with the quantum hypothesis of Planck and Einstein to make the first quantitative model of the atom. Bohr pictured the hydrogen atom as a single positively charged proton with a size of roughly 10^{-15} m, with the electron moving around the proton in a circular orbit with a diameter of 10^{-10} m. To determine the structure, Bohr introduced four new hypotheses. He assumed that not all circular orbits for the electron were possible but that the only allowed orbits were those for which the orbital angular momentum of the electron was an integer multiple of the Planck constant divided by 2π. The allowed orbits are called *orbitals*. The orbital angular momentum L of an object is a measure of its rotational inertia and for a particle in a circular orbit is equal to the product of the mass, the velocity, and the radius of the orbit. It is represented by a vector perpendicular to the plane of the orbit with a magnitude equal to mvr:

$$L = mvr = n\hbar$$

Bohr also assumed, in contradiction to the predictions of classical electromagnetic theory, that while the electron was moving around the proton in one of the allowed orbits, and undergoing classical centripetal acceleration, its total energy was constant and it did not radiate electromagnetic energy. Bohr postulated instead that the electron radiated electromagnetic energy only when it passed from one allowed circular orbit to a second allowed circular orbit with lower energy, and that the frequency of the radiation was given by the equation:

$$f = \frac{E_1 - E_2}{h}$$

Here, E_1 and E_2 are the total energies when the electron is in each of the allowed orbits, and h is the Planck constant. By using classical mechanics to calculate the total energy in each of the orbits, Bohr obtained for the frequency of the radiation the formula:

$$f = Rc\left(\frac{1}{n_1^2} - \frac{1}{n_2^2}\right)$$

where n_1 and n_2 are positive integer quantum numbers that characterize the angular momentum of the two allowed orbits. This expression is the same as that given by the Balmer formula, with the Rydberg constant given in terms of independently known quantities.

Bohr also postulated, in violation of classical physics, that there was a lowest energy state or orbital, and no orbital existed at any lower energy. In terms of the above expression, n_1 and n_2 cannot be less than 1. The lowest energy orbital is called the *ground state*.

Thus, the *angular momentum* (and *energy*) of an electron in an atom are quantized. This semi-classical Bohr theory was very successful and predicted a variety of spectral series in hydrogen and ionized helium, where $n_1 = 1, 2, 3, \ldots$, and n_2 takes on integer values greater than n_1. The theory was subsequently refined by Sommerfeld and Wilson in order to treat properly the three-dimensional nature of the problem and to take account of special relativity.

The refinements of the theory showed that the angular momentum of the atom was space-quantized. Classically, the angular momentum vector points in a fixed direction in an inertial coordinate system, and it can assume any orientation with respect to the reference axis. The quantization of the angular momentum introduced in the Bohr theory predicts that the angular momentum vector can only assume a small number of orientations with respect to the reference axis. This quantization of the angular momentum was confirmed by Stern and Gerlach through experiments in which an inhomogeneous magnetic field was used to deflect a beam of silver atoms. They found that the deflection of the beam was quantized.

In 1924, Louis de Broglie introduced the hypothesis that the electron behaved like a wave with a wavelength λ that depended on its momentum p through the relationship:

$$\lambda = h/p = h/mv$$

De Broglie found that he could understand the Bohr orbits in terms of circles, each of whose circumference was an integral number of wavelengths. This hypothesis led Erwin Schroedinger to the invention in 1926 of a partial differential equation, now called the Schroedinger equation, for the description of the wave behavior of the electron. Schroedinger found that those solutions for which

the amplitude of the wave function remained finite around the nucleus for all time predicted for the electron a discrete set of possible energies, which were the same as those predicted by the Bohr theory. The solutions of the Schroedinger equation are now called the *wave functions* for the electron. The value of the wave function at any point is in general a complex number, and its absolute value squared gives the relative probability of the electron being found at a given point in space and time if a measurement is made.

More precise measurements of the radiation emitted by atoms showed that many of the spectral lines were not one line but several lines close together. The additional structure is called the *fine structure*. It was found that this structure could be understood if the electron were not a simple point-like particle but a small spinning sphere with an angular momentum that could only have two directions with respect to a given reference axis. The electron could either have a component of angular momentum $\hbar/2$ or $-\hbar/2$ along the reference axis. This internal angular momentum is now referred to as the *spin angular momentum*, and an electron is said to have spin $1/2$. This spin can be thought of as a classical angular momentum of an extended object, but in fact the electron is apparently a point, and its spin is an intrinsic quantum feature of it, like its mass and charge. Thus, *spin* is quantized, too.

Speaking classically, since the electron is charged, its rotation produces a current, and this current results in a magnetic moment. The electron magnetic moment μ_e is about twice as large as one would expect classically for a rotating uniformity charged sphere. The moment is said to be anomalous:

$$\mu_e = g\mu_B$$

g is the g-factor of the electron, nearly but not exactly equal to 2. g is currently the most accurately measured physical quantity, measured to within 8 parts in 10^{12}.

The interaction of the magnetic moment with the magnetic field in the rest frame of the electron—produced by the motion of the electron in the electric field of the nucleus—causes an energy level to split into several closely spaced energy levels. This is called the *fine structure of the spectral lines*. The energy of the electron is different when the spin points in the same direction as the orbital angular momentum than when it points in the opposite direction.

Hyperfine splitting of energy levels is similar, but is due to the changes in energy depending on the relative orientations of the electron spin and nuclear spin. Typically, hyperfine splitting is much smaller than fine structure splitting.

In 1928, P. A. M. Dirac discovered a generalization of the Schroedinger equation that takes proper account of special relativity. Dirac found that this equation required that the electron have spin $1/2$ and predicted that the magnetic moment would be exactly twice as large as one would expect classically for a rotating uniformity charged sphere. This theory also predicted that there was a positive partner to the electron that had the same mass and spin angular momentum. This particle, which is now called the *positron*, was discovered in cosmic rays by Carl Anderson in 1932. The Dirac theory was a major triumph for relativistic quantum mechanics.

The other characteristic property of the electron is that no two electrons can occupy the same quantum state. This was first hypothesized by Wolfgang Pauli and is now called the Pauli exclusion principle. It is a characteristic of all particles called *Fermions*, with intrinsic spin $1/2$. The Pauli principle is the major ingredient for explaining the periodic chart of the elements.

The alpha-particle scattering experiments of Rutherford and his collaborators and the measurements of the characteristic X-rays of the elements by Moseley, which were carried out in the early part of the twentieth century, showed that each element could be characterized by the charge Z of the nucleus and that the number of electrons in the outer orbits was equal to the charge of the nucleus. Measurements of the mass of the nuclei, using mass spectrometers invented by Thomson and refined by Aston, Dempster, and others, showed that for each light element the mass M_{nucl} of the nucleus was roughly twice the mass of the protons M_p in the nucleus ($M_{nucl} = 2M_p$). For heavier elements, the factor of 2 becomes larger, in the range 2.1–2.4. There are generally several nuclei with the same Z and a different total mass. Nuclei with the same Z and different masses are called *isotopes*. It was first hypothesized that the nucleus contained electrons that compensated the charge from some of the protons. This hypothesis was contradicted by quantum mechanics and by the total angular momentum observed experimentally for some nuclei. It was speculated by Rutherford that there was a heavy neutral particle called the *neutron*. In 1932, James Chadwick discovered the neutron and showed that it had a mass that was very slightly greater than that of the proton.

The entire atom is composed of protons, neutrons, and electrons. The nucleus of a typical atom consists of Z protons and N neutrons; it is assigned the atomic number Z and mass number $A = Z + N$. M_{nucl} is proportional to A, and M_p is proportional to Z. In general, for any given Z, there are several isotopes with different values of N and A. The nucleus is a roughly spherical object with a radius equal to $1.2A^{1/3} \times 10^{15}$ m.

The electrons occupy the region around the nucleus and are contained in a sphere with a diameter of about 10^{-10} m, which depends on Z (bigger for bigger Z). For neutral

TABLE I Fundamental Physical Constants and Units

Constant	Symbol	Value	Unit
Speed of light	c	299 792 458	m/sec
Magnetic constant	μ_0	$4\pi \times 10^{-7}$	N/A^2
Electric constant	ε_0	$8.854\,187\,817\ldots \times 10^{-12}$	F/m
Gravity constant	G	$6.673\,(10) \times 10^{-11}$	$m^3/(kg\ sec)$
Planck constant	h	$6.626\,068\,76\,(52) \times 10^{-34}$	J sec
$h/2\pi$	$\hbar$	$1.054\,571\,596\,(82) \times 10^{-34}$	J sec
Elementary charge	e	$1.602\,176\,462\,(63) \times 10^{-19}$	C
Bohr magneton	μ_B	$927.400\,899\,(37) \times 10^{-26}$	J/T
Nuclear magneton	μ_N	$5.050\,783\,17\,(20) \times 10^{-27}$	J/T
Fine structure constant (inverse)	$1/\alpha$	137.035 999 76 (50)	
Rydberg constant	R_∞	10 973 731.568 549 (83)	1/m
Hartree	E_h	$4.359\,743\,81\,(34) \times 10^{-18}$	J
Electron mass	m_e	$9.109\,381\,88\,(72) \times 10^{-31}$	kg
Electron g-factor	g	2.002 319 304 373 7 (82)	
Proton mass	m_p	$1.672\,621\,58\,(13) \times 10^{-27}$	kg
Neutron mass	m_n	$1.674\,927\,16\,(13) \times 10^{-27}$	kg
Avogadro constant	N_A	$6.022\,141\,99\,(47) \times 10^{23}$	1/mole
Faraday constant	F	96 485.341 5 (39)	C/mole
Molar gas constant	R	8.314 472 (15)	J/(mole K)
Boltzmann constant	k	$1.380\,650\,3\,(24) \times 10^{-23}$	J/K

Unit of measurement	Quantity measured	
Meter	Length	m
Kilogram	Mass	kg
Second	Time	sec
Newton	Force	$N = kg\ m/sec^2$
Amp	Electric current	A = C/sec
Farad	Capacitance	F
Joule	Energy	$J = kg\ m^2/sec^2$
Coulomb	Electric charge	C
Tesla	Magnetic field	T
Mole	Number of particles	mole
Kelvin	Temperature	K

Adapted from www.physics.nist.gov/cuu/index.html.

atoms, the number of electrons equals Z, and by the Pauli exclusion principle, the chemical behavior is determined by the number of (outer) electrons. Since the atoms are named by their chemical behavior, Z and the name of the element are redundant.

The science of atomic physics can be divided into two very closely linked portions: formulating theories to explain observations and actually making the observations (measurements). A substantial debate is ongoing as to which comes first and which is more important. The measurements are often directed towards measuring the values of fundamental physical constants of nature, which are included in the theories. Some of the most important constants are given in Table I. The number in the parentheses at the end of most values is the one standard deviation uncertainty or error in the measurement. For example, 1.23 (4) means 1.23 ± 0.04, and 1.2345 (67) means 1.2345 ± 0.0067. Constants without an error, such as the speed of light, are defined exactly in the scheme of constants and have zero uncertainty. α and g have no units. One can see that Newton's Universal Constant of Gravitation, G, is by far the least accurately known fundamental physical constant, although historically it was the first to be recognized.

II. THE HYDROGEN ATOM

The simplest atom is the hydrogen atom. It consists of a single proton and electron. Figure 1 shows the energy levels for the hydrogen atom predicted by the Schroedinger equation with the addition of the electron magnetic

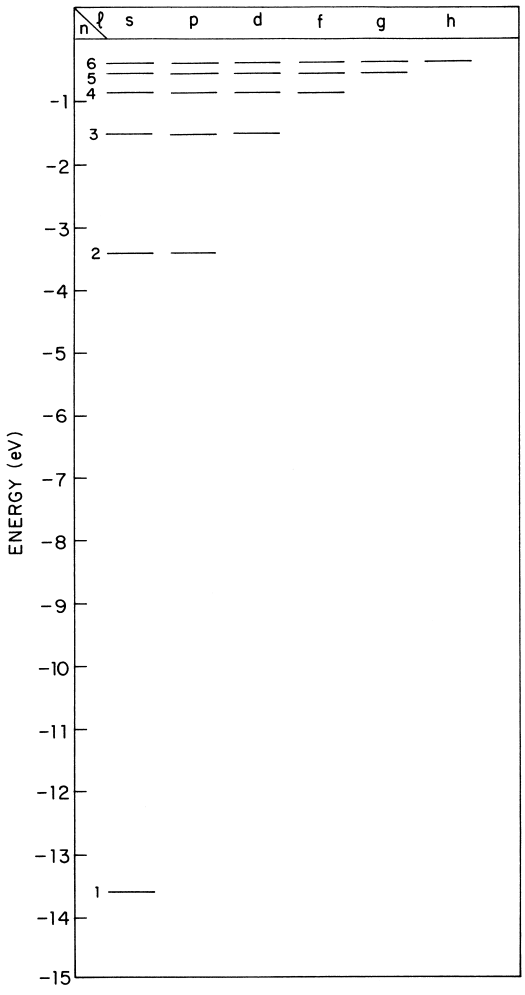


FIGURE 1 The energy levels of the hydrogen atom.

moment and the inclusion of special relativity to second order. The energies of H atom states are by convention shown as negative, below $E = 0$, which is the ionization limit where the electron is removed from the atom. Lower in energy means more tightly bound to the atom.

The energy levels are characterized by the energy E, the principal quantum number n, the orbital angular momentum l, the spin angular momentum s, and the total angular momentum j. It is convenient to express the angular momentum in units of $\hbar = h/2\pi$; we will use this unit throughout the rest of this article. The wave function for a particular state is sometimes referred to as the atomic orbital for that state. Lower-case letters are used to designate the orbitals for a single electron, and uppercase letters are used to designate the angular momentum states of the whole atom, which may in general be due to the sum of the orbital and spin angular momenta of several electrons. The following notation is used to characterize the angular momentum of each quantum state:

$$2s + 1_{L_J}$$

where L is the orbital angular momentum and has the possible values 0, 1, 2, 3, ... It is customary to use the letters S, P, D, F, G, H, etc. to designate the respective angular momenta. The first four letters in the series stand for sharp, principal, diffuse, and fine, which were historical descriptions of spectral lines recorded on photographic plates. For each angular momentum L there are $2L + 1$ substates, which are characterized by the projection M of the angular momentum along the z axis, according to the rules of quantum mechanics. For angular momentum L, M can assume the values:

$$-L, -(L-1), \ldots, 0, \ldots, L-1, L$$

For one non-relativistic electron in the absence of a magnetic field, these levels of different L and M all have the same energy and are said to be degenerate. Quantum mechanically, the angular momentum can be pictured as a vector of length $\sqrt{l(l+1)}$, which has a component m along the z axis and lies in a cone with its axis along the z axis. The relationship of the classical angular momentum and the quantum mechanical representation is shown in Fig. 2. The cone represents the equally probable orientations of the angular momentum, and the angular momentum vector is at rest at an indeterminate position in the conical surface. This is shown graphically for $l = 2$ in Fig. 3.

For a single electron, the spin angular momentum has a component of $1/2$ or $-1/2$ along the z axis. The quantum mechanical magnitude is $\sqrt{3}/2$. In general, the total spin S is the vector sum of the spins of several electrons. For each spin S there are $2S + 1$ sublevels. The multiplicity $2S + 1$ is written as a superscript in front of the letter designating the orbital angular momentum. The most common cases are $S = 0, 1/2, 1, 3/2$, called singlet, doublet, triplet, quartet, respectively ($2S + 1 = 1, 2, 3, 4$). The names refer

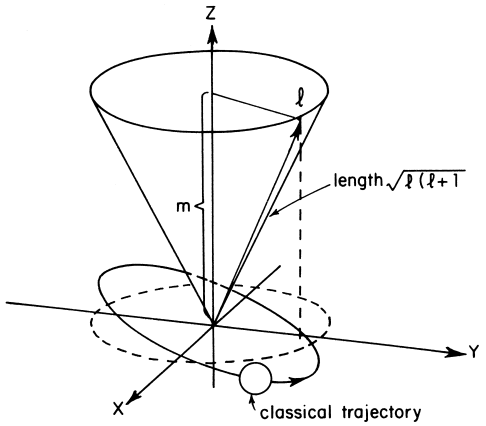


FIGURE 2 The relationship between the classical angular momentum and the quantum-mechanical representation.

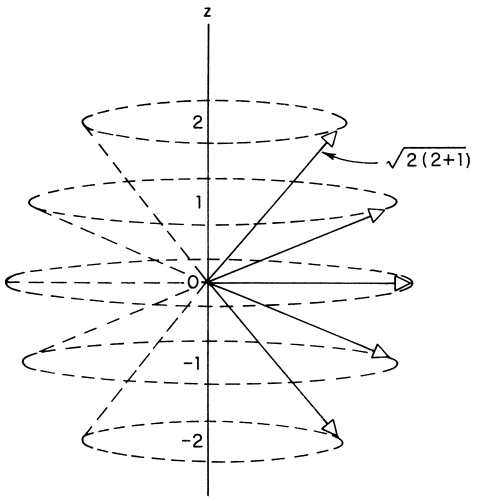


FIGURE 3 The allowed orientations for angular momentum $l = 2$.

to the fact that spectral lines are fine-structure split into that many components, depending on the spin. Hydrogen atoms with one electron are always spin doublets. To obtain the total angular momentum, one adds quantum mechanically the spin angular momentum and the orbital angular momentum. That is,

$$J = L + S$$

The possible values for the total angular momentum J are

$$|L - S|, \ldots, |L + S|$$

The total angular momentum can have either half integral or integral values. For example, for an electron in the 2p state, $L = 1$, $S = 1/2$, $2S + 1 = 2$, and

$$J = 1 + 1/2 = 3/2, \quad \text{or} \quad 1 - 1/2 = 1/2$$

Thus, the 2p electron gives rise to the doublet $(2p)^2P_{1/2}$ and $(2p)^2P_{3/2}$ states. Due to the interaction of the magnetic moment of the electron with the magnetic field produced by the motion of the electron about the proton, the energy of the $^2P_{1/2}$ and $^2P_{3/2}$ states will be different (fine structure splitting). The resultant energy levels for the $n = 2$ and $n = 3$ states of hydrogen are shown in Fig. 4. The near degeneracy of the states with the same total angular momentum is true for both the Schroedinger theory and the Dirac theory. It is due to the inverse square law behavior of the Coulomb force law. This fine structure splitting of the different L orbitals for a given n is much larger in multielectron atoms than in hydrogen, due to the interaction of several electrons.

The wave function of an atom can also be characterized by its parity. The parity is defined by the change in the sign of the wave function when the coordinate system is inverted and (x, y, z) is replaced by $(-x, -y, -z)$. That is,

$$\psi(-x, -y, -z) = \pm\psi(x, y, z)$$

If the multiplying factor is $+1$ or -1, the parity of the wave function is said to be even or odd, respectively. For a given angular momentum state, all the substates have the same parity; the hydrogen atom states with even angular momentum have even parity, and states with odd angular momentum have odd parity.

Higher-order electrodynamic corrections cause a decrease of the effective Coulomb force when the electron is near the proton and result in a smaller binding energy for the S states. This shift in the energy levels is called the *Lamb shift*, denoted by script $\mathscr{L}$ in Fig. 4. One-electron states with the same n and different J and L ought to have the same energy (they ought to be degenerate in energy), according to Schroedinger and Dirac theory. They almost do, but electron magnetic moments and higher-order relativistic effects break the degeneracy and cause, for example, $^2P_{3/2}$ to rise in energy relative to $^2P_{1/2}$ and $^2S_{1/2}$. The Lamb shift also makes the binding energy for the $3^2P_{3/2}$ level slightly smaller than that for the $3^2D_{3/2}$ level, so these energy levels are also no longer degenerate. The Lamb shift comes about in a way similar to the anomalous g value of the electron.

The square of the absolute value of the normalized wave function for each of the quantum states gives the probability per unit volume for finding the electron at each point in space. Figure 5 shows the spatial probability for the 1s, 2s, $2p_0$, $2p_{+1}$, $2p_{-1}$ orbitals, where the subscript denotes the z component of the angular momentum. The s orbitals are spherically symmetric with the greatest probability for the

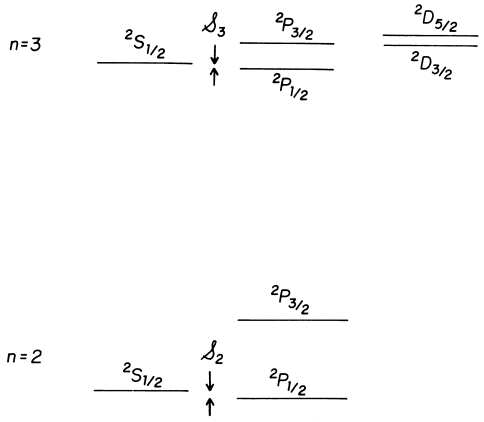


FIGURE 4 The energy levels for the $n = 2$ and $n = 3$ states of hydrogen predicted by the Dirac theory and the Schroedinger theory with inclusion of the electron magnetic moment and special relativity to order $(v/c)^4$. The Lamb shift $\mathscr{L}$ shifts the binding energy for levels of the same J, so they are no longer degenerate.

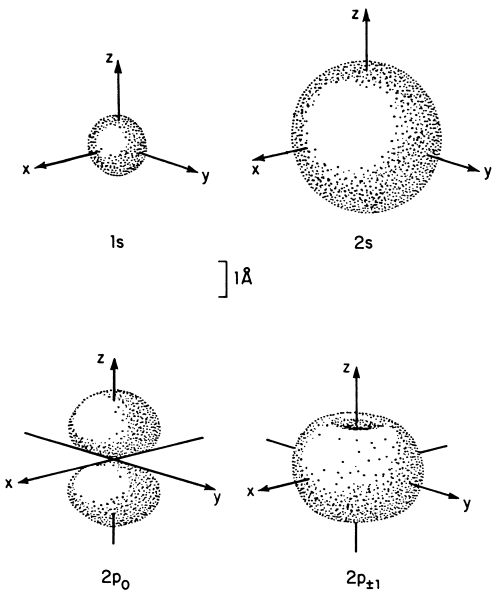


FIGURE 5 Boundary surface plots of the probability density for the 1s, 2s, $2p_0$, and $2p_{\pm 1}$ states of the hydrogen atom. The boundary surface excludes all space in which $|\psi|^2$ is less than one tenth of its maximum value. [Adapted from Bockoff, F. J. (1969). "Elements of Quantum Theory," Addison-Wesley, Reading MA.]

electron to be close to the nucleus. The p orbitals have an anisotropic distribution. The distributions for the p_{+1} and p_{-1}, orbitals are consistent with what one would expect for an electron moving in a circular orbit in the x, y plane, taking into account the fact that the angular momentum vector lies in a cone with a half-angle of 45°. The distribution for the p_0 orbital is not easily visualized in terms of an electron moving in a circular orbit.

The radial probability density $|\psi_{nl}|^2 = P_{nl}$ for the electron in the nl atomic orbital to be in a spherical shell with radius between r and $r + dr$ is given by integrating the probability per unit volume over the volume enclosed between spheres of radii r and $r + dr$. Figure 6 shows the radial probability density for the $n = 1$, 2, and 3 states of hydrogen. The dashed line shows the radius of the corresponding orbit in the Bohr theory.

III. THE HELIUM ATOM

The nucleus of the helium atom consists of two protons and two neutrons. It is sometimes referred to as the alpha particle and was first observed in the decay of heavy radioactive nuclei. The energy levels for He^+, which is a helium nucleus with only one bound eletron, are similar to those for hydrogen, with a larger binding energy due to the higher Z for the nucleus. To a good approximation the energy levels are given by:

$$E_{nl} = -RcZ^2/n^2$$

To obtain the energy levels for He one adds a second electron to the He^+ ion. Since the second electron cannot be in the same quantum state and thus have the same quantum numbers as the first electron, it must be either in a different orbital state or in a different spin state. Since the spin has little effect upon the energy levels, a lower energy is obtained by putting the second electron in the same orbital state as the first electron but with a different value for the projection of the spin angular momentum along the z axis. Thus, the quantum numbers (n, l, m_l, m_s) for the first electron will be $(1, 0, 0, \frac{1}{2})$ and the quantum numbers for the second electron will be $(1, 0, 0, -\frac{1}{2})$. The z component of the total spin angular momentum, which is the sum of the z components of the spin angular momenta of the two electrons, will be zero; it can be shown that the total spin angular momentum will also be zero. The binding energy for the second electron will be smaller than that for the first electron because the first electron tends to shield the

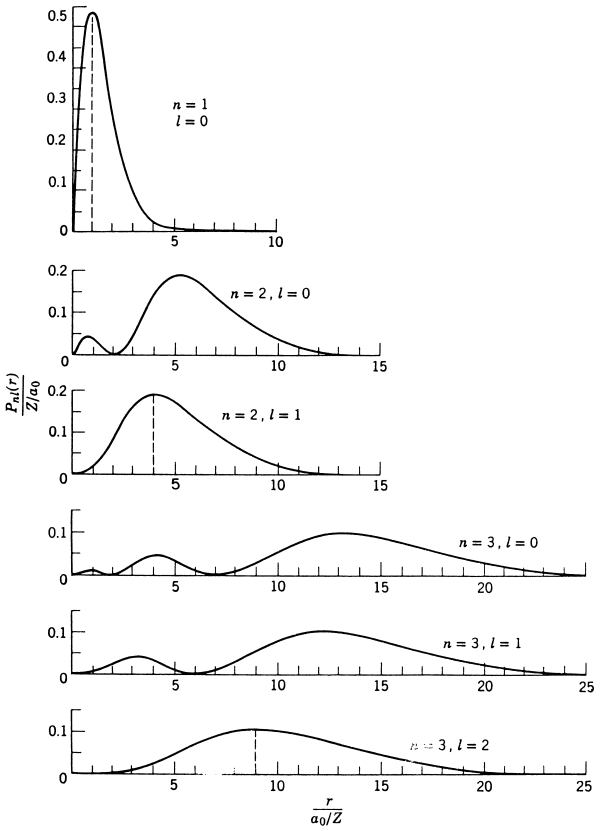


FIGURE 6 The radial probability density for the electron in a one-electron atom for $n = 1$, 2, 3 and all the possible values of l. The dashed lines shows the radius of the corresponding circular orbit in the Bohr theory. [Adapted from Eisberg, R., and Resnick, R. (1985). "Quantum Physics of Atoms, Molecules, Solids, Nuclei, and Particles," John Wiley & Sons, New York.]

nucleus and thus reduce the effective Coloumb field seen by the second electron.

In building up the states for the helium atom, we spoke of the electrons as if they were distinguishable and there were a first electron and a second electron. In reality, the electrons are indistinguishable and we cannot determine which is the first electron and which is the second. The Pauli exclusion principle, which requires that no two electrons occupy the same quantum state, can be reformulated in terms of a requirement that the wave function describing the electrons in any quantum-mechanical system be antisymmetric with respect to the exchange of any two electrons. The corresponding wave function for the ground state of helium is

$$\psi(\mathbf{r}_1, \mathbf{r}_2) = \psi_{10}(\mathbf{r}_1)\psi_{10}(\mathbf{r}_2)\sqrt{\tfrac{1}{2}}\left[\chi_1\left(\tfrac{1}{2}\right)\chi_2\left(-\tfrac{1}{2}\right) - \chi_1\left(-\tfrac{1}{2}\right)\chi_2\left(\tfrac{1}{2}\right)\right]$$

Here, $\psi_{10}(\mathbf{r}_i)$ is the orbital wave function of the ith electron in the ground state and $\chi_i(m_s)$ is the spin wave function for the ith electron with component m_s along the z axis. This wave function retains the indistinguishability of the two electrons. It says that the electrons occupy the same orbital state and that one electron has spin component $+\frac{1}{2}$ along the z axis and the other electron has spin component $-\frac{1}{2}$ along the z axis. It does not say which electron has which spin component.

To obtain the first excited state of helium, one can add the second electron in either a 2s or 2p state of the helium ion with the spin vectors of the two electrons adding to either 0 or 1. Since the two electrons now occupy different orbital states, there is no restriction on the spin states. There are four possible states for the helium atom:

$$(1s)(2s)^1S \qquad (1s)(2s)^3S$$
$$(1s)(2p)^1P \qquad (1s)(2p)^3P$$

where the S and P refer, respectively, to the total orbital angular momentum of the two electrons and the leading superscript 1 or 3 uses the multiplicity $2S + 1$ to designate the total spin. The states are referred to, respectively, as singlet and triplet states. For each of these states one can further add the spin and orbital angular momenta to obtain the total angular momentum. The complete list of possible states is

$$(1s)(2s)^1S_0 \qquad (1s)(2s)^3S_1$$
$$(1s)(2p)^1P_1 \qquad (1s)(2p)^3P_0, {}^3P_1, {}^3P_2$$

For the singlet states, the spin function is antisymmetric under the interchange of the two electrons and the orbital wave function is symmetric; for the triplet states, the spin wave function is symmetric under the interchange of the two electrons and the orbital wave function is antisymmetric. The net result is that in the singlet state the electrons are in general close together and the Coulomb repulsion is larger. This results in a smaller binding energy for the singlet states. The electrons act as if they were subject to a force that depends on the relative orientation of the spins. This quantum-mechanical effect is sometimes referred to as due to the exchange force; it has no classical analog. The electrons are said to be correlated, and the difference in energy of the singlet and triplet states is due to the difference in the Coulomb repulsion between the two electrons due to the electron correlation.

It requires a detailed variational calculation to predict the order of the energy levels. Figure 7 shows the observed energy-level diagram for helium. The parity of the wave function is determined by inverting the coordinates of both electrons.

The interaction that causes a helium atom in an excited state to decay through the emission of photons is due primarily to the interaction between the electromagnetic field and the charge of the electron. The interaction of the electromagnetic field with the spins is much smaller. The net result is that there is little coupling between the

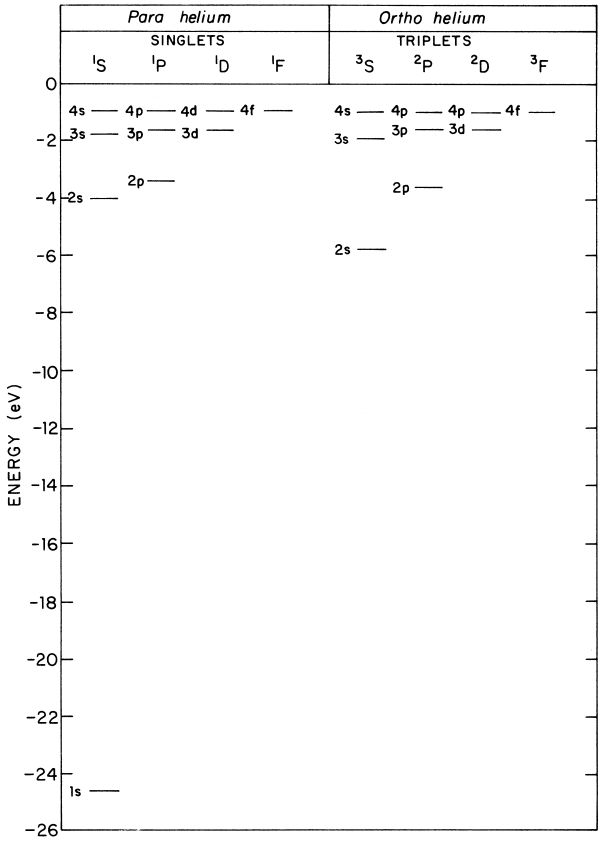


FIGURE 7 The energy levels for the helium atom.

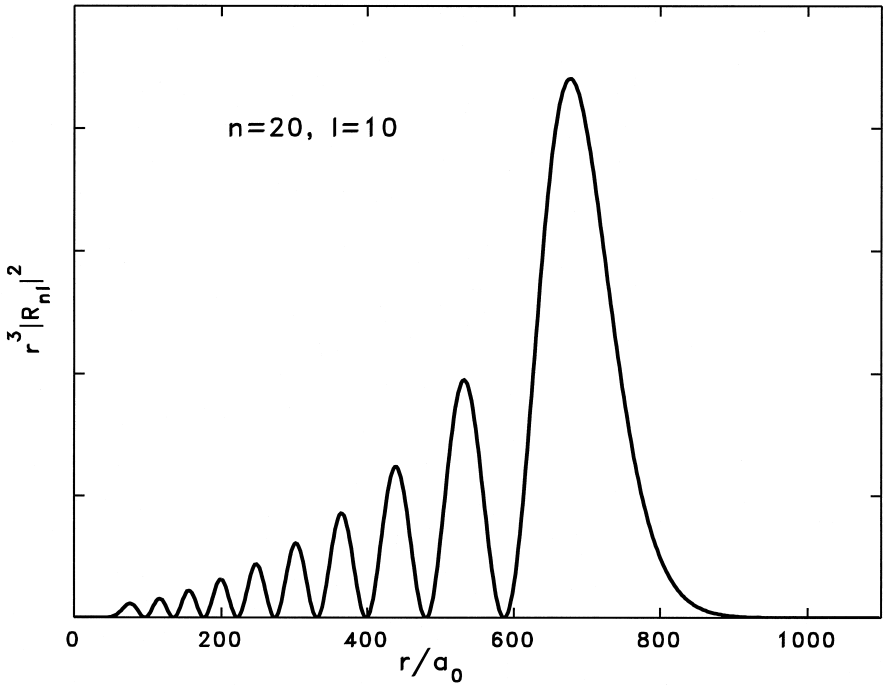


FIGURE 8 Probability density $|\Psi|^2 r^3 = |R_{nl}(r/a_0)|^2 r^3$ for Rydberg H atom, $n = 20$, $l = 10$. (Factor of r^3 takes into account the larger three-dimensional volume at larger radius, to show the probability of finding the electron at a particular r.) Compare with Fig. 6; recall the size of the H atom with $n = 1$ is about $r/a_0 = 1$ (a_0 is the Bohr radius, 5.29×10^{-11} m).

singlet and triplet states. Helium acts as if there were two different species, one with spin 0 and one with spin 1. They are sometimes referred to, respectively, as *para*- and *ortho*-helium. A similar behavior is observed for molecular hydrogen, which has energy levels very similar to those for helium.

A further result of this weak coupling is that the $(1s)(2s)^3S_1$ state of helium has a very small probability for decay to the singlet ground state. It is said to be metastable. In the absence of external perturbations, it decays by single-photon magnetic dipole radiation with a lifetime of 0.841×10^4 sec. This is to be compared with a lifetime of 20 msec for the $(1s)(2s)^1S_0$ state, which decays by the emission of two electric dipole photons. Normally, an allowed one-photon electric dipole decay has a lifetime on the order of nanoseconds.

IV. RYDBERG AND EXOTIC ATOMS

Certain highly excited atoms, with one or more electrons at very high energy, just below the ionization potential energy needed to tear the electron off, are called *Rydberg atoms*. The Rydberg electron is bound to the atom, but just barely so. The principal quantum number n is a measure of the excitation of the Rydberg atom; atoms with n up to 600 or more have been observed. (In principle, n can go up to infinity before the electron becomes unbound or ionized; however, then the Rydberg atom becomes extraordinarily fragile and very difficult to measure.)

Rydberg atoms have a number of interesting properties. Since the electron is nearly unbound (the binding energy drops as n^{-2}), it moves quite far away from the nucleus, as shown in Fig. 8. The size of a Rydberg atom goes as n^2, and its cross-section goes as n^4. For very high n values, the electron orbital radius can be several microns, almost macroscopic in size. The distance and weak interaction of the Rydberg electron with the nucleus mean that all Rydberg atoms are very similar to H atoms. Even with an extended and more complex core such as Na^+, or even with a molecular core such as H_2^+, the distant electron "sees" a point source of positive charge to a good approximation, just as the electron in an hydrogen atom does. Thus, most of the quantum mechanics mathematical apparatus and notation developed for the hydrogen atom can be used. The radiative lifetime of a Rydberg electron is calculated according to the usual hydrogen atom electric dipole matrix elements, and varies as n^3. That is, as n goes up, the Rydberg electron becomes less and less likely to radiate. This is explained physically as an isolation of the Rydberg electron from the charge center of the nucleus, so it acts more and more like an isolated free electron,

which does not radiate. The polarizability of the "floppy" Rydberg atom can be very large, and goes as n^7.

Since the Rydberg electron interacts weakly with the rest of the atom, perturbation theory methods can be used to calculate various properties that would be impossible to calculate for a low-lying, strongly interacting electron. The Rydberg electron can act as a sensitive probe of various core properties such as polarizability and quadrupole moment of the ion core.

The weak interaction between the Rydberg electron and its core can allow a relatively slow transfer of energy between the two. When an excited core transfers energy to the Rydberg electron, usually giving it enough to be ionized, the process is called *autoionization*. In this, the core loses the energy the Rydberg electron gains. When a Rydberg electron loses energy, which is transferred to the core, the process is called *dielectronic recombination*. This is the most important mechanism in plasmas whereby free electrons and ions combine to form a normal hot gas. These two process are the time reversal of each other, but are described by the same mathematics.

Since n is very high for Rydberg states, and l and m can take on a large number of values, typically a very large number of Rydberg states are available, all at nearly the same energy close to but just below the ionization potential. (Autoionizing states, counting the core energy, actually lie above the ionization potential.) There are so many states close in energy that usually they overlap in energy and interfere with each other in a quantum mechanical way, leading to very complex situations and to a so-called quasi-continuum of states.

An "atom" of positronium is formed by an electron and its antiparticle, a positron. Although the two eventually annihilate each other (via the overlapping of their wave functions), they can live for up to 10^{-7} sec, orbiting each other very much like in a hydrogen atom, except the reduced mass is half that of a normal hydrogen atom's electron. While positronium exists, it can absorb and emit photons with a spectrum similar to atomic hydrogen, except all wavelengths are doubled relative to atomic hydrogen. The lifetime of the positron is sensitive to the details of the wave function and so can probe the inside solid-state systems of the wave function. The quantum mechanical state labels of He apply to positronium.

A muonic atom is formed by a normal atom with one electron replaced by a negative muon, which is very similar to an electron but weighs 207 times as much. The muon in an atom has a wave function and transitions just as the electron does, but the much higher mass means that the energies are higher and the wave functions are "tighter" (occupy less space). In fact, a significant fraction of the muonic wave function exists inside the nucleus of the muonic atom, thus muonic atoms are used to probe the exact spatial distribution of mass and charge of the nucleus, especially near the edge of the nucleus. Muonic molecules exist, with the muon pulling two nuclei very close together in a chemical bond. Unfortunately, thermalized muons are difficult to produce, and muons are unstable and only live 2 μsec before decaying, so muon catalyzed fusion of hydrogen nuclei has been observed but is not efficient.

An antiproton and a positron form the exotic atom of antihydrogen. This atom has been formed and detected at the high energies commensurate with the formation of the antiproton. A great deal of work is ongoing to slow down the antihydrogen to thermal or less energies and even to laser trap them. An atom consisting of a normal He^{++} nucleus, an antiproton, and an electron (antiprotonic helium) has been spectroscopically measured, as it is easier to form than plain antihydrogen. From the measured Rydberg constant of antiprotonic helium, the antiproton mass has been measured to be the same as that of a normal proton to within 3 ppm.

V. COMPLEX ATOMS

The scheme used to construct the energy levels of helium can be generalized and used to construct the energy levels of more complex atoms with many electrons. One adds electrons one at a time, with each electron placed in the unoccupied orbital with the lowest energy in accord with the Pauli exclusion principle. To first order it can be assumed that the electrons are independent and that the order for filling the levels is the same as that for hydrogen. When all the orbitals for a given angular momentum with different z components of angular momentum are filled, one obtains a spherically symmetric configuration with total angular momentum equal to zero, which is referred to as a *closed shell*. Helium, which has two electrons in the 1s orbital, is the first atom with a closed shell. The ground state of helium has total orbital angular momentum equal to zero, total spin angular momentum equal to zero, and total angular momentum equal to zero. The second closed shell is for two electrons in the 2s orbital; this is the beryllium atom. The third closed shell is for six electrons in the 2p orbital; this is the neon atom. The complete configuration of neon is $1s^2 2s^2 2p^6$. The closed shells that correspond to the filling of all the levels with the same principle quantum number in hydrogen are major closed shells that correspond to the rare gases. This accounts for their inertness and failure to be chemically active.

As Z increases, the order of filling of electrons in energy levels is modified from that for the energy levels of hydrogen, due to the Coulomb interaction between the electrons. Figure 9 shows the empirical order of filling

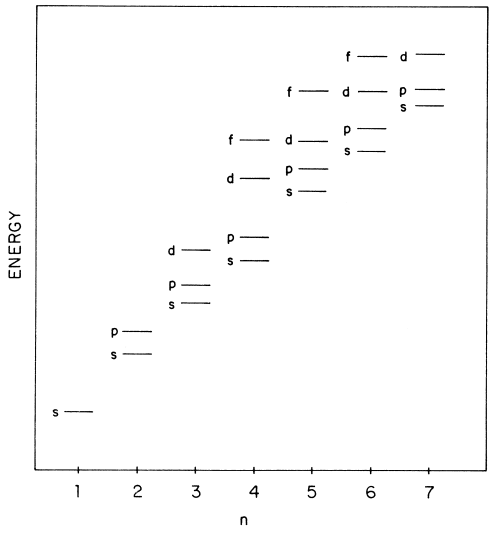


FIGURE 9 The order of filling of the levels for many-electron atoms.

of the levels. Table II gives the ground-state orbital electronic configuration, that is, the single electron orbitals that are filled for that atom. All the rest of the (higher energy) orbitals exist but are not filled. Since the electrons are not really single electrons, as is implicitly assumed in column four of Table II, the table gives the ground-state designation (in the format $^{2S+1}L_J$, LS coupling) for the entire number of electrons, whose various angular momenta are coupled to form single S, L, and J values. Table II also gives for each element the first ionization potential, which is the energy required to remove the most weakly bound (highest energy or outer) ground-state electron. The more tightly bound atoms have higher ionization potentials.

It has not been possible to solve exactly the Schroedinger equation for a many electron system. Techniques such as the Hartree consistent field method have been developed to take into account the interaction between the electrons. The Hartree method uses a wave function that is the product of single-particle wave functions. A guess is made of the wave function for each electron, and these wave functions are used to calculate the electrostatic potential due to the charge distribution of the electrons. These potentials are averaged over angles and summed over all the electrons but one. This summed potential is used in the Schroedinger equation for the remaining electron to solve for the wave function for that electron. This procedure is repeated for each electron, and a new wave function for each electron, and a new wave function for each electon is obtained. The calculation is iterated until there is no change in the wave functions. Figure 10 shows the calculated wave functions for the argon atom. A more elaborate procedure called the Hartree–Fock method has been developed to take proper account of the Pauli principle.

To obtain the angular momentum wave function of a complex atom, one must add together the spin and orbital angular momenta of the individual electrons outside the closed shells to obtain the total angular momentum of the atom. Two approximate coupling schemes are used, Russell–Saunders (or LS) coupling and jj coupling. In Russell–Saunders coupling, the orbital angular momenta of the individual electrons are added together to obtain the total orbital angular momentum of the complex atom, and the spin angular momenta of the individual electrons are added together to form the total spin angular momentum. The total orbital angular momentum is then added to the total spin angular momentum to obtain the total angular momentum. According to Hund's rule, the state with the highest multiplicity is the lowest energy state. In jj coupling, for each electron the orbital and the spin angular momentum are added together to form the total angular momentum for that electron. The total angular momenta of the individual electrons are then added together to obtain the total angular momentum for the atom. Low-Z atoms are best described by Russell–Saunders coupling, and high-Z atoms by jj coupling. Helium provides a good example of Russell–Saunders coupling. The parity of the wave function for a many-electron atom is determined by inverting the coordinates of all the electrons. The rule for determining parity is given in Footnote b of Table II. As the electrons in an atom become excited to

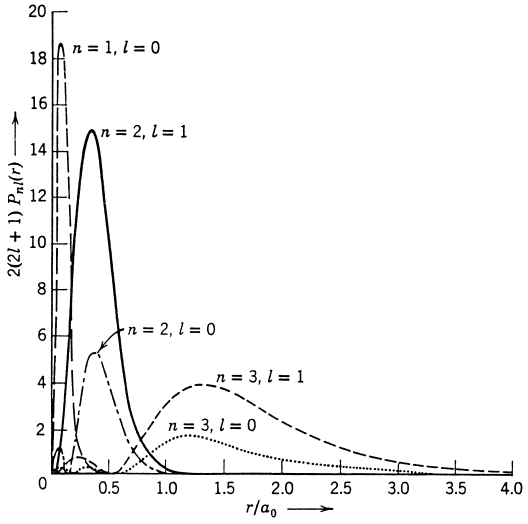


FIGURE 10 The Hartree theory radial probability densities for the filled quantum states of the argon atom. [Adapted from Eisberg, R., and Resnick, R. (1985). "Quantum Physics of Atoms. Molecules, Solids, Nuclei, and Particles," John Wiley & Sons, New York.]

TABLE II Ground Electronic Configurations and Ionization Energies for the Elements

Z	Abbr.	Atom	Orbital electronic configuration[a]	Ground-state designation	Ionization potential (eV)
1	H	Hydrogen	1s	$^2S_{1/2}$	13.5984
2	He	Helium	$1s^2$	$^1S_0$	24.5874
3	Li	Lithium	[He] 2s	$^2S_{1/2}$	5.3917
4	Be	Beryllium	[He] $2s^2$	$^1S_0$	9.3227
5	B	Boron	[He] $2s^2$ 2p	$^2P_{1/2}$	8.2980
6	C	Carbon	[He] $2s^2$ $2p^2$	$^3P_0$	11.2603
7	N	Nitrogen	[He] $2s^2$ $2p^3$	$^4S_{3/2}$	14.5341
8	O	Oxygen	[He] $2s^2$ $2p^4$	$^3P_2$	13.6181
9	F	Fluorine	[He] $2s^2$ $2p^5$	$^2P_{3/2}$	17.4228
10	Ne	Neon	[He] $2s^2$ $2p^6$	$^1S_0$	21.5646
11	Na	Sodium	[Ne] 3s	$^2S_{1/2}$	5.1391
12	Mg	Magnesium	[Ne] $3s^2$	$^1S_0$	7.6462
13	Al	Aluminum	[Ne] $3s^2$ 3p	$^2P_{1/2}$	5.9858
14	Si	Silicon	[Ne] $3s^2$ $3p^2$	$^3P_0$	8.1517
15	P	Phosphorus	[Ne] $3s^2$ $3p^3$	$^4S_{3/2}$	10.4867
16	S	Sulfur	[Ne] $3s^2$ $3p^4$	$^3P_2$	10.3600
17	Cl	Chlorine	[Ne] $3s^2$ $3P^5$	$^2P_{3/2}$	12.9676
18	Ar	Argon	[Ne] $3s^2$ $3p^6$	$^1S_0$	15.7596
19	K	Potassium	[Ar] 4s	$^2S_{1/2}$	4.3407
20	Ca	Calcium	[Ar] $4s^2$	$^1S_0$	6.1132
21	Sc	Scandium	[Ar] 3d $4s^2$	$^2D_{3/2}$	6.5615
22	Ti	Titanium	[Ar] $3d^2$ $4s^2$	$^3F_2$	6.8281
23	V	Vanadium	[Ar] $3d^3$ $4s^2$	$^4F_{3/2}$	6.7462
24	Cr	Chromium	[Ar] $3d^5$ 4s	$^7S_3$	6.7665
25	Mn	Manganese	[Ar] $3d^5$ $4s^2$	$^6S_{5/2}$	7.4340
26	Fe	Iron	[Ar] $3d^6$ $4s^2$	$^5D_4$	7.9024
27	Co	Cobalt	[Ar] $3d^7$ $4s^2$	$^4F_{9/2}$	7.8810
28	Ni	Nickel	[Ar] $3d^8$ $4s^2$	$^3F_4$	7.6398
29	Cu	Copper	[Ar] $3d^{10}$ 4s	$^2S_{1/2}$	7.7264
30	Zn	Zinc	[Ar] $3d^{10}$ $4s^2$	$^1S_0$	9.3942
31	Ga	Gallium	[Ar] $3d^{10}$ $4s^2$ 4p	$^2P_{1/2}$	5.9993
32	Ge	Germanium	[Ar] $3d^{10}$ $4s^2$ $4p^2$	$^3P_0$	7.8994
33	As	Arsenic	[Ar] $3d^{10}$ $4s^2$ $4p^3$	$^4S_{3/2}$	9.7886
34	Se	Selenium	[Ar] $3d^{10}$ $4s^2$ $4p^4$	$^3P_2$	9.7524
35	Br	Bromine	[Ar] $3d^{10}$ $4s^2$ $4p^5$	$^2P_{3/2}$	11.8138
36	Kr	Krypton	[Ar] $3d^{10}$ $4s^2$ $4p^6$	$^1S_0$	13.9996
37	Rb	Rubidium	[kr] 5s	$^2S_{1/2}$	4.1771
38	Sr	Strontium	[Kr] $5s^2$	$^1S_0$	5.6949
39	Y	Yttrium	[Kr] 4d $5s^2$	$^2D_{3/2}$	6.2171
40	Zr	Zirconium	[Kr] $4d^2$ $5s^2$	$^3F_2$	6.6339
41	Nb	Niobium	[Kr] $4d^4$ 5s	$^6D_{1/2}$	6.7589
42	Mo	Molybdenum	[Kr] $4d^5$ 5s	$^7S_3$	7.0924
43	Tc	Technetium	[Kr] $4d^5$ $5s^2$	$^6S_{5/2}$	7.28
44	Ru	Ruthenium	[Kr] $4d^7$ 5s	$^5F_5$	7.3605
45	Rh	Rhodium	[Kr] $4d^8$ 5s	$^4F_{9/2}$	7.4589
46	Pd	Palladium	[Kr] $4d^{10}$	$^1S_0$	8.3369
47	Ag	Silver	[Kr] $4d^{10}$ 5s	$^2S_{1/2}$	7.5762

Continues

TABLE II (*continued*)

Z	Abbr.	Atom	Orbital electronic configuration[a]	Ground-state designation	Ionization potential (eV)
48	Cd	Cadmium	[Kr] $4d^{10}\ 5s^2$	$^1S_0$	8.9938
49	In	Indium	[Kr] $4d^{10}\ 5s^2\ 5p$	$^2P_{1/2}$	5.7864
50	Sn	Tin	[Kr] $4d^{10}\ 5s^2\ 5p^2$	$^3P_0$	7.3439
51	Sb	Antimony	[Kr] $4d^{10}\ 5s^2\ 5p^3$	$^4S_{3/2}$	8.6084
52	Te	Tellurium	[Kr] $4d^{10}\ 5s^2\ 5p^4$	$^3P_2$	9.0096
53	I	Iodine	[Kr] $4d^{10}\ 5s^2\ 5p^5$	$^2P_{3/2}$	10.4513
54	Xe	Xenon	[Kr] $4d^{10}\ 5s^2\ 5p^6$	$^1S_0$	12.1298
55	Cs	Cesium	[Xe] 6s	$^2S_{1/2}$	3.8939
56	Ba	Barium	[Xe] $6s^2$	$^1S_0$	5.2117
57	La	Lanthanum	[Xe] 5d $6s^2$	$^2D_{3/2}$	5.5769
58	Ce	Cerium	[Xe] 4f 5d $6s^2$	$^1G_4$	5.5387
59	Pr	Praseodymium	[Xe] $4f^3\ 6s^2$	$^4I_{9/2}$	5.473
60	Nd	Neodymium	[Xe] $4f^4\ 6s^2$	$^5I_4$	5.5250
61	Pm	Promethium	[Xe] $4f^5\ 6s^2$	$^6H_{5/2}$	5.582
62	Sm	Samarium	[Xe] $4f^6\ 6s^2$	$^7F_0$	5.6436
63	Eu	Europium	[Xe] $4f^7\ 6s^2$	$^8S_{7/2}$	5.6704
64	Gd	Gadolinium	[Xe] $4f^7$ 5d $6s^2$	$^9D_2$	6.1501
65	Tb	Terbium	[Xe] $4f^9\ 6s^2$	$^6H_{15/2}$	5.8638
66	Dy	Dysprosium	[Xe] $4f^{10}\ 6s^2$	$^5I_8$	5.9389
67	Ho	Holmium	[Xe] $4f^{11}6s^2$	$^4I_{15/2}$	6.0215
68	Er	Erbium	[Xe] $4f^{12}6s^2$	$^3H_6$	6.1077
69	Tm	Thulium	[Xe] $4f^{13}\ 6s^2$	$^2F_{7/2}$	6.1843
70	Yb	Ytterbium	[Xe] $4f^{14}\ 6s^2$	$^1S_0$	6.2542
71	Lu	Lutetium	[Xe] $4f^{14}$ 5d $6s^2$	$^2D_{3/2}$	5.4259
72	Hf	Hafnium	[Xe] $4f^{14}\ 5d^2\ 6s^2$	$^3F_2$	6.8251
73	Ta	Tantalum	[Xe] $4f^{14}\ 5d^3\ 6s^2$	$^4F_{3/2}$	7.5496
74	W	Tungsten	[Xe] $4f^{14}\ 5d^4\ 6s^2$	$^5D_0$	7.8640
75	Re	Rhenium	[Xe] $4f^{14}\ 5d^5\ 6s^2$	$^6S_{5/2}$	7.8335
76	Os	Osmium	[Xe] $4f^{14}\ 5d^6\ 6s^2$	$^5D_4$	8.4382
77	Ir	Iridium	[Xe] $4f^{14}\ 5d^7\ 6s^2$	$^4F_{9/2}$	8.9670
78	Pt	Platinum	[Xe] $4f^{14}\ 5d^9$ 6s	$^3D_3$	8.9587
79	Au	Gold	[Xe] $4f^{14}\ 5d^{10}$ 6s	$^2S_{1/2}$	9.2255
80	Hg	Mercury	[Xe] $4f^{14}\ 5d^{10}\ 6s^2$	$^1S_0$	10.4375
81	Tl	Thallium	[Xe] $4f^{14}\ 5d^{10}\ 6s^2$ 6p	$^2P_{1/2}$	6.1082
82	Pb	Lead	[Xe] $4f^{14}\ 5d^{10}\ 6s^2\ 6p^2$	$^3P_0$	7.4167
83	Bi	Bismuth	[Xe] $4f^{14}\ 5d^{10}\ 6s^2\ 6p^3$	$^4S_{3/2}$	7.2856
84	Po	Polonium	[Xe] $4f^{14}\ 5d^{10}\ 6s^2\ 6p^4$	$^3P_2$	8.417 (?)
85	At	Astatine	[Xe] $4f^{14}\ 5d^{10}\ 6s^2\ 6p^5$	$^2P_{3/2}$	(?)
86	Rn	Radon	[Xe] $4f^{14}\ 5d^{10}\ 6s^2\ 6p^6$	$^1S_0$	10.7485
87	Fr	Francium	[Rn] 7s	$^2S_{1/2}$	4.0727
88	Ra	Radium	[Rn] $7s^2$	$^1S_0$	5.2784
89	Ac	Actinium	[Rn] 6d $7s^2$	$^2D_{3/2}$	5.17
90	Th	Thorium	[Rn] $6d^2\ 7s^2$	$^3F_2$	6.3067
91	Pa	Protactinium	[Rn] $5f^2$ 6d $7s^2$	—[b]	5.89
92	U	Uranium	[Rn] $5f^3$ 6d $7s^2$	—[b]	6.1941
93	Np	Neptunium	[Rn] $5f^4$ 6d $7s^2$	—[b]	6.2657
94	Pu	Plutonium	[Rn] $5f^6\ 7s^2$	$^7F_0$	6.0262
95	Am	Americium	[Rn] $5f^7\ 7s^2$	$^8S_{7/2}$	5.9738

Continues

TABLE II (*continued*)

Z	Abbr.	Atom	Orbital electronic configuration[a]	Ground-state designation	Ionization potential (eV)
96	Cm	Curium	[Rn] $5f^7$ $6d$ $7s^2$	$^9D_2$	5.9915
97	Bk	Berkelium	[Rn] $5f^9$ $7s^2$	$^6H_{15/2}$	6.1979
98	Cf	Californium	[Rn] $5f^{10}$ $7s^2$	$^5I_8$	6.2817
99	Es	Einsteinium	[Rn] $5f^{11}$ $7s^2$	$^4I_{15/2}$	6.42
100	Fm	Fermium	[Rn] $5f^{12}$ $7s^2$	$^3H_6$	6.50
101	Md	Mendelevium	[Rn] $5f^{13}$ $7s^2$	$^2F_{7/2}$	6.58
102	No	Nobelium	[Rn] $5f^{14}$ $7s^2$	$^1S_0$	6.65
103	Lr	Lawrencium	[Rn] $5f^{14}$ $7s^2$ 7p (?)	$^2P_{1/2}$ (?)	4.9 (?)
104	Rf	Rutherfordium	[Rn] $5f^{14}$ $6d^2$ $7s^2$ (?)	$^3F_2$ (?)	6.0 (?)

[a] Configurations with an odd number of odd l (p or f) electrons have odd parity. All other configurations have even parity.

[b] These atoms cannot be correctly described by an *LS* coupled configuration.

Adapted from "Ground Levels and Ionization Energies for the Neutral Atoms," NIST website www.physics.nist.gov/PhysRefData/IonEnergy/tb1New.html.

higher energy and different wavefunctions, the parity will change.

There is additional structure in the energy levels of many atoms due to the interaction between the electrons and the magnetic and quadrupole moments of the nucleus. For these cases the nucleus has angular momentum and one must add the nuclear angular momentum to the electronic angular momentum to obtain the total angular momentum. Measurements of the hyperfine structure are used to determine the magnetic and quadrupole moments of nuclei.

VI. INTERACTION OF ATOMS WITH RADIATION

The interaction of an atom with an external electromagnetic field results in the emission and absorption of radiation. An atom in an excited state will decay exponentially with a mean life that depends on the available energy and the character of the initial and final states. The dominant form of radiation is electric dipole radiation; it takes place only between states with opposite parity.

In order to describe the interaction of atoms with the radiation field, it is necessary to quantize the electromagnetic field. The universe is pictured as a large cubic box with reflecting walls, and the set of solutions used to describe the electromagnetic field is taken to be the set of plane waves that satisfy the boundary conditions for a box with perfectly reflecting walls. The electromagnetic field is then expanded in a Fourier series in terms of these plane waves. The field can be specified by giving the amplitude of each of the Fourier components. For each mode there are two polarizations with the polarization vectors perpendicular to the direction of propagation of the plane wave. Due to the form of Maxwell's equations, the time dependence of the amplitude of these individual modes is the same as that for a simple harmonic oscillator whose frequency is the frequency of the Fourier component. A familiar example of a simple harmonic oscillator is a weight at the end of a spring where the restoring force due to the spring is proportional to its extension from the equilibrium configuration. To quantize the electromagnetic field, one uses the same technique used to quantize the simple harmonic oscillator involving so-called raising and lowering operators. [*See* QUANTUM MECHANICS.] This quantization gives for each mode a set of equally spaced energy levels whose separation is $h\nu$, where ν is the frequency of the Fourier component. These energy levels are depicted in Fig. 11. Each excitation corresponds to one photon. If for the oscillator with frequency ν, direction of propagation $\mathbf{k}$, and polarization $\mathbf{e}$, the nth energy level is occupied, then there are n photons with frequency ν and polarization $\mathbf{e}$ propagating in direction $\mathbf{k}$.

In the electric dipole approximation, the Hamiltonian interaction between the electrons in the atom and the electromagnetic field has the form:

$$\mathcal{H}_I = -\sum_i e\mathbf{r}_i \cdot \mathbf{E}(\mathbf{r}_i)$$

where $\mathbf{r}_i$ is the coordinate of the ith electron and $\mathbf{E}(\mathbf{r}_i)$ is the external electromagnetic field at the position of the

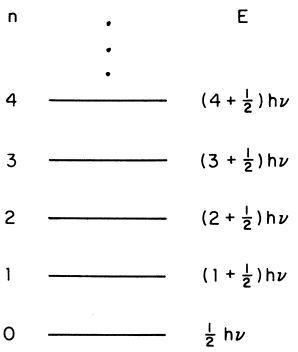


FIGURE 11 The energy levels for the simple harmonic oscillator. $\frac{1}{2}h\nu$ is the zero point energy.

ith electron. The interaction of an atom with the radiation field results in the emission and absorption of radiation. If an atom is in an excited state and there are no photons present, the interaction will cause the atom to decay to a lower state, with a characteristic mean life dependent on the energy difference between the levels and the wave functions describing the initial and final states of the atom. This emission of radiation in the absence of external radiation is called *spontaneous emission*.

Electric dipole transitions take place only between states of opposite parity, and only certain changes in the quantum numbers between the initial and final states are allowed. The selection rules are summarized in Table III. The probability for the atom to be in the upper state decreases exponentially with time. The radiation emitted by an atom is not monochromatic but has a frequency distribution given by the equation

$$I(\nu) = I_0\left(\frac{\gamma}{4\pi^2}\right)\frac{1}{(\nu-\nu_0)^2+(\gamma/4\pi)^2}$$

This distribution was first discussed by Lorentz, and it is called the Lorentzian line profile. The mean life τ for decay is related to the width of the distribution by the equation:

$$\tau = \frac{1}{\gamma}$$

The Lorentzian line profile can be explained by a simple classical picture: the excited electron is treated as sinusoidal oscillator whose amplitude is exponentially decreasing, as shown in Fig. 12. This amplitude is given by the formula:

$$A(t) = A_0\cos(\nu t)\,e^{-(\gamma/4\pi)t}$$

The squared Fourier transform of this amplitude is simply the Lorentz form in Fig. 12. (The reverse is also true—the Fourier transform of the Lorentzian is an exponentially decaying sinusoidal.) Of course, a real quantum atom does not gradually decay, but rather decays suddenly, with the exact time of decay given by an exponential decay distribution for an ensemble of atoms.

If an atom is struck by a plane electromagnetic wave whose frequency ν is related to the difference in energy of the two levels E_1 and E_2 through the Bohr relation:

$$\nu = (E_1 - E_2)/h$$

and there is an allowed dipole transition connecting the two levels, then an atom in the lower state can be excited into the higher state with the absorption of a photon or an atom in the upper state can be stimulated to make a transition to the lower state with the emission of a photon with the same frequency, polarization, and direction as the incident electromagnetic wave. These two processes are referred to, respectively, as *absorption* and *stimulated emission*. It is the latter process that makes possible the laser.

The interaction of the atom with the electromagnetic field also produces corrections to the energy levels. These corrections arise through higher-order processes in which virtual photons are emitted and reabsorbed or a photon changes for a short time into a virtual electron–positron pair. The understanding and calculation of these energy shifts is one of the major triumphs of quantum electrodynamics (QED). The first convincing experimental observation of the need for such corrections was the measurement of the Lamb shift in hydrogen. According to the Dirac theory, in the $n=2$ state of hydrogen the $(2s)^2S_{1/2}$ and $(2p)^2P_{1/2}$ state have exactly the same energy. In an experiment carried out at Columbia University by Willis Lamb and his coworkers, it was found that the $(2s)^2S_{1/2}$ state was less bound than the $(2p)^2P_{1/2}$ state by $\frac{1}{10}$ of the fine structure splitting between the $(2p)^2P_{1/2}$ and $(2p)^2P_{3/2}$ states (see Fig. 4). It was shown later that this shift agrees with the predictions of quantum electrodynamics. The

TABLE III Selection Rules for Electric Dipole Radiation

For a single electron
$\Delta l = \pm 1$
$\Delta m = 0, \pm 1$
$\Delta j = 0, \pm 1$ but not $j = 0$ to $j = 0$
$\Delta m_j = 0, \pm 1$ but for $\Delta j = 0$, not $m_j = 0$ to $m_j = 0$
Parity must change
For a many-electron configuration
$\Delta L = 0, \pm 1$ but not $L = 0$ to $L = 0$
$\Delta M_L = 0, \pm 1$
$\Delta J = 0, \pm 1$ but not $J = 0$ to $J = 0$
$\Delta M_J = 0, \pm 1$ but for $\Delta J = 0$, not $M_J = 0$ to $M_J = 0$
Parity must change

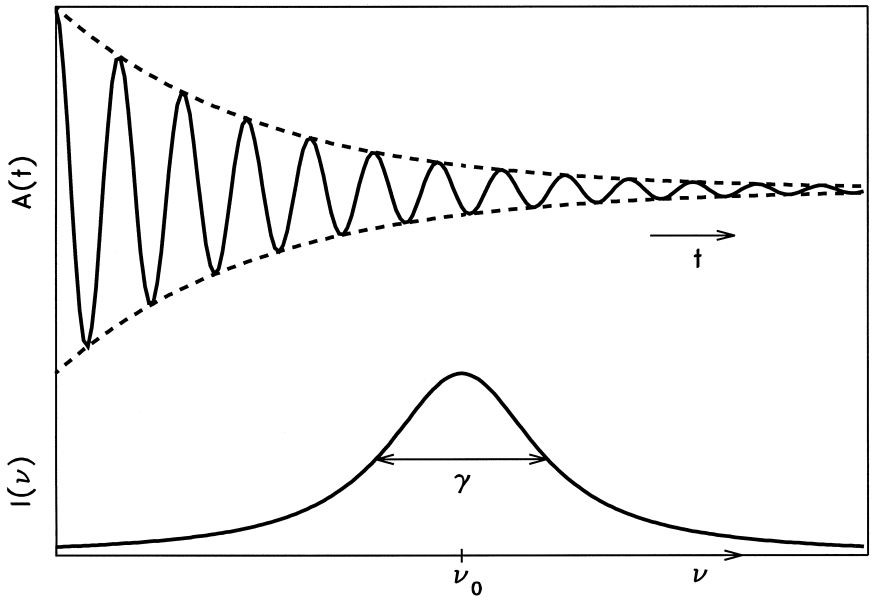


FIGURE 12 Exponentially decaying sinusoidal with frequency ν_0 and decay constant (as a function of t) and Lorentzian line shape with center frequency ν_0 and width (as a function of ν). Each is the Fourier transform of the other.

electron in the $(2s)^2S_{1/2}$ state spends more time near the proton than an electron in the $(2p)^2P_{1/2}$ state. Due to the emission and reabsorption of virtual photons, the electron is pushed away from the proton so that its interaction with the Coulomb field due to the proton is weaker. This decreases the binding energy for the $(2s)^2S_{1/2}$ state more than for the $(2p)^2P_{1/2}$ state and results in a shift of the $(2p)^2S_{1/2}$ state upward from the $(2p)^2P_{1/2}$ state. The Lamb shift has now been measured to be 1057.8446 (29) MHz within 3 ppm, with satisfactory agreement between experiment and theory.

The most precisely measured higher-order effect due to the electromagnetic field is the correction to the magnetic moment of the electron. The magnetic moment of the electron differs by a small amount from the value predicted by the Dirac theory due to the interaction with the electromagnetic field.

Monumental calculations by Kinoshita and others of up to eighth-order radiative and self-energy QED corrections to the electron in the virtual radiation field give a value of the ratio g, the electron magnetic moment to the Bohr magneton, to be

$$\frac{g-2}{2} = 1\ 159\ 652\ 201\ (27) \times 10^{-12}$$

The experimental value of the g-factor, measured by Dehmelt and others by trapping and following for months a single electron in a magnetic trap, is

$$\frac{g-2}{2} = 1\ 159\ 652\ 187\ (4) \times 10^{-12}$$

which is in outstanding agreement with theory (with the theory about seven times less accurate).

SEE ALSO THE FOLLOWING ARTICLES

ATOMIC SPECTROMETRY • COLLISION-INDUCED SPECTROSCOPY • NUCLEAR PHYSICS • QUANTUM MECHANICS • QUANTUM OPTICS • QUANTUM THEORY

BIBLIOGRAPHY

Atkins, P. W. (1974). "Quanta: A Handbook of Concepts," Oxford Univ. Press, London.

Atkins, P. W., and Friedman, R. S. (1997). "Molecular Quantum Mechanics," Oxford Univ. Press, London.

Bates, D. R., and Bederson, B. (1975–1985). "Advances in Atomic and Molecular Physics," Vols. 11–20, Academic Press, New York.

Bernath, P. F. (1995). "Spectra of Atoms and Molecules," Oxford Univ. Press, London.

Bethe, H. A., and Salpeter, E. (1957). "Quantum Mechanics of One- and Two-Electron Atoms," Springer-Verlag, Berlin/New York.

Bockoff, F.J. (1969). "Elements of Quantum Theory," Addison-Wesley, Reading, MA.

Corney, A. (1977). "Atomic and Laser Spectroscopy," Oxford Univ. Press, London.

Cowan, R. D. (1981). "The Theory of Atomic Structure and Spectra," University of California Press, Berkeley, CA.
Eisberg, R., and Resnick, R. (1985). "Quantum Physics of Atoms, Molecules, Solids, Nuclei, and Particles," Wiley, New York.
Liboff, R. L. (1998). "Introductory Quantum Mechanics," Addison-Wesley, Reading, MA.
Lindgren, I., and Morrison, J. (1982). "Atomic Many-Body Theory," Springer-Verlag, Berlin/New York.
NIST Atomic Physics Database. (2000). "www.physics.nist.gov/PhysRefData/contents.html," National Institute of Standards and Technology, Gaithersburg, MD.
Van Dyck, R. S., Jr., and Fortson, E. N., eds. (1984). "Atomic Physics," Vol. 9, World Scientific, Singapore.
Weissbluth, M. (1978). "Atoms and Molecules," Academic Press, New York.



Collider Detectors for Multi-TeV Particles

C. W. Fabjan
CERN

I. Overview
II. Collider Detector Components: A Primer
III. Designing a Collider Detector
IV. Constructing a Collider Detector
V. Closing Remarks

GLOSSARY

Accelerator, circular A machine which increases the kinetic energy of particles (e.g., electrons, protons, and their antiparticles); magnetic fields are used to guide them on a circular path many times through the same accelerating system.

Accelerator, linear A machine which accelerates particles in a straight line.

Calorimeter A particle detector which measures the energy of an elementary particle by absorbing the particle and converting its energy into a measurable signal.

Collider, circular A machine in which two circular accelerators are combined to accelerate and store beams moving in opposite directions. It is used to produce very high energy phenomena in the collisions between particles in the beams.

Collider, linear A machine consisting of two linear accelerators which accelerate beams in opposite directions.

Collider detector (CD) A complex instrument used to measure the momenta and energies of the particles produced in a collider collision. Layers of detectors dedicated to specific measurement tasks surround the collision point.

Electron volt (eV) Unit of energy used in high-energy and accelerator physics. It is defined as the kinetic energy imparted to an electron passing through a potential difference of 1 V.

Hadron electron ring accelerator (HERA) An electron–proton collider, 6.3 km in circumference, at Deutsches Elektronen-Synchrotron (DESY) in Germany.

Higgs particle Putative particle introduced to describe one mechanism by which elementary particles acquire masses in their interaction with the all-pervasive Higgs field.

Large electron–positron (LEP) collider The world's largest particle accelerator, 27 km in circumference, at the European Organization for Nuclear Research, or CERN.

Large hadron collider (LHC) A particle accelerator under construction in the tunnel of the LEP machine at

CERN. The LHC will collide proton beams with a combined energy of 14 TeV.

Relativistic heavy-ion collider (RHIC) A circular particle accelerator, 3.8 km in circumference, at the U.S. Department of Energy's Brookhaven National Laboratory. It provides collisions between protons and between ions as heavy as gold.

Standard model (SM) A model that describes our present understanding of nature's electroweak and strong forces. It is beautiful, in agreement with almost all observations, but not the ultimate truth. Arguably, new physics horizons will be opened when the TeV frontier is reached.

Tevatron A particle accelerator at the U.S. Department of Energy's Fermi National Accelerator Laboratory (Fermilab), 6.3 km in circumference, which can accelerate protons to energies of almost 1 TeV. It is also used as a proton–antiproton collider.

Tracking detectors Instruments which measure the trajectories of charged particles with a spatial resolution ranging from 0.01 to 1 mm. Typically, these detectors consist of thin layers of noble-gas mixtures or silicon, in which thin wires or metal electrodes collect the ionization produced by the passage of the particle signaling the particle position.

COLLIDER DETECTORS for multi-TeV particles are complex instruments that detect and measure simultaneously the parameters of thousands of particles in a detector volume of 10,000–20,000 m^3. This splash of particles is produced in ultra-energetic collisions between two protons in counterrotating beams in a particle collider. The next generation of these instruments, at present under construction, aims at the study of physics phenomena at energies of 10^{11} to 10^{12} eV or more, such as the mechanism generating the masses of fundamental particles, the concept unifying the fundamental forces, or the process through which the matter–antimatter symmetry was violated during the first microsecond after the big bang. These instruments are composed of several layers of detectors, each with a specific measurement task, and are subdivided into $\sim 10^8$ detection cells. Collisions exhibiting novel physics signals are expected to occur only rarely, typically at the rate of 1 in 10^{10} collisions or less; therefore, these instruments also push the data rates to limits exceeding 10^{15} bits/sec. Worldwide collaborations of thousands of physicists and engineers are building research facilities, in concert with industry, to develop these instruments.

I. OVERVIEW

Exploring the structure of matter at the microscopic scale, exposing the laws of physics which shape the evolution of matter in our universe, grasping the forces which hold particles together: these are some of the aims of elementary particle physics. Several tools are used in this research, but foremost among them are particle accelerators called colliders and their associated experimental apparatus (collider detectors) which open the way into this invisible and microscopic world.

This overview will set the stage: The collider machines are being developed to create this new (physics) world; collider detectors (CDs) live with them in a symbiotic relationship. They "see" the new physics, capturing the fleeting signals and transforming them into bytes—terabytes—of information. Thousands of scientists around the globe are decoding this data in search of deeper or new insights into our physics understanding.

The "actors" in a CD will be a number of different detectors, each playing a specialized measurement role. In Section II we will highlight features of modern instrumentation, which will set the foundation for designing a collider detector (Section III). Finally, in Section IV we will turn the blueprint into physics and engineering reality.

A. What Is a Collider?

A collider accelerates and stores beams of elementary particles (e.g., protons or electrons). Two beams, traveling at relativistic speed close to the speed of light in opposite directions, are guided to intersect inside the center of the CD, and particles will, with a certain probability, collide. In such an interaction the energy carried by the particles will be concentrated in a tiny volume, which will create a state of very high energy density. This state will materialize into a variety of particles (pions, kaons, protons, etc.) to be captured and identified in the CD. Occasionally, these signals may reflect a hitherto unknown state of matter or a novel fundamental particle.

In such a collision the particles also "feel" each other intimately and sense each other's internal structure. Much as an electron microscope reveals the structure of matter at the 10^{-8}-cm scale, the most energetic particle colliders probe matter at the 10^{-17}-cm scale.

In short, the higher the energy of the colliding beams, the higher the microscopic resolution, the energy density, and the sensitivity for signals of new physics. The energy of these particle beams is measured in electron volts (eV), with present colliders operating between giga electron volts ($1\ \text{GeV} = 10^9$ eV) and tera electron volts ($1\ \text{TeV} = 10^{12}$ eV). In today's colliders different stable particles—electron–positron (e^-e^+), proton–proton (pp), proton–antiproton ($p\bar{p}$), electron–proton (ep), ion–ion—are being used, the choice depending foremost on the research emphasis. Equally decisive, however, is the technology available to accelerate and store the particles.

B. Why Colliders?

In the collision of two particles of energies (momenta, masses) $E_1(\bar{p}_1, m_1)$ and $E_2(\bar{p}_2, m_2)$ the total center-of-mass energy (E_{CMS}) can be expressed in the form

$$E_{\mathrm{CMS}} = \left[(E_1 + E_2)^2 - (\bar{p}_1 + \bar{p}_2)^2\right]^{1/2}.$$

For two relativistic, counterrotating beams colliding head-on,

$$E_{\mathrm{CMS}} \approx [2E_1 \cdot E_2 \cdot (1 + \beta_1\beta_2)]^{1/2}$$

or

$$E_{\mathrm{CMS}} \approx 2E \quad \text{for} \quad E_1 = E_2, \beta_1 = \beta_2 \approx 1.$$

Contrast this result with the energy E_{CMS} made available in the collision of an accelerated beam of particles, energy E_1, colliding with nucleons, mass m_{N}, on a stationary target ($E_2 = m_{\mathrm{N}}$, $\beta_2 = 0$):

$$E_{\mathrm{CMS}} \approx \left[2E_1 \cdot m_{\mathrm{N}}c^2\right]^{1/2} \approx [2E_1(\mathrm{GeV})]^{1/2},$$

with $m_{\mathrm{N}}c^2 \approx 1$ GeV. It would need an energy $E_1 = 2E^2/m_{\mathrm{N}}c^2$ to match the E_{CMS} achievable in a collider with two beams of energy E. As an example, the 27-km-long large hadron collider (LHC) ring would turn into an impossibly large 200,000-km accelerator ring.

Obviously, if energy is at a premium, one should build a collider. These machines are honed to extend the energy frontier for physics research; see Table I. A few explanations will help an individual to understand the shorthand of this table. Perhaps the most striking feature is the difference in E_{CMS} between the e^+e^- collider, or large electron–positron (LEP) collider, and the $p\bar{p}$ collider (Tevatron), and the future LHC. Although the LEP collider ring has a circumference of 27 km and a bending radius $r = 3100$ m, the synchrotron radiation loss dW/dt,

$$dW/dt \sim (E/mc^2)^4/r^2,$$

is so high that for electrons, with their small mass m, it is unrealistic to go beyond the LEP collider energy with circular machines. Today, higher energies can be reached only with a pp($p\bar{p}$) machine for which synchrotron losses are not significant. The LHC, under construction in the LEP collider tunnel, will be able to handle beams 70 times the LEP collider energy.

Dramatically new ways are being pursued to prepare colliders with energies beyond those of the LEP collider and the LHC. Any future energy increase in e^+e^- machines will require novel linear collider geometries. The pp route implies very high field superconducting magnet technologies and very large (100-km) circumference rings. A remarkable compromise between these two extremes could be a $\mu^+\mu^-$ collider, which promises the

TABLE I Today's Energy Frontier of Collider Detectors[a]

	Laboratory/Instrument					
	BNL	CERN		DESY	Fermiab	
	RHIC	LEP collider	LHC	HERA	Tevatron 1	Tevatron 2
Colliding particles	pp Au–Au	e^+e^-	pp Pb–Pb	ep	$p\bar{p}$	$p\bar{p}$
Beam energy (GeV)	250 (proton) 100/nucleon (Au)	≤104	7,000 (proton) 2,760/nucleon (Pb)	28 + 920	900	1000
E_{CMS} (GeV)	500 (pp) 200 (NN)	≤208	14,000 (pp) 5,520 (NN)	320	1800	2000
Collision rate (Hz)	$<10^6$ (p) $\sim 10^3$ (Au)	<10	10^9 (p) 10^4 (Pb)	Few 10^2	$\sim 10^6$	$\sim 10^7$
Discovery reach, (GeV)	10 GeV/fm^3 in $\sim 10^3$ fm^3	208	5,000	100	<300	~500
Major physics emphasis (examples)	Quark–gluon plasma	Standard model (SM)	Origin of particle masses	Proton structure	SM top quark	Top quark
		Higgs search	Physics beyond the SM	Particle searches		Search for symmetry breaking

[a] BNL, the U.S. Department of Energy's Brookhaven National Laboratory; RHIC, relativistic heavy-ion collider; CERN, the European Organization for Nuclear Research; LEP, large electron–positron; LHC, large hadron collider; DESY, the German laboratory Deutsches Elektronen-Synchrotron; HERA, hadron electron ring accelerator; Fermilab, the U.S. Department of Energy's Fermi National Accelerator Laboratory; NN, nucleon–nucleon.

attractiveness of a circular e^+e^- machine without its bane of synchrotron radiation.

Table I also shows that pp machines can be used to collide fully ionized nuclei, up to Pb^{92+}. A dedicated nuclear collider, the relativistic heavy-ion collider (RHIC), has been constructed to study nuclear matter under extreme conditions (e.g., in the quark–gluon plasma phase). Starting in 2006 this program will also be pursued at the LHC.

C. Collider Detectors

Table I also hints at an intimate relation between the type of collider and its collision rate, which determines the key features of the CD.

In every e^+e^- annihilation all the energy is converted into the physics state under study. For this reason the discovery reach for new particles of mass M approaches approximately $M \approx E_{\mathrm{CMS}}$, with rather clean signals and little disturbing background.

The contrast to pp($p\bar{p}$) machines is striking. Protons are made of three quarks, held together with gluons, which share in a fluctuating way the energy of the proton. These pp($p\bar{p}$) colliders are actually quark–quark (quark–antiquark) and gluon–gluon colliders at reduced energy. In most cases the quarks collide at glancing angles, transferring little energy. Only rarely, typically in 1 collision in 10^6 or less, is there a head-on encounter with most of the energy of both particles made available. In pp collisions, the effective energy reach for discovery is therefore much smaller than E_{CMS}. Furthermore, most of the time much energy is squandered to produce the familiar collision products, which is a big nuisance. The new, exciting physics happens excruciatingly rarely. Some of the "most wanted" new particles might be discovered only at the rate of 1 in 10^{14}.

In these colliders the particles in the beams are grouped into bunches: at the LHC, every 25 nsec, pairs of intense proton bunches will sweep through each other; this produces typically 20 collisions and some 3000 particles.

For these reasons CDs at pp colliders, such as the LHC, must function at very high collision rates, must select the rare interesting physics phenomena with lightning speed, and need to measure them in the presence of thousands of background particles. These are the challenges for the new generation of TeV collider detectors under construction for the LHC, as will be described in the following sections.

II. COLLIDER DETECTOR COMPONENTS: A PRIMER

The particles created in the wake of a collision are kinematically described by their momentum $\bar{p}$ and energy E. While in principle a measurement of $\bar{p}$ and E of all the particles produced characterizes the original state, sometimes more specific quantities, such as the identity or mass of a particle, are measured with specialized instrumentation.

We present the principal detector components used for these measurements and argue that the laws of physics shape the basic configuration and dimensions of a CD, conceptually shown in Fig. 1.

How are momenta and energies measured?

A. Momentum Measurement of Charged Particles

For the momentum measurement there is only one recipe: immerse the volume around the collision in a magnetic field, and instrument this volume with tracking detectors, such that the curved particle trajectory may be reconstructed and hence its momentum inferred. The minimal requirements for such a measurement are as follows:

- At least three position measurements must be made along the trajectory to deduce the curvature.
- Tracking detectors which disturb the trajectory ever so gently must be used to minimize instrumental "blurring" of the trajectory.

Given the radius of curvature r [m] of a particle with momentum p [GeV/c] and unit charge in a magnetic field B [T], with

$$r \approx 10p/3B,$$

one derives an approximate expression for the momentum accuracy $\delta p/p$ of such a detector:

$$\frac{\delta p}{p} = \frac{\sigma}{\sqrt{N_{\mathrm{m}}}} \cdot \frac{80p}{3L^2B}, \tag{1}$$

where σ [m] is the accuracy in the position measurement of the tracking detector and N_{m} counts the number of measurements along the track. Including the spectrometer length L [m] and the magnetic field strength B [T], there are four parameters available for the spectrometer optimization.

There are two main classes of tracking detectors used in CDs. They are discussed next.

1. Solid-State Detectors

Solid-state detectors (see Fig. 2) provide a signal by collecting the charge liberated in the passage of the particle through a semiconductor. Suitably implanted electrodes, typically strips, apply an electric drift field, in which the ionization charges are collected and induce a detectable signal. The workhorse for this type of detector is silicon

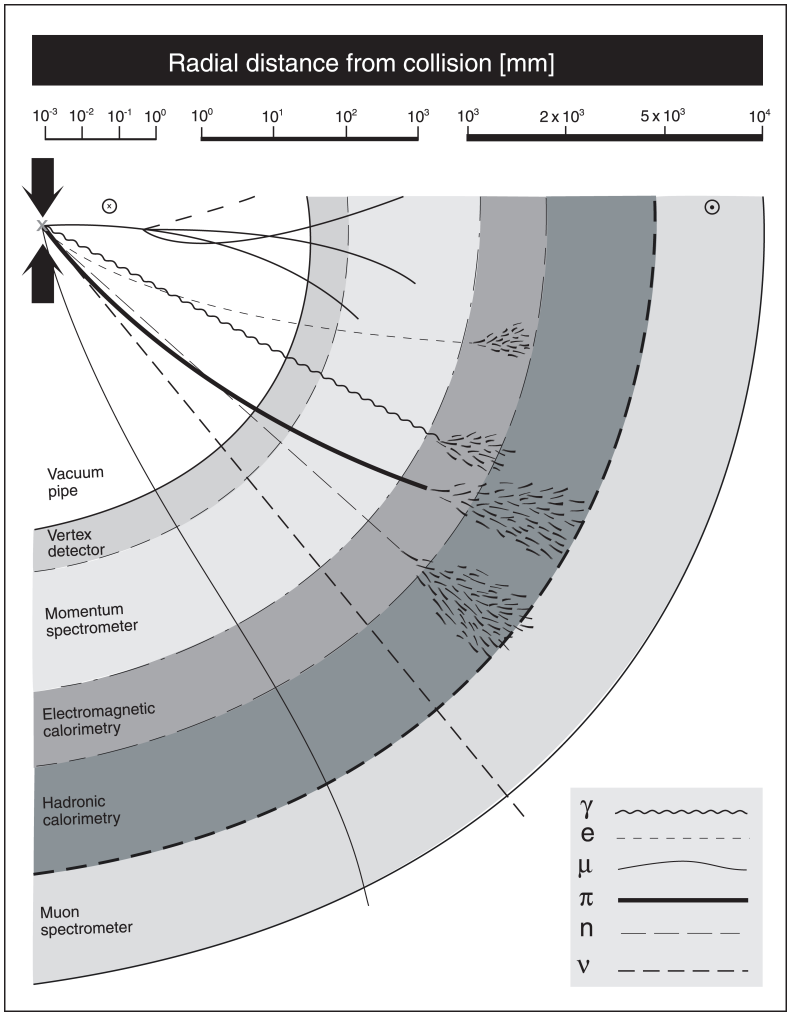


FIGURE 1 Cross section through a conceptual configuration of a collider detector. Like nested Russian dolls, a sequence of detectors enclose the collision point. In each detector layer, the particle is subjected to a specific measurement, shedding, layer by layer, its physics information. Note the three logarithmic scales of the radial dimension.

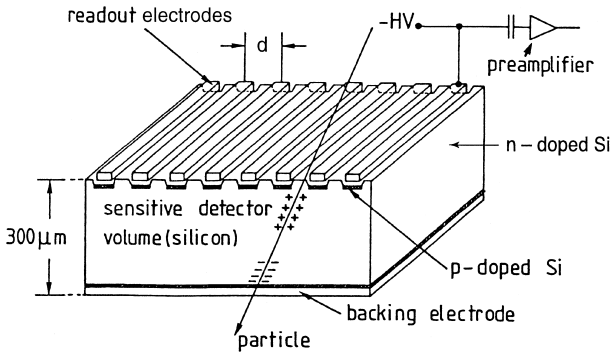


FIGURE 2 Schematic view of a silicon microstrip detector (not to scale). The charges liberated in the passage of the charged particle are collected at the readout electrodes, from which the position is derived. The distance (d) between strips is typically in the range of 20–50 μm. HV.

(Si), which has profited from the technological advances of the computer-chip industry.

It is easy to estimate the strength of the signal. In Si, an energy of $\epsilon \sim 3.6$ eV is required to move an electron into the conduction band. Industry can process, relatively easily, Si disks 300 μm thick. With an average energy loss of $\sim$4 MeV/cm, about 30,000 electron–hole pairs are created, which is large compared with the effective thermal noise $N_e \sim 1000$ e in the preamplifier, attached to electrodes. If each electrode is connected to an electronic channel to register the signal, the spatial resolution σ is comparable to the strip spacing d: $\sigma = d/\sqrt{12}$.

Typical sizes of such detectors range from 5×5 to $\sim 10 \times 10$ cm^2. A modern development is the pixel detector, made possible through the prowess of the electronics industry. A checkerboard of electrodes is used, with

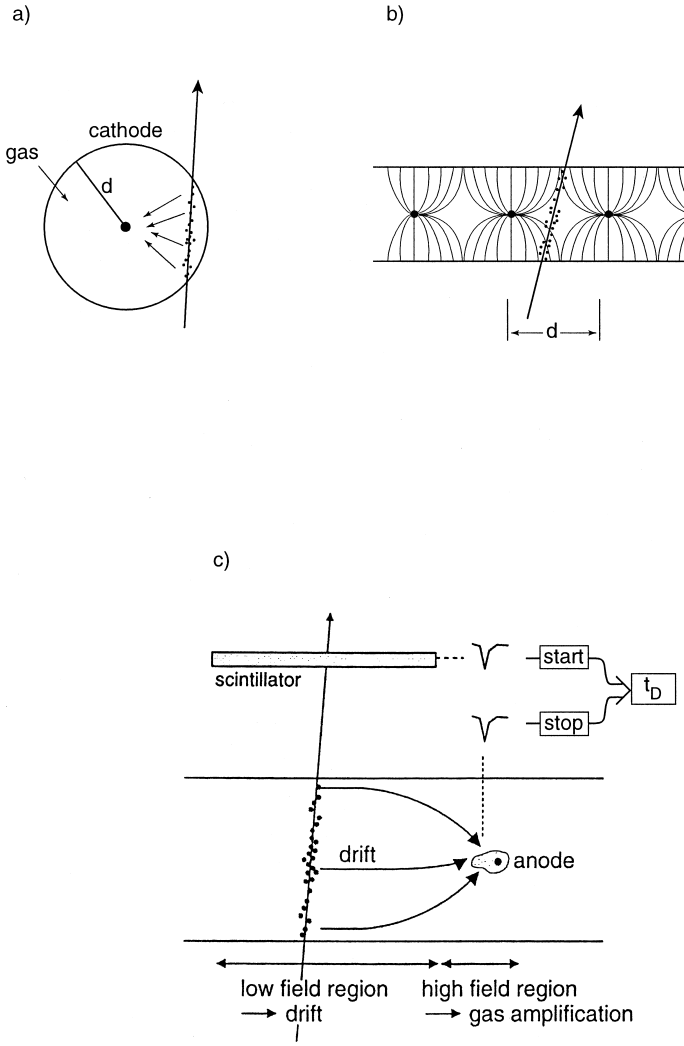


FIGURE 3 Three generations of wire-chamber tracking detectors: (a) proportional tube (also called, when operated at very high gas amplification, the Geiger-Müller counter); (b) multiwire proportional chamber. Registration of the ionization electrons measures the track with a resolution $\sigma = d/\sqrt{12}$. Measuring the drift time interval t_D improves the spatial accuracy; typical values in a drift chamber (c) are $\sigma \sim 100$ μm.

dimensions of $\sim 50 \times$ (few) hundred square microns, allowing a two-dimensional measurement capability. Modern versions have the readout electronics piggybacked to and matching the size of the tiny electrodes. Such detectors allow tracks to be followed with a precision of ≤ 20 μm even in collisions with hundreds or thousands of tracks. They are the key component to identifying particles which decay after a few tens or hundreds of microns, such as the enigmatic beauty particles.

2. Gaseous Detectors

The second category of tracking detector works on ionization produced in layers of suitable gas mixtures. Again, electrodes collect the charge for further signal processing. The primary signal is very feeble: in 1 cm of argon at atmospheric pressure, a favorite detector gas, only about 100 electron–ion pairs are created, which is not enough for convenient processing. However, a clever trick is used: in sufficiently strong electric fields, the ionization electrons can gain enough energy on their migration to ionize in turn the gas, which initiates an avalanche of electron–ion pairs. Physicists have found gas mixtures in which this avalanche amplification can, in a controlled way, reach values up to 10^6, which makes the signal processing straightforward. The ancestor of such devices is the Geiger-Müller counter (Fig. 3a). A thin wire, d 50 μm, is the collecting anode. Thanks to the characteristic cylindrical electric

field, $E \sim V_o/r$, conditions for avalanche amplification can be adjusted to occur within a few microns of the wire surface.

Experimentation in particle physics was revolutionized in the late 1960s with the advent of the multiwire proportional chamber (MWPC; see Fig. 3b), a deceptively simple extension of the proportional tube. Simple, robust, and cheap, it allows large detector areas to be instrumented with high-quality tracking detectors and tracking information to be processed at megahertz event rates.

The drift chamber (Fig. 3c) is an ingenious derivative. A measurement of the drift time interval t_D of the ionization electrons between creation and arrival at the anode wire gives the distance $d = v_D t_D$, if the drift velocity v_D is known. In suitable gas mixtures typical velocities are $v_D \sim 1$–5 cm/μsec. The achievable spatial resolution is ~ 100 μm.

The invention of the MWPC and the drift chamber was the start of an explosive development of gaseous wire chambers: it is relatively easy to shape electric fields and cajole the electrons to drift in line with the demands of physicists. The pinnacle in purity and elegance is the time projection chamber (Fig. 4). A gas-filled cylinder, with a central electrode, drifts the electrons to and projects them onto the detector disks at the end, which are wire chambers. They provide a two-dimensional view of the projected tracks, with the drift time giving the third dimension. It is an ideal imaging tracker for high-multiplicity topologies occurring at low event rates. This technique was optimal for and reached its apogee in CDs at the LEP collider. It is also the technique of choice for tracking in experiments at ion colliders (RHIC, LHC heavy ions).

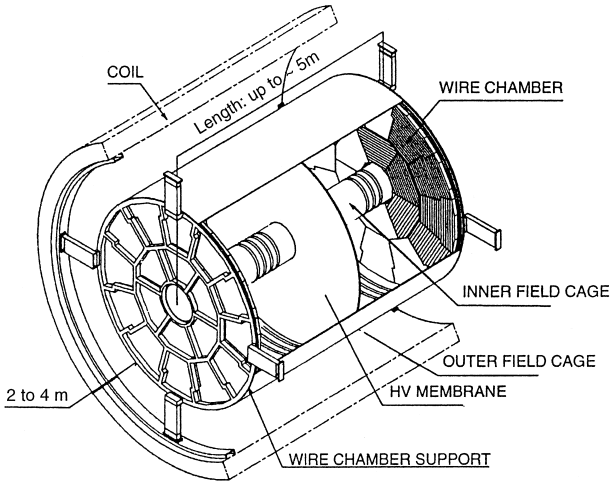


FIGURE 4 Schematic view of a time projection chamber. The large cylinder acts as a drift chamber, projecting the track ionization onto the wire chambers at the two end faces.

B. Energy Measurement

Energy measurement of particles is a second, complementary technique. In contrast to the gentle interactions in tracking detectors, the energy measurement relies on complete absorption of the particles. Massive detectors are instrumented, in which the particles interact either electromagnetically or by means of the strong (nuclear) interaction. Sufficiently energetic particles (E 100 MeV) produce a cascade of secondary particles with increasingly lower energy. These instruments are called calorimeters in analogy to the instruments which measure the total absorption of mechanical or chemical energy through a temperature rise. For calorimeters employed in CDs, an intermediate step in the absorption process—ionization or excitation of the detector molecules—is used to derive a conveniently measurable signal. On average this signal S is proportional to the number of cascade particles N produced, which in turn is proportional to the incident particle's energy, $S = aN = bE$. In individual measurements, the number N fluctuates because the cascade is a statistical sequence of approximately independent collisions. The fluctuations ΔN of cascade particles, $\Delta N \sim \sqrt{N}$, will generally determine the ultimate limit to the accuracy (resolution) of the energy measurement. The relative resolution,

$$\Delta E/E \sim \frac{\Delta N}{N} \sim \frac{\sqrt{N}}{N} \sim 1/\sqrt{N} \sim 1/\sqrt{E},$$

improves with increasing energy, which makes calorimetry more precise than momentum spectroscopy for energies around and beyond $\sim$100 GeV. A second attractive and distinctive feature is the capability of measuring neutral particles (high-energy gammas, neutrons, etc.). This is a further reason why calorimetry is essential in modern particle physics, where increasingly we need to obtain rather complete information about a collision event. Two categories of calorimeters are being used. They are described next.

1. Electromagnetic Calorimeters

For the energy measurement of electrons, positrons, and gammas the absorption proceeds essentially by means of the electromagnetic interaction. Bremsstrahlung of photons from electrons and positrons, and e^+e^- production in the interaction of gammas in the Coulomb field of the nuclei of the detector material are the principal processes driving and propagating the electromagnetic (e.m.) cascade. These instruments are therefore called electromagnetic calorimeters. A pictorial view of such a particle shower is shown in Fig. 5.

The correlation between the measured energy and the number of cascade electrons and positrons is very tight,

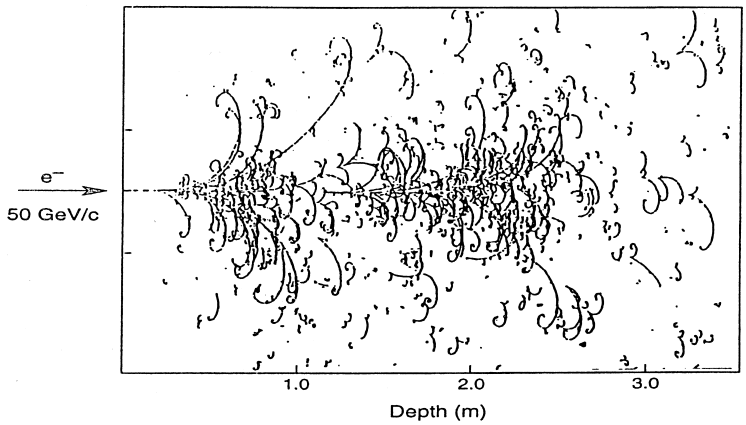


FIGURE 5 A bubble-chamber picture of an electromagnetic shower. The chamber was filled with a liquid Ne–H_2 mixture, which has a long characteristic absorption length for electrons. It was immersed in a 3-T magnetic field. The curved lines are the images of electron and positron tracks in the electromagnetic cascade. In collider detectors, materials for which this absorption length is at least 10 times shorter are chosen.

which results in a potentially very high quality energy measurement; the practical limit to the energy resolution in e.m. calorimeters is at the level of $\sigma(E)/E \quad 0.001/\sqrt{E(\text{GeV})}$. That is, for a 1-GeV electron (gamma) the intrinsic resolution can be as good as 1% (i.e., 10 MeV). At 100 GeV it could reach 0.1% (i.e., 100 MeV), although in present-day detectors instrumental effects mask this stupendous precision.

Practical examples of such high-quality instruments are certain types of crystals, in which the e^+e^- cascade excites the crystal molecules with subsequent emission of light flashes. A famous example of such a crystal is bismuth germanate, which was developed for an LEP collider experiment. It has become the detector of choice for positron-emission tomography (PET) used in medical diagnostics.

2. Hadron Calorimeters

Hadron calorimetry is more complex. In the nuclear and strong interactions a very wide spectrum of cascade particles is produced: part of the energy is channeled into energetic hadron production, and part is deviated into nuclear collisions accompanied by soft neutron and photon emission, nuclear breakup, spallation, and so forth. This multitude and complexity of reaction channels severely distrubs the energy–signal correlation and hence the intrinsic resolution is much worse. In the best devices, in which certain tricks are used to improve this correlation, the intrinsic resolution is measured to be at the level of $\sigma(E)/E \sim 0.2/\sqrt{E(\text{GeV})}$, although in many practical detectors the resolution is closer to $\sigma(E)/E \sim 0.5/\sqrt{E(\text{GeV})}$. Nevertheless, at the high-energy frontier where typical energies are well beyond 100 GeV, these devices have a very respectable accuracy at the percent level.

3. Calorimeter Dimensions

In both types of calorimeters the instrumental dimensions are determined by a characteristic mean free path between the collisions in the cascade. For the e.m. devices this length X_0 (radiation length) is determined by the density of atomic electrons; one finds approximately

$$X_0 \approx 180A/Z^2[\text{g cm}^{-2}].$$

For lead, with density $\delta = 11.4$ g cm^{-3}, one obtains $X_0 \approx 5.5$ g cm^{-2}; hence X_0 (Pb) ≈ 0.5 cm. Approximately 50 X_0 (i.e., 25 cm Pb) are enough to absorb 99% of a 100-GeV particle.

For hadrons, the density of nuclear scatterers determines the nuclear interaction length λ. For Pb again, we find $\lambda(\text{Pb}) = 194$ g cm^{-2} or $\sim$17 cm. Because of the higher multiplicity in secondary-particle production, fewer mean free paths are needed for complete absorption. Approximately 10 λ are needed to absorb a 100-GeV hadron: hadron calorimeters therefore typically have a depth of $\sim$2 m:

4. Calorimeter Optimization

We have broadly sketched the physics underlying the energy measurement. Practical devices need to be finely instrumented to extract a signal: it may be ionization charge left in a suitable detector medium (e.g., liquid argon), or it may be a light flash in a plastic scintillator or a scintillating crystal. Such instruments may be sensitive in the

full detector volume (e.g., a crystal) or divided into separate layers of absorber and instrumentation (sampling calorimeters). It is the choice of this instrumentation which will tailor the calorimeter to a specific measurement requirement. This quest for optimization has led many research groups over the past 20 yr to explore an incredible range of instrumentation possibilities. In particular, the next generation of CDs is characterized by excellent e.m. calorimetry, combined with hadron calorimetry optimized for very good high-energy measurements.

C. Particle Identification and Event Topologies

Inspection of Fig. 1 also indicates that the combined p and E measurements contain a wealth of—for the physics program—essential additional information on the types of particles produced:

1. *Short-lived particles*. Particles containing a charm or a beauty quark are short-lived, as is the heavy tau-lepton: they decay within the beam pipe. Reconstruction of the decay vertices with high-resolution tracking detectors allows this very interesting group of particles to be identified.
2. *Electrons and photons*. These particles deposit their energy in the e.m. calorimeter. A matching track would identify an electron; the neutral photon remains invisible in the tracking detectors.
3. *Hadrons*. Most of the energy is deposited in the hadronic calorimeter. Information on the charge is obtained from the tracker.
4. *Jets*. Free quarks do not exist but manifest themselves as "jets" of hadrons, measured both in the tracker and in the calorimeter.
5. *Muons.* These are as essential as electrons. Because they weigh approximately 200 times more than electrons they cannot be absorbed in a calorimeter and are therefore identified as charged particles, traversing the detector and leaving only a relatively faint signal of ionization.
6. *Neutrinos.* Although they traverse the detectors, almost never leaving any direct signal, their production can still be inferred. They do carry with them E and $\bar{p}$ and reveal themselves, provided all the other particles are adequately well measured. In practice, the energy component projected on a plane transverse to the collision axis, E_T, reveals most tellingly the missing energy carried away by the neutrino.

D. Performance Limits of Detectors

In Table II, the performance limits of the various detectors are summarized. These limits of individual detector performance set the scale and determine the performance of the CD.

We have gained an understanding about the basic detector components. The physics of the detection processes imposes a natural arrangement and sets the scale for the size of the detector components, as illustrated with a few benchmark numbers:

1. *Momentum measurement*:
   a. The typical magnetic field reached with superconducting magnets is $B \sim 2$ T.
   b. The typical spatial resolution with Si detectors is $\sigma = 20\ \mu$m.
   c. A 1% precision measurement at 100 GeV/c would require the following:
      (1) Four layers of position measurement with $\sigma = 20\ \mu$m (alternatively, 100 drift-chamber measurements with $\sigma \approx 100\ \mu$m).
      (2) Length of spectrometer $L \approx 100$ cm.

TABLE II Performance Limits of Collider Detector Components

Measurement	Detector	Limit to performance	Process limiting performance	Practical performance
Position of charged particles	Silicon-strip detectors	σ 5 μm	Spread of ionization electrons along track	$\sigma = 10$–$20\ \mu$m
	Drift chamber	σ 50 μm	Diffusion of ionization electrons	$\sigma = 50$–$150\ \mu$m
			Thermal noise in electronics	
Energy measurement of electrons or photons	Crystals	$\sigma \sim 10$ MeV at 1 GeV	Signal fluctuations Signal sampling	$\sigma = 10$–100 MeV at 1 GeV
	Fine sampling calorimetry	$\sigma \sim 1$ GeV at 100 GeV		$\sigma \approx 1$ GeV at 100 GeV
of hadrons	Sampling calorimeters	$\sigma \sim 3$–5 GeV at 100 GeV	Fluctuations in absorption process	$\sigma \sim 5$ GeV at 100 GeV

2. *Energy measurement* (surrounding the momentum spectrometer, starting with the thin e.m. calorimeter):
   a. For 100-GeV photon or electron absorption, approximately 50 X_0 are needed, which require typically 50 cm.
   b. For hadron containment, $E \sim$ few hundred GeV; 12–14 λ are required. In relatively compact instruments values of $\lambda \sim 20$ cm can be reached: a total of 250–300 cm are required.
3. *Muon measurement* (performed in the last instrumentation layer surrounding the hadron calorimeter; very large magnetic-field volumes are required):
   a. Technologically practical field levels in air are limited to 0.5–1 T; they reach 1.8 T in saturated iron.
   b. Typically 2–5 m of momentum spectroscopy are needed.

In conclusion, it is clear that the laws of physics shape the basic configuration of a collider detector: A "Russian doll" sequence of application-specific detectors enclose the collision point, peeling off the physics information carried by the particle, step-by-step; see once more Fig. 1.

In the next section, we will show how the aims of a specific physics research program and the ingenuity and tastes of the researchers lead to a design of such a research facility, the collider detector.

III. DESIGNING A COLLIDER DETECTOR

A panoply of measurement techniques have been developed to track and capture particles. The laws of physics shape the generic CD. We are now equipped to partake in this exciting intellectual adventure: the creation of a new detector and research facility. Intellect, as well as experience seasoned with emotion, shapes the design, as will be illustrated with the largest effort undertaken to date: the LHC general-purpose CD facilities.

A. Intellect: Design Shaped by Physics Potential

Research since the 1970s, and in particular during the 1990s at the LEP collider, has culminated in and supported the Standard Model of Electroweak and Strong Interactions (SM) as a remarkably good approximation of the world experienced so far. However, the clarity and precision of the results is casting a shadow of new physics not yet observed directly or contained in the SM:

- According to the SM a new, fundamental field (the Higgs field) permeates all space. Particles acquire masses through interaction with this field. If this is true, one particle, the Higgs particle, will be observable at the LHC. Research at the LEP collider has placed a lower bound on the Higgs mass of $M_H > 121\ m_p$ (m_p is the proton mass) and an upper bound of, very likely, $M_H < 200\ m_p$.
- Nature has operated with, until now, mysterious violations of symmetries. One such mechanism created an imbalance between matter and antimatter, such that a minute amount of matter, 1 quark in 10^9, survived. Stars, galaxies, and life owe their existence to this tiny violation.
- Neutrinos appear to have small masses, which the SM has not predicted. Experiments must show the way out of this conundrum.
- Our universe is dominated by invisible dark matter. Circumstantial evidence favors the existence of unknown elementary particles, as postulated, for example, in one popular extension of the SM: supersymmetry. If this is true, supersymmetric particles should be discovered at the LHC.

This catalog is incomplete, but it gives a flavor of the breadth of physics to be studied and provides the yardstick to gauge the performance of the CD design. As an example, the various signatures through which a Higgs particle might reveal its existence are shown in Fig. 6. Books have been written, journals filled, and conferences organized to develop, debate, and anticipate this physics. The consequences and challenges for the CD are synthesized in Table III.

Measurement accuracies are typically a factor of 2–5 better than for CDs at the LEP collider. However, the dramatically novel feature is operation at a collision rate of 10^9 sec^{-1}, which is needed to observe the new phenomena with adequate significance. As each collision typically produces $\sim$100 particles, the CD must see approximately 10^{15} particles for each detected Higgs particle: if the proverbial haystack is made up of some 10^7 straws, then searching for the Higgs particle is like looking for a needle in not 1 but 10^8 of these haystacks.

B. Experience and Emotion: Design Driven by Physicists

More than 2000 physicists and engineers are mobilized to design and construct an LHC CD. Early in the program the important decision was made to construct two such general facilities in view of the enormous physics stakes. This would ensure competition, cross-fertilization of ideas

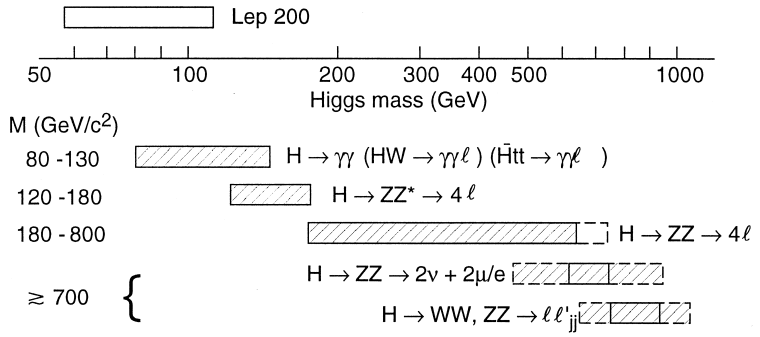


FIGURE 6 Diagrammatic representation of the experimentally most suitable decay modes in the search for the Higgs particle at the large hadron collider (LHC), as a function of the Higgs mass. The large electron–positron (LEP) collider experiments have provided a lower bound of 114 GeV/c^2.

and methods, and cross-checking on the quality of the detectors and the results.

However, with the physics program, the experimental requirements, and the measurement techniques all in the public domain, might the two CDs not look like two clones? The answer is an emphatic no: scientists, not computers with a universal operating system, are designing the research facilities. Two examples may illustrate the process.

High-quality muon spectroscopy is recognized as a cornerstone of a CD. These measurements must be done in the last shell outside the hadron calorimeter, therefore dominating the overall size and cost. The two collaborations, ATLAS (a toroidal LHC apparatus) and CMS (compact muon solenoid), have chosen very different solutions for the muon spectrometer magnet (Fig. 7). ATLAS opted for three dedicated toroidal magnetic fields. Muons are measured twice: in the inner tracker and in the toroidal muon field. An audacious design was developed, with eight 25-m-long superconducting coils forming the central toroid. CMS pursued a one-magnet concept: a high-field (4-T) and large ($R = 3$ m) solenoid encloses the central tracker and calorimetry. Muons are measured with high precision in the central tracker and are reconfirmed in the return yoke of the magnet, albeit with considerably less accuracy. Although compact and elegant, it is somewhat more risky than the ATLAS approach, as the high-quality muon measurement must succeed in the presence of thousands of charged particles flooding the central tracker.

The equally important measurement of electrons and photons is another example. Members of ATLAS were experienced with calorimetry based on the liquid-argon ionization technique. The prime motivation was stability and uniformity of measurement response. They decided to "breed" it into a form able to survive the ferocious LHC environment. A totally new accordion geometry (Fig. 8) had to be invented. CMS boasted experts who had participated in developing the revolutionary crystal calorimeter built for the L3 experiment at the LEP collider. The elegant, high-performance method used for L3 had to be ruggedized for LHC survival. A remarkable worldwide collaboration between particle physicists, material scientists, crystallographers, and industry succeeded in growing a novel crystal, $PbWO_4$, fit for LHC research.

TABLE III Discovery Physics and Corresponding Measurements at the LHC

Discovery signal	Measurement	Measurement accuracy	Signal rate per collision
Higgs $\rightarrow \gamma\gamma$	$m_H < 2m_Z$: precise γ energy measurement	1% at 100 GeV	$\sim 10^{-13}$
$\rightarrow$ ZZ	$m_H > 2m_Z$: momentum of e and μ; ν (= missing energy)	~2% at 100 GeV/c	$\sim 10^{-11}$
Other mechanisms of symmetry breaking	Precision μ, e spectroscopy at large momenta	~10% at 1 TeV	$\sim 10^{-14}$
Supersymmetric particles	Jets of particles; missing energy; e, μ	Energy with ~1% at 1 TeV	$\sim 10^{-13}$
New gauge bosons	e, μ; ν (= missing energy)	Missing energy with few % at 1 TeV	$\sim 10^{-14}$
Quark constituents	Precision jet measurements	~1% linearity up to few TeV	$\sim 10^{-14}$

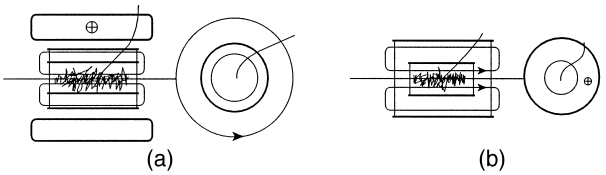


FIGURE 7 Conceptual muon spectrometer magnet configurations adopted by the two LHC collaborations: (a) ATLAS (a toroidal LHC apparatus) configuration and (b) CMS (compact muon solenoid) configuration. For both experiments, a longitudinal view (left) and a transverse view (right) are shown. Arrows indicate the direction of the magnetic-field lines. Note the different curvature of the muon tracks (long heavy line) in the two projections. For ATLAS only the central toroidal magnet is drawn.

These examples show two groups of physicists with two different stategies to meet the same challenge. Ultimately it is scientists—teams with different experiences, know-how, and tastes—who drive the final decisions. Without these strong, personal involvements the CDs would never be built.

C. Complexity of the Design

Much as the physical size of the CD is a direct consequence of the physics and technology of the detection process, its complexity is driven by the physics research. Complexity reflects the number of independent detection cells (channels), the associated signal-processing rate, and the staggering level of event selectivity.

1. Number of Detection Cells

The central tracker is constructed of enough independent detector channels such that the probability of simultaneous occupancy of one element by more than one particle is very small (e.g., of the order of 0.1%). The consequence of this approach is that a few 10^6 channels are needed.

Different considerations prevail in the vertex detector, whose role is the reconstruction of decay vertices produced by telltale short-lived particles. The decay topology must be reconstructed in three dimensions with a resolution of $10 \times 10 \times 50\ \mu m^3$; silicon detectors are divided into pixels of typically $50 \times 300\ \mu m^2$. Three layers surrounding the 30-cm-long collision zone are divided into a total of $\sim 10^8$ pixel elements.

In the calorimetry, the concept of very fine subdivision is limited by the natural scale of the transverse dimensions of the particle cascade, and low occupancy cannot be achieved. Remember, however, that most collisions are glancing ones, transferring little energy to the collision products: a redeeming feature. These thousands of low-energy particles just produce a background

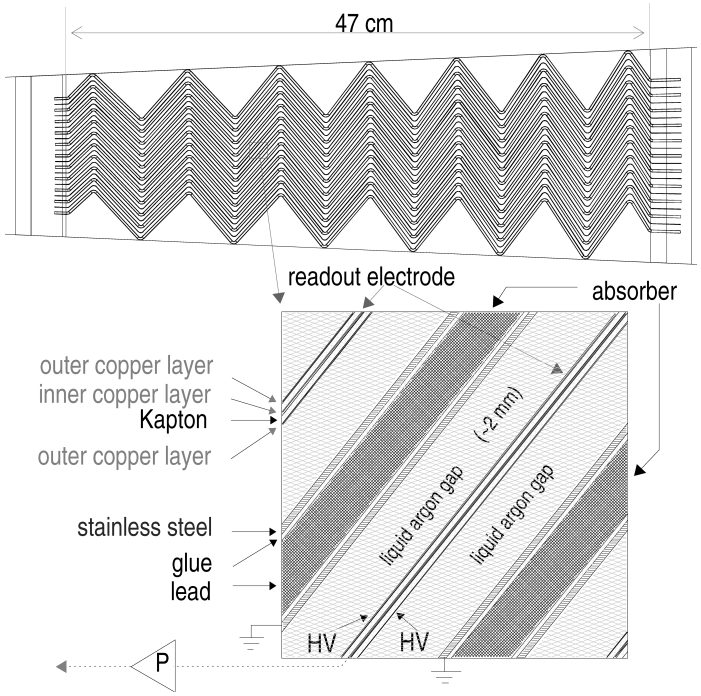


FIGURE 8 (Top) The accordion structure of the absorber plates of the ATLAS Pb–liquid-argon electromagnetic calorimeter. The particle enters from the left. (Bottom) Details of the electrode structures needed to collect the ionization charge.

noise and do not significantly disturb the measurement of the valuable energetic particles producing a strong signal.

Dividing the e.m. calorimeter into cells of natural dimensions results in typically ~2000 cells/m^2, subdivided longitudinally a few times. A few 10^5 channels will suffice for such a detector. Similar arguments lead to a subdivision of the hadronic calorimeter into approximately 10^4 channels.

For the muon system, considerations on occupancy similar to those for the inner tracker also apply. Therefore, approximately 0.5×10^6 tracking cells are needed.

2. Signal-Processing Rate and Event Selectivity

In an LHC detector, every 25 nsec more than 2000 particles will be produced. What is more, the collision products induce an intense level of background radiation comparable to, sometimes much greater than, the particle rates. The signals left in the various detectors will linger for up to a few hundred nanoseconds. At any given moment there are several waves of event information racing through the CD and its signal-processing system. Several strategies are needed to digest this simultaneity of event information.

A new form of signal processing has to be used to deal with this staggering amount of information (Fig. 9). The primary signals of hundreds of such events are stored sequentially in "pipelines." Detectors combined with ultrafast signal processors decide on the potential physics interest of each event, typically in less than 2 μsec. After this initial preselection the compacted data rate is reduced to ~10^{13} bits/sec, which corresponds to the data volume of approximately 10^8 simultaneous telephone calls. Subsequently more refined algorithms are applied to this data to retain events of potential physics interest. We expect that only 1 event in 10^7 collisions (i.e., 100 events/sec) will meet such event-selection criteria and will be archived. These CDs require communication links and a processing power similar to those of a large telecommunications company.

3. Engineering Complexity

The CDs are the size of a six-story office building. They are composed of some 20,000 detector elements, constructed with 0.01- to 1-mm tolerances. Many of these units are supported by lightweight frames, familiar in the space industry. The detector components, together with the required services, are being fitted together with millimeter clearances; they are surveyed with tens of thousands of optical rays, which locate them in the 20,000-m^3 detector volume with an accuracy of 10–30 μm.

The 10^8 signal sensors, some of which work at cryogenic temperatures, are exposed to a ferocious radiation

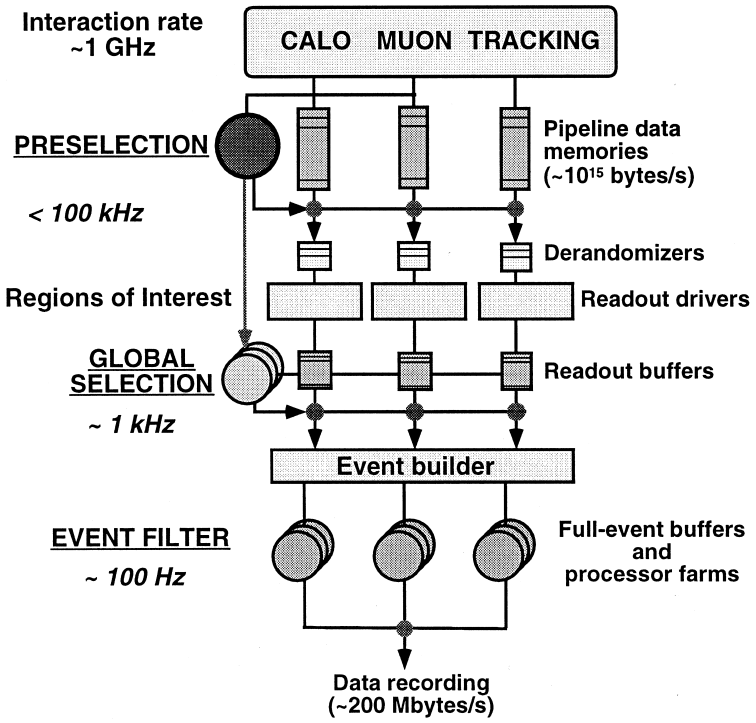


FIGURE 9 Conceptual diagram of the data-processing and event-selection system. Signals from each of the 10^8 detector channels are processed through one of these pipelines needed to store the data awaiting a selection decision.

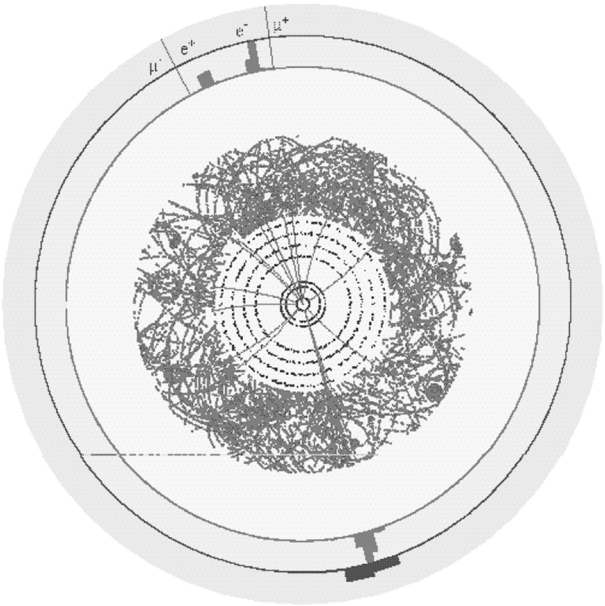


FIGURE 10 Computer simulation of an H → ee$\mu\mu$ event in the ATLAS detector. The electrons are identified as charged particles in the inner tracker with a matching energy deposit in the electromagnetic calorimeter. The muon tracks are clearly visible in the outer spectrometer. Although thousands of particles crowd the tracking detectors, the electron and muon tracks can be identified and measured in the computer reconstruction.

environment of up to 10^6 particles/cm^2 sec reminiscent of certain military applications. Many of the detectors are practically inaccessible and are therefore designed to reliability standards of the space industry.

These detectors use military- and space-type technologies, and engineering methods usually associated with big space projects, but are constructed with university personnel and budgets. Remarkably, the cost of the LHC experimental program has not shown the same level of inflation as performance and specifications: it is comparable to the past LEP collider program.

D. The Design in Virtual Reality

Long before any metal is cut or any amplifier produced, the CD is given a rigorous "health check" in computer-simulated reality. Fortunately, the developments in hardware detectors have been paralleled by equally intensive collaborations to provide the necessary simulation tools. Suffice it to say that we have learned to simulate the detector performance, the influence of engineering choices, and the optimization of materials, services, and so forth at least to the level at which these features influence the final detector performance. These complex detectors simply could not be built without these faithful tools. The importance of such tools can perhaps best be gauged by realizing that the most modern version, dubbed GEANT, an object-oriented code in C^{++}, is being developed by more than 100 physicists and informatics engineers over several years in a worldwide collaboration. Besides the world's particle physics laboratories, the European Space Agency (ESA), NASA, and medical establishments are using this code. Applying the more reliable GEANT estimates on radiation doses in tumor therapy, doctors estimate that approximately 10,000 more lives per year will be saved in the United States alone. As a preview, the GEANT simulation of an event pp → particles + Higgs → particles + 2e + 2μ is shown in Fig. 10.

IV. CONSTRUCTING A COLLIDER DETECTOR

In the previous sections the basic design features of the new generation of CDs, and the motivation for them, were explained. But can such a physics and detector concept be transformed into engineering reality? Today we know that the answer is yes, for the following reasons:

- During the 1990s many groups developed these novel detectors and demonstrated their performance in LHC-like beam tests.
- The groups have learned to address the new management aspects of a project of such unprecedented size.

The evolution of scale and complexity of the new CDs is indicated in Table IV.

In addition to the technological and engineering complexity (see Section III.C.3), several new issues have to be successfully addressed:

TABLE IV Scale and Complexity of Collider Detectors

	Tevatron	LEP collider	LHC
Measurement volume of CD [m^3]	2000	2000	10,000/20,000
Measurement channels	Few × 10^6	Few × 10^6	10^8
Interaction rate [sec^{-1}]	10^6–10^7	10	10^9
Size of collaboration			
Physicists/engineers	500	500	2000
Institutions	50	50	150
Countries	15	20	50
Time scale (concept to first data) [yr]	∼10	∼10	∼20

TABLE V Tools for Managing the Construction of Collider Detectors

Tool	Purpose
Computer-aided design (CAD) systems[a]	• Construct virtually the complete CD • Check fit of all components • Design layout of cables, cryogenic lines, pipes, other services (total length ~100,000 km)
Virtual-reality software	Simulate dynamic procedures: • Installation of detector components • Access to components
Engineering data management system (EDMS)	Central depository of all information:[b] • Text files • Drawings • Videos Total data volume of documentation for one CD: Tbyte
Project management tools	Variety of programs used for • Approval process for drawings and documents • Procurement • Scheduling • Project progress monitoring • Production database of quality control and detector operational data

[a] The centrally used CAD must interface with the different CADs used around the world.

[b] To provide a worldwide collaboration with the unique detector baseline design.

- *Technology*. The detectors are pushing the limits of realizable technologies. Never before was performance demanded at the physics limits for such large systems, nor was industry asked in many-year-long collaborations to develop technologies to LHC demands. One major stimulus of this development is the healthy fight for funds, in which particle physics is joined by other sciences (biology, astronomy, space-base experiments). Never before have funding agencies obtained such an outstanding performance/price ratio.
- *Time scale*. The increase in the time scale is significant. On a purely technical level, new concepts of documentation need to be implemented to retain the memory of the CD design, construction, and modification over several decades.
- *Globalization*. The scale of these instruments is such that the responsibility for a given subdetector (e.g., the tracker) is typically shared by 10 countries, with their own national priorities and funding-agency conditions.

Fortunately, new worldwide networked tools have become available to support this globalization; see Table V. The toolbox is the World-Wide Web: a CERN invention, originally developed to help in the analysis of physics data.

For the LHC, construction of the detector components, the experimental infrastructure, and underground halls has started. A computer-generated view of the two LHC detectors, as they will be in place in 2005, is shown in Figs. 11 and 12.

V. CLOSING REMARKS

A collaboration of physicists, engineers, and technicians, in concert with industrial partners, conceives, designs, and builds these physics cathedrals. With these instruments, new worlds are explored, the laws of physics are extended, and a deeper understanding of our universe, our own origin, is reached. Like all human expeditions to new frontiers it is a fascinating amalgam of human curiosity, intellect, and passion. Like past explorers we reach these new frontiers with the most appropriate, state-of-the-art

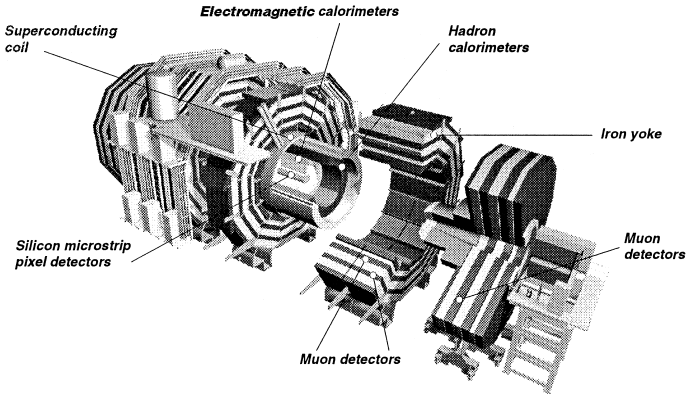


FIGURE 11 A three-dimensional exploded view of the CMS facility, under construction by a worldwide collaboration. In operation it will have a length of almost 30 m and a diameter of 15 m.

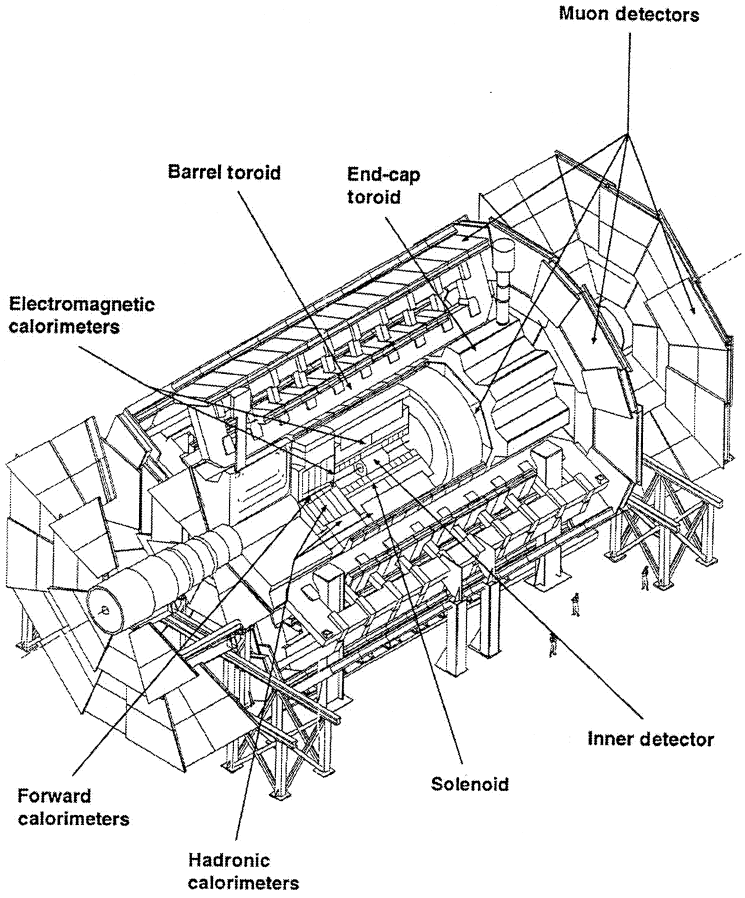


FIGURE 12 A three-dimensional view of the ATLAS facility. The total length of the apparatus is close to 50 m, the diameter is approximately 22 m, and the total weight exceeds 7000 tons.

technologies. As befits the twenty-first century, this expedition to the beginning of our universe is a global voyage of spaceship Earth.

ACKNOWLEDGMENTS

I am grateful for advice and critical comments from W. Blum, N. Ellis, F. Gianotti, and T. S. Virdee. I am indebted to the CERN desktop publishing service for the thoughtful editing of the manuscript and for very professionally transcribing my notes and sketches into printable form.

SEE ALSO THE FOLLOWING ARTICLES

ACCELERATOR PHYSICS AND ENGINEERING • ATOMIC AND MOLECULAR COLLISIONS • COLLISION-INDUCED SPECTROSCOPY • ENERGY TRANSFER, INTRAMOLECULAR • NUCLEAR PHYSICS • PARTICLE PHYSICS, ELEMENTARY

BIBLIOGRAPHY

ATLAS (1994). "Technical Proposal for a General-Purpose pp Experiment at the Large Hadron Collider at CERN," CERN/LHCC/94-43, CERN, Geneva. (Electronic link to CERN Library existing.)

Brandt, D. *et al*. (2000). "Accelerator physics at LEP," *Rep. Prog. Phys*. **63,** 939–1000.

CMS (1994). "The Compact Muon Solenoid, Technical Proposal," CERN/LHCC/94-38, CERN, Geneva. (Electronic link to CERN Library existing.)

Ellis, N., and Virdee, T. S. (1994). "Experimental challenges in high-luminosity collider physics," *Annu. Rev. Nucl. Part. Sci*. **44,** 609–653.

Fabjan, C. W. (1994). LHC: Physics, machine, experiments. *In* "AIP Conference Proceedings 342" (A. Zepeda, ed.), Am. Inst. of Phys., New York. (Electronic link to CERN Library existing.)

Lefèvre, P., and Petterson, T., eds. (1995). "The Large Hadron Collider," CERN/AC/95-05 (LHC), CERN, Geneva. (Electronic link to CERN Library existing.)

Virdee, T. S. (1990). "Experimental techniques," *Rep. CERN 99-04*. (Electronic link to CERN Library existing.)

Williams, H. H. (1986). "Design principles of detectors at colliding beams," *Annu. Rev. Nucl. Part. Sci*. **36,** 361–417.



Collision-Induced Spectroscopy

Lothar Frommhold
University of Texas

I. Brief Overview
II. Collision-Induced Absorption
III. Collision-Induced Light Scattering
IV. Virial Expansions
V. Significance for Science and Technology
VI. Conclusions

GLOSSARY

Absorption coefficient The natural logarithm of the ratio of incident and transmitted intensity divided by the optical path length, $\alpha(\omega) = \log_e(I_0/I)/L$. It is a function of frequency ω, temperature, and gas density.

Allowed and forbidden transitions Atomic and molecular systems exist in a variety of states. Spectroscopic lines arise from transitions between such states, but not all possible transitions are "allowed" for emission or absorption of a photon. Selection rules exist which determine which of the transitions are optically allowed. Forbidden transitions may, however, take place without the emission or absorption of a photon, e.g., in collisional interactions.

Bound vs free van der Waals systems A certain, usually small fraction of the atoms or molecules ("monomers") of virtually any gas exist as van der Waals molecules, i.e., systems of two (or more) monomers, bound together by the weak van der Waals intermolecular forces ("dimers," "trimers," . . .). Below, we will be concerned mainly with complexes of two (or more) unbound monomers which exist only for the very short duration of a fly-by encounter. Free and bound van der Waals systems have many properties in common, but only the latter possess the relative stability of a molecule. Properties of bound and free van der Waals systems will be referred to as supramolecular properties.

Dipole Molecules have a permanent dipole moment if the centers of positive and negative charge do not coincide. Dipole moments can also be induced by external fields (polarization) or momentarily by collisional interactions.

Frequency units Frequency f is measured in cycles per second, or hertz (Hz). Spectroscopists have good reasons to express frequencies ν in cycles per centimeter, or wavenumbers (cm^{-1}), so that $\nu = f/c$, with c being the speed of light in vacuum. Theorists usually prefer angular frequencies $\omega = 2\pi f$; units are radians per second.

Infrared spectroscopy Molecular spectra are observed at frequencies ranging roughly from the microwave region ($\approx 10^{10}$ Hz) to the soft X-ray region ($> 10^{16}$ Hz) or more. While the high-frequency spectra arise from electronic transitions that are of lesser interest here, the low-frequency spectra occur in the microwave,

infrared, and visible region of the electromagnetic spectrum and arise typically from internal rotation and vibration and from molecular encounters.

Raman spectroscopy If monochromatic light falls on a molecular target, polarization and light scattering results. The scattered light is basically of the same frequency as the incident light, but modulated with the rotovibrational frequencies characteristic of the molecule.

Spectroscopic notation Examples are $S_0(0)$ and $Q_3(1)$. The meaning of $X_n(j)$ is as follows: $j = 0, 1, \ldots$ is the rotational quantum number of the initial state; the subscript $n = v' - v$ is the difference of the vibrational quantum numbers of initial (v) and final (v') vibrational state; and X stands for one of these letters O, P, Q, R, S, etc., each specifying a different rotational transition: $j' - j = -2, -1, 0, 1, 2$, etc., respectively. We note that the subscript n is often omitted when it is clear what vibrational band it refers to.

Stimulated emission For every photon absorption process an inverse process exists, called stimulated emission. Stimulated emission is important when the population of states of sufficiently high energy are significant. For example, in the far infrared, where photon energies are comparable to the mean thermal energy of collisional pairs of molecules, stimulated emission forces the absorption to zero as frequencies approach zero.

Virial expansion At gas densities that are substantially smaller than those of liquids and solids, certain properties of a gas may be described by a power series of density. The leading term is typically linear in density and reflects the contributions of the (non-interacting) monomers. The next term is quadratic in density and reflects the induced contributions of exactly two interacting monomers, etc. The most familiar virial expansion is that of the equation of state, which relates pressure, volume, and temperature of a *real* (as opposed to an *ideal*) gas. There are, however, several other and, for our present focus, more relevant examples of virial expansions that are related to the dielectric properties of gases, to be mentioned below in some detail.

I. BRIEF OVERVIEW

Rarefied molecular gases absorb and emit electromagnetic radiation if the individual molecules are *infrared-active,* i.e., if the structure of individual molecules is consistent with the existence of an electric dipole moment. Homonuclear diatomic molecules (H_2, N_2, ...) are infrared-inactive, but characteristic rotovibrational absorption bands and certain continuous spectra exist in the rarefied gases composed of polar molecules (HCl, NO, ...). Compressed molecular gases, on the other hand, show quite generally a variety of *additional* absorption bands—even if the individual molecules are infrared-inactive. These are the *collision-induced* absorption spectra that arise from fluctuating dipole moments induced momentarily when molecules collide. Collision-induced dipole moments are of a *supramolecular* nature; they are properties of *complexes* of two or more interacting molecules and are foreign to the individual molecules of the complex, as long as these are separated most of the time from all the other molecules by distances amounting to several molecular diameters or more—as may be thought of being the case in rarefied gases.

Similarly, if monochromatic laser light is incident on a molecule, the molecule is polarized by the electric field. This field-induced dipole will emit (or "scatter") radiation of the frequency of the incident light. It will also emit at other frequencies that are shifted relative to the laser frequency by certain rotovibrational transition frequencies of the individual molecules, if the molecule is Raman-active, i.e., if the invariants of the polarizability tensor are nonzero for certain rotational and/or vibrational transitions of the molecule. Compressed gases show quite generally a variety of *additional* Raman bands, the collision-induced Raman bands, even if the individual molecules are Raman-inactive. For example, in rarefied monatomic gases, the scattered laser light will be strictly at the laser frequency; no shifted Raman lines or bands exist. However, in the compressed rare gases, Raman continua exist which are due to collisional and, to some extent, to bound van der Waals pairs (and, at higher gas densities, triples, ...) of interacting atoms. Collision-induced Raman spectra of the common gases are well known and will be discussed in greater detail below.

II. COLLISION-INDUCED ABSORPTION

A. A Discovery

In his famous dissertation, J. D. van der Waals argued compellingly in 1873 that the forces between two molecules of a gas must be repulsive at near range and attractive at larger separations. Ever since, theorists conjectured the existence of what today would be called *van der Waals molecules* or *dimers*, that is, weakly bound systems of two argon atoms (Ar_2) in a gas consisting almost purely of argon atoms (Ar) or of two oxygen molecules $(O_2)_2$ bound together by the weak van der Waals forces in a gas that otherwise consists purely of O_2 molecules; etc. In spectroscopic laboratories around the world efforts ensued to discover characteristic dimer bands and thus demonstrate directly the existence of dimers. However, it took almost 100 years of dedicated

research efforts before some such dimer bands of the elusive van der Waals molecules were actually discovered. Today, we know that such dimers exist in small concentrations in virtually all gases, almost without exception.

In 1949 H. L. Welsh and his associates had hoped to demonstrate the existence of dimers in compressed oxygen gas but failed—just like all other efforts elsewhere had failed at the time. However, in the course of that work a new and arguably more significant type of absorption spectra was discovered instead: the much stronger and truly universal spectra of *unbound* pairs of molecules. In other words, fluctuating dipoles induced in collisionally interacting, free molecules, e.g., O_2–O_2 [to be distinguished from the bound pairs $(O_2)_2$], etc., actually absorb more radiation and absorb over a greater frequency band than the bound dimers do. One simple reason for the stronger absorption of light by collisional pairs is that, typically in compressed gases, at any instant one counts many more collisional pairs than bound dimers. In short, collision-induced absorption of compressed oxygen gas was discovered. In quick succession similar absorption bands were seen in virtually all common molecular compressed gases—a truly universal, new, *supramolecular* spectroscopy was thus discovered: collision-induced absorption.

This discovery of collision-induced absorption was accomplished at infrared frequencies, where the rotovibrational bands of the common molecules typically are found. The new spectroscopy is, however, not limited to such frequencies; it is now known to extend from the microwave region throughout the infrared and well into the visible—and in a few known cases actually beyond.

A quantitative knowledge of the absorption of light by the Earth's atmosphere is essential to scientists, especially to astronomers who need to correct their observational data for such absorption as much as possible. Since the atmosphere absorbs very little visible light, in 1885 Janssen attempted to measure absorption by the atmospheric gases in a high-pressure cell. He found a number of absorption bands of oxygen, unknown from previous studies conducted at much lower gas densities. Absorption in these bands could be enhanced by the addition of nitrogen, but pure nitrogen did not show any absorption bands in the visible and near ultraviolet regions of the electromagnetic spectrum. A telling feature of these new absorption bands is that the absorption coefficient of pure, pressurized oxygen increases with increasing gas density as the *square* of density when the expectation at the time would have been a linear dependence; the enhancement of the absorption coefficient by the addition of nitrogen was found to be proportional to the product of O_2 and N_2 densities. These density dependences suggest a kind of absorption that requires two interacting O_2 molecules, or an O_2–N_2 pair, as opposed to just one O_2 or one N_2 molecule for every absorption process. The new absorption bands had early on been called *interaction-induced* bands by some prominent spectroscopists. However, the process seemed somewhat mysterious because this type of absorption appeared to be limited to situations involving oxygen; others of the most common gases do not have such striking pressure-induced bands in the visible region of the spectrum. Today, we understand that these interaction-induced absorption bands of oxygen are collision-induced bands involving *electronic* (as opposed to purely rotovibrational) transitions of the O_2 molecules; further details may be found below.

B. Monatomic Gases

1. Pure Monatomic Gases

Collision-induced absorption by pure monatomic gases has not been observed. Of course, it is clear that collisional *pairs* of like atoms cannot develop a collision-induced dipole, owing to their inversion symmetry which is inconsistent with the existence of a dipole moment. However, triatomic and higher complexes of like atoms theoretically could absorb infrared radiation, but apparently these absorption coefficients are so small that thus far a measurement has been impossible. Even at the highest densities, e.g., in liquefied rare gases, only a very small upper limit of the infrared absorption coefficient could be established for a few rare gases. Pure monatomic gases are probably the only gases that do not show significant collision-induced absorption at any frequency well below X-ray frequencies.

2. Mixtures

However, *mixtures* of monatomic gases absorb in the far infrared. Whereas like collisional pairs such as Ar–Ar do not support a dipole moment, dissimilar pairs such as He–Ar generally do. The collisional complex of two dissimilar atoms lacks the inversion symmetry that precludes the existence of an electric dipole moment. Absorption by dissimilar pairs is now well known, even if gas densities are well below liquid state densities.

At a given frequency, the intensity of a beam of light falls off exponentially with increasing path length x,

$$I(x) = I_0 \, \mathrm{e}^{-\alpha x},$$

if absorption occurs (Lambert's law). The measurement determines the absorption coefficient α as function of frequency, temperature, and the densities ρ_1, ρ_2, of atoms of both species (assuming a binary mixture). The so-called absorption spectrum α is preferably presented by the "normalized" absorption coefficient, $\alpha/(\rho_1\rho_2)$, which is invariant under variation of the densities ρ_1, ρ_2, as long as

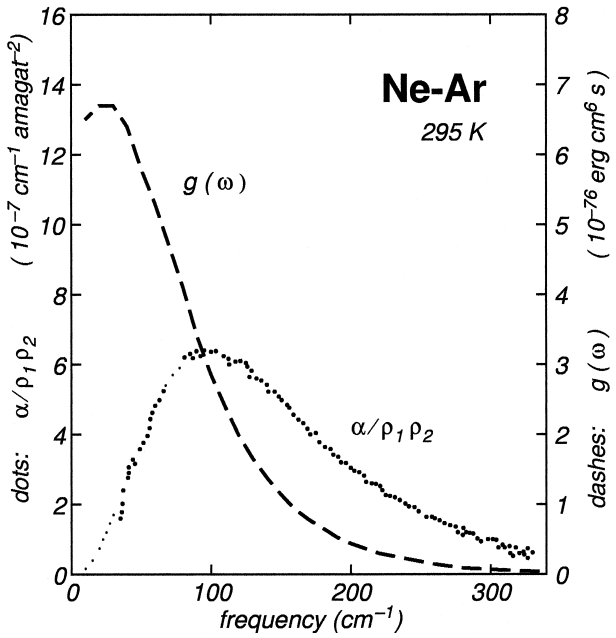


FIGURE 1 The dots represent the measurement of the binary collision-induced translational absorption coefficient of the mixture of neon and argon gas at 295 K by Bosomworth and Gush [(1965). *Can. J. Phys.* **43**, 735]. The dashed curve is the so-called spectral density function derived from the above by correcting the measurement for stimulated emission; it is a diffuse "line" that may be thought to be centered at zero frequency.

no three-body or higher order interactions interfere with the measurement (i.e., at intermediate gas densities that are well below liquid state densities). An example of the absorption spectrum of mixtures of neon and argon is shown in Fig. 1. Densities employed in that measurement were in the order of roughly 10 amagats (1 amagat $\approx 2.7 \times 10^{19}$ atoms/cm^3). The frequencies are shown in wavenumber units (cm^{-1}); they are in the far infrared region of the electromagnetic spectrum and are commensurate with the reciprocal time scales of the atomic collisions. The mean relative speed of the Ne–Ar pair at the temperature of 295 K is about 750 m/s, and the size of the collision diameter for Ne–Ar collisions is in the $\approx 3 \times 10^{-11}$ m range, so that the duration of an average Ne–Ar collision amounts to $\Delta t \approx 4 \times 10^{-14}$ s. According to Heisenberg's uncertainty relation,

$$\Delta t \, \Delta\omega > 0.5, \tag{1}$$

the spectral frequency band $\Delta\omega$ of the average collision amounts to at least $\Delta\omega \approx 1 \times 10^{13}$ rad/s, or >70 cm^{-1}. Actually, absorption over a much greater range of frequencies was observed (Fig. 1). The estimated spectral width in wavenumber units, $\Delta\omega/(2\pi c) \approx 70$ cm^{-1}, is actually very close to the half-width of the so-called spectral density function $g(\omega)$ which is also sketched in Fig. 1 (dashed curve). The spectral density function is obtained by dividing the measured, normalized absorption coefficient $\alpha/(\rho_1\rho_2)$ by the photon energy $\hbar\omega$ and by $[1 - \exp(-\hbar\omega/kT)]$ to correct for stimulated emission. Both factors force the absorption to zero as the frequencies approach zero. The spectral density function may be considered (half of) a spectral line centered at zero frequency, with a half-width of roughly 70 cm^{-1} as was estimated from Heisenberg's uncertainty relation.

The translational absorption band is the only collision-induced absorption spectrum in mixtures of monatomic gases at frequencies well below the ultraviolet.

According to quantum mechanics, the absorption of a photon corresponds to a transition of the colliding pair from a state of relatively low energy of relative motion to a higher such state. The spectrum shown in Fig. 1 is therefore called a "translational" spectrum.

3. Ternary and Many-Body Interactions

In the measurement (Fig. 1), the density variation of the normalized absorption coefficient $\alpha/(\rho_1\rho_2)$ was carefully checked and found to be independent of either density. This density invariance indicates the binary nature of the collision-induced spectrum in the range of gas densities employed in that measurement. We mention that if the density of either gas is further increased, a point is reached where the three-body interactions—and eventually many-body interactions—manifest themselves by a breakdown of the invariance of the normalized absorption coefficient. In a few cases ternary absorption spectra have actually been separated from the binary contributions for detailed analyses. At even higher densities (≈liquid state densities), true many-body effects control the spectroscopy. All these many-body spectra differ from the binary spectra in characteristic ways; the spectra observed at the highest densities must be considered superpositions of supramolecular spectra of binary, ternary, . . . , many-body systems.

C. Molecular Gases

In molecular (as opposed to monatomic) gases much richer collision-induced absorption spectra are observed in a number of spectral bands, because of the increased degrees of freedom of the collisional complexes. Any collisional pair possesses the degrees of the translational motion and an associated kinetic energy of relative motion. Additional degrees of freedom and thus of energies are associated with rotational and vibrational motion of one or more molecules of the complex, if molecular collisions are considered. Accordingly, photons are absorbed over a much greater range of frequencies, in the vicinity of the various

rotovibrational bands of the molecules, and at sums and differences of such rotovibrational frequencies, if two or more molecules interact. In other words, the energy of a collisional pair involving at least one molecule is given by the sum of translational and rotovibrational energies of the interacting molecules. Absorption of a photon corresponds to a transition of the transient "supermolecule" from a state of lower to one of higher energy, and because there are typically many different rotovibrational states accessible, photons of varying energy (frequency) can be absorbed by the pair. One speaks of collision-induced rotovibrational spectral bands. It is important to remember that these induced bands are observable even if the corresponding bands of the individual molecules are "forbidden," i.e., if the individual molecules are infrared-inactive at such frequencies.

1. Rotovibrational Spectra

As an example, Fig. 2 shows the collision-induced fundamental band of the H_2 molecule in compressed hydrogen gas, which is "forbidden" in isolated (i.e., non-interacting) H_2 molecules. A photon is absorbed in the near infrared, in the broader vicinity of the transition frequency from the ground state to the lowest vibrationally excited state ($v = 0 \rightarrow 1$), which occurs at about 4155 cm^{-1}. Especially at the lower temperatures, three broad lines labeled $Q(1)$, $S(0)$, and $S(1)$ are noticed which correspond in essence to transitions of the H_2 molecule involving rotational quantum numbers $j = 1 \rightarrow 1$, $j = 0 \rightarrow 2$, and $j = 1 \rightarrow 3$, subject to the vibrational transition $v = 0 \rightarrow 1$, combined with a change of the translational energy of relative motion of the collisional pair of H_2 molecules. These lines are very diffuse (i.e., not "sharp," like a bright, thin line of light against a dark background, which elsewhere would be called a spectroscopic line), reflecting the short duration of the collisional encounter [Heisenberg's uncertainty relation, Eq. (1)].

Especially at the higher temperatures in Fig. 2, other lines may be discovered in careful analyses of the measurements, such as lines associated with higher rotational states ($j = 2, 3, \ldots$). Perhaps more surprisingly, so-called double transitions can also be discovered (but are not immediately discernible in Fig. 2) which correspond to

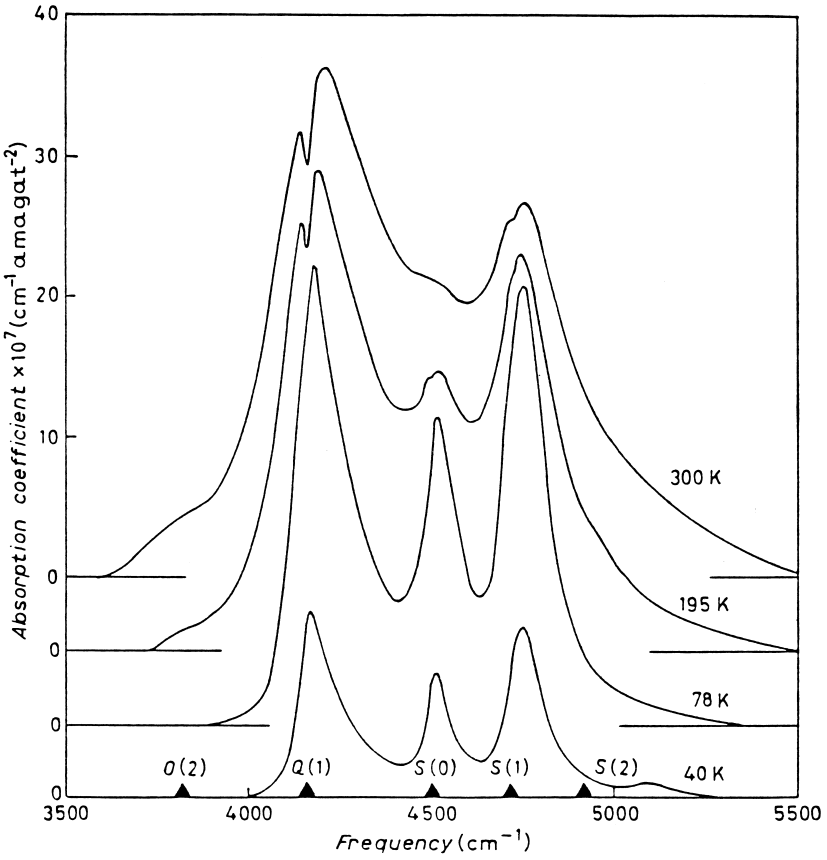


FIGURE 2 The collision-induced absorption spectrum of gaseous hydrogen of moderate density at infrared frequencies near the fundamental band of H_2. [After Hunt, J. L., and Welsh, H. L. (1964). *Can. J. Phys.* **42,** 873.]

simultaneous rotovibrational transitions in both H_2 molecules, e.g., a purely rotational transition in one molecule and a purely vibrational transition in the other. The latter cause absorption near sums and differences of the rotovibrational transition frequencies of the H_2 molecule, duly broadened by the short duration of the collision.

At the four temperatures shown in Fig. 2 the spectra are quite similar, except that at the higher temperatures the lines are broader, reflecting the decreasing duration of the average collision with increasing temperature, i.e., with increasing relative speed of the collisional encounters, Eq. (1).

The three diffuse lines seen so clearly in the Fig. 2 are called the Q and S lines of the H_2 molecule. However, one must keep in mind that none of these lines can be due to single, non-interacting H_2 molecules, owing to the inversion symmetry of H_2 which is inconsistent with the existence of an electric dipole moment. Hydrogen molecules are infrared-inactive, and rarefied hydrogen gas shows virtually no absorption at such frequencies. The spectrum shown (along with others shown below in Figs. 3, 4, and 9) are observable in compressed gases only and are collision induced. Significantly, there must be at least one interacting atom or molecule nearby for these lines to appear. An important point to be made here is that this type of rotovibrational absorption spectra occurs universally in virtually all molecular gases and in mixtures of atomic and molecular gases. Moreover, the absorption of pairs of molecules is

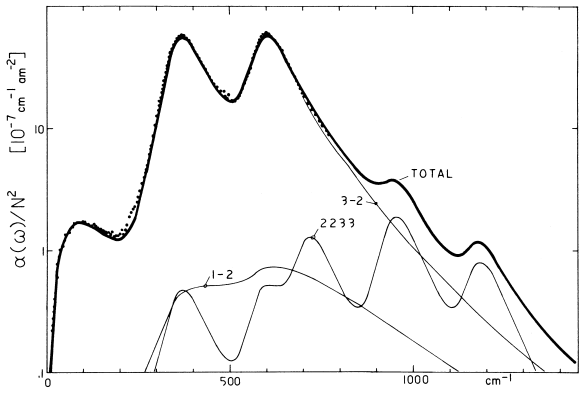


FIGURE 3 The binary collision-induced rototranslational absorption spectrum of hydrogen at the temperature of 77.4 K. The dots represent the measurement. The heavy solid curve represents a calculation based on first principles. That result is actually composed of contributions from mainly three different induced dipole components, which are sketched lightly and marked 1-2, 2233, 3-2. The main features are the rotational $S_0(0)$ and $S_0(1)$ lines centered at 354 and 585 cm^{-1}. A translational peak near 100 cm^{-1}, arising from orientational transitions, is also discernible. The theoretical structures at 940 and 1170 cm^{-1} arise from double transitions $S_0(0) + S_0(1)$ and $2S_0(1)$. [After Meyer, W., Frommhold, L., and Birnbaum, G. (1989). *Phys. Rev. A* **39,** 2434.]

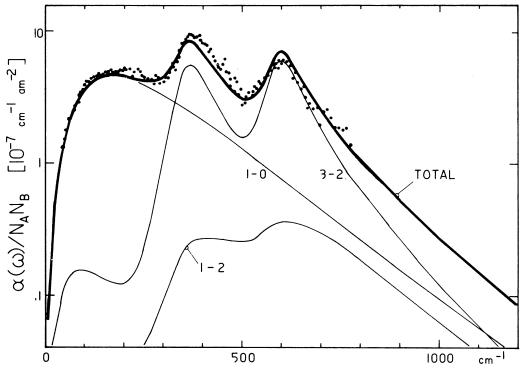


FIGURE 4 The enhancement of the binary collision-induced rototranslational absorption spectrum of hydrogen obtained by the addition of helium to hydrogen at the temperature of 77.4 K. The dots represent the measurement. The heavy solid curve represents a calculation based on first principles. The spectra of several contributing dipole components (marked 1-2, 1-0, 3-2) are sketched lightly. The $S_0(0)$ and $S_0(1)$ lines of H_2 are seen near 354 and 585 cm^{-1}. A translational spectrum much stronger than in Fig. 3 is also noticeable, which arises from the dissimilar nature of the collision partners. [After Meyer, W., and Frommhold, L. (1986). *Phys. Rev. A* **34,** 2237.]

not limited to the fundamental band of H_2 (or of any other molecule in collisional interaction with atoms or another molecule). Instead, it occurs also at the other vibrational bands, e.g., the overtone bands where vibrational quantum numbers v change by 2 or some larger integer, at frequencies in the near infrared and even the visible, and in the rototranslational band in the far infrared (see Figs. 3, 4, and 9).

2. Mixtures of Gases

The collision-induced spectral features may quite generally be enhanced by admixtures of other atomic or molecular gases. In binary gas mixtures, say of hydrogen and helium, at gas densities where binary interactions prevail, one may quite naturally distinguish the supramolecular spectra of H_2–H_2 and H_2–He pairs; absorption by the former varies as the hydrogen density squared and absorption by the latter varies as the product of hydrogen and helium densities. If binary mixtures of molecular gases, say of hydrogen and nitrogen, three types of collision-induced spectra can be distinguished: those of H_2–H_2, H_2–N_2, and N_2–N_2 pairs, again on the basis of their density dependences. Mixtures of more than two gases will quite generally show contribution of all possible pairs, and at higher densities triples, . . . that are consistent with the existence of electric dipole moments during interaction. An example of such enhancement of collision-induced absorption in the mixture of hydrogen and helium is given below (Fig. 4).

3. Rototranslational Spectra

Figure 3 shows the collision-induced absorption of compressed hydrogen gas in the far infrared (dots), i.e., at much lower frequencies than the absorption spectra shown in Fig. 2. As we noted above, in the discussion related to Fig. 1, as the frequencies approach zero, absorption falls off to zero for several reasons, one of them being stimulated emission. The rapid increase of absorption with frequency increasing from zero and the first broad peak correspond to translational absorption. The next two peaks near 354 and 585 cm^{-1} are the collision-induced $S_0(0)$ and $S_0(1)$ lines of H_2. The remaining two broad peaks are double rotational transitions of the type $S(0)+S(1)$ and $S(1)+S(1)$, combined with a change of the translational state of the pair *by absorption of a single photon*—truly a supramolecular feature. We note that besides the measurement (dots), a calculation based on first principles is shown in Fig. 3; theory reproduces all aspects of the measurement closely and on an absolute frequency scale and with precision.

If helium is added to the hydrogen, collision-induced absorption is enhanced in proportion to the helium partial density. The enhanced absorption is also proportional to the hydrogen density, indicating as its origin the H_2–He pair. Figure 4 shows the enhancement spectrum, which is recorded in the same frequency range as Fig. 3. The absorption by H_2–He pairs is again zero at zero frequency, rises to a first translational peak and two further peaks, the $S_0(0)$ and $S_0(1)$ lines of H_2, and falls off at higher frequencies; note the absence of double transitions in this case. Another noteworthy fact is the very strong translational peak which is characteristic of enhancement spectra of virtually all mixtures of gases.

4. Many-Body Effects

The collision-induced absorption spectra mentioned thus far were all obtained at densities substantially less than liquid state densities; the observed absorption is primarily due to exactly two interacting atoms or molecules. Of course, when we go to densities approaching liquid or solid state densities (e.g., several 100 amagats), molecular complexes of more than two molecules will also shape the observable spectra. While initially collision-induced absorption becomes rapidly more important with increasing density, many-body contributions will quickly modify the shapes and (normalized) intensities, relative to the binary spectra shown above.

One such many-body effect that is very striking actually appears at moderate or even small gas densities: the so-called intercollisional interference. However, at low density that effect is limited to a small number of relatively narrow frequency regions so that the binary character of the bulk of low-density spectra is undeniable. Specifically, dipoles induced in subsequent collisions tend to have a significant anticorrelation, resulting in partial cancellations. At the higher photon frequencies, this phase relationship is scrambled and amounts to virtually nothing, but at the lowest frequencies absorption dips results, e.g., at zero frequency and at the Q and (less strikingly) S line centers. Such dips are discernible in Fig. 2 at the higher temperatures. The dips may be viewed as inverted Lorentzian line profiles which broaden roughly in proportion to the density variation (see Fig. 5). The intercollisional interference dips are often narrow in comparison to the collision-induced Q or S lines on which they reside, because their width is given by the reciprocal *mean time between collisions*, when the widths of the intracollisional lines are given by the reciprocal *mean duration* of a collision. Under conditions where binary spectra can be recorded, the former is generally substantially longer than the latter. (However, at liquid densities the intercollisional dips may be very broad.) The intercollisional dip is a true many-body feature.

Figure 6 attempts to demonstrate the similarities and differences of the hydrogen spectra of highly compressed gas, of the liquid gas, and of the solid gas. Whereas the curve labeled *50 atm, 78 K* is still similar to the 78 K curve shown in Fig. 2, the uppermost curve in Fig. 6 was recorded at a much higher pressure (4043 atm) than the 300 K curve of Fig. 2. The intercollisional dip is now much broader (from the points marked Q_p to Q_r); the normalized absorption coefficient $\alpha/(\rho_1\rho_2)$ is no longer invariant under density variation. When we compare the spectra of the gas at 78 K with those of the liquid (at 17.5 K) and solid (at 11.5 K), we notice not only a sharpening of the lines with decreasing temperature, but a few new features may also be noticed. The spectrum of the solid shows broad bands that arise from combination tones of molecular frequencies and lattice frequencies (phonon spectra). The long extension of the phonon spectra toward higher frequencies is probably due to multiple phonon generation. Three weak double transitions of the type S_1+S_0 are also discernible. At higher spectral resolution than was employed in Fig. 6, the $S(0)$ and $S(1)$ groups show weak single transitions, $S_1(0)$ and $S_1(1)$, and much stronger double transitions of the type $Q_1(j)+S_0(j)$, with $j=0$ and 1, and the Q branch shows fine structures related to orientational transitions of two ortho-H_2 molecules.

5. Infrared-Active Gases

The emphasis thus far was on infrared-inactive gases. Spectral lines that are allowed in the individual molecules will at high gas densities be accompanied by a broad collision-induced background. In such a case, it will often be difficult to separate allowed and induced components

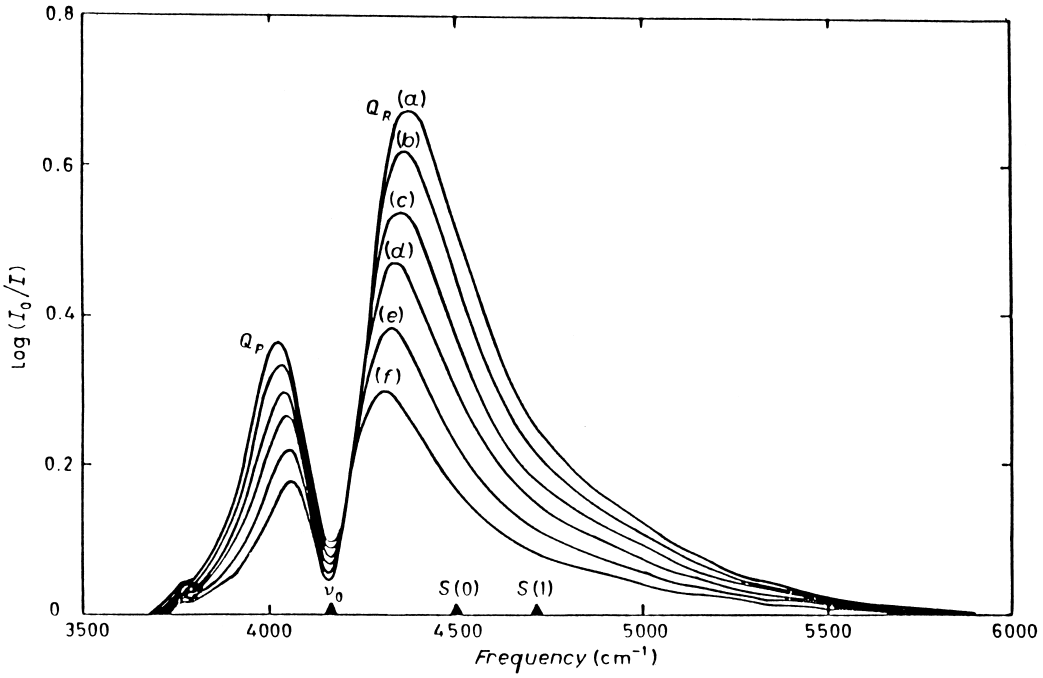


FIGURE 5 The intercollisional dip in the *Q* branch of H_2, here due to H_2–He collisions, at various densities. The absorption pathlength was 4 cm, the temperature was 298 K, and the mixing ratio of hydrogen and helium was fixed to 1:18. Helium densities from (a) to (f) are 1465, 1389, 1304, 1204, 1088, and 950 amagats. [After Gush, H. P., et al. (1960). *Can. J. Phys.* **38,** 180.]

if their intensities differ widely. Nevertheless, a number of very careful measurements exist in deuterium hydride (HD) and in mixtures of HD with rare gases at densities where binary interactions prevail.

The HD molecule is infrared-active, because the zero-point vibrational motion of the proton is slightly greater than that of the deuteron—a non-adiabatic effect. As a consequence, one side of the HD molecule is more

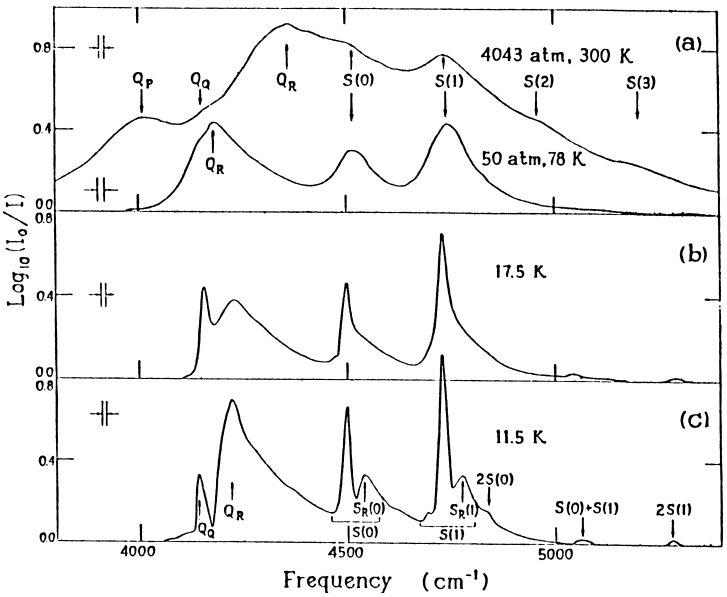


FIGURE 6 The collision-induced absorption spectrum of hydrogen in the fundamental band of the H_2 molecule in (a) the compressed gas, (b) the liquid gas, and (c) the solid gas. [After Hare, W. F. J., and Welsh, H. L. (1957). *Can. J. Phys.* **36,** 88.]

positively charged than the other; thus, a weak dipole moment exists—rotovibrational molecular bands thus occur. Collision-induced bands also occur, basically at the same frequencies as the allowed lines but more diffuse. The two dipole components are of comparable magnitude so that quite striking interference effects result. We note that the permanent dipole moments of the more common polar molecules are generally much stronger than the induced dipoles so that such interferences are harder to detect. Nevertheless, clear indications of such an interference have been discovered in the far wings of allowed spectral lines in compressed gases.

6. Electronic Collision-Induced Spectra

All spectra mentioned thus far arise from supramolecular transitions that leave the electronic states of the collisional partners unchanged. The possibility of the electronic state changing is, of course, finite if the photon energies are sufficiently large. Collision-induced electronic spectra typically occur at higher photon energies (e.g., in the ultraviolet region of the electromagnetic spectrum) than we considered above. However, a few of the common molecules possess electronic states with excitation energies low enough so that collision-induced spectra in the infrared, visible, and near ultraviolet region occur.

Absorption of a photon $\hbar\omega$ by an interacting molecular pair $A + B$ can symbolically be written as

$$A + B + \hbar\omega \rightarrow A' + B' + \Delta E,$$

where ΔE represents a change of the (e.g., translational) energy of the pair and primes indicate a possible change of the internal states of A and/or B. We may distinguish two different supramolecular collision-induced absorption processes: one affects the rotovibrational energies of A, B only and the other involves electronic transitions. Electronic collision-induced absorption typically requires visible and ultraviolet photons. Rotovibrational collision-induced absorption occurs normally in the microwave and infrared regions of the spectrum. The boundaries are, of course, not rigid and depend very much on the specific systems under consideration.

Collision-induced rotovibrational transitions are often studied at atmospheric and higher densities. Collision-induced electronic transitions, on the other hand, are usually studied at much lower densities, where absorption would be difficult to measure. In that case, other detection schemes, such as the ensuing product fluorescence, are employed.

As mentioned above, it has long been established that compressed oxygen has several diffuse absorption bands in the visible and infrared that are unknown from studies of rarefied oxygen and air (Janssen 1885). The intensities of these bands vary as density squared, indicating absorption by *pairs* of O_2 molecules. These bands are now understood to correspond to the "forbidden" electronic transitions from the ground state, $X^3\sum_g^-$, to the electronically excited states $a^1\Delta_g$ and $b^1\sum_g^+$ of O_2; simultaneous vibrational transitions of one or both of the interacting O_2 molecules may also occur in the process. Individuals who have handled liquid air are familiar with the blue tint of liquid oxygen which is caused by these electronic collision-induced absorption bands in the red portion of the visible spectrum. Simultaneous electronic transitions in *both* interacting O_2 molecules are also observed at shorter wavelengths. Similar bands of O_2–N_2 pairs have long been known, too. A few other molecules are known to have quite analogous, collision-induced electronic absorption bands in the visible and near ultraviolet regions of the spectrum in a compressed gas environment.

7. Line Profiles

The individual line shapes of the collision-induced spectra have previously been described by a Lorentzian profile, as for pressure-broadened lines. The width of the former is approximately given by the mean reciprocal duration of a collision, the ratio of root mean square speed v and range ρ of the induction mechanism (which is roughly of the order of the collision diameter, $\rho \approx \sigma$). In the low-density limit, v/ρ is density independent. However, the wings of collision-induced lines are roughly exponential and fall off much faster than the Lorentzian profile. Collision-induced spectral features are strikingly asymmetric (due to the principle of detailed balance), whereas Lorentzian profiles are not. For all these reasons, generalized spectral model profiles have been found that describe collision-induced profiles quite well. They are analytical and nearly as simple as the Lorentzian function, but without its limitations. These usually include the Lorentzian profile as a limiting case.

8. Charged Systems

So far, the collisional systems considered were all neutral. It is noteworthy that, in collisions involving one electrically charged particle, induction occurs, which in fact is much stronger than the quadrupole-induced dipoles discussed above. For example, induced spectra involving ion-neutral collisions and electron-neutral collisions (polarization bremsstrahlung) have been observed. It turns out that much stronger allowed electronic spectra usually prevail under conditions in which significant concentrations of free electrons or charged particles are observed. These often mask the weaker, induced spectra.

9. Nonlinear Interactions

With the advent of the laser it became possible to investigate non-linear interactions of light with matter. We just mention hyper Raman, stimulated Raman, coherent anti-Stokes Raman gain (SRG) spectroscopies, along with various other multi-photon processes, which have widened the scope of molecular spectroscopy dramatically. Most of these also have a collision-induced counterpart, much as the cases discussed above.

10. van der Waals Molecules

van der Waals forces are weak when compared to chemical forces that bind the common molecules—so weak that most dimer–monomer collisions will destroy (dissociate) the dimer. There is a certain destruction rate of dimers which balances the formation rates and keeps bound dimer concentrations typically at a low level. The average lifetime of bound dimers amounts, in many cases, to just a few mean free times between monomer collisions, or roughly $\approx 10^{-9}$ s in air at standard temperature and pressure; with increasing density and temperature the lifetimes decrease correspondingly. For comparison, we note that a collisional complex may be thought to "exist" for the duration of a fly-by collision, or roughly 10^{-12} s or so, nearly independent of density. According to Heisenberg's uncertainty relation [Eq. (1)], typical dimer bands may be much "sharper" than the collision-induced lines, if they are not severely pressure broadened.

The infrared spectra of van der Waals molecules are due to the same collision-induced dipole moments that generate the Q, S, etc. lines of the collision-induced spectral profiles shown above. Dimer features thus appear near the centers of these ("forbidden") lines in the collision-induced spectra, i.e., near zero frequency (where they are difficult to record), and at the Q, S, etc. lines of the (forbidden) rotovibrational bands where the dimer features were actually discovered. We note that collision-induced Raman spectra show similar dimer signatures in the vicinity of the Rayleigh line and at the other line centers discernible in the collision-induced Raman spectra.

van der Waals dimer bands are known to be highly susceptible to pressure broadening; this is one reason why it took so long to actually record dimer bands. Moreover, if one wants to resolve dimer bands spectroscopically, relatively high spectral resolution must be employed. We note that the spectra shown above were recorded with low resolution and relatively high pressures so that the dimer features seem absent. However, in recent years in several collision-induced vibrational spectral bands, dimer bands could be recorded and analyzed, and invaluable new knowledge concerning intermolecular interactions has been obtained in this way. The structures attributed to the $(H_2)_2$ dimer, which are indicated by the small rectangle near the $S_0(0)$ line center in Fig. 9 below, were seen in the Voyager spectra of Jupiter and Saturn, in which the conditions are more favorable than in the older laboratory measurements (low density and long absorption path lengths). For most binary systems, similar dimer structures must be expected. Spectral transitions involving dimers are from bound-to-bound and bound-to-free states of the pair, while typical collision-induced spectra correspond to free-free transitions of the pair.

D. Collision-Induced Emission

Any gas that absorbs electromagnetic radiation will also emit; supramolecular absorption and supramolecular emission are inseparable. However, cold gases will emit in the far infrared, which will often go unnoticed. However, the emission spectra of the outer planets are well known (see Fig. 9 below). Striking supramolecular emission occurs in hot and dense environments, e.g., shockwaves, "cold" stars, etc.

E. Collision-Induced Dipoles

In most cases when absorption or emission of electromagnetic radiation occurs, one can identify an electric dipole moment that is responsible for such spectroscopic processes. Rotating or vibrating electric dipoles emit (and absorb) at the frequencies of rotation and vibration; translationally accelerated charges (dipoles) emit a continuum. Emission and absorption take place in *transitions* between certain quantum states. For example, molecules with a permanent electric dipole moment, such as HCl or H_2O, emit in transitions between rotovibrational states if certain selection rules (i.e., conservation of energy and angular momentum) are satisfied. Even if no permanent dipole moment exists, molecules may emit in transitions between *electronic* states, but these transitions typically require more energy than the rotovibrational ones: a higher photon energy for absorption and higher excitation energy (e.g., higher temperatures) for emission. At room temperature, at frequencies in the infrared and in rarefied gases, the common homonuclear diatomic gases (hydrogen, nitrogen, etc.) do not undergo electronic transitions; no absorption is observed because their inversion symmetry is inconsistent with the existence of a permanent dipole moment.

Supramolecular systems, on the other hand, usually do possess a "permanent" dipole moment during their short lifetime. Four mechanisms are known that induce an electric dipole moment in two or more interacting molecules: (1) multipole induction, (2) exchange force

interactions, (3) dispersion interaction, and (4) molecular frame distortion—the same mechanisms that are familiar from the studies of the intermolecular forces.

Multipole induction arises from the fact that all molecules are surrounded by an electric field. While molecules are electrically neutral, the electric field surrounding each molecule is set up by the internal electronic and nuclear structure of the molecule. It may be described by a multipole expansion, i.e., by a superposition of dipole, quadrupole, octopole, ... fields. For example, the monopole and dipole terms are zero for all neutral homonuclear diatomic molecules; in this case the lowest order multipole is a quadrupole. When two such molecules interact, the collisional partners are polarized and thus possess momentarily—for the duration of the collision—dipole moments that interact with electromagnetic radiation. In the case of pure compressed hydrogen gas, quadrupolar induction provides nearly 90% of the total induced absorption. Since the quadrupole field rotates with the molecule, collision-induced rotational S lines are quite prominent in the spectra of compressed hydrogen and, of course, of all similar molecules.

Exchange forces control the repulsive part of the intermolecular interactions. In a collision at near range, when the electronic charge clouds of the collisional partners overlap, a momentary redistribution of electric charge occurs that is caused by electron exchange (Pauli exclusion principle). Especially when dissimilar atoms or molecules are involved, a dipole moment results from this redistribution. In most cases, the partner with fewer electrons temporarily assumes a positive charge, and the other assumes a negative one. This mechanism is usually the dominant one when dissimilar particles collide (as is the case for the spectra shown in Figs. 1 and 4). Exchange force-induced dipoles in molecules can also have a certain anisotropy of quadrupolar or higher symmetry. Examples of spectral features induced by anisotropic overlap are shown in Figs. 3 and 4 (components marked 1-2) and also implicitly in most of the other figures.

Dispersion forces control the attractive part of the intermolecular interactions. Over moderately wide separations, atoms or molecules interact through dispersion forces that are of an electric nature and arise from electronic intercorrelation. For dissimilar pairs, these are associated with a dipole moment whose asymptotic strength is proportional to the inverse seventh power of the intermolecular separation, and the polarity is typically the opposite of the overlap-induced dipole. The dispersion dipole is usually weaker than multipole-induced and overlap-induced dipoles, but is usually discernible in discriminating analyses.

Molecular frame distortion by collisions may break temporarily the high symmetry many molecules possess. For example, the unperturbed CH_4 molecule, owing to its tetrahedral symmetry, has a zero dipole moment, even though the C–H bond is strongly polar: the vector sum of these four dipoles is zero as long as the exact tetrahedral symmetry persists. However, the momentary displacement of one of the hydrogen atoms by a collision will immediately produce a non-zero dipole moment which then may interact with radiation.

Typically, three or all four of the interaction-induced dipole moments are present at the same time. They cause collision-induced absorption, as well as the absorption of bound van der Waals systems. The point to be made here is that supramolecular complexes may have properties very different from those of the (non-interacting) monomers, such as a dipole moment even if the non-interacting constituents may be without.

We note that collision-induced dipoles are roughly one to three orders of magnitude weaker than the permanent dipoles of ordinary molecules. Correspondingly, the collision-induced spectra of individual pairs of molecules are typically much less intense than analogous spectra of ordinary molecules. Nevertheless, at high density the spectra can be intense. Observable intensities, when integrated over a line or a spectral band, are proportional to the number of molecules, that is, to the density, if the spectra of polar molecules are considered. For binary-induced spectra, on the other hand, integrated intensities are proportional to the number of *pairs* in a sample. If N molecules exist in a box, we have $N(N-1)/2$ pairs. Since, in all practical cases, N is a very large number, N and $(N-1)$ are nearly indistinguishable, and we can approximate the number of pairs by $N^2/2$. This quadratic density dependence may, at high densities, generate substantial intensities from pairs of molecules even if the individual collision-induced dipole moments are weak.

1. Ternary Dipoles

Recently, semi-empirical models of the most significant dipole components of three interacting homonuclear diatomic molecules have been obtained. Ternary-induced dipoles are the vector sum of the pairwise-additive dipoles (which are often well known) and the (previously essentially unknown) irreducible ternary dipoles. The model is consistent with three different experimental manifestations of the irreducible dipole component, in this case of compressed hydrogen gas: the third virial coefficient of the collision-induced, integrated absorption spectrum of the fundamental band of H_2; the triple transition $3Q_1$ by absorption of a single photon at 12,466 cm^{-1}; and certain features of the intercollisional dip of the Q_1 line, all of which are significantly shaped by the irreducible induced dipole.

When two molecules collide at near range, exchange forces redistribute electronic charge. As a result, a collision-induced dipole may thus be created, along with higher electric multipole moments. The strongest of these is the exchange-force-induced quadrupole moment. In the electric field of that quadrupole a third molecule will be polarized and an irreducible ternary dipole is thus created. There are several other mechanisms that contribute some to the irreducible dipole moment, but the exchange quad rupole-induced dipole (EQID) appears to be by far the most significant contribution to the irreducible ternary-induced dipole of three H_2 or similar molecules.

2. Dipoles and Dense Matter

In dense systems (e.g., in liquids and solids) the three-body and probably higher order cancellations due to destructive interference are most important. We distinguish between two components of translational spectra: one due to the diffusive and the other to the oscillatory ("rattling") motions of the molecules in a liquid. The latter is the analog of the intercollisional spectrum and consists of a dip to very low intensities near zero frequencies. While translational spectra of monatomic liquids are rather well understood, this cannot be said of those of the molecular liquids.

3. Electronic-Induced Dipoles

The induction proceeds, for example, via the polarization of molecule B in the multipole field of the electronically excited atom A, by long-range transition dipoles which vary with intermolecular separation R as R^{-3} or R^{-4}, depending on the symmetry of the electronic excited state involved. At near range, a modification of the inverse power dependence due to electron exchange is often quite noticeable, much as this is known for the rotovibrational collision-induced spectra discussed above.

III. COLLISION-INDUCED LIGHT SCATTERING

If a molecular substance is irradiated by the intense monochromatic light of a laser, light of the same frequency ν_0 is scattered (the Rayleigh line). Besides, at lower frequencies ($\nu < \nu_0$, Stokes wing), as well as at higher ones ($\nu > \nu_0$, anti-Stokes wing), other lines may appear that are shifted relative the incident frequency by certain rototranslational molecular transition frequencies. Many such states normally exist, certainly for the Stokes wing. This basically describes the Raman spectrum of ordinary molecules.

If laser light is directed at a sample of a rarefied monatomic (rather than a molecular) gas, no Raman spectra besides the Rayleigh line are observed, owing to the absence of rotovibrational states. However, at high enough densities, continuous Stokes and anti-Stokes wings appear that are collision induced. These arise because the polarizability of two interacting atoms differs slightly from the sum of polarizabilities of the (non-interacting) atoms: the collisional complex possesses an *excess* polarizability which generates these wings. The wings are continuous because the translational energies of relative motion of the two atoms are continuous, i.e., unlike the rotovibrational energies of bound diatomic molecules which are quantized like all periodic motion. Collision-induced Raman spectra of the rare gases were predicted and soon after demonstrated in actual measurements by G. Birnbaum and associates in 1967.

It has long been known that two types of light scattering by gases must be distinguished; so-called polarized scattering maintains the polarization of the incident laser beam, but depolarized scattering does not. Depolarized scattering arises from the anisotropic part of the polarization tensor of a molecule; it rotates the polarization plane of the incident laser beam randomly so that nearly no memory of the polarization of the incident beam remains in the scattered light at a given frequency.

Figures 7 and 8 show the Raman spectra of two interacting He atoms as obtained at gas densities corresponding to 10 or 20 times their density at standard temperature and pressure (1 atm and 273 K). Similar spectra are obtained in a gas of the rare isotope, ^{3}He. In an actual measurement the depolarized (Fig. 7) and the polarized (Fig. 8) spectra are superimposed and must be separated artificially, on the basis of their different polarizations, which introduces considerable uncertainty in the weaker of the two (Fig. 8). Similar collision-induced Raman spectra are known for the other rare gases and for mercury vapor, another monatomic gas.

The profiles of the Raman spectra, Figs. 7 and 8, resemble the spectral density function $g(\omega)$ seen in collision-induced absorption (Fig. 1). In fact, Heisenberg's uncertainty principle, Eq. (1), will directly relate the observed half-widths of these profiles and the mean duration of the collision, just as this was pointed out above in Fig. 1. Note that the range of spectroscopic interaction of the trace is shorter than that of the anisotropy of the polarizability tensor, so that the polarized spectra are actually more diffuse than the depolarized ones, under otherwise comparable conditions. Furthermore, we note that the reader may find the profiles shown in Figs. 7 and 8 to look different from the dashed profile in Fig. 1. The apparent differences of shape are, however, largely due to the logarithmic intensity

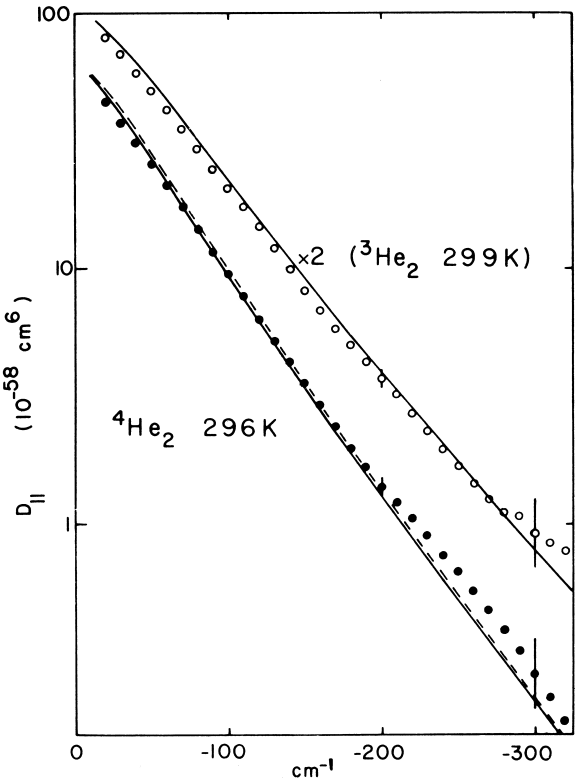


FIGURE 7 The Stokes side of the binary collision-induced depolarized Raman spectrum of the helium diatoms of the abundant and the rare isotope of helium at room temperature. The intense Rayleigh line (at zero frequency shift) was suppressed. Dots and circles represent the measurement; the solid lines represent a calculation based on the fundamental theory. For clarity of the display, the spectrum of the rare isotope was multiplied by a factor of 2. [After Dacre, P. D., and Frommhold, L. (1982). *J. Chem. Phys.* **76,** 3447.]

scales used in Figs. 7 and 8, when a linear scale was used in Fig. 1.

Similar work with compressed *molecular* gases demonstrates that the collision-induced Raman spectra are as universal as their collision-induced absorption counterpart discussed above. In molecular gases, additional rotovibrational lines and bands appear, much like those seen in the infrared, regardless of whether the gases are Raman-active or not.

Just as in the case of collision-induced absorption, collision-induced Raman spectra of binary and many-body complexes differ significantly. In the low-density limit in Raman-inactive gases, the binary collision-induced spectra are dominant, and the intensities vary as the square of density. At increasing densities a point is reached where many-body cancellations of the binary intensities are observed, owing to ternary (and higher order) interactions. At still higher densities, and in the case of liquids and solids, many-body collisional complexes contribute to the observable spectra which further modifies the induced spectra seen in the low-density limit.

A. Collision-Induced Polarizabilities

All molecules are polarized in an external electric field, i.e., the field pulls the electrons slightly off to one side and the nuclei to the other so that a "field-induced dipole moment" **d** results,

$$\mathbf{d} = \mathbf{A} \cdot \mathbf{F}, \tag{2}$$

where **F** is the electric field strength. If an alternating field (laser) is used, the field-induced dipole moment oscillates and thus radiates at the same frequency as the incident light (Rayleigh scattering). Since the polarizability **A**, a 3×3 tensor with two invariants, trace, and anisotropy, depends on the molecular orientation and the nuclear structure of the molecule, the scattered light is also modulated by certain molecular rotation and vibration frequencies. These modulations make up the familiar Raman spectra of the ordinary (unperturbed) molecules.

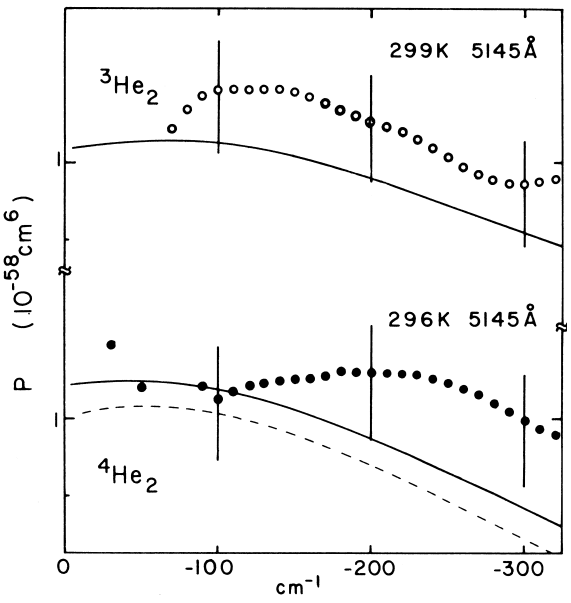


FIGURE 8 The Stokes side of the collision-induced polarized Raman spectrum of the helium diatoms at room temperature. The intense Rayleigh line has been suppressed. Dots and circles represent the measurements. The size of the error bars give an impression of the uncertainty of the measurement, which was substantial, owing to the low scattered intensities in the presence of the depolarized background. The solid lines are calculations based on the fundamental theory. [After Dacre, P. D., and Frommhold, L. (1982). *J. Chem. Phys.* **76,** 3447.]

Supramolecular Raman spectra are generated by the *excess* of the polarizability invariants of the interacting complex over the sum of the corresponding invariants of the non-interacting (i.e., widely separated) members of the supramolecular complex. A simple example will illustrate this: The atoms of the monatomic gases are completely isotropic; the field-induced dipole will always be exactly parallel to the applied field, $\mathbf{d} \parallel \mathbf{F}$, so that the anisotropy (which would rotate the dipole relative to the field) is zero—as long as the atoms are sufficiently spaced apart (rarefied gas). At higher densities, one cannot ignore the likelihood of another atom B sailing by close enough so that the local field at the position of atom A is perturbed: it is now the vector sum of the external field and the electric field set up by the dipole induced in atom B. For example, the resulting pair dipole moment and thus the polarizability is typically greater than the sum of atomic polarizabilities if the internuclear axis is parallel to the field; it is smaller if the internuclear axis is perpendicular to the field. At the other orientations, the plane of polarization of the scattered light is actually rotated relative to that of the incident beam. In other words, the anisotropy of the atom pair is no longer zero and a collision-induced depolarized Raman spectrum exists. To a large extent, this is simply due to the fact that the effective exciting field in the former case is enhanced by the vicinity of the dipole in the collisional partner, while in the latter case it is weakened (this is the dipole-induced-dipole interaction, often abbreviated DID). To some lesser extent, electronic overlap at near range and dispersion forces at more distant range have an effect on the polarizability of the pair, and the DID model must be considered an approximation that describes roughly 90% of the pair polarizability of most systems of interest; it is thus quite useful.

The field-induced, excess dipole moment of the pair goes from very small values at large initial separations to a maximum at the point of nearest approach, whereupon it falls off again. During this time, the dipole is oscillating with the very high frequency of the exciting field. The Raman spectra (e.g., the Fourier transform of the autocorrelation function of the time-varying excess dipole moment) consist of the Rayleigh line and the Stokes and anti-Stokes wings. While the Rayleigh line arises mostly from the scattering of monomers so that its width is not affected by the short lifetime of the collisional complexes, the Stokes and anti-Stokes wings are of a width consistent with the reciprocal lifetime of the collisional complex.

We note that the pressure-induced depolarization of scattered light by dense, isotropic gases is closely related to the mechanism that produces the depolarized collision-induced spectra.

IV. VIRIAL EXPANSIONS

A. Equation of State

A gas in thermal equilibrium obeys the equation of state,

$$p = kT\left\{\frac{N_A}{V} + \frac{B(T)}{V^2} + \frac{C(T)}{V^3} + \cdots\right\},$$

where p, T, V, k designate pressure, temperature, molar volume, and Boltzmann's constant, respectively; N_A is Avogadro's number, and $B(T)$, $C(T)$, ... are the second, third, ... virial coefficients of the equation of state. The infinite series to the right is called a virial expansion.

In the limit of a highly diluted gas ($V \to \infty$), the right-hand side of this expression is equal to ρkT, where $\rho = N_A/V$ is the density of the gas. This is the *ideal gas approximation* of the equation of state which approximates the gas as a collection of non-interacting point particles—which is quite a reasonable model for rarefied gases. The second and higher terms in the virial expansion represent the effects of the intermolecular interactions. In particular, the second virial coefficient $B(T)$ expresses the effect of the strictly binary interactions upon the pressure of the gas. The third coefficient $C(T)$ describes the effect of the ternary interactions, and so forth. With increasing gas densities (e.g., in compressed gases) these virial coefficients become more and more important. At densities as high as those of liquids, the virial expansion becomes meaningless. Under such conditions every particle of the fluid is in simultaneous interaction with quite a number of other particles nearby.

The second virial coefficient is expressed in terms of the pair-interaction potential $V(R)$,

$$B(T) = -2\pi N_A^2 \int_0^\infty \left(\mathrm{e}^{-V(R)/kT} - 1\right) R^2\, dR.$$

Similarly, $C(T)$ can be expressed in terms of ternary interactions, of both the pairwise and the irreducible kind. Valuable information on intermolecular interactions has been obtained from measurements of the virial coefficients. The second, third, ... virial coefficients are functions of temperature only and can be calculated in terms of the interactions of two, three, ... molecules in the volume V. In other words, the N_A-body problem of the imperfect gas has been reduced to a series of one-, two-, three-body, ... problems which are much more tractable than the very difficult many-body problem of an amorphous fluid.

B. Other Virial Expansions

It is plausible that besides the equation of state there should be other thermodynamic functions, i.e., other properties

of matter that may be described by superposition of the effects of unitary, binary, ternary, ..., molecular interactions, that is, by a virial expansion. In fact, the discussion of collision-induced spectral intensities above suggests such a virial expansion of the observed line shapes and the integrated absorption coefficients.

C. Collision-Induced Spectra

Rarefied gases interact with electromagnetic radiation in proportion to density variations: if at a given frequency absorption exists, the absorption coefficient will, in general, double if gas densities (pressures) are doubled. However, with increasing number density of the gas, as we approach roughly $\approx 1\%$ of the liquid state densities, supramolecular absorption, emission, and light scattering becomes increasingly important: the contributions from *pairs* of molecules will increase as density squared; molecular *triples* contribute proportional to density cubed; etc. In other words, at not too high densities and at most (but not all!) frequencies of a given collision-induced band, a virial expansion of spectral intensities is possible and permits a separation of the contributions of monomers, dimers, trimers, ... It is clear that with increasing gas densities the collision-induced contributions must be of increasing importance—regardless of whether the dimers, trimers are van der Waals molecules or collisionally interacting complexes with fluctuating dipole moments.

The leading term of any virial expansion is due to the non-interacting monomer contributions. Contrary to the leading coefficient of the equation of state, the corresponding spectroscopic first virial coefficient vanishes if the molecules are infrared- or Raman-inactive; in that case the virial series starts with the second spectroscopic coefficient, expressible in terms of the pair-interaction potential and the induction operator (i.e., induced dipole or induced polarizability invariants), depending on whether we consider absorption and emission or the Raman process.

We note that certain sum formulas, e.g., the integrated intensity of a collision-induced band, may at intermediate densities be represented by a virial expansion.

D. Dielectric and Refractive Virial Expansions

Other equilibrium properties of gases and liquids are known that possess a virial expansion and are intimately related to the collision-induced spectroscopies. The density dependence of the relative dielectric constant ϵ of a gas is given by the Clausius-Mossotti equation,

$$\frac{\epsilon - 1}{\epsilon + 2} = \frac{A_\epsilon(T)}{V} + \frac{B_\epsilon(T)}{V^2} + \frac{C_\epsilon(T)}{V^3} + \cdots,$$

where A_ϵ, B_ϵ, ... are the first, second, ... dielectric virial coefficients. The dielectric coefficient provides a measure of the polarization of matter, and in principle there are two mechanisms that can be distinguished: orientation of existing (permanent) dipoles in the external electric field and generation of dipoles by field-induced polarization, $\mathbf{d} = A\mathbf{F}$. In other words, the first virial coefficient is given by the Debye expression,

$$A_\epsilon(T) = \frac{4\pi N_A}{3(4\pi\epsilon_0)}\left(A + \frac{d_p}{3kT}\right),$$

which is the sum of the field-induced and the permanent dipole contribution. Accordingly, the second dielectric virial coefficient, which represents the leading term describing the lowest order deviations from the dielectric ideal gas behavior, is written as

$$B_\epsilon(T) = \frac{2\pi N_A^2}{3\Omega(4\pi\epsilon_0)} \times \iint \left[A_{12}(R) + \frac{d_{12}^2}{3kT}\right] \mathrm{e}^{-V(R)/kT}\, d^3R\, d^2\omega_{12}.$$

The integration is over all positions and orientations of molecule 2 relative to molecule 1 (here assumed to be identical). The quantity Ω is defined by $\int d^3R\, d^2\omega_{12} = \Omega V$. The excess isotropic polarizability of the pair, $A_{12}(R)$, also called the collision-induced trace of the pair polarizability, is a function of intermolecular separation R. Similarly, the squared collision-induced dipole moment, d_{12}, more precisely the *excess* of the squared dipole moment of the pair above those of the non-interacting molecules, is also a function of separation and orientation.

Static dielectric properties are measured with static electric fields. If the frequencies of alternating fields approach those of visible light, the refractive index n, which is related to the dynamic (frequency-dependent) dielectric constant by $\epsilon = n^2$, becomes important. In this case an equation that is completely analogous to the Clausius-Mossotti expression, and which is commonly called the Lorentz-Lorenz equation, is of interest. It has formally the same virial coefficients as the former, only the polarizabilities are now measured at the frequencies of visible light. Furthermore, when at high frequencies the rotational inertia of polar molecules does not permit the molecules to orient fast enough in response to the applied alternating field, the orientational terms simply disappear from the expression for $B_\epsilon(T)$.

The isotropic pair polarizability A_{12} mentioned here (the trace of the pair polarizability tensor) is the very same quantity that controls the purely polarized collision-induced Raman process. Furthermore, the collision-induced dipole moments, d_{12}, whose squares occur in the

expression for the second virial dielectric coefficient, are clearly the same that cause collision-induced absorption (and emission), and intimate relationships exist between the second virial spectroscopic coefficient of absorption and the part of the dielectric coefficient $B_\epsilon(T)$ that arises from the orientational dependence.

E. Kerr Constant

If a uniform, strong electric field is applied to a fluid, it becomes birefringent. In that case the refractive index is no longer isotropic (independent of the direction of propagation of light). Instead, we distinguish refractive indices $n_\parallel$ and $n_\perp$ for propagation of light in a direction parallel and perpendicular to the electric field vector, which increasingly differ as the field strength increases (Kerr effect). Every substance has a Kerr constant K which determines how much $n_\parallel$ and $n_\perp$ will differ for a given field strength, F,

$$K = \lim_{F\to 0}\left\{\frac{6n(n_\parallel - n_\perp)V}{(n^2+2)^2(\epsilon+2)^2F^2}\right\}.$$

This effect is related to the optical anisotropy of molecules. In optically anisotropic gases, the molar Kerr constant varies linearly with gas density. However, as the density is increased, collision-induced optical anisotropies arise, which can be accounted for by a virial expansion such as

$$K = A_K(T) + B_K(T)\frac{1}{V} + C_K(T)\frac{1}{V^2} + \cdots.$$

The $A_K(T)$, $B_K(T)$, ... are the first, second, ... virial Kerr coefficients. A_K is the ideal gas value of the molar Kerr constant; for monatomic gases it is related to the hyperpolarizability. The second Kerr coefficient is given by

$$B_K(T) = \frac{8\pi^2 N_A^2}{405kT(4\pi\epsilon_0)}\int_0^\infty \beta_v(R)\beta_0(R)e^{-V(R)/kT}R^2\,dR,$$

where $\beta_0(R)$ and $\beta_v(R)$ are the (nearly equal) collision-induced anisotropies at zero frequency and the frequency v of the incident light, respectively. It is exactly this same collision-induced anisotropy that generates the depolarized collision-induced Raman spectra.

V. SIGNIFICANCE FOR SCIENCE AND TECHNOLOGY

A. Molecular Physics

Throughout this article numerous remarks have been made that illustrate the significance of the collision-induced spectroscopies for the study of molecular interactions. We summarize these by stating that complete binary spectra can be reproduced in all detail by a rigorous, quantum mechanical procedure if two functions of the molecular interaction are known: the intermolecular interaction potential and the pair induction operator (i.e., the collision-induced dipole surface if infrared absorption and emission are considered, or the collision-induced polarizability invariants for polarized or depolarized, collision-induced Raman spectra, respectively). Inversely, we may say that measurements of such collision-induced spectra *define* these functions of the interaction, certainly if accurate spectra over a wide frequency and temperature range are obtained. At present, no entirely satisfactory procedure is known to obtain these functions from spectroscopic measurements. Nevertheless, reasonably successful inversions of measurements exist which have generated valuable information concerning intermolecular interactions.

While ternary spectra are known and valuable pioneering work with various third virial coefficients exists, the precision of the data is usually somewhat limited. Consequently, the analyses have not always been as discriminating as one would like. The problem is, of course, the presence of the dominating binary process, combined with the contributions from four-body (and higher) interactions. A subject of considerable interest, namely, the separation of the irreducible parts of the three-body interactions from the pairwise component, has recently shown great promise.

B. Atmospheric Sciences

The interest of the planetary scientist in collisional absorption comes as no surprise. The most abundant molecules and atoms in space are non-polar (H_2, H, and He). In the dense and cool regions of space (i.e., in planetary atmospheres and "cool" stars), the most significant spectroscopic signatures that can be observed in the infrared are of the collision-induced nature. The atmospheres of the outer planets are opaque in the far infrared because of collision-induced absorption of H_2–H_2 and H_2–He pairs.

The exploration of the solar system by infrared spectroscopy has been an exciting and productive area of astronomical research, especially in recent times. Major goals are the detection of atmospheric constituents and their elemental and isotopic abundance ratios, which are so important to the understanding of the evolutionary process of the solar system, the establishment of thermal properties (i.e., brightness temperature and effective temperature), and the vertical thermal structure of the atmospheres ($p-T$ profiles). In all these endeavors collision-induced spectroscopy is an essential ingredient. While we have emphasized the hydrogen (H_2–H_2) collision-induced spectral contributions, other systems such as H_2–He, H_2–CH_4, and H_2–N_2 are also important under many conditions. The two systems mentioned last are, for example, of interest in

Titan's atmosphere, which like the Earth's atmosphere is composed primarily of nitrogen. It has, therefore, received much attention in recent years. The related spectra of the van der Waals dimers of such systems [e.g., $(H_2)_2$ and H_2N_2] appear to be of considerable interest for astrophysics for the same reasons.

The atmospheres of the outer planets are thought to be of a composition that resembles that of the primordial solar nebula. Therefore, the study of the composition might provide important answers to the ancient scientific problem of the origin of the solar system. Collision-induced spectroscopy can help to determine one of the most important parameters of primordial matter, the helium-to-hydrogen abundance ratio, which may be determined from the distinct features of the collision-induced spectra of H_2–H_2 and H_2–He and of the associated $(H_2)_2$ dimer spectra. (It is noteworthy that the H_2–He system is one of the few that do not form a bound dimer.)

Figure 9 shows the far infrared part of the emission spectra of the north equatorial region of Jupiter, recorded by the Voyager I IRIS spectrometer during the 1979 fly-by encounter. The frequency axis (abscissa) ranges from 200 to 600 cm^{-1}. Intensities are measured in units of *brightness temperature*, i.e., the temperature a black body has that emits the same intensity at the given wavelength. It is clear that high brightness temperature corresponds to high emission intensity. However, one should keep in mind the fact that brightness temperature is a highly non-linear measure of intensity.

At low frequencies, we notice structures that have been identified as molecular bands of the NH_3 molecule. Similar bands of other molecules, such as CH_4, occur at higher frequencies (>600 cm^{-1}) and are not shown in the figure. These strong (allowed) bands come from the deep interior of the atmospheres where temperatures are high. Of special interest here are the broad, relatively unstructured regions of the smallest intensities, extending from about 250 to beyond 600 cm^{-1}. These are the collision-induced rotational absorption lines of H_2 molecules that are collisionally interacting with other H_2 molecules or with He atoms. (These lines are the same shown in Figs. 3 and 4.) We note that helium atoms are the second most abundant species after H_2 in the atmosphere. Contrary to the NH_3 and CH_4 bands mentioned, these lines are forbidden in the non-interacting H_2 molecules. The broad dips of the emission spectra are centered near 354 and 585 cm^{-1}. The H_2 $S_0(0)$ and $S_0(1)$ lines are actually "dark fringes" in the thermal emission spectrum of Jupiter; their origin is completely analogous to the well-known dark Fraunhofer lines in the solar spectrum. The relatively cool outer regions of Jupiter's atmosphere are opaque at these frequencies, owing to collision-induced absorption; emitted radiation reflects the temperatures of these opaque regions. In contrast to the Fraunhofer lines, the collision-induced features are very broad because of the short durations of typical H_2–H_2 and H_2–He collisions [Eq. (1)]. We note that the hydrogen densities in the outer regions where collision-induced absorption takes place amount to about 0.4 amagat; the mean absorption pathlength amounts to roughly 20 km.

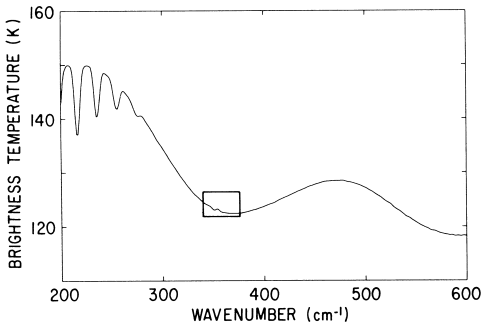


FIGURE 9 Emission spectrum of Jupiter's north equatorial belt obtained with the Voyager 1 IRIS far infrared spectrometer in the fly-by mission. The relatively sharp, striking structures at the lowest frequencies are NH_3 bands. The collision-induced $S(0)$ line of the H_2 molecule ranges from roughly 280 to 420 cm^{-1} as a broad, inverted feature. A similar dark and broad feature at higher frequencies is partially discernible, which is due to the $S(1)$ line of H_2, with a center near 585 cm^{-1}. The small rectangle near the center of the $S(0)$ line points out an interesting structure arising from bound-to-free transitions involving the van der Waals molecule $(H_2)_2$. [After Frommhold, L., Samuelson, R., and Birnbaum, G. (1984). *Astrophys. J.* **283,** L82.]

C. Astrophysics

In 1952, a few years after the discovery of collision-induced absorption, G. Herzberg pointed out the collision-induced $S_3(0)$ overtone structure of hydrogen in the spectra of Uranus and Neptune. This was the first direct evidence for the existence of molecular hydrogen (H_2) in the atmospheres of the outer planets, which consist of roughly 90% H_2 molecules! This direct detection of H_2 had to await the discovery of collision-induced absorption.

D. Applied Sciences

The liberation of observational data for astronomy, satellite-supported meteorology, and remote atmospheric sensing from the aggravating influence of the Earth's atmosphere has been a classical problem in the applied sciences. Precise quantitative knowledge of the coefficients of continuous absorption, especially in the far wings of spectral lines, and of their temperature dependence is indispensable for the solution of the inverse problem in satellite meteorology and weather prediction. The inverse problem attempts to reproduce accurately the distribution curves

of physical parameters of the atmosphere from measurements of spectral composition and emission. For these tasks the collision-induced spectra of the atmospheric constituents (e.g., N_2, O_2, and H_2O) are essential.

The propagation of laser beams through the atmosphere is affected by atmospheric extinction from the scattering and absorption of light, both of which have a significant collision-induced component. Long-range monitoring of various physical and chemical parameters of the atmosphere (LIDAR) is a promising new direction in science and engineering; it is affected by collisional spectroscopies. To some extent all laser communication and information transmission systems, locating and telemetering systems, and mapping and navigational systems require access to quantitative data describing the effect of a dense atmosphere on the parameters of laser beams, which serve as the carriers of information.

Photoattenuation at wavelengths in the extreme red wings of resonant lines of electronic transitions have a strong collision-induced component. The degree of attenuation increases rapidly with increasing temperature, which has a detrimental effect on the performance of gas lasers. Since every scattering process has a stimulated counterpart, the coefficient of collision-induced scattering is likely to increase with increasing laser power (stimulated collision-induced scattering) to severely limit the highest attainable internal power density of high-power lasers. Collision-induced dipoles are known to be the prime cause of far-wing absorption of radiation in excimer lasers and multiphoton processes. Volumetric heating of non-polar gases, liquids, and even solids is possible by utilizing collision-induced absorption lines of the systems involved. Other applications in laser physics and chemistry have been proposed that attempt to control collisional processes and involve collision-induced spectroscopic transitions and lasers.

Frozen deuterium-tritium mixtures may be used as nuclear fuel for inertial confinement fusion reactors. Collision-induced, vibrational-rotational spectra of liquid and solid mixtures of deuterium are known, which are the isotopic analogs of the hydrogen spectra shown in Fig. 5. However, new infrared lines in the tritiated solid hydrogens below about 11 K were observed, which are due to tritium molecules perturbed by the electrostatic field of nearby ions that were formed by the beta rays of a decaying tritium nucleus. This is a form of collisional induction by a charged particle. The new lines are apparently those of the fundamental band but are shifted by the strong field of the electric monopoles (Stark shift).

In recent years a considerable technological interest in the non-linear optical properties of liquids has evolved. It centers around the third-order susceptibility, which controls many aspects of optical signal processing, image processing, stimulated scattering, and so on. Molecular susceptibility is related to the polarizabilities that determine the Rayleigh and induced Raman spectra of the fluids.

VI. CONCLUSIONS

The examples of collision-induced spectra shown in this article are chosen for their relative simplicity. The spectra were those of complexes of atoms and simple molecules, recorded under well-defined laboratory conditions and reproduced from the fundamental theory with precision for a demonstration of the basic principles involved. These choices, however, do not indicate the scope of collisional induction, which actually encompasses (1) quite large molecules as well as the smallest ones; (2) virtually any gas or mixtures of gases, liquids, and solids; (3) spectra in virtually any frequency band of the electromagnetic spectrum, up to X-ray frequencies; and (4) optical phenomena observable at any temperature, from near absolute zero to tens of thousands of kelvin.

Collision-induced spectroscopy is the extension of the spectroscopy of ideal gases to one of real gases and to important aspects of the condensed state. It is thus a very practical science that continues to provide new understanding of molecular interactions. Almost from the moment of their discovery, the collision-induced spectroscopies have had an enormous impact in astrophysics and other disciplines. Their significance for science and technology seems to be ever increasing. The field is diverse and has prospered through the furtherance of many disciplines and technologies. Not only has the full extent of microwave, infrared, and Raman spectroscopy with low and high resolutions been mobilized, but these techniques had to be paired with other advanced technologies (e.g., ultrahigh pressure capabilities and laser and cryogenic technologies) before the now familiar, very general statements concerning the collision-induced spectroscopies could be made. New theoretical thinking, combining the elements of statistical mechanics, liquid state theory, thermodynamics, quantum chemistry, and molecular dynamics studies, had to be developed and supported by modern supercomputers for the simulation of measurements and quantitative tests of the assumptions made. Perhaps because of the great diversity of interests and resources that have been essential for all work in the collision-induced spectroscopies, only recently have a few major attempts been known to review the field and to collect the existing knowledge in a few conference proceedings and monographs. These are quoted below.

Under conditions that are not favorable for the occurrence of electronic or molecular spectra, that is, at low temperatures and if non-polar molecules are considered, collision-induced spectra can be quite prominent,

especially at high gas densities, in liquids and solids. Best known are the rototranslational absorption spectra in the far infrared and microwave regions of the nonpolar gases and liquids; vibrational absorption bands in the near infrared, analogous Raman spectra, especially of the Raman-inactive gases; and various simultaneous transitions in pairs and triples of interacting molecules. Collisionally induced spectra are ubiquitous in dense environments in almost any gas; the only exception of such absorption spectra seem to be the pure monatomic gases.

Collision-induced spectra and the spectra of van der Waals molecules (of the same monomeric species) are due to the same basic dipole induction mechanism. An intimate relationship of the induction mechanisms responsible for the collision-induced spectra with the dielectric virial properties of matter exists.

SEE ALSO THE FOLLOWING ARTICLES

ATOMIC AND MOLECULAR COLLISIONS • ENERGY TRANSFER • INFRARED SPECTROSCOPY • MICROWAVE SPECTROSCOPY, MOLECULAR • PLANETARY ATMOSPHERES • RAMAN SPECTROSCOPY

BIBLIOGRAPHY

Frommhold, L. (1994). "Collision-induced Absorption of Gases," Cambridge Univ. Press, Cambridge.

Birnbaum, G., Borysow, A., and Orton, G. S. (1996). "Collision-induced absorption of H_2–H_2 and H_2–He in the rotational and fundamental bands for planetary applications," *Icarus* **123,** 4.

Borysow, A., Jørgensen, U. G., and Zheng, C. (1997). *Astron. Astrophys.* **324,** 185.

Herman, R. M., ed. (1999). "Spectral Line Shapes," Vol. 10, Am. Inst. Physics, Woodbury, NY. See the articles in that volume by Borysow, A., Le Duff, Y., Moraldi, M., Tipping, R. H., and Zoppi, M., and others.

Tabisz, G. C., and Neuman, M. N., ed. (1995). "Collision- and Interaction-Induced Spectroscopy," NATO ASI Series C, Vol. 452, Kluwer Academic, Dordrecht, Boston, London.

Bibliographies of collision-induced absorption and light scattering exist which are being updated every few years; these are quoted in the literature mentioned above.



Mössbauer Spectroscopy

Gilbert R. Hoy
Old Dominion University

I. Basic Theory
II. Mössbauer Spectra
III. Applications
IV. Exotic Applications
V. Future Possibilities

GLOSSARY

Cross section (σ) In a scattering (or absorption process, the cross section is defined as the ratio of number of particles scattered (or absorbed) per unit time (I) to the number of particles per unit area per unit time (J_0) in the incident beam. Thus $\sigma = I/J_0$ has the units of area.

Debye temperature In solid-state physics, a frequently used approximation to describe the acoustic mode lattice vibrations is the Debye model. The Debye temperature is a parameter of the model which gives the energy of the most energetic lattice vibration in temperature units.

Electric field gradient tensor (EFG) Second rank tensor describing the second spatial derivative of the electric potential (V). The usual notation is to set $V_{xy} = \partial^2 V/(\partial x \partial y)$ and similarly for the other components. In the principal axis system of the EFG, $V_{xx} + V_{yy} + V_{zz} = 0$. In addition, these components are conventionally chosen such that $|V_{zz}| \geq |V_{yy}| \geq |V_{xx}|$.

Gamma rays Electromagnetic waves emitted by certain nuclei in an excited state as they decay to a lower energy state.

Internal conversion Excited state nuclei which decay to the ground state by emission of the gamma ray can usually also decay by the emission of an inner-shell atomic electron. This mechanism called internal conversion is one in which the excitation energy is delivered directly to the electron without emitting a gamma ray.

Nuclear magnetic moment (μ) Nuclear parameter which describes the strength and circulation of the infinitesimal electrical currents in a nucleus and is equal to $g\mu_N I$. The g factor (g) describes the magnetic dipole character of the nuclear state, I is the nuclear spin, and μ_N is a constant ($\approx 5.0508 \times 10^{-27}$ joules/tesla) called the nuclear magneton.

Nuclear parity Parity is a parameter that arises because when we observe most physical phenomena, we cannot tell the difference between looking directly at the phenomenon or viewing it in an arbitrarily oriented mirror. (Not true for the so-called "weak" interactions.) This invariance with respect to inversion of the coordinate system leads to the concept of conservation of parity. Thus, quantum mechanical descriptions of the state of a nucleus contain its parity which can be + (even) or − (odd).

Nuclear quadrupole moment (Q) Nuclear parameter describing the axially symmetric deviation from spherical shape of the positive charge distribution in the nucleus. If the charge distribution is elongated at the

poles (prolate), sometimes called a cigar shape, $Q > 0$. If the charge distribution is compressed at the poles (oblate), sometimes called a pancake shape, $Q < 0$.

Nuclear spin (I) Intrinsic angular momentum possessed by a nucleus in a particular ground or excited state. Quantum mechanics restricts the possible values to be either integer or odd half-integer multiples of $\hbar$ (Planck's constant divided by 2π).

Phonon In the quantum mechanical formulation of lattice vibrations in a solid, the "quantized" lattice vibrations are called phonons.

Photon In the quantum mechanical formulation of the emission of electromagnetic radiation, the "quantized" electromagnetic waves are called photons.

Recoil-free fraction (f) The probability that a gamma ray will be emitted or absorbed without nuclear recoil.

THE MÖSSBAUER EFFECT (ME) is the recoil-free emission and/or recoil-free absorption of gamma radiation. In 1957 Rudolph L. Mössbauer, while still a graduate student, discovered the effect which bears his name, for which he received the Nobel Prize for physics in 1961. The surprising aspect of this effect is that it seems to require the overthrow of one of the most fundamental laws of physics, namely the conservation of linear momentum. In this process a nucleus, usually in its first excited state, decays by the emission of a gamma ray that carries all of the energy initially in the system. In order for this to be possible, the nucleus cannot recoil, because if it did, some of the initial energy would be carried away by the recoiling nucleus. By considering the problem in more detail, one realizes that the nucleus is in an atom which, in most applications, is bound in a solid. As the nucleus tries to recoil during the emission of the gamma ray, the atomic electrons, which interact with the nucleus, also interact with the solid in which they are embedded. If the solid is sufficiently rigid, a large mass of the solid participates in this recoil event. This mass can be so large that the actual recoil energy becomes vanishingly small. Imagine, by comparison, firing a loaded rifle, first by holding it against your shoulder and second by burying the butt of the rifle in a large mass of hardened concrete.

The remarkable utility of this effect is due to the fact that the spectral energy distribution of the gamma rays emitted in this recoil-free process has a width (Γ) essentially governed by the lower limit of Heisenberg's time-energy uncertainty principle: $\Gamma\tau \geq \hbar$, where Γ is the energy spread of the photon distribution and τ is the lifetime of the excited state. Thus, we have an experimental technique which is capable of resolving energies with unprecedented precision. The probability of observing such a recoil-free event will be inversely related to the energy in the excited state nucleus (E_0) and the solid's temperature (T).

I. BASIC THEORY

From the quantum mechanical point of view, the recoil-free emission and absorption of gamma radiation is described by saying that phonons are not created or destroyed during the process. In this way, the state of the lattice is unchanged. A question may immediately arise concerning the extremely narrow energy spread (Γ) claimed for the photon distribution. There will be lattice vibrations in the solid at any temperature. (Recall that even at a temperature of absolute zero, the solid undergoes so-called "zero-point" motion.) Therefore, the emitting nucleus is moving. One might expect "Doppler" broadening as found in gaseous systems when viewing atomic electromagnetic transitions. This effect is not observed in our case because the period of the lattice vibrations is much smaller than the lifetime (τ) of the excited nuclear level. Thus, the velocity of the source of the radiation averages to zero. Even though this first-order Doppler shift vanishes, the same cannot be said for the second-order Doppler shift. This effect is briefly discussed in the section on the isomer shift.

The emission and absorption of gamma radiation is described in quantum mechanics by saying that a photon is emitted or absorbed. The ground and excited states of a nucleus are described by their respective energies, spins, parities, quadrupole moments, and magnetic moments. In general, the nucleus will interact with its environment, the influence of which will be due primarily to the atomic electrons because of their proximity, and secondarily to the rest of the solid and any externally applied fields. This interaction is described in quantum mechanics by introducing a Hamiltonian operator ($\mathbf{H}$). The resulting energy levels of the nucleus can be determined by solving the time-independent Schrödinger equation (TISE), $\mathbf{H}\psi = E\psi$, where ψ represents the nuclear states and E corresponds to the allowed energy levels. Normally one must solve this equation for both the excited and ground nuclear states. Finally, the transition probabilities for gamma-ray emission or absorption between the excited and ground nuclear states must be calculated.

In order to be more specific about these matters, consider the most commonly used Mössbauer nuclide, namely ^{57}Fe. Figure 1 shows an energy level diagram for this case. The ground state spin (I_g) is $\frac{1}{2}\hbar$ and the state has a magnetic moment, but the quadrupole moment is zero. The excited state spin (I_e) is $\frac{3}{2}\hbar$, and hence the state has nonzero quadrupole and magnetic moments. If the environment is such that it does not interact with the nucleus, the excited

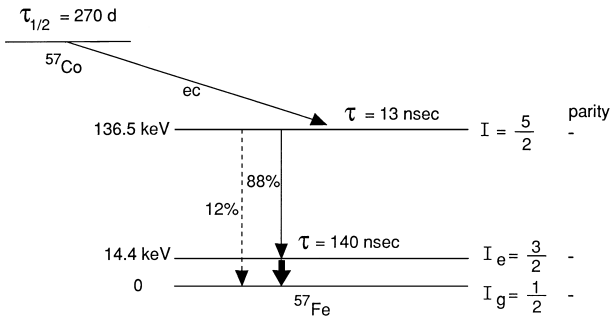


FIGURE 1 The decay scheme of ^{57}Co showing the parameters for the Mössbauer-effect nuclide, ^{57}Fe. ^{57}Co decays by electron capture to the second excited state of ^{57}Fe which decays most often by a gamma–gamma cascade. The last transition from the 14.4-keV level to the ground state, indicated by the heavy arrow, is the most famous Mössbauer transition.

state has a fourfold energy degeneracy corresponding to the four possible orientations of the nuclear spin in space. Similarly, under these conditions, the nuclear ground state will have a twofold energy degeneracy. These degeneracies can be partially or completely removed by interaction with the environment which is described by the appropriate Hamiltonian (**H**). Using the matrix representation of quantum mechanics, the TISE becomes an eigenvalue–eigenvector problem. Consider the $I_e = \frac{3}{2}\hbar$ first excited state of the ^{57}Fe nucleus. Any general state ψ can be represented as

$$\psi = a_1\left|\tfrac{3}{2}\quad\tfrac{3}{2}\right\rangle + a_2\left|\tfrac{3}{2}\quad\tfrac{1}{2}\right\rangle + a_3\left|\tfrac{3}{2}\quad-\tfrac{1}{2}\right\rangle + a_4\left|\tfrac{3}{2}\quad-\tfrac{3}{2}\right\rangle, \tag{1}$$

where the notation $|II_z\rangle$ is used for the individual components, and the a_1's are the amplitudes for the components. (Note: The wave functions, or eigenstates, are usually normalized so that $\Sigma_i|a_i|^2 = 1$.) I_z is the projection of the nuclear spin along the selected z axis, i.e., the axis of quantization. The matrix representation of the Hamiltonian **H** is found by forming the array of matrix elements, $\mathbf{H}_{ij} = \langle\frac{3}{2}\ m_i|\mathbf{H}|\ \frac{3}{2}m_j\rangle$, where i and j take on the values 1, 2, 3, and 4 corresponding to the four allowed values of I_z. Using this approach the TISE takes the form

$$\begin{bmatrix} H_{11} & H_{12} & H_{13} & H_{14} \\ H_{21} & H_{22} & H_{23} & H_{24} \\ H_{31} & H_{32} & H_{33} & H_{34} \\ H_{41} & H_{42} & H_{43} & H_{44} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_3 \end{bmatrix} = E \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix}, \tag{2}$$

where the energy eigenvalues E are found by diagonalizing the Hamiltonian matrix, and the eigenvectors, or eigenstates, are found using standard matrix techniques. [Viewed from a slightly different point of view, Eq. (2) is simply a set of four coupled equations.] A similar procedure is used for the ^{57}Fe nuclear ground state where the Hamiltonian matrix is now only 2×2.

The transition probabilities, which give a relative measure of the amount of gamma radiation emitted or absorbed per unit solid angle in a particular transition, are proportional to the square of the absolute value of the relevant vector spherical harmonic (representing the photon) and the square of a so-called Clebsch–Gordon coefficient. (In a complicated case, a sum of such quantities occurs.) The Clebsch–Gordon coefficient arises because of the need to conserve angular momentum. The initial and final nuclear states have angular momenta: their spins (I_g, I_e). The photon itself has an angular momentum (L) whose projection along the quantization axis is often written as M. The symbol for the Clebsch–Gordon coefficient is often written $\langle I_g m_g L M; I_e m_e\rangle$. One may think of this by saying that the initial angular momentum (I_e) must equal the vector sum of the final angular momenta ($I_g + L$). For the case under consideration, the transition is predominantly classified as magnetic dipole (M1), and the Clebsch–Gordon coefficients take the form $\langle\frac{1}{2}m_g 1 M; \frac{3}{2}m_e\rangle$. Table I gives values for various quantities corresponding to the ^{57}Fe case.

In general, any measurable quantity obtained from utilizing the ME will consist of the product of two parts; namely, a nuclear part and an environmental (or field) part. A nuclear physicist will be most interested in determining

TABLE I Quantities for the ^{57}Fe Case

Excited nuclear level ($\frac{3}{2}m_e$)	Ground nuclear level ($\frac{1}{2}m_g$)	Clebsch-Gordon coef. squared	Vector spherical harmonic absolute value squared
$\frac{3}{2}$	$\frac{1}{2}$	1	$\frac{3}{2}(1+\cos^2\theta)$
$\frac{3}{2}$	$-\frac{1}{2}$	0	—
$\frac{1}{2}$	$\frac{1}{2}$	$\frac{2}{3}$	$\frac{3}{2}(\sin^2\theta)$
$\frac{1}{2}$	$-\frac{1}{2}$	$\frac{1}{3}$	$\frac{3}{4}(1+\cos^2\theta)$
$-\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{3}{4}(1+\cos^2\theta)$
$-\frac{1}{2}$	$-\frac{1}{2}$	$\frac{2}{3}$	$\frac{3}{2}(\sin^2\theta)$
$-\frac{3}{2}$	$\frac{1}{2}$	0	—
$-\frac{3}{2}$	$-\frac{1}{2}$	1	$\frac{3}{4}(1+\cos^2\theta)$

nuclear parameters and hence must know the value of the field part. On the other hand, a solid-state scientist must know the nuclear parameters in order to extract solid-state information. For either case, it has not always been necessarily true that the needed quantities were known. This has produced interesting multidisciplinary scientific interactions.

A. Recoil-Free Fraction (f)

The size of the recoil-free fraction determines the magnitude of the ME itself. If one attempts to calculate this factor from first principles, it is necessary to have a model describing the possible lattice vibrations, i.e., the phonon distribution for the particular solid. This is clearly a formidable task. The model used most often to describe the acoustic vibrations of a solid is called the Debye model, which uses a simplified approximation for the phonon spectrum. Using that assumption, the recoil-free fraction can be calculated and is given by the following result:

$$f_{\mathrm{D}} = \exp\left\{-\frac{6E_{\mathrm{R}}}{k\Theta_{\mathrm{D}}}\left[\frac{1}{4}+\left(\frac{T}{\Theta_{\mathrm{D}}}\right)^2\int_0^{\Theta_{\mathrm{D}}/T}\frac{x\,\mathrm{d}x}{(e^x-1)}\right]\right\}, \tag{3}$$

where E_{R}, the recoil energy of a free nucleus of mass m, equals ${E_0}^2/(2mc^2)$ and Θ_{D} is the Debye temperature. From this equation one can see that f_{D} will decrease as E_0 and the sample temperature T increase. The recoil-free fraction can be expressed in simpler forms when $T=0$, $T \gg \Theta_{\mathrm{D}}$, or $T \ll \Theta_{\mathrm{D}}$.

There is another, more general, form for the recoil-free fraction,

$$f = e^{-\langle x^2\rangle/(\lambda/2\pi)^2} \tag{4}$$

where $\langle x^2\rangle$ is the mean-square displacement of the nucleus, and λ is the wavelength of the emitted or absorbed gamma ray. One can see from this formulation that an additional complication arises in anisotropic solids in which the mean-square displacements in different directions are unequal. This effect is even observable in polycrystalline samples and is called the Goldanski–Karyagin effect.

B. Nuclear Resonance Cross Section

The probability of nuclear resonance absorption and scattering are conventionally expressed in terms of cross sections and hence have the dimensions of an area. These can be written

$$\sigma_{\mathrm{abs}}(E) = \sigma_0\frac{\Gamma\Gamma_\gamma}{4(E-E_0)^2+\Gamma^2} \tag{5}$$

and

$$\sigma_{\mathrm{scatt}}(E) = \sigma_0\frac{\Gamma_\gamma^2}{4(E-E_0)^2+\Gamma^2}, \tag{6}$$

where Γ is the total width of the absorption transition ($\Gamma=\Gamma_\gamma+\Gamma_{\mathrm{ec}}$), Γ_γ is the partial width associated with the gamma-ray process, Γ_{ec} is the partial width due to internal conversion, and σ_0 is the maximum resonance cross section given by

$$\sigma_0 = [\lambda^2/(2\pi)]\frac{2I_{\mathrm{e}}+1}{2I_{\mathrm{g}}+1}, \tag{7}$$

where λ is the wavelength of the gamma ray. After a little manipulation σ_{abs} can be written as

$$\sigma_{\mathrm{abs}} = \sigma_0'\frac{(\Gamma/2)^2}{(E-E_0)^2+(\Gamma/2)^2}, \tag{8}$$

where now, σ_0' is

$$\sigma_0' = 2\pi(\hbar c/E_0)\left[\frac{2I_{\mathrm{e}}+1}{2I_{\mathrm{g}}+1}\right]\frac{1}{1+\alpha} \tag{9}$$

and α, the internal conversion coefficient, is defined such that $(1+\alpha)=(\Gamma/\Gamma_\gamma)$. Notice the characteristic energy dependence in Eq. (8) due to the resonance process. This energy dependence has a so-called Breit–Wigner or Lorentzian line shape.

C. Hyperfine Interactions

Strictly speaking, the hyperfine interaction is defined as the interaction between the nucleus and the atomic electrons. In the field of Mössbauer spectroscopy, the hyperfine interaction is usually split up into several components for ease of discussion. Effects produced by applying an external magnetic field are classified under the title of the nuclear Zeeman effect.

1. Isomer Shift (δ)

The isomer (or chemical) shift is one of the quantities that can be obtained using Mössbauer spectroscopy. As mentioned above, such quantities have two contributions. In this case, the relevant nuclear part depends on the nuclear radii of the ground and excited states involved in the transition. The environmental part is the electron density that the atomic electrons, usually s electrons, have at the nucleus. (Relativistically, the p orbitals also make a contribution. Relativistic effects become more important as the atomic number Z increases. The situation is further complicated by the shielding effects of p and d orbitals on s orbitals.) Since it is not uncommon for valence electrons to be in s orbitals, one can see why this quantity has also been known as the chemical shift. The isomer shift can be understood by considering how the nuclear energy levels

are shifted by the electrostatic interaction between these electrons and a positively charged nucleus of finite radius compared with a positively charged point nucleus. The isomer shift is seen in experiments in which recoil-free gamma radiation from a source is absorbed recoillessly by resonant nuclei in an absorber. The type and condition of the solid containing the source nuclei need not be identical to that containing the absorber nuclei. The value of the isomer shift is given by

$$\delta = \frac{2\pi}{3} e^2 Z \left\{ |\phi_s(0)|^2 - |\phi_a(0)|^2 \right\} \left\{ \left\langle R_e^2 \right\rangle - \left\langle R_g^2 \right\rangle \right\}, \quad (10)$$

where Ze is the nuclear charge, $|\phi(0)|^2$ is the electron density at the nucleus, subscripts s and a refer to the source and absorber, respectively, and $\langle R^2 \rangle$ is the mean square radius of the nuclear positive charge distribution.

In order to obtain the total isomer shift (δ_T), the effects of the second-order Doppler shift (SOD) must be included: $\delta_T = \delta + \text{SOD}$. The SOD is due to the fact that the source and absorber nuclei are in motion. When a moving source emits resonant radiation, there is a frequency shift relative to a stationary observer as a result of the relativistic time dilation. The size of the SOD depends on the nature and temperature of the solids containing the source and absorber nuclei.

2. Quadrupole Interaction

A nucleus having a spin and a nonzero quadrupole moment will experience an orientational torque if located at a position where there is an electric field gradient. This interaction will remove some energy degeneracy in the excited and ground state nuclear levels. In the simplest cases, the splitting depends on the absolute value of the projection of the spin on the quantization axis. If the nuclear spin is less than or equal to $\frac{1}{2}\hbar$, the nuclear quadrupole moment will be zero. If the environment surrounding the nucleus possesses cubic symmetry, the electric field gradient at the nucleus is zero.

The Hamiltonian governing this interaction, written in the principal axis system of the EFG, is given by

$$\mathbf{H}_Q = \frac{e^2 q Q}{4I(2I-1)} \left[3I_z^2 - I^2 + \tfrac{1}{2}\eta \left(I_+^2 + I_-^2 \right) \right], \quad (11)$$

where $eq = V_{zz}$ (q is often called the electric field gradient), η (the asymmetry parameter) $= (V_{xx} - V_{yy})/V_{zz}$, and $I_\pm = I_x \pm I_y$. (Notice that, by convention, $0 \le \eta \le 1$; since Laplace's equation applies, only two components q and η are needed to specify the EFG.)

In attempts to calculate the EFG parameters, a separation is usually made between the contributions: one being due to the atomic electrons not in spherically symmetric distributions and the other due to the remaining lattice. The problem of calculating the EFG components at the nucleus is further complicated by the existence of shielding or antishielding effects produced by the subsequent distortion of atomic electrons in closed shells (so-called Sternheimer factors).

For the example of ^{57}Fe, this interaction results in the nuclear excited state splitting into two doubly degenerate energy levels. The ground state for this case is unsplit. Solving the TISE using the quadrupole Hamiltonian $\mathbf{H}_Q$ gives the energy separation of the excited state levels,

$$\Delta E_Q = \frac{e^2 q Q}{2} \left[1 + \frac{\eta^2}{3} \right]^{\frac{1}{2}} \quad (12)$$

Because of the form of $\mathbf{H}_Q$, the excited nuclear energy states must, in general, be linear combinations of $|\frac{3}{2}\ \frac{3}{2}\rangle$ and $|\frac{3}{2}\ -\frac{1}{2}\rangle$ or $|\frac{3}{2}\ \frac{1}{2}|$ and $|\frac{3}{2}\ -\frac{3}{2}\rangle$. If the asymmetry parameter η is zero, the nuclear excited states can be labeled $|\frac{3}{2}\ \pm\frac{3}{2}\rangle$ and $|\frac{3}{2}\ \pm\frac{1}{2}\rangle$, while the nuclear ground states are labeled $|\frac{1}{2}\ \pm\frac{1}{2}\rangle$.

3. Magnetic Interaction

The nucleus, through its magnetic moment, also interacts with the electronic spin and orbital distributions. In general, this hyperfine interaction can be written as

$$\mathbf{H}_m = \mathbf{I} \cdot \mathbf{A} \cdot \mathbf{J}, \quad (13)$$

where now the eigenstates of the system must include both nuclear and atomic states. $\mathbf{A}$ is the hyperfine interaction tensor, and $\mathbf{J}$ is the total angular momentum of the atom or ion containing the Mössbauer nucleus. For a solid containing iron transition elements, it is not unusual for the orbital angular momentum to be partially or completely "quenched," i.e., the average value of each of the three components of the total orbital angular momentum equals zero. This is not true in the rare earth elements where $\mathbf{J}$ is a good quantum number.

Results from electron-spin resonance measurements using iron transition elements as well as other elements have lead to the concept of an effective spin (S) for the ion. The value of $\mathbf{S}$ is determined by setting the degeneracy of the ionic ground state, in the absence of any magnetic field effects, equal to $2S+1$. (On occasion $\mathbf{S}$ is equal to the actual spin.) Using this approach the Hamiltonian can be rewritten as

$$\mathbf{H}_m = \mathbf{I} \cdot \mathbf{A}' \cdot \mathbf{S}, \quad (14)$$

where a frequency used scheme can be represented by states labeled $|II_zSS_z\rangle$.

Under conditions in which the electronic spin ($\mathbf{S}$) or total angular momentum ($\mathbf{J}$) is fluctuating rapidly compared to the motion of the nuclear spin ($\mathbf{I}$), the nucleus

only senses the average value of the electronic angular momentum. This case can be treated by introducing an effective magnetic field ($\mathbf{B}_e$). The interaction takes the same form as that of a nucleus in an externally applied magnetic field, i.e., the nuclear Zeeman effect for which

$$\mathbf{H}_m = -\mu \cdot \mathbf{B}_e. \tag{15}$$

However, for those cases in which an external magnetic field is applied, the field at the nucleus is usually not simply equal to the value of the external field. There are three cases that arise, depending on whether the sample material is diamagnetic, paramagnetic, or magnetically ordered in some fashion (i.e., ferromagnetic, antiferromagnetic, ferrimagnetic, etc). Usually, the effective magnetic field $\mathbf{B}_e$ is written as

$$\mathbf{B}_e = \mathbf{B}_{ext} + \mathbf{B}_{fc} + \mathbf{B}_{sd} + \mathbf{B}_{orb}, \tag{16}$$

where $\mathbf{B}_{ext}$ is the externally applied magnetic field neglecting demagnetizing and Lorentz fields, $\mathbf{B}_{fc}$ is the Fermi contact term described below, $\mathbf{B}_{sd}$ is the electronic spin-dipole contribution from the surrounding ions (in cubic symmetry this term is zero), and $\mathbf{B}_{orb}$ is the magnetic field due to any resultant orbital angular momentum in the ion.

The Fermi contact field ($\mathbf{B}_{fc}$) arises from a resultant spin density (either up or down) at the nucleus. This effect is thus due to *s* electrons. Even in the case of spin-paired *s* electrons, their spatial distribution in the atom is affected by electrons in incomplete shells. For example, in an iron transition ion, the *s* electrons having their spins parallel to the resultant 3*d* spin will be "attracted" to the 3*d* spin, while the *s* electrons antiparallel to the 3*d* spin will be "repelled." The net effect of this "core-polarization," since the *s* electrons are at the nucleus, is that the magnetic field experienced by the nucleus can be orders of magnitude larger and in the opposite direction from that of the applied magnetic field in certain cases.

D. Time-Dependent Hyperfine Fields

The ME can be used to study the time dependence of the nuclear environment. Consider the case when the hyperfine interaction can be represented in terms of an effective magnetic field $\mathbf{B}_e$, as in Eq. (15). The orientation of $\mathbf{B}_e$ is influenced by spin–spin and spin–lattice interactions in the solid. Suppose, for example, that the phonons interact with the atomic electrons such that the effective magnetic field $\mathbf{B}_e$ at the nucleus flips between "up" and "down" orientations. The relevant times for this discussion are the relaxation (or flipping) time τ_R and the Larmor precession time τ_L of the nuclear magnetic moment in the magnetic field $\mathbf{B}_e$. If $\mathbf{B}_e$ stays in one orientation long enough for the magnetic moment to precess in the field, i.e., if $\tau_R \gg \tau_L$, then the nucleus "knows" that it is in a magnetic field. If $\mathbf{B}_e$ is flipping rapidly, i.e., if $\tau_R \ll \tau_L$, and spends equal times in the up and down positions, the nucleus will "think" the magnetic field is zero. If $\tau_R \ll \tau_L$, but the magnetic field spends more time in either the up or down position, the nucleus will experience the average magnetic field that is smaller than the magnitude of $\mathbf{B}_e$.

There are many different types of time-dependent hyperfine interactions that can be studied. These include time-dependent electric field gradients, isomer shifts, and magnetic interactions. Furthermore, time-dependent recoil-free fractions and even atomic diffusion processes can be studied using the ME.

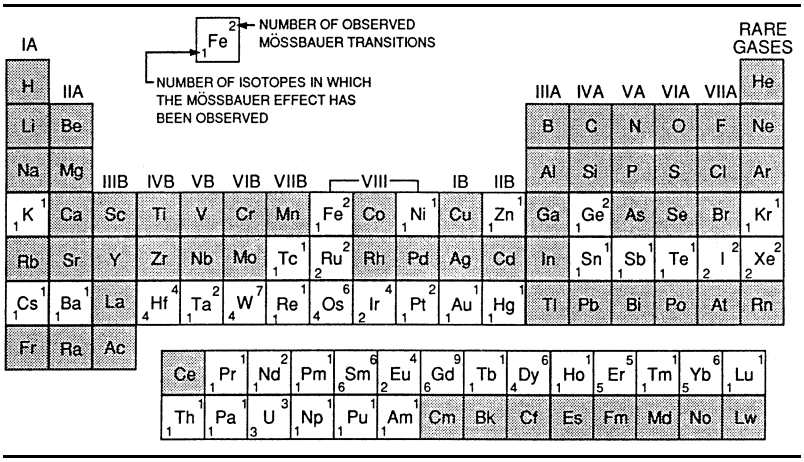
E. Conversion-Electron Mössbauer Spectroscopy (CEMS)

Excited state nuclei which decay to the ground state by emission of the gamma ray usually can also decay by internal conversion, i.e., the emission of an inner-shell atomic electron. Thus, the presence of a nuclear excited state can be signaled by the subsequent emission of a gamma ray or conversion electron. If one is using the detection of the conversion electron to indicate that recoil-free nuclear resonance absorption (i.e., ME) has occurred, a different region of the sample is being studied than is the case using ordinary Mössbauer spectroscopic methods. This is because the range of the conversion electron in the solid is in general much shorter than the range of the gamma ray in that solid, and the electron must leave the solid in order to be detected. Thus, the use of conversion-electron Mössbauer spectroscopy allows the study of surface layers. If, in addition to simply detecting the presence of the conversion electron, one also measures its energy, it is possible to investigate surface layers at different depths inside the solid. This technique is called depth-selective conversion-electron Mössbauer spectroscopy (DCEMS). Since the internal-conversion process can involve conversion electrons from the K, L, and M shells as well as the subsequent Auger electrons, the electrons coming from the surface of the material are widely distributed in energy. Of course the electrons having the greatest initial energy can penetrate through larger distances in the solid and still be detected. Therefore, the detection of conversion electrons together with their energy permits the probing of the environmental effects at different layers inside the solid.

F. Mössbauer Nuclei

Because of the special conditions inherent in the above discussion, the ME will not occur for all excited state nuclei. The nuclei for which the ME has already been observed

TABLE II Mössbauer Periodic Table



are given in Table II. Even though the examples given below use ^{57}Fe, one should realize there are a number of other possibilities (see Table II).

II. MÖSSBAUER SPECTRA

The ME has been used to study a large number of scientific and technological problems. The particular experimental technique that is used in any given case depends upon the scientific problem under consideration. Each experimental procedure is designed to obtain a Mössbauer spectrum from which physically meaningful results, relevant to the problem being studied, may be extracted. The most frequently used experimental arrangement is the transmission geometry (see Fig. 2).

In transmission geometry a radioactive source containing the relevant excited-state nuclei is usually placed on the end of a rod which is connected to an electromechanical drive mechanism. The absorber containing the same nuclide in the ground state is interposed between the source and the gamma-ray detector. The source is normally driven in the so-called constant-acceleration mode in which the source spends equal times in equal velocity intervals. The gamma-ray detector is set to signal the arrival of gamma rays of the correct energy selected by a single-channel analyzer. The logic output of the single-channel analyzer is fed into the input of a multichannel analyzer operating in the multichannel-scaling mode. The multichannel-scaling timing must be synchronized with the movement of the source such that when the source is moving at a particular velocity, the multichannel analyzer will open a particular channel. This will give rise to a spectrum in the multichannel analyzer in which the number of counts transmitted through the absorber will be stored as a function of the velocity of the source. Since the source is moving, the energy of the emitted gamma ray will be Doppler shifted according to the relation, $E = (v/c)E_0$. When the Doppler shifted energy of the source gamma radiation matches a possible absorptive transition in the absorber, a drop in the counting rate will appear at the detector. The resulting spectrum stored in the multichannel analyzer provides information about the nuclear spins, parities, quadrupole, and magnetic moments, as well as the fields produced at the nucleus by the atomic electrons, the solid-state environment, and any externally applied fields.

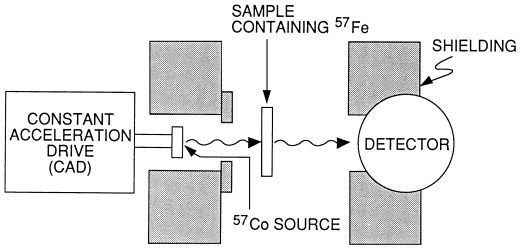


FIGURE 2 A schematic representation of an experimental arrangement using the ^{57}Fe nuclide in the transmission geometry. The source is usually connected to an electromechanical drive (CAD) which moves with constant acceleration in one direction and returns with equal and opposite acceleration. The gamma-ray detector, for these low-energy photons, is usually a proportional counter.

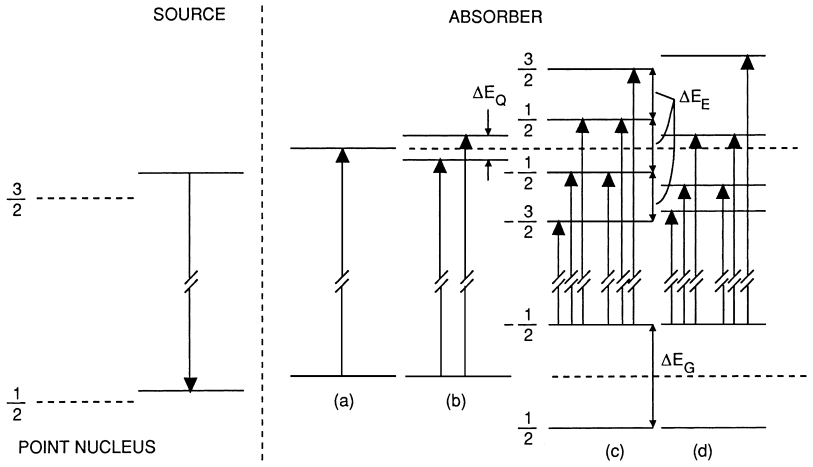


FIGURE 3 Possible energy level schemes for the nuclide, ^{57}Fe. On the left side, the assumed energy levels for nuclei in the source are shown. The source is assumed to have a single-line emission line shape. On the right side, the energy levels in the absorber are shown corresponding to four different types of hyperfine interactions: (a) isomer shift relative to the source, (b) quadrupole interaction (ΔE_Q is the quadrupole splitting), (c) magnetic interaction (ΔE_G is the ground-state splitting and ΔE_E is the excited-state splitting), and (d) combined quadrupole and magnetic interactions. (Note: the energy spacing between the ground and excited nuclear levels is not to scale.)

Figure 3 shows several possible energy level schemes for the excited and ground nuclear states using the example nuclide, ^{57}Fe. On the far left of the figure, energy levels are schematically represented assuming that the nuclei are just points. The true levels are shifted because the ground and excited nuclear states have finite nuclear radii. Assume that in the source the ^{57}Fe nuclei are not subjected to an interaction that removes any energy degeneracy. This source, therefore, emits monochromatic gamma radiation: the only spread is due to the Heisenberg time–energy uncertainty relation. On the right side of the figure, the possible nuclear energy levels in the absorber are depicted corresponding to four different cases. In Fig. 3a we assume there is only a shift due to the finite nuclear radii and an electron charge density at the nuclei which is different from that in the source. In Fig. 3b–d we imagine an absorber in which the ^{57}Fe nuclei experience an electric field gradient, a magnetic interaction, and a combined electric field gradient and magnetic interaction, respectively.

In Fig. 4 possible Mössbauer transmission spectra, corresponding to the four situations presented in Fig. 3, are shown. Since the nearly monochromatic source radiation is Doppler shifted to sweep over all the absorptive energy transitions in the absorber, there will be a drop in the counting rate at the detector when the energies match. Figure 4a shows the result assuming a "single-line" source and a single-line absorber having different electron densities at the ^{57}Fe nuclei. The consequences of a quadrupole interaction in the absorber are shown in Fig. 4b. Assume that our absorber is a polycrystalline sample and the microcrystals are randomly oriented (thereby neglecting so-called "texture" effects). Assume further that the EFG parameters are such that $eq > 0$ and $\eta =$ zero. Under these conditions, the higher energy level in the excited state shown in Fig. 3b can be labeled $|\frac{3}{2}\ \pm\frac{3}{2}\rangle$ and the lower $|\frac{3}{2}\ \pm\frac{1}{2}\rangle$. It is clear that there will be two dips in the Mössbauer transition spectrum for this case, but what will be their sizes? In order to predict their sizes, we need to calculate their transition probabilities. We can use the information in Table I for this purpose. Since the absorber is assumed to be polycrystalline, the absolute value squared of the vector spherical harmonics (column 4 in Table I) must be averaged over all solid angles. The normalization of these vector spherical harmonics has been chosen so that the average for each is simply equal to 1. Thus, it is only the value listed in column 3 that determines the relative transition probability for each transition. Considering all the allowed transitions, it is easy to see that the sizes of the two dips will be equal for this case. Similarly, for the case corresponding to a simple magnetic interaction, as in Figs. 3c and 4c, the intensity ratios for the six transitions are 3:2:1:1:2:3. If we assume a simple case of combined quadrupole (Fig. 3b) and magnetic (Fig. 3c) interactions in which the major axis of the EFG tensor and the magnetic field are collinear, the result is shown in Fig. 4d. For this case, the shifts of the nuclear energy levels are particularly easy to deduce, and the intensity ratios of the six dips are the same as for Fig. 4c.

As an example of time-dependent effects, consider the case corresponding to Figs. 3c and 4c, but now suppose that the effective magnetic field produced by the atomic system at the nucleus is flipping between up and down

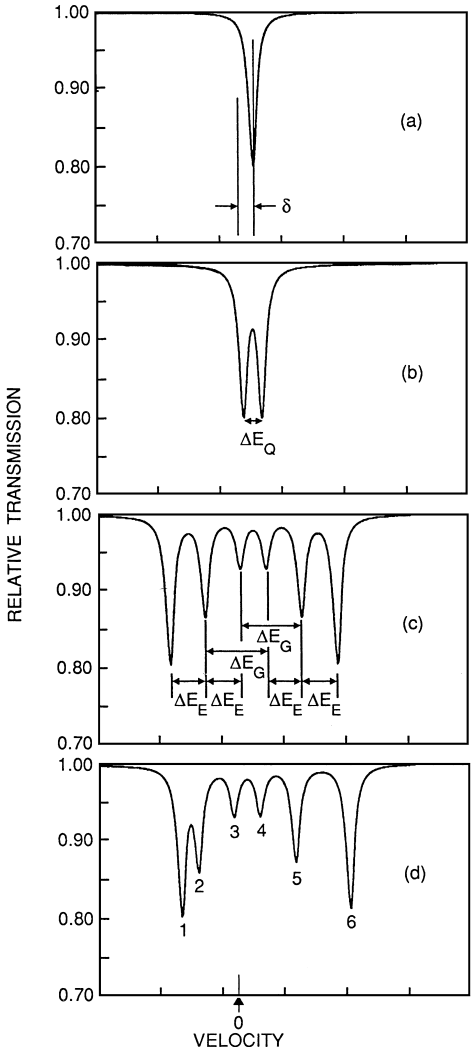


FIGURE 4 Hypothetical Mössbauer transmission spectra are shown corresponding to the cases depicted in Fig. 3: (a) the isomer shift, (b) the quadrupole interaction, (c) the magnetic interaction, and (d) the combined quadrupole and magnetic interactions. In (a) the isomer shift is indicated by δ. The quadrupole splitting is labeled in (b) by ΔE_Q. For the case of a simple magnetic hyperfine structure, the ground-state splitting ΔE_G and excited-state splittings ΔE_E are shown in (c). Part (d) depicts the results for a combined quadrupole and magnetic hyperfine structure in the simple case when the major axis of the EFG, in its principal axis system, is collinear with the direction of the magnetic field.

orientations as a consequence of interactions with other spins and/or lattice vibrations. Figure 5 shows results corresponding to different flipping rates, Ω. (Here Ω is measured in energy units having values compared with the ground-state splitting in the static magnetic field (ΔE_G).) As the flipping rate increases from zero, the spectral dips broaden and the spectrum collapses. Finally, when the flipping rate is sufficiently rapid, the spectrum becomes a single dip corresponding to no magnetic field.

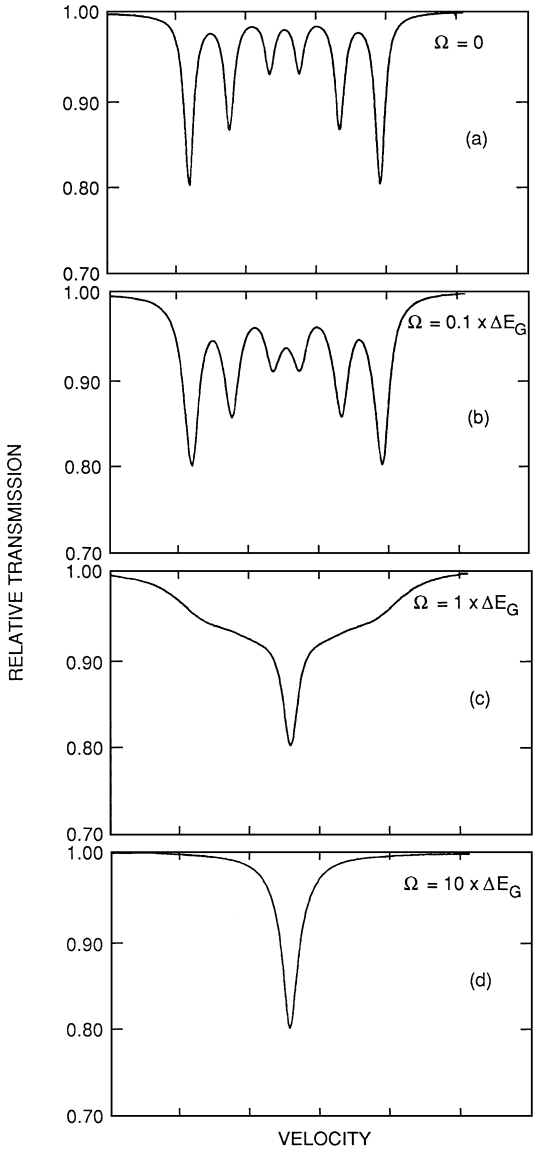


FIGURE 5 Hypothetical Mössbauer transmission spectra showing the effect of an effective magnetic field which is assumed to be flipping between "up" and "down" orientations at the locations of the ^{57}Fe nuclei in the absorber. The magnetic field is assumed to spend equal times in the up and down orientations. The source is assumed to emit a "single line." The flipping rate Ω is given in energy units scaled by the ground-state splitting (ΔE_G) in the fixed magnetic field. In (a) the effective magnetic field is not flipping. In (b), (c), and (d), the flipping rate is increasing and hence the spectrum broadens and collapses as expected. Notice that even at the highest flipping rate shown, the spectrum is not yet completely "motionally narrowed."

Figure 6 shows a schematic representation of the scattering geometry. An interesting aspect of experiments done in this geometry is that Rayleigh scattering, the prompt elastic scattering of photons from the atom as a whole, is coherent with the nuclear resonant scattering.

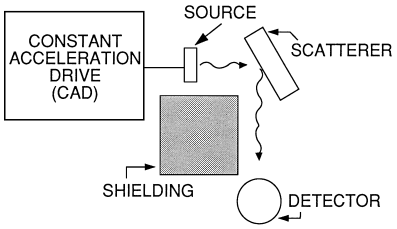


FIGURE 6 A simple schematic diagram of the scattering geometry.

Furthermore, when the scatterer resonantly absorbs the recoil-free radiation from the source, it is often impossible to localize the energy excitation on one particular nucleus. This fact, coupled with utilization of a single-crystal scatterer, results in the formation of a collective coherent nuclear excitation. The subsequent decay of this specially prepared state is different from the decay of a single excited nucleus (see Section IV).

Figure 7 shows a schematic representation of the conversion-electron Mössbauer spectroscopy (CEMS) technique. In this case, the Mössbauer spectra show peaks instead of dips when there is nuclear resonant absorption in the sample. This is because the detection of the conversion electron signals the event. For experiments in which the sample is at room temperature, the electron detector is often a gas flow proportional counter containing He plus 10% CH_4 as a quench gas. This detector cannot work well at liquid He temperatures and in such applications is replaced by channeltrons or microchannel plates. If one is interested in doing depth-selective conversion-electron Mössbauer spectroscopy (DCEMS), the energy of the selected conversion electron must be measured. This can be done by incorporating a magnetic or electrostatic analyzer in the detection system.

Figure 8 shows the selective excitation double Mössbauer spectroscopy (SEDM) technique. In such experiments the source is driven at constant velocity in order to match a particular absorptive transition in the scatterer. The subsequent re-emitted radiation is studied by interposing a resonant absorber, driven in the constant acceleration mode, in front of the gamma-ray detector. Using this technique one can, for example, study Rayleigh scattering in the presence of nuclear-resonant scattering. Furthermore, time-dependent effects in the scatterer can be investigated using this procedure.

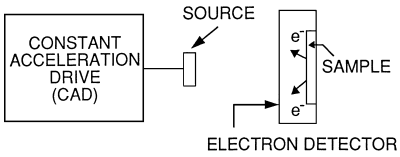


FIGURE 7 A simple schematic diagram of the conversion-electron Mössbauer spectroscopy (CEMS) technique. The detector in this case is often a He + 10% CH_4 flow proportional counter. Notice that the gamma rays need to go through the detector undetected. The nuclear resonance absorption in the sample must occur significantly near the sample surface, which is inside the detector, so that the conversion electrons can get out and be detected.

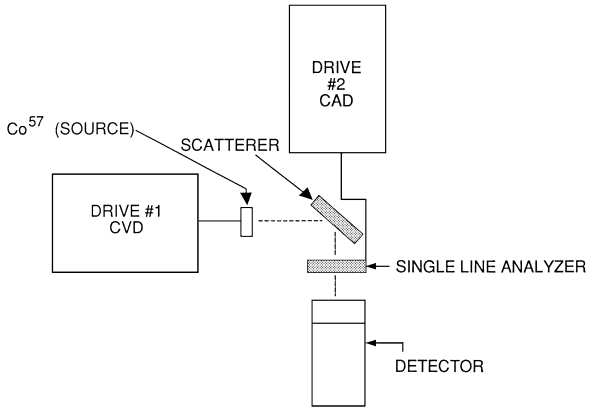


FIGURE 8 A simple schematic diagram of the selective excitation double Mössbauer (SEDM) spectroscopy technique. In this case the source is driven at constant velocity (CVD) chosen to correspond to a particular transition of interest in the scatterer. Of course, provision must be made for the drive to return to its original position and the process repeated. Electronic control is utilized so that no data are recorded during the return phase. The scattered radiation is "analyzed" to determine its spectral content by a single-line analyzer driven in the constant acceleration mode (CAD).

There are a number of other experimental techniques used in conjunction with the ME, e.g., radio frequency and acoustic modulation, synchroton radiation, ion implantation, and coulomb excitation. However, this section will be concluded by considering only one more; namely, coincidence-Mössbauer spectroscopy, also known as time-differential Mössbauer spectroscopy (TDMS). Figure 9 shows a schematic representation of such a setup using the nuclide ^{57}Fe. In this case one can think of doing a normal transmission experiment, but now individual spectra are sorted according to the time the 14.4-keV excited nuclear state lives. By adjusting the electronics, one can choose to record a Mössbauer spectrum for which the excited-state nuclear (14.4 keV) level decays at times between t and $t + \Delta t$. The resulting emission line shape is thus modified from the usual Lorentzian one of natural linewidth. This fact, together with the passage of the radiation through a resonant absorber, can result in unusual Mössbauer spectra (see Section IV).

III. APPLICATIONS

Mössbauer spectroscopic studies have been applied to an enormous number of problems in science and technology.

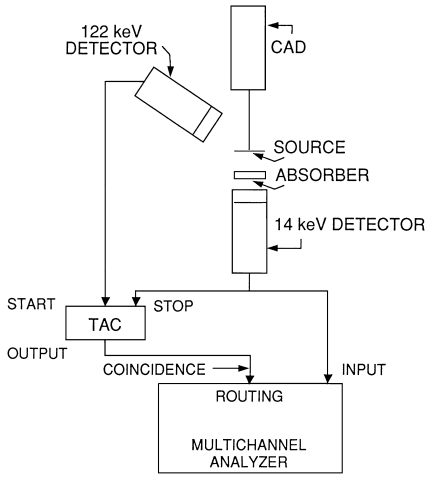


FIGURE 9 A simple schematic diagram of the time-differential Mössbauer spectroscopy (TDMS) technique. In this case, transmission Mössbauer spectra can be obtained for which the source ^{57}Fe nuclei live for a time between t and $t+\Delta t$. This is accomplished by detecting the 122 keV gamma ray (see Fig. 1), which "time marks" the formation of the 14.4-keV excited state. The 122-keV signal is sent to the "start" input of a time-to-amplitude converter (TAC). The signal from the subsequent 14.4-keV detector, indicating the transmission through the absorber, is not only sent to the input of the multichannel analyzer, but also to the "stop" input of the TAC. The output of the TAC produces a pulse whose height is proportional to the time delay between the start and stop signals. The multichannel analyzer memory can be divided so when an input pulse arrives it is "routed" into a certain portion of the memory corresponding to emission of the 14.4-keV photon between a time t and $t+\Delta t$ after its formation. The information about where to route the 14.4-keV input signal is contained in the height of the "coincident" TAC output pulse. Clearly, many electronic details have been omitted from this schematic figure.

This is partly due to the ubiquitous nature of iron in our universe. Problems ranging in scope from fundamental questions in physics, chemistry, and biology to applied studies of steels and coal, and even analyses of pigments and clays used in works of art have been examined using these methods. Any attempt to summarize the full range of applications of the ME cannot be done in these pages. Therefore, the examples selected below are necessarily somewhat subjective.

Results presented in Fig. 10 show an ME study of long-range magnetic order associated with the magnetic phase transition in the antiferromagnet FeF_2 using the transmission geometry. These researchers were able to measure not only the sublattice magnetization but also the so-called critical exponent describing how the sublattice magnetization depends on temperature near the transition temperature.

Another interesting ME study in the transmission geometry is shown in Fig. 11. Here, Mössbauer spectra are shown using deuterated ferric, aluminum ammonium alum, $NH_4(Fe,AI)(SO_4)_2{\cdot}12D_2O$. In this case, the ferric ion ground state has a total angular momentum $S=(5/2)\hbar$, and the crystalline field splits the ground ionic level into three Kramers doublets. In this compound the spin–spin interaction is weak, and when the sample temperature is lowered, the spin–lattice interaction is also weak. Therefore, the ^{57}Fe nuclei see a fixed effective magnetic field due to each Kramers doublet. (Note that the hyperfine structure associated with the $\pm\frac{1}{2}$ Kramers doublet is a bit complicated.) The relative contribution of each subspectrum to the total Mössbauer spectrum is determined by the sample temperature and crystal-field splittings of the Kramers doublets.

An example of an unusual and surprising time-dependent hyperfine structure occurs in ferric hemin (see Fig. 12). Observe the result associated with the ferric ion in this case when the sample is at liquid helium temperature. Notice that there is no effective magnetic field structure, in contrast with the example just discussed. Thus, we conclude that the ionic spin system is relaxing rapidly. However, as the temperature is raised, the spectrum broadens and smears out characteristic of a decreasing relaxation rate. This surprising behavior is explained by noting that the Kramers doublet corresponding to the $S_z=\pm\frac{1}{2}\hbar$ is the lowest ionic level. At low temperature this level has the highest population. At higher temperature when the upper Kramers doublets are populated, spin relaxation becomes increasingly difficult because the levels corresponding to $S_z=\pm\frac{3}{2}\hbar$ and $\pm\frac{5}{2}\hbar$ require a large change in angular momentum when the ionic spin flips up and down.

IV. EXOTIC APPLICATIONS

The wealth of research using the ME in "exotic" applications is also impressive, and here again, it is impossible to review the subject completely.

The development of methods which use synchrotron radiation in conjunction with ME has produced some very interesting results. Synchrotron radiation is produced when electrons, traveling at relativistic speeds, are accelerated in circular orbits. This characteristic radiation is highly polarized and quite directional. When such radiation impinges on a single-crystal scatterer containing resonant nuclei, it is possible to establish a coherent collective nuclear excited state in the scatterer, as mentioned previously. The subsequently emitted radiation has unusual properties. Figure 13 shows the results of such an experiment. Since the synchrotron is pulsed, the coherent collective nuclear excited states are formed at a particular time. One can then measure the number of re-emitted photons as a function of time after the formation of the

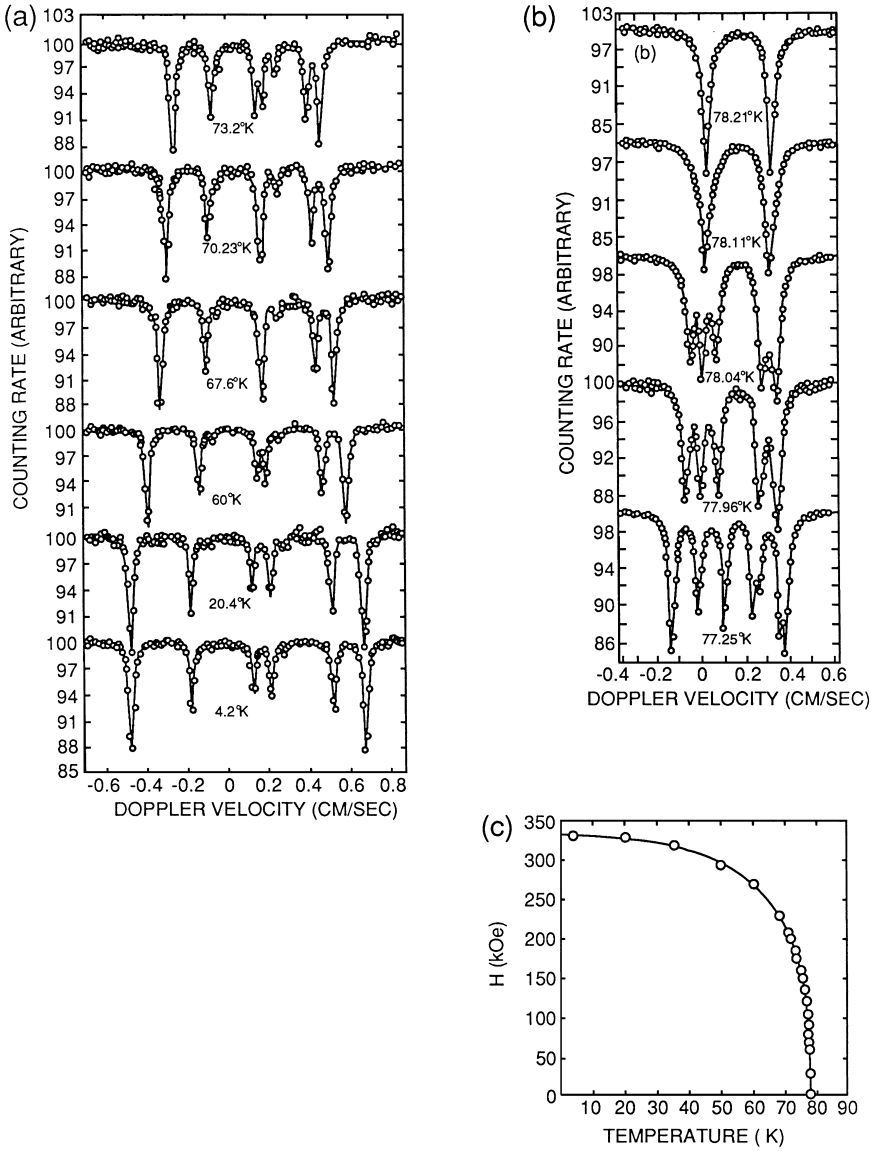


FIGURE 10 An ME study, in the transmission geometry, of the antiferromagnetic phase transition in FeF_2: (a) hyperfine structure at temperatures well below the Néel temperature, (b) hyperfine structure at temperatures near the Néel temperature, and (c) the deduced values of the sublattice hyperfine effective magnetic field as a function of temperature. Notice the precision of the results. [From Wertheim, G. K., and Buchanan, D. N. E. (1967). *Phys. Rev.* **161,** 478.]

states. The resulting measured lifetime of such a state is not the ordinary lifetime of the excited state nucleus but is, in fact, shorter. In addition, Fig. 13 shows a "quantum beat" structure resulting from the interference of radiation coming from the four hyperfine transitions present in the single-crystal $FeBO_3$ scatterer when excited by the synchrotron radiation.

Figure 14 shows an application of TDMS to the study of electronic charge states in ^{57}Fe following electron capture of the parent nuclide ^{57}Co. This electron capture not only results in the formation of the ^{57}Fe 136 keV nuclear excited state (see Fig. 1) but also a highly excited ionic charge state due to Auger processes. This highly charged ionic state will evolve into the appropriate charge state or states determined by the chemistry and imperfections of the sample. If this evolution occurs on a time scale comparable with the nuclear first excited state lifetime, TDMS techniques can detect these changes.

Perhaps the most exotic application of the ME is in the study of the gravitational red-shift. According to Einstein's Principle of Equivalence, it is not possible to distinguish between observations made in the presence

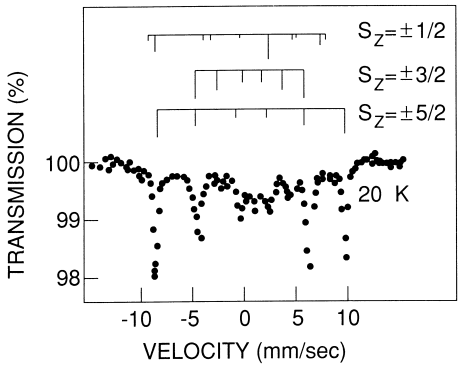


FIGURE 11 An ME study, in the transmission geometry, of deuterated ferric, aluminum ammonium alum, $NH_4(Fe,Al)(SO_4)_2{\cdot}12D_2O$. At low temperature (20 K) the Mössbauer spectrum is a superposition of the three hyperfine patterns arising from each of the three ionic Kramers doublet levels. In this case the effective magnetic fields are practically static. [From Campbell, L. E., and DeBenedetti, S. (1968). *Phys. Rev.* **167,** 556.]

of a uniform gravitational field and those made without a gravitational field but, instead, viewed in a reference frame moving with acceleration **g**, the local gravitational constant, in the opposite direction. Heroic ME experiments investigating this effect have been performed. Imagine, for example, a situation on earth in which the source is located a distance h below an identical nuclear resonant absorber, where "below" is determined by the direction of **g**. The radiation is emitted by the source, but now, to account for the gravitational field, the absorber may be thought of as accelerating upward, away from the source. Thus, the absorber "sees" radiation shifted to lower frequencies (red shifted); i.e., the emitted photons have a

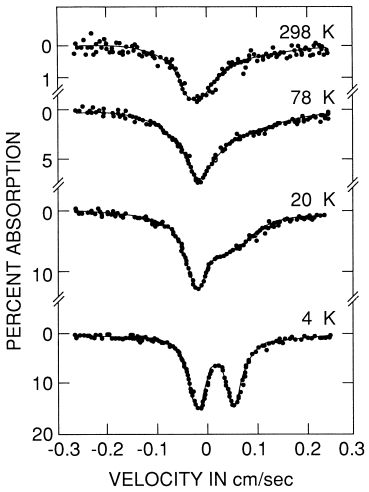


FIGURE 12 An ME study of ferric hemin as a function of temperature in the transmission geometry. These results show a slowing down of the relaxation rate as the temperature increases. [From Blume, M. (1967). *Phys. Rev. Lett.* **18,** 305.]

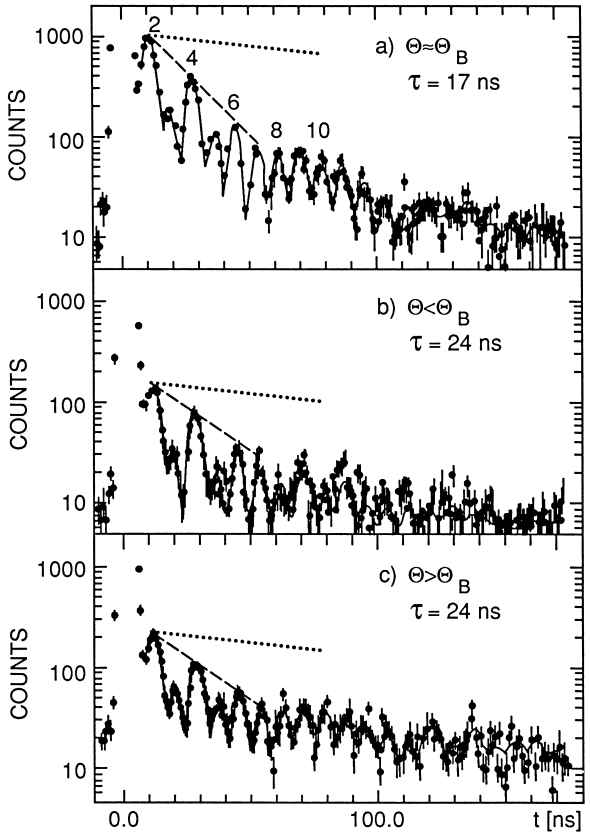


FIGURE 13 Synchrotron radiation is scattered from a single crystal of $^{57}FeBO_3$. Since the synchrotron radiation is pulsed, it is possible to measure the lifetime of excited state formed in the scatterer. The re-emitted radiation is detected: (a) exactly on the Bragg angle, (b) slightly below the Bragg angle, and (c) slightly above the Bragg angle. The dotted lines show the normal decay rate for the 14.4-keV ^{57}Fe excited state. The dashed lines show the enhanced decay rate (lifetime τ). In addition, they observed the effect of "quantum beats." [From van Bürck, U., Mössbauer, R. L., Gerdau, E., Rüffer, R., Hollatz, R., Smirnov, G. V., and Hannon, J. P. (1987). *Phys. Rev. Lett.* **59,** 355.]

lower energy $\Delta E = E_0(gh/c^2)$ when they reach the absorber. Of course, if the source were above the absorber, there would be a corresponding "blue shift." Such a shift, confirming Einstein's prediction, has been observed.

V. FUTURE POSSIBILITIES

The limiting resolution of ME experiments is, of course, related to the natural linewidth of the excited nuclear level undergoing recoil-free emission and absorption. Since nuclear excited states having longer lifetimes (τ) have narrower linewidths (Γ), it is interesting to speculate about the ultimate limitations in using the ME when τ increases. The ME has already been clearly observed in ^{67}Zn ($\tau = 13.2$ μsec) and ^{181}Ta ($\tau = 9.8$ μsec). Experiments

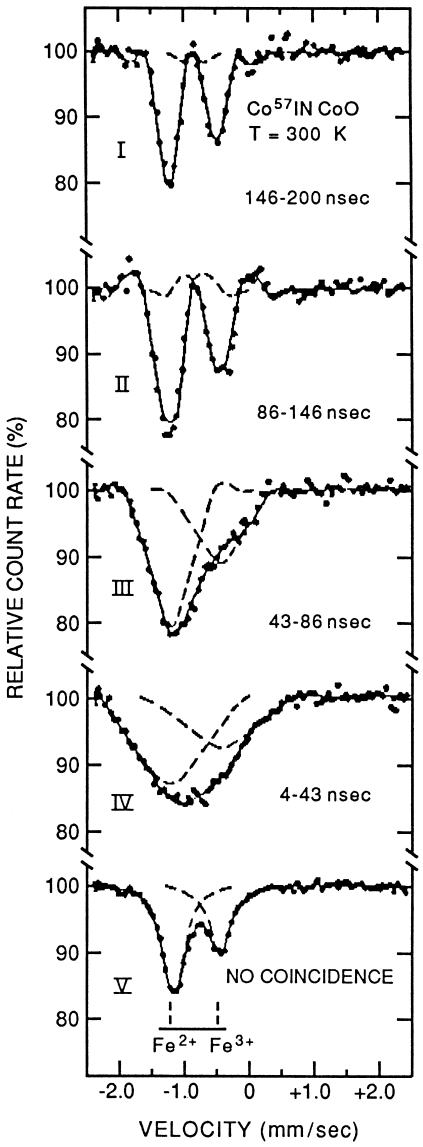


FIGURE 14 An ME source study of CoO using the time-differential Mössbauer spectroscopy (TDMS) technique at 300 K. The unsplit absorber is potassium ferrocyanide. Mössbauer spectra are collected in which the first excited nuclear level of ^{57}Fe in the source (CoO) decays during four different time intervals. The relevant time interval is labeled for each spectrum. Notice that at early times the spectrum is broad while for later times the spectra narrow. This is a consequence of the time–energy uncertainty principle. The purpose of these experiments is discussed in the text. [From Triftshäuser, W., and Craig, P. P. (1967). *Phys. Rev.* **162,** 274.]

have been, and continue to be, done looking for the ME in ^{107}Ag and ^{109}Ag ($\tau \approx 40$ sec). The inhomogeneous fields that must occur to some degree even in carefully prepared samples make such observations difficult, if not impossible. Furthermore, gravitational effects may play a role in the ME search for such ultranarrow transitions.

The use of synchrotron radiation in ME experiments will no doubt continue to be an important field of research. In fact, the rather new and relatively unexplored field of gamma-ray optics should provide opportunities for discovering new and exciting physics. This is particularly true due to technological advances in the production of nearly perfect single crystals. These advances have already resulted in the observation of effects in X-ray diffraction attributed to multiple scattering processes. These multiple scattering processes occurring in single crystals are explained theoretically by the so-called dynamical theory of X-ray diffraction. An experimental manifestation of this theory is the anomalous transmission of X radiation through an extremely thick single crystal (the Borrmann effect). A relatively small group of researchers are even looking into the possible development of gamma-ray lasers.

SEE ALSO THE FOLLOWING ARTICLES

ELECTRON SPIN RESONANCE • GAMMA-RAY SPECTROSCOPY • NUCLEAR MAGNETIC RESONANCE • QUANTUM MECHANICS • RADIATION PHYSICS • RELATIVITY, GENERAL • X-RAY, SYNCHROTRON RADIATION, AND NEUTRON DIFFRACTION

BIBLIOGRAPHY

Balko, B., Cohen, L., and Sparrow, D. A., eds. (1988). "Gamma-Ray Lasers," Pergamon Press, Oxford.

Cohen, R. L., ed. (1976, 1980). "Applications of Mössbauer Spectroscopy," Vols. I and II, Academic Press, New York.

Cohen, S. G., and Pasternak, M., eds. (1973). "Perspectives in Mössbauer Spectroscopy," Plenum Press, New York.

Coyne, L. M., and McKeever, S. W. S., eds. (1999). "Spectroscopic Characterization of Minerals and Their Surfaces," ACS Symposium Series, American Chemical Society, Washington, DC.

Dickson, D. P. E., and Berry, F. J., eds. (1986). "Mössbauer spectroscopy," Cambridge University Press, Cambridge.

Gonser, U., ed. (1975, 1981). "Mössbauer Spectroscopy" and "Mössbauer Spectroscopy II," Springer-Verlag, New York.

Kagan, Yu. M., and Lyubutin, I. S., eds. (1985). "Applications of the Mössbauer Effect," Vols. 1–5. Gordon and Breach Science Pubs., New York.

Long, G. J., and Stevens, J. G., eds. (1986). "Industrial Applications of the Mössbauer Effect," Plenum Press, New York.

Long, G. J., eds. (1984, 1987, 1989). "Mössbauer Spectroscopy Applied to Inorganic Chemistry," Vols. 1, 2, and 3, Plenum Press, New York.

McCammon, C., Hutchison, M., and Harris, J. (1997). "Ferric iron content of mineral inclusions in diamonds from Sao Luiz: A view into the lower mantle," *Science* **278,** 434–436.

Shenoy, G. K., and Wagner, F. E., eds. (1978). "Mössbauer Isomer Shifts," North-Holland, New York.

Solomon, E. I., and Hodgson, K. O., eds. (1998). "Spectroscopic Methods in Bioinorganic Chemistry," ACS Symposium Series, American Chemical Society, Washington, DC.

Stevens, J. G. (compiled by) (1981). *In* "Handbook of Spectroscopy" (J. W. Robinson, ed.), Vol. III, p. 403, CRC Press, Boca Raton, Florida.

Stevens, J. G., and Shenoy, G. K., eds. (1981). "Mössbauer Spectroscopy and Its Chemical Applications," American Chemical Society, Washington, DC.

Stevens, J. G., and Stevens, V. E., eds. (1968–1990). "Mössbauer Effect Data Index," Plenum, New York.

Thosar, B. V., Iyengar, P. K., Srivastava, J. K., and Bhargava, S. C., eds. (1983). "Advances in Mössbauer Spectroscopy," Elsevier, New York.



Multiphoton Spectroscopy

Y. Fujimura
Tohoku University

S. H. Lin
Academia Sinica

I. Introduction
II. Theory
III. Experimental Methods
IV. Characteristics and Spectral Properties
V. Applications
VI. Summary

GLOSSARY

Above-threshold ionization Ionization of atoms and molecules by absorption of more photons than the minimum needed for ionization. The spectrum of photoelectrons that are emitted by above-threshold ionization consists of a series of bands separated by the photon energy applied.

Doppler-free multiphoton spectroscopy Multiphoton spectroscopy aims at eliminating inhomogeneous Doppler broadening.

High-order harmonic generation Conversion process of multiphotons absorbed by atoms and molecules into emission of a series of photons with odd harmonic frequencies of the driving laser field. High-order harmonic generation is being investigated as a promising approach to the development of a compact, coherent, soft X-ray radiation source.

Ion-dip spectroscopy A high-resolution multiphoton spectroscopy based on competition between ionization and stimulated emission (or stimulated absorption).

Multiphoton absorption Simultaneous absorption of multiple numbers of photons by materials under the irradiation of laser light with high intensities.

Multiphoton ionization mass spectroscopy Multiphoton ionization method combined with mass detection. This spectroscopy allows us to identify molecules by the optical spectrum related to the resonant intermediate state as well as by their mass.

Polarization dependence Dependence of the cross sections of multiphoton transitions on photon polarizations of laser used.

Resonance enhancement A drastic increase in the ability of the multiphoton process observed when laser is tuned and the energy of the laser approaches that of a real intermediate state. See Multiphoton absorption.

Rydberg states States of a valence electron orbiting about a positively charged core consisting of the nucleous and inner electrons in atoms or molecules. Jumps of an electron between Rydberg states are called Rydberg transitions. The transition frequency, ω, is given by the following:

$$\omega_n = I_\mathrm{p} - R/(n-\delta)^2,$$

where I_p is the ionization potential, R the Rydberg constant, n the principal quantum number of the Rydberg

electron, δ the quantum defect, and $R/(n-\delta)^2$ the term values. The classic orbit radius of the Rydberg electron increases as n^2, while the orbital velocity decreases as n^{-1}.

MULTIPHOTON SPECTROSCOPY consists of the simultaneous interaction of atoms or molecules with two or much more number of photons. That is, it is spectroscopy with the use of multiphoton transitions, or, more generally, the spectroscopic research field of the interaction between matter and two or more photons.

I. INTRODUCTION

In Fig. 1, several typical multiphoton processes are shown. The result of the material–multiphoton interaction is usually detected through direct absorption, fluorescence, ionization current, or a photoelectron detection system. The excited-state structures of these materials in gases, liquids, or solids, such as electronic, vibrational, or rotational states or fine structure, which are not found in ordinary single-photon spectroscopy because of their difference in selection rules and low transition intensity, can be seen in a wide frequency range from lower electronic excited states to ionized continua.

Multiphoton spectroscopy requires an intense light source. The first experimental observation of the simplest multiphoton transition, two-photon absorption of an Eu^{2+}-doped CaF_2 crystal in the optical region by Kaiser and Garrett (1961), was made possible only after a high-power monochromatic ruby laser was developed as the intense incident light source, although the possibility of simultaneous two-photon absorption or stimulated emission was pointed out in 1931 by Goeppert-Mayer.

The main reasons for wide interest in the multiphoton spectroscopy are due to the advent of dye lasers for tunability and of multiphoton ionization technique for detecting information from the excited state created by the multiphoton excitation.

The tunability of dye lasers is particularly important for multiphoton excitation because one can obtain an excitation source by using only a single-frequency laser beam rather than the two or more lasers of different frequencies. The multiphoton ionization technique consists of collecting free electrons produced by the multiphoton ionization process after irradiation by a tunable laser pulse, amplifying ion currents, and recording the signal as a function of the laser frequency. In general, even ion currents of only a few charges per second can be detectable. Therefore, by using this method, one can detect and characterize extremely small amounts of atoms or molecules, even in a rarefied gas. The sensitivity exceeds that of fluorescence and other detections. The multiphoton ionization technique is also important in practical applications such as isotope separation, laser-induced fusion, and the dry etching process.

A Ti:Sapphire laser makes it possible to generate pulses whose intensity is stronger than 10^{13} W/cm^2 in an ultrashort time. Application of such intense laser pulses to atoms and molecules is expected to open up new fields of study on multiphoton processes, such as high-order harmonic generation, above-threshold ionization, and above-threshold dissociation. These cannot be explained by using a simple perturbative treatment. Nonperturbative

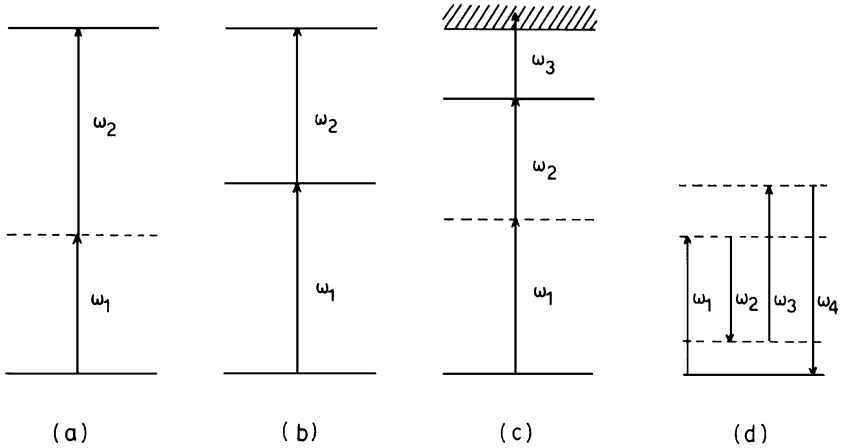


FIGURE 1 Several multiphoton processes seen in atoms and molecules: (a) a nonresonant two-photon absorption process; (b) a resonant two-photon absorption process; (c) a two-photon resonant three photon ionization; and (d) a four-wave mixing process. Solid lines and broken lines represent real and virtual states, respectively; ω_i denotes photon frequencies.

treatments should be used to explain the mechanisms of such multiphoton processes. A direct method for solving the time-dependent Schrödinger equation as well as other theoretical methods is being developed.

Multiphoton transitions related to the multiphoton spectroscopy have several characteristic features: laser intensity dependence, resonance enhancement, polarization dependence, and so on. For example, the transition probability of the nonresonant two-photon absorption process shown in Fig. 1a with $\omega_1 = \omega_2$, $W^{(2)}_{i\to f}$, can be written as:

$$W^{(2)}_{i\to f} = I^2\sigma^{(2)}/(\hbar\omega_R)^2 \tag{1}$$

where $\sigma^{(2)}$, I, and ω_R denote the cross section for the two-photon absorption, the laser intensity, and the laser frequency, respectively. Equation (1) indicates that the two-photon transition probability is proportional to the square of the laser intensity applied. This is called the formal intensity law. If no saturation of photon absorption takes place during the multiphoton processes, the order or the multiphoton transition can be experimentally determined from measuring the slope of log–log plots of the transition probability as a function of the laser intensity as shown in Fig. 2:

$$\ln W^{(2)}_{i\to f} = 2\ln I + C \tag{2}$$

for the nonresonant two-photon process.

Multiphoton spectroscopy usually utilizes resonance enhancement; that is, a dramatic increase in the multiphoton transition ability can be seen when the exciting laser is tuned and its frequency approaches a real intermediate electronic state called a resonant state. In this case, the level width of the resonant state plays a significant role in determining the transition ability. It is well known that photons can be regarded as particles of mass 0 and spin 1. Polarization dependence of multiphoton processes is associated with the spin angular momentum. Polarization dependence and symmetry selection rules of multiphoton transitions are of great importance in characterizing the multiphoton transition process and in determining the symmetry of the states relevant to the transitions. For example, for a two-photon transition of a molecule with a center of symmetry, the initial and final states have the same parity, which is in contrast to the parity selection rule of one-photon spectroscopy governed by the opposite parity. Therefore, one- and two-photon spectra are complementary for measuring vibronic states of the molecule. This relationship between one- and two-photon spectroscopic techniques is similar to that between the infrared (IR) absorption governed by the opposite parity and Raman spectroscopy by the same parity. These characteristic features mentioned briefly are described in detail in Section IV after the theoretical treatment and experimental techniques for the multiphoton spectroscopy are introduced. Not only structures but also dynamic behaviors in electronically excited states of atoms and molecules are widely studied using the multiphoton spectroscopic methods. Typical examples of such applications of multiphoton spectroscopy are presented in Section V.

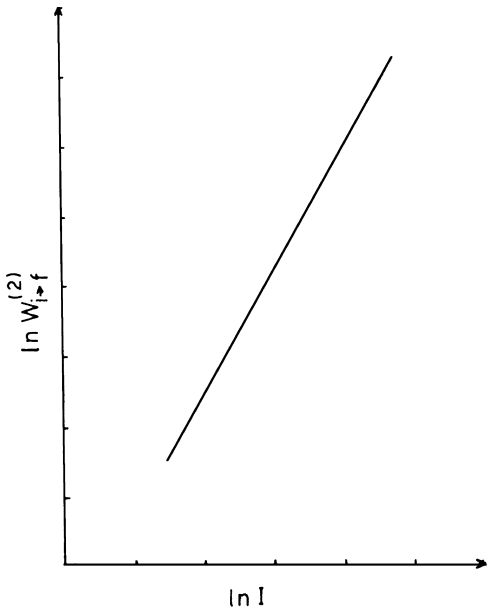


FIGURE 2 The formal intensity law for a nonresonant two-photon process. Here I and $W^{(2)}_{i\to f}$ denote the laser intensity and two-photon transition probability from the i to the f state, respectively.

II. THEORY

In order to analyze multiphoton spectra, various theoretical treatments based on the ordinary time-dependent perturbation method, the Green's function method, the density matrix method, the susceptibility method, and so on have been applied to deriving the expressions for the multiphoton transition probability. A direct method has also been used for solving the time-dependent Schrödinger equation for an atom or molecule interacting with intense laser pulses. In this section these theoretical treatments are considered, focusing on their advantages and on the restrictions on their application to the multiphoton process.

A. General Considerations

The theory of multiphoton processes can be developed based on the semiclassical or quantum-mechanical formalisms. Here the quantum-mechanical theory in which the photon field is described in the second quantization is presented. In some cases the semiclassical formalism in

which the photon field is treated classically and matter–radiation field interaction is characterized by a time-dependent behavior is convenient.

The Hamiltonian of the total system consisting of atoms or molecules plus radiation field is written as follows:

$$H = H_S + H_R + V = H_0 + V, \tag{3}$$

where H_S is the Hamiltonian of the particles, H_R the Hamiltonian of the free radiation field, and V the interaction between them. The radiation Hamiltonian can be written in terms of photon creation and annihilation operators as follows:

$$H_R = \sum_{\mathbf{k}} \hbar\omega_k \left(b_{\mathbf{k}}^\dagger b_{\mathbf{k}} + \frac{1}{2} \right), \tag{4}$$

where $\mathbf{k}$ specifies both the wave vector and polarization, $\omega_k = c|\mathbf{k}|$ the angular frequency of the $\mathbf{k}$th photon mode, and $b_{\mathbf{k}}^\dagger$ and $b_{\mathbf{k}}$ are the creation and annihilation operators of the photon, respectively. The unperturbed Hamiltonian of the total system satisfies the following:

$$H_0|I\rangle = \varepsilon_I |I\rangle, \tag{5}$$

where

$$\varepsilon_I = \varepsilon_i + \sum_{\mathbf{k}} \left(n_{\mathbf{k}} + \frac{1}{2} \right) \hbar\omega_k, \tag{6}$$

in which ε_i is an eigenvalue of H_s and $n_{\mathbf{k}}$ is the number of photons with frequency ω_k. The eigenstate of the unperturbed Hamiltonian H_0, $|I\rangle$, can be written as the product of that of the system $|i\rangle$ and that of the radiation field $|n(k_1)n(k_2)\cdots\rangle$.

The interaction Hamiltonian V originates from the coupling of the vector potential $\mathbf{A}(\mathbf{r})$ with moving charged particles with mass m_j and electric charge e_j and is given by the following:

$$V = \sum_j \left(-\frac{e_j}{m_j c} \right) \mathbf{P}_j \times \mathbf{A}(\mathbf{r}_j) + \sum_j \left(\frac{e_j^2}{2m_j c^2} \right) \mathbf{A}^2(\mathbf{r}_j), \tag{7}$$

where the vector potential is written as follows:

$$\mathbf{A}(\mathbf{r}) = \sum_l \left(\frac{\hbar}{2\varepsilon_0 L^3 \omega_l} \right)^{1/2} \mathbf{e}_l \left\{ b_l \exp(i\mathbf{k}_l \cdot \mathbf{r}) + b_l^\dagger \exp(-i\mathbf{k}_l \cdot \mathbf{r}) \right\}, \tag{8}$$

where L^3 is a cubic box of the photon field and $\mathbf{e}$ the polarization unit vector of the photon.

The interaction Hamiltonian, Eq. (7), can be written in terms of the multipole expansion as follows:

$$V = V_d + V_q + \cdots, \tag{9}$$

where V_d and V_q denote the electric dipole and the electric quadrupole interactions, respectively, and they are given by the following:

$$V_d = -e\mathbf{r} \cdot \mathbf{E} \tag{10}$$

and

$$V_q = -\frac{e}{2} \sum_{ij} \mathbf{Q}_{ij} \nabla_j E_i, \tag{11}$$

in which $\mathbf{Q}$, the quadrupole dyadic, is defined by the following:

$$\mathbf{Q}_{ij} = \mathbf{r}_i \mathbf{r}_j - \tfrac{1}{3}\mathbf{r}^2 \delta_{ij} \quad (\mathbf{r}_i = x, y, z). \tag{12}$$

Since the wavelength of photons of optical frequencies is much larger than atomic or molecular dimension, the spatial dependence of the photon field can be neglected in calculating the transition probabilities (dipole approximation). The first term of the interaction Hamiltonian V_d makes a significant contribution to ordinary multiphoton transitions, although effects of the quadrupole interaction term have been observed in some cases. In the dipole approximation the interaction Hamiltonian can be expressed as the following:

$$V = -ie \sum_l \left(\hbar\omega_l / 2\varepsilon_0 L^3 \right)^{1/2} \mathbf{r} \cdot \mathbf{e}_l \left(b_l - b_l^\dagger \right) \tag{13}$$

B. Ordinary Time-Dependent Perturbation Theory

In the ordinary time-dependent perturbation theory, the first-order transition probability per unit time from I to F states, $W_{I\to F}^{(1)}$, is given by the following:

$$W_{I\to F}^{(1)} = \frac{2\pi}{\hbar} |V_{FI}|^2 \delta(E_F - E_I) \tag{14}$$

and the mth-order transition probability per unit time, $W_{I\to F}^{(m)}$, is given by the following:

$$W_{I\to F}^{(m)} = \frac{2\pi}{\hbar} \left| \sum_{M_1} \sum_{M_2} \cdots \sum_{M_m} \frac{V_{FM_m} \cdots V_{M_2 M_1} V_{M_1 I}}{\hbar\omega_{M_m I} \cdots \hbar\omega_{M_1 I}} \right|^2 \times \delta(E_F - E_I), \tag{15}$$

where $M_1 \cdots M_m$ specify the intermediate states for the multiquantum transition. The Dirac δ function $\delta(E_F - E_I)$ expresses the energy conservation between initial and final states of the transition. It should be noted that Eqs. (14) and (15) have been derived in the long time limit $t \to \infty$. These expressions are usually called Fermi's golden rule for the transition probability.

The expression for the multiphoton transition probability can be derived by using Fermi's golden rule. The expression for the m-photon (quantum) transition probability comes from that of the mth-order transition probability in Fermi's golden rule. For example, the expression for a two-photon absorption probability from initial state i to the final state f of the system of interest, $W_{i\to f}^{(2)}$, induced by irradiation of two types of the laser light with

frequencies ω_l and $\omega_{l'}$ and polarization unit vectors $\mathbf{e}_l$ and $\mathbf{e}_{l'}$, takes the following form:

$$W^{(2)}_{i\to f} = \frac{2\pi}{\hbar^2}\sum_l\sum_{l'}\left(\frac{e^2 n_l\omega_l}{2\varepsilon_0 L^3}\right)\left(\frac{e^2 n_{l'}\omega_{l'}}{2\varepsilon_0 L^3}\right) \times |S_{fi}(\omega_l\mathbf{e}_l, \omega_{l'}\mathbf{e}_{l'})|^2\delta(\omega_{fi}-\omega_l-\omega_{l'}), \quad (16)$$

where

$$S_{fi}(\omega_l\mathbf{e}_l, \omega_{l'}\mathbf{e}_{l'}) = \sum_m\left[\frac{\mathbf{e}_{l'}\cdot\mathbf{R}_{fm}\mathbf{e}_l\cdot\mathbf{R}_{mi}}{\omega_{mi}-\omega_l} + \frac{\mathbf{e}_l\cdot\mathbf{R}_{fm}\mathbf{e}_{l'}\cdot\mathbf{R}_{mi}}{\omega_{mi}-\omega_{l'}}\right], \quad (17)$$

in which $\mathbf{R}_{fm}$ is the transition matrix element defined by $\langle f|\mathbf{r}|m\rangle$, and ω_{mi} the frequency difference between m and i levels.

The summations over the photon modes in Eq. (16) can be replaced by integrations after taking $L\to\infty$ according to the following equation:

$$\sum_l \to \frac{L^3}{8\pi^3 c^3}\int_0^\infty d\omega_l\omega_l^2\int_{\Omega_l} d\Omega_l \quad (18)$$

together with $n_l\to n(\omega_l)$, where Ω_l is a solid angle about the polarization vector $\mathbf{e}_l$. The resulting expression for the two-photon transition probability is given by the following:

$$W^{(2)}_{i\to f} = 2\pi(2\pi\alpha)^2 I_1 I_2\omega_1\omega_2|S(\omega_1\mathbf{e}_1, \omega_2\mathbf{e}_2)|^2 \times\delta(\omega_{fi}-\omega_1-\omega_2), \quad (19)$$

where I_1 and I_2 are the photon fluxes in units of numbers of photon per second per area and I_i is given by the following:

$$I_i = \frac{1}{8\pi^3 c^2}\int d\omega_i\omega_i^2 n(\omega_i)\int_{\Omega_i} d\Omega_i. \quad (20)$$

In Eq. (19), α denotes the fine-structure constant and is given by $\alpha = e^2/(4\pi\varepsilon_0\hbar c)\simeq 1/137.04$.

An expression for an m-photon transition probability can be obtained from the Fermi's golden rule in a similar manner. The resulting expression explains the formal intensity law and is valid for nonresonant multiphoton transitions. In order to study resonance effects and saturation phenomena, one has to utilize other theoretical treatments taken into account an infinite perturbation procedure.

C. Green's Function (Resolvent) Method

The Green's function method can be applied to the interpretation of optical Stark effects and also resonance effects. The Green's function (resolvent) is defined by the following:

$$(E-H)G(E) = 1 \quad (21)$$

The transition amplitude from I to F states, $U_{FI}(t)$, is expressed in terms of the time-independent Green's function as follows:

$$U_{FI}(t) = \langle F|\exp(-itH/\hbar)|I\rangle = \frac{1}{2\pi i}\int dE\exp\left(\frac{-iEt}{\hbar}\right)G_{FI}(E), \quad (22)$$

where $G_{FI}(E)$ is the matrix element of the Green's function.

The transition probability per unit time is given by

$$W_{I\to F} = \lim_{t\to\infty}\frac{d}{dt}|U_{FI}(t)|^2. \quad (23)$$

The matrix elements of the Green's function are evaluated by using the Dyson equation,

$$G(E) = G^0(E) + G^0(E)VG(E), \quad (24)$$

where the zero-order Green's function $G^0(E)$ satisfies the following:

$$(E-H_0+i\eta)G^0(E) = 1, \quad (25)$$

with $\eta\to 0^+$.

One of the merits of using the Green's function method is that the effect of level width Γ_M and level shift D_M of the intermediate states of M in the expression for the multiphoton transition probability can easily be taken into account. The width and shift originate from the interaction of atoms or molecules with the photon field and/or the heat bath. For example, for two-photon processes such as two-photon absorption, and Raman scattering, after utilizing the Dyson equation in evaluating the relevant matrix elements of the Green's function, the matrix element $G_{FI}(E)$ can be expressed as follows:

$$G_{FI}(E) = \sum_M\frac{G^0_{FF}(E)V_{FM}V_{MI}G^0_{II}(E)}{E-E^0_M-\Lambda_{MM}(E)}, \quad (26)$$

where energy-dependent self-energy $\Lambda_{MM}(E)$ can be written as follows:

$$\Lambda_{MM}(E) = D_M(E) - (i/2)\Gamma_M(E), \quad (27)$$

in which the level shift $D_M(E)$ and the width $\Gamma_M(E)$ are given by the following:

$$D_M(E) = P\sum_B\frac{|\langle B|V|M\rangle|^2}{E-E^0_B}, \quad (28)$$

where P denotes the principal part and $\Gamma_M(E)$ is defined by the following:

$$2\pi\sum_B|\langle B|V|M\rangle|^2\delta\big(E-E^0_B\big), \quad (29)$$

respectively. The term B appearing in Eqs. (28) and (29) excludes the initial, final, and intermediate states and denotes the states combined with the intermediate states

through the system–photon field interaction and/or the system–perturber interactions. From Eqs. (26), (22), and (23), an expression for the two-photon transition probability $W^{(2)}_{I\to F}$ can be derived. The resulting expression is identical to that derived in the ordinary perturbation approach, except for the inclusion of the level shifts and the widths of the intermediate states.

The Green's function method just described can be applied to the multiphoton processes in the low-temperature case. In order to take into account temperature effects on the multiphoton processes, other methods, such as the temperature-dependent Green's function or the density matrix method, are commonly used.

D. Density Matrix Method

The density matrix method is widely applied to investigation of mechanisms of multiphoton transitions and to derivation of an expression for the nonlinear susceptibility for nonlinear optical processes of atoms and molecules in the presence of the heat bath. Collision-induced multiphoton transitions that are induced by an elastic interaction between the system and the heat bath during the photon absorption, sometimes referred to as optical collisions, can be explained by using the density matrix method.

The density matrix for the total system, including the heat bath and the photon field $\rho(t)$, is defined by the following:

$$\rho(t) = \sum_i N_i |\psi_i(t)\rangle\langle\psi_i(t)|, \tag{30}$$

where $\psi_i(t)$ is the wave function for the ith quantum system of the total system at time t and N_i the weighting factor. The time evolution of the total system is determined by the Liouville equation for the density matrix as follows:

$$i\hbar\frac{\partial}{\partial t}\rho(t) = [H, \rho(t)], \tag{31}$$

where the square-bracketed term denotes the commutator.

Let us consider a contribution of a collision-induced two-photon absorption by irradiation of two kinds of lasers with a near-resonance frequency ω_1 between initial and resonant states and with frequency ω_2 to the rate constants. In describing the near-resonant two-photon absorption processes, it is necessary to obtain the fourth-order solution of the Liouville equation of Eq. (31). The perturbative solution can be written as follows:

$$\rho(t) = (i\hbar)^{-4}\int_{-\infty}^{t} dt_1 \int_{-\infty}^{t_1} dt_2 \int_{-\infty}^{t_2} dt_3 \int_{-\infty}^{t_3} dt_4 \times [V(t_1), [V(t_2), [V(t_3), [V(t_4), \rho(-\infty)]]]], \tag{32}$$

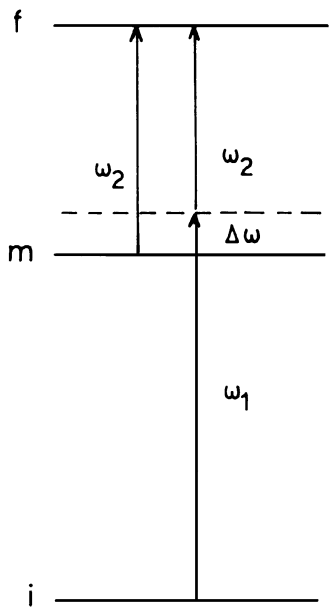


FIGURE 3 Three-level model for a near-resonant two-photon absorption. The detuning frequency is defined by $\Delta\omega = \omega_1 - \omega_{mi}$, where ω_{mi} is the frequency difference between the resonant and initial levels.

where $V(t)$ denotes the system–photon field interaction Hamiltonian in the interaction picture and $\rho(-\infty)$ is the density matrix in the initial state. After tracing out over the photon-field variables and the heat-bath variables, the diagonal matrix element representing the final state density of the system $\rho^{(S)}_{ff}(t)$ in a three-level model shown in Fig. 3 can be expressed as follows:

$$\rho^{(S)}_{ff}(t) = 2\,\mathrm{Re}\{A_{ff}(t) + B_{ff}(t) + C_{ff}(t)\}, \tag{33}$$

where

$$A_{ff}(t) = \frac{1}{\hbar^4}\int_{-\infty}^{t} dt_1 \int_{-\infty}^{t_1} dt_2 \int_{-\infty}^{t_2} dt_3 \int_{-\infty}^{t_3} dt_4 \times \langle E_1^{(+)}(t_2)E_1^{(-)}(t_4)\rangle\langle E_2^{(+)}(t_1)E_2^{(-)}(t_3)\rangle \times g_{fm}(t_1 - t_2)g_{fi}(t_2 - t_3)g_{mi}(t_3 - t_4) \tag{34a}$$

$$B_{ff}(t) = \frac{1}{\hbar^4}\int_{-\infty}^{t} dt_1 \int_{-\infty}^{t_1} dt_2 \int_{-\infty}^{t_2} dt_3 \int_{-\infty}^{t_3} dt_4 \times \langle E_1^{(+)}(t_3)E_1^{(-)}(t_4)\rangle\langle E_2^{(+)}(t_1)E_2^{(-)}(t_2)\rangle \times g_{fm}(t_1 - t_2)g_{mm}(t_2 - t_3)g_{mi}(t_3 - t_4)] \tag{34b}$$

and

$$C_{ff}(t) = \frac{1}{\hbar^4}\int_{-\infty}^{t} dt_1 \int_{-\infty}^{t_1} dt_2 \int_{-\infty}^{t_2} dt_3 \int_{-\infty}^{t_3} dt_4 \times \langle E_1^{(+)}(t_3)E_1^{(-)}(t_4)\rangle\langle E_2^{(+)}(t_2)E_2^{(-)}(t_1)\rangle \times g_{fm}(t_1 - t_2)g_{mm}(t_2 - t_3)g_{mi}(t_3 - t_4) \tag{34c}$$

In Eq. (34), $\langle \cdots \rangle$ denotes the photon-field correlation function. The matrix element g, for example, $g_{mi}(t_\alpha - t_\beta)$, is that of the time evolution operator for the density matrix representing the system and heat bath and can be expressed phenomenologically as follows:

$$g_{mi}(t_\alpha - t_\beta) = \exp[-i(t_\alpha - t_\beta)\omega_{mi} - |(t_\alpha - t_\beta)|\Gamma_{mi}], \quad (35)$$

where ω_{mi} is the frequency difference of the system between m and i states and Γ_{mi} is the dephasing constant relevant to these states. The structure of the dephasing constant can be clarified by using the density matrix method combined with the projection operator or the cumulant expansion technique. In the Markoff approximation, the dephasing constant is given by the following:

$$\Gamma_{mi} = \tfrac{1}{2}(\Gamma_{mm} + \Gamma_{ii}) + \Gamma_{mi}^{(d)}, \quad (36)$$

where $\Gamma_{mm}(\Gamma_{ii})$ represents the population decay constant of the $m(i)$ state, with $\Gamma_{mi}^{(d)}$, in which $m \neq i$ is the pure dephasing constant originating from the elastic interaction between the system and the heat bath.

In qualitatively understanding the mechanism of multiphoton processes it is convenient to use a diagrammatic representation of time evolution of the ket and bra vectors. The diagrammatic representation in the case of the two-photon absorption is shown in Fig. 4. Figures 4a–4c correspond to the time evolution in Eqs. (34a), (34b), and (34c), respectively. The lower and upper lines represent the time evolution of the ket $|\rangle$ and bra $\langle|$ vectors of the system, respectively. Time develops from the left-hand to right-hand sides. The wavy lines represent the system–photon interaction. The dotted points indicate the interaction points. In the diagram in Fig. 4a, two wavy lines with ω_1 and ω_2 overlap each other during time $t_2 - t_3$. This corresponds to a simultaneous two-photon process. Figures 4b and 4c describe sequential two-photon processes in which interactions of the system with the photon fields 1 and 2 are independent of each other.

The two-photon transition probability per unit time, $W_{i\to f}^{(2)}$, is defined as follows:

$$W_{i\to f}^{(2)} = \lim_{t\to\infty} \frac{d}{dt}\rho_{ff}^{(S)}(t). \quad (37)$$

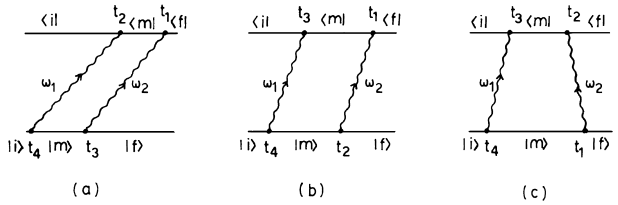


FIGURE 4 Diagramatic representations of the ket and bra vectors for a near-resonant two-photon absorption: (a), (b), and (c) correspond to the time evolution of the density matrix in Eqs. (34a), (34b), and (34c), respectively.

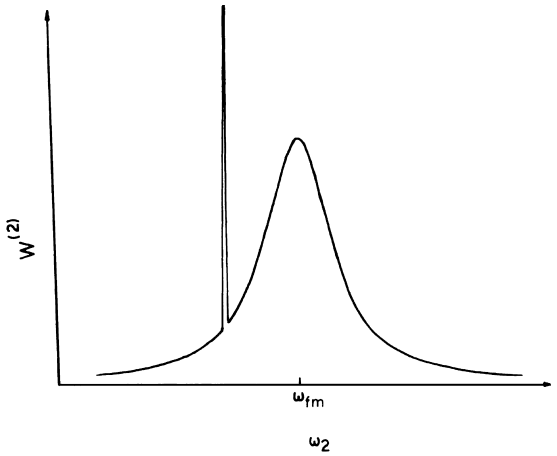


FIGURE 5 A line shape of a two-photon absorption as a function of the second laser frequency ω_2. The broad band centered at $\omega_2 = \omega_{fm}$ originates from the collision-induced sequential mechanism and the sharp band at $\omega_2 = \omega_{fm} - \Delta\omega$ from the coherent two-photon mechanism. The band width of the former is mainly characterized by the dephasing constant due to the collision between the system and the perturbers and that of the latter by the laser band width.

For an idealized steady-state laser excitation characterized by a negligibly small band width for both lasers, the transition probability per unit time takes the following form:

$$W_{i\to f}^{(2)} = \propto \frac{1}{(\omega_{mi} - \omega_1)^2 + \Gamma^2}\{\pi\delta(\omega_f - \omega_i - \omega_1 - \omega_2) + \frac{\Gamma_{mi}^{(d)}}{\Gamma_{mm}}\frac{\Gamma}{[(\omega_f - \omega_m - \omega_2)^2 + \Gamma^2]}. \quad (38)$$

In deriving Eq. (38), $\Gamma_{fm} = \Gamma_{mi} = \Gamma$ and $\Gamma_{fi} = 0$ have been assumed for simplicity. The first term in Eq. (38) represents the coherent two-photon transition, and the second term the collision-induced (sequential) two-photon transition due to the system–heat bath elastic interaction. The latter transition rate constant is proportional to pressure of the perturbers added. A line-shape function of the resonant two-photon absorption as a function of the frequency of the second laser ω_2 is drawn schematically in Fig. 5 to demonstrate contribution of the collision-induced two-photon transition to the total line shape. The broad line shape represents that for the collision-induced sequential two-photon transition and the sharp line shape that of the simultaneous process.

E. Susceptibility Method

The susceptibility method is widely applied to the explanation of nonlinear optical phenomena, such as harmonic generation, sum- and difference-frequency generation, stimulated scattering, and multiphoton absorption, that originate from nonlinear interaction between a coherent

laser field and material. The macroscopic polarization of the material induced by the incident radiation field P can be expanded as follows:

$$\mathbf{P} = \chi^{(1)} \times \mathbf{E} + \chi^{(2)} : \mathbf{EE} + \chi^{(3)} \vdots \mathbf{EEE} + \cdots, \quad (39)$$

where $\mathbf{E}$, the incident radiation field, is expressed as follows:

$$\mathbf{E} = \frac{1}{2} \sum_i \{\mathbf{e}_i E_i \exp[i(\mathbf{k}_i \times \mathbf{r} - \omega_i t)] + \text{c.c.}\} \quad (40)$$

and $\chi^{(n)}$ is called the nth-order susceptibility. For isotropic materials characterized by inversion symmetry, the lowest nonlinear susceptibility $\chi^{(2)}$ vanishes, and the nonlinearity in these materials is usually described in terms of the third-order nonlinear susceptibility $\chi^{(3)}$, neglecting higher order ones. The electric field $\mathbf{E}$ originating from the nonlinear polarization satisfies the Maxwell wave equation as follows:

$$\nabla \times \nabla \times \mathbf{E} + \frac{1}{c^2} \frac{\partial^2 \mathbf{E}}{\partial t^2} = -\frac{1}{\varepsilon_0 c^2} \frac{\partial^2 \mathbf{P}}{\partial t^2}. \quad (41)$$

The real physical fields contain both positive and negative frequency components. Depending on the choice of these components, specific optical nonlinear processes are described. Typical third-order optical processes, third-harmonic generation, sum-frequency generation, and coherent anti-Stokes Raman scattering (CARS) are schematically shown in Fig. 6.

For a polarization at the sum frequency $\omega_4 = \omega_1 + \omega_2 + \omega_3$, in the presence of the three kinds of lasers with amplitudes E_1, E_2, and E_3, the electric field amplitude E_4 at distance l can be expressed as follows:

$$E_4 = \chi^{(3)} E_1 E_2 E_3 l[\exp(i \Delta k\, l) - 1]/\Delta k, \quad (42)$$

where Δk, called the phase match parameter, is given by $\Delta k = k_1 + k_2 + k_3 - k_4$. The intensity of the net field I_4 can be written as follows:

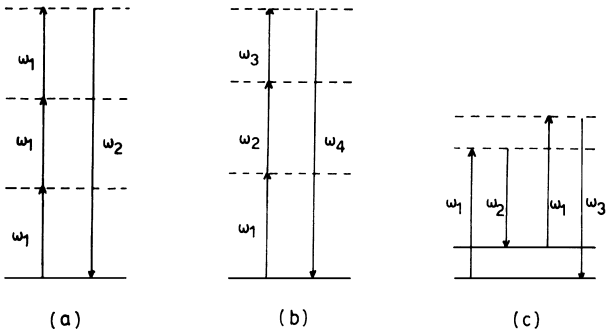


FIGURE 6 Schematic energy level diagrams for third-order optical processes: (a) third-harmonic generation, $\omega_2 = 3\omega_1$; (b) sum-frequency generation, $\omega_4 = \omega_1 + \omega_2 + \omega_3$; (c) coherent anti-Stokes Raman scattering (CARS), $\omega_3 = 2\omega_1 - \omega_2$.

$$I_4 = \frac{256\pi^4 \omega_1 \omega_2 \omega_3 \omega_4^3}{c^8 k_1 k_2 k_3 k_4} \left|\chi^{(3)}\right|^2 I_1 I_2 I_3 l^2 \times [\sin(\Delta k\, l/2)/\Delta k\, l/2]^2, \quad (43)$$

in which $k_i = \varepsilon_i^{1/2} \omega_i / c$. Here the dielectric constant $\varepsilon_i = 1 + 4\pi\chi^{(1)}$ is equal to the square of the index of the refraction n_i. So far the nonlinear wave mixing processes have been treated in the classical method. It has been shown that the signal is proportional to the products of the intensity of each incident laser field, and for the signal to be observed the phase matching condition $\Delta k = 0$ has to be satisfied.

In order to investigate the frequency dependence of the material, especially resonance enhancement of nonlinear process, it is necessary to clarify the structure of the third-order nonlinear susceptibility. This can be carried out by using the semiclassical method, in which the radiation field is treated classically and the material system quantum-mechanically in solving equation of motion for the density matrix, Eq. (31). The resulting polarization is given in terms of the expectation value of the dipole moment $\boldsymbol{\mu} = e\mathbf{r}$ as follows:

$$\mathbf{P} = N \,\text{Tr}(\rho \boldsymbol{\mu}), \quad (44)$$

where ρ is a solution of Eq. (31) in the steady-state condition.

A two-photon absorption can be treated as a third-order process in which two lasers at ω_1 and ω_2 excite atoms or molecules of the materials from the i to f states. The nonresonant two-photon absorption coefficient is linearly proportional to the imaginary part of the third-order nonlinear susceptibility $\chi''^{(3)}$, which is expressed as follows:

$$\chi''^{(3)} = |S_{fi}|^2 g(\hbar\Delta\omega)(N_i - N_f). \quad (45)$$

The line shape function $g(\hbar\Delta\omega)$ is given as follows:

$$g(\hbar\Delta\omega) = \frac{\hbar\Gamma_{fi}}{\pi\left[(\hbar\Delta\omega)^2 + (\hbar\Gamma_{fi})^2\right]}, \quad (46)$$

with $\Delta\omega = \omega_f - \omega_i - \omega_1 - \omega_2$ and $N_i(N_f)$ the number of atoms or molecules in the initial (final) state.

F. Direct Method for Solving the Time-Dependent Schrödinger Equation

In treating multiphoton processes of atoms or molecules induced by an intense ultrashort pulsed laser, it is convenient to use the semiclassical treatment of system–radiation interactions. The Hamiltonian is given as follows:

$$H(t) = H_s + V_{sr}(t), \quad (47)$$

where H_s is the system Hamiltonian and $V_{sr}(t)$ is the interaction Hamiltonian. The interaction Hamiltonian is expressed in the dipole approximation as follows:

$$V_{\mathrm{sr}}(t) = -e\mathbf{r}\cdot\mathbf{e}\,E(t)\sin(\omega_{\mathrm{R}}t + \varphi(t)), \qquad (48)$$

where $E(t)$ is the pulse-envelope function, ω_{R} is the central frequency of the laser pulse, and $\varphi(t)$ is the phase which generally depends on time.

The time-dependent Schrödinger equation is expressed as follows:

$$i\hbar\frac{\partial\Psi(t)}{\partial t} = H(t)\Psi(t), \qquad (49)$$

where $\Psi(t)$ is the wave function of the system at time t. The probability amplitude $c_m(t)$ for a bound eigenstate m is obtained by projecting $\Psi(t)$ on the bare eigenstate of the free field system, as long as the field is not too strong, as follows:

$$c_m(t) = \langle m|\Psi(t)\rangle. \qquad (50)$$

The population at time t is given by taking the absolute square of the amplitude. The nonlinear scattering light spectrum $\sigma(\omega)$ related to harmonic generation is obtained by taking the square of the Fourier transform of time-dependent acceleration as follows:

$$\alpha(t) = \left\langle\frac{d^2x}{dt^2}\right\rangle = \left\langle\Psi(t)\left|\frac{d^2x}{dt^2}\right|\Psi(t)\right\rangle$$

as

$$\sigma(\omega) = \left|\frac{1}{T}\int_0^T dt\,\exp(-i\omega t)\alpha(t)\right|^2. \qquad (51)$$

There are two methods for evaluating $\Psi(t)$. One is to utilize the expansion of spatially delocalized bases with time-dependent coefficients. The other consists of using a grid representation. In the former expansion method, the time-dependent Schrödinger equation is transformed into a system of coupled first-order differential equations for the time-dependent coefficients. For the motion of a spatially localized electronic wavepacket created by an ultrashort, intense laser field pulse, the description of the delocalization brings about a poor convergence in certain cases. In the grid representation method, split operator techniques combined with a fast Fourier transform have been successfully applied to nuclear wavepacket dynamics. The time-dependent electronic wave packets are evaluated using an analytical potential model for Coulomb interactions because of its long range and the singularity at the origin. A dual transformation method has recently been developed as an efficient grid method for accurately treating electronic dynamics. This consists of wave functions that are set at zero at the Coulomb singularity point and the introduction of a new scaled coordinate in which the unit is small near the nuclei and large at longer distances.

III. EXPERIMENTAL METHODS

Since two-photon absorption is a second-order process, it is rather weak at the moderate light intensities available from tunable dye lasers. This demands a sensitive technique for detection of only a few two-photon absorption events in the sample. In principle, higher light intensities could be used, but then higher order processes become more probable and the measured spectrum may be a superposition of two- and three-photon spectra. Even ionization and fragmentation of the molecules is possible at high light intensities, and in this case also the multiphoton spectrum of a fragment may be superposed. For this reason in most cases one should refer to highly sensitive detection techniques and moderate light intensities rather than high light intensities and nonsensitive detection technique. First, a series of sensitive detection techniques are discussed in this section.

A. Measurement of the Photon Absorption Due to a Two-Photon Absorption Process

The measurement of absorption contains two steps: first the light power in front of the sample has to be measured and the light power after passing the sample. The detection limit is strongly dependent on the special technique used in the experiment, the integration time, and so on. Usually, one is not able to detect differences in light intensities less than 0.1%. Qualitatively, it is clear that real absorption measurements in two-photon spectroscopy are only possible for samples of high density (e.g., liquids and solids). It is not a feasible method for gasphase spectroscopy, since it is impossible to get the laser light highly focused over a long absorption path length. The first absorption measurement in two-photon spectroscopy was done using the combination of a ruby laser and continuum flash lamp. The two-photon absorption was monitored on an oscilloscope as a short dip in the transmitted flash-lamp light intensity that coincides with the laser pulse. For this experiment, an accurate overlap of both light beams over a long distance is necessary in order to get a high level of absorption.

A special setup was developed by Hopfield *et al.*, who used a crystal that acts as a light guide for both light beams.

In a more recent work, a combination of a high-power pulsed dye laser and a fixed-frequency continuous-wave (cw) Kr^+ ion laser was used to detect absorption differences as small as 0.1%. Even though in this way a spectrum is hard to measure, very accurate absolute values of the two-photon absorption cross section were obtained for diphenylbutadiene at a special wavelength. In addition to the spectrum, the absolute two-photon cross section yields

further arguments whether the two-photon absorption is pure electronic or is vibrationally induced.

A very sensitive technique for detection of weak absorption is the intracavity absorption technique. Here the weakly absorbing sample is placed in the cavity of a dye laser. As a consequence, the dye-laser emission is quenched at those wavelengths where the material absorbs. In this way an absorption as small as 10^{-4} can be detected in a time interval. This technique is appropriate to detect small amounts of material with sharply structured spectra rather than to measure a complete spectrum. Two-photon absorption in anthracene solution was detected by intracavity absorption.

B. Detection of Fluorescence after Two-Photon Excitation of Molecules

The most convenient and sensitive method for the measurement of two-photon spectra is the detection of the fluorescence of the excited molecules. As a consequence of a two-photon absorption, most molecules emit a photon at about twice the energy of the absorbed photons, as shown in Fig. 7. These emitted photons are detected with high sensitivity. A typical setup is shown in Fig. 8. The light from a tunable dye laser is focused within the fluorescence cell containing the molecular gas or the solution. In the small focus, simultaneous absorption of two (visible) photons takes place and ultraviolet (UV) photons are emitted with a quantum yield typical for the molecule under investigation. The emitted UV photons are observed with a high-gain photomultiplier. Special filters are used

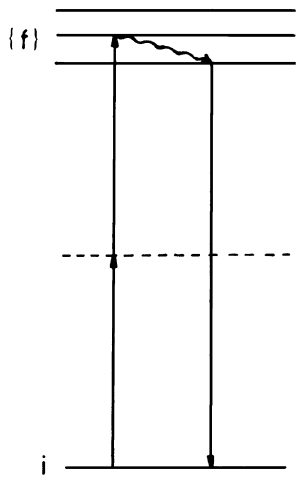


FIGURE 7 Principle of a nonresonant two-photon absorption measurement by proving the photon emission. The straight lines with arrow denote the photon absorption and emission processes, and a wavy line represents a relaxation process from the level excited to the lowest one in the final electronic state.

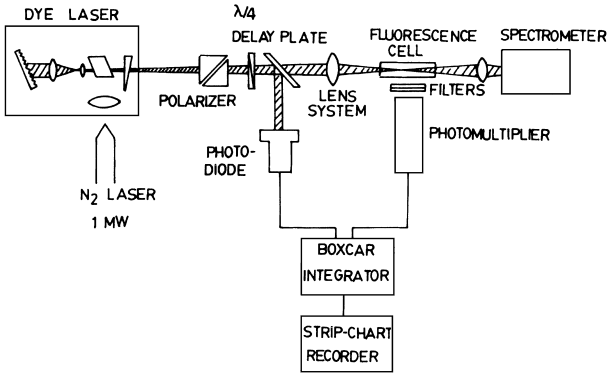


FIGURE 8 Experimental setup of the fluorescence detection after a two-photon excitation of a gaseous sample.

in order to discriminate them from the intense exciting visible light. This can easily be accomplished since there usually is a large frequency shift between the exciting visible and the emitted UV light in two-photon spectroscopy. This greatly improves the signal-to-noise ratios relative to one-photon excitation and represents a side benefit of this detection technique. In the gas and in solution of low concentration there is no reabsorption of the emitted photons; however, this can be a problem in pure liquids and in crystals with a large absorption cross section. The signal of the photomultiplier is then fed into a boxcar integrator and integrated there within a short time interval of 10 nsec to some picoseconds, depending on the lifetime of the fluorescence. The integrated signal then is recorded on a strip-chart recorder. Finally, when continuously scanning the wavelength of the dye laser, one obtains the two-photon excitation spectrum of the molecule under investigation. For very weak signals and for time-resolved fluorescence measurements, the boxcar integrator may be replaced by a transient digitizer and a data-processing system.

The solid angle for observation of fluorescence is about 6×10^{-1} sr in this setup. The detection limit of the setup is reached when the signal-to-noise ratio in the measured spectrum is better than 1:1. For a conventional photomultiplier, the recording of two-photon spectra is possible when 200 photons are transmitted from the focus for a single exciting laser pulse.

For molecules with a fluorescence quantum yield of unity, this means that some 200 two-photon absorption events should have taken place during the laser pulse in order to be able to measure a two-photon excitation spectrum. Unfortunately, the fluorescence quantum yield of most molecules is smaller than 1, this being a fundamental disadvantage of the fluorescence detection method. In this case the detection by resonance-enhanced multiphoton ionization might be useful, as discussed in the next

section. It is also possible to detect phosphorescence if the molecules undergo a fast intersystem crossing process into the triplet system. For detection of vibronic states with high excess energies above S_1, it might be useful to use a high-pressure buffer gas producing a fast collisional deactivation of the excited levels down to the vibrationless ground state. At high pressures this collisional deactivation might compete with the internal radiationless process, and fluorescence from the thermalized S_1 with laser fluorescence quantum yields is observed. By this method, vibronic states with excess energy as large as 6000 cm^{-1} above the vibrationless electronic S_1 state have been observed.

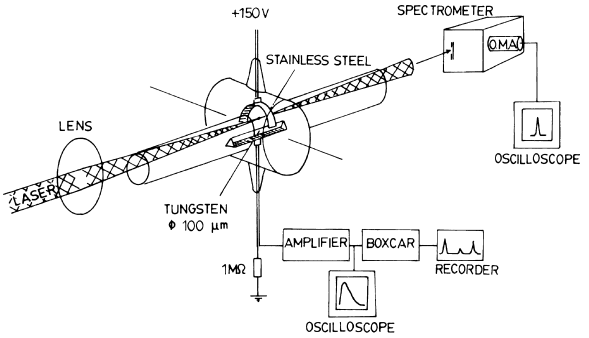


FIGURE 9 Ionization cell and the setup for recording multiphoton intermediate-state spectra of gaseous samples.

C. Detection of Two- or Three-Photon Absorption by Multiphoton Ionization of Molecules

After a molecule has been excited in an intense light field by two- or three-photon absorption, there is a high probability of absorbing further photons, which finally result in an ionization of the molecule. Another sensitive detection technique for two- or three-photon absorption processes in molecular gases is based on the subsequent ionization of the molecule after the two- or three-photon excitation.

The interesting feature of the multiphoton ionization from the spectroscopic point of view is the resonance enhancement by resonant intermediate states. Since the ionization efficiency is strongly enhanced when the photon energy comes to resonance with real intermediate states, a wavelength scan of the laser leads to a modulation of the ion current, which reflects the spectrum of the intermediate states. Thus it is possible to measure the intermediate-state (two- or three-photon) spectrum by measuring the ion currents. This method is used to measure two-photon spectra of polyatomic molecules by a three- and four-photon ionization.

1. The Ionization Cell

In Fig. 9, the ionization cell is shown in detail. The laser is focused into the cell containing the molecular gas at a typical pressure of a few torr. Very common is a device with a thin wire that is axially positioned in a cylindrical metal plate biased with a positive voltage of some 100 V. The potential drives the free electrons produced by the multiphoton ionization process to the electrode with positive charge. If enough voltage is applied between the electrodes, and if the particle density is sufficient, charge amplification by collisions can take place, increasing the detectability of the electrons. If the gas pressure is very low, the addition of a buffer gas is necessary in order to get charge multiplications in the ionization cell. The voltage produced by the current at a 1-MΩ impedance is then amplified in a preamplifier by one order of magnitude fed into a boxcar integrator and then integrated with a gate width of some 10 μsec. The modulation of the current as a function of wavelength is then recorded on a strip-chart recorder and reflects the intermediate-state resonance. Without a charge amplification in the ionization cell, the empirically found detection limit is about 1000 ions produced within one laser pulse.

2. Ion Detection in a Mass Spectrometer

Ion detection in an ionization cell is the simplest method and is very sensitive. There is, however, no information about the type of ions produced in the multiphoton ionization process.

In order to shed light on the ionization process, it is useful to detect the ions in a mass spectrometer, which allows one to determine the mass of the ions. The scheme of the setup for mass-selective ion detection is shown in Fig. 10. The tunable laser light or the frequency-doubled light is focused into an effusive molecular beam close to the aperture of the nozzle. The ions produced by the multiphoton ionization process are withdrawn through an

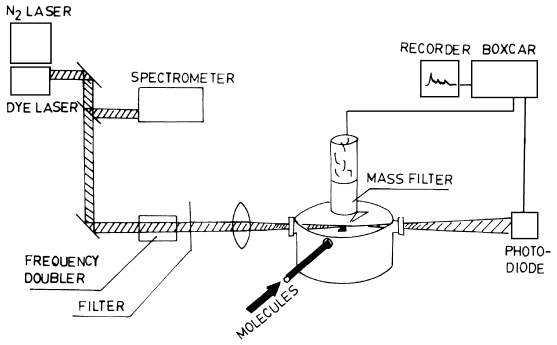


FIGURE 10 Experimental setup for multiphoton ionization mass spectroscopy.

ion lens system into the mass filter, mass-analyzed, and finally recorded with an ion multiplier. Ion detection in a mass spectrometer clearly reveals that in a typical two-photon experiment, molecules are not only excited to the two-photon state but molecules are ionized, and a rich pattern of fragment ions is formed. Therefore, in a two- or, especially, three-photon experiment with highly intense laser light, we must check whether these fragments may influence the measured two- or three-photon spectrum. This is certainly of importance for the three- and four-photon ionization experiments.

The ions produced in a multiphoton ionization process have particular features: they are produced within a short time interval of some nanoseconds during the laser pulse in the small volume of about 10^{-5} cm^3 given by the focus of the laser light. This spotlike type of pulsed ion sources makes feasible a special type of mass spectrometer. Mass detection for these types of ion sources can be achieved in a more appropriate manner by a time-of-flight mass analyzer.

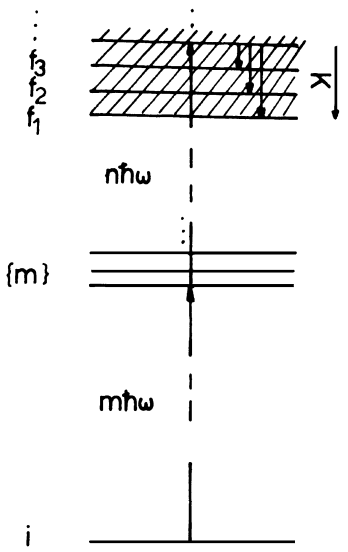


FIGURE 11 Principle of a multiphoton ionization photoelectron spectroscopy; K denotes the photoelectron energy.

D. Photoacoustic Detection Method

In this section, the photoacoustic detection method is briefly discussed. This method is principally different from the most convenient methods discussed in Sections III.B and III.C. Fluorescence detection and, to some degree, ionization detection decrease in sensitivity when a fast competing intra- or intermolecular relaxation takes place from the multiphoton excited level. In this case there is a dissipation of the energy selectively released in the excited level into thermal energy. After pulsed excitation there is a rapid conversion of absorbed energy into pressure fluctuations, which then can be detected by a microphone. This means that photoacoustic spectroscopy is based on the detection of those effects that are loss channels in fluorescence and ionization detection. Apparently, photoacoustic detection technique should be principally suitable for observing two-photon spectra in weakly fluorescing materials.

For the time being, there have been only a few successful attempts to measure multiphoton spectra of molecules by the photoacoustic methods. The two-photon spectrum of liquid benzene has been published with the main vibronic bands and a value for the two-photon absorption cross section.

E. Detection of Photoelectrons

This method is a combination of a photoelectron spectroscopic technique and a multiphoton spectroscopic technique with a multiphoton ionization laser system. The photoelectron intensity is measured as a function of the kinetic energy of electrons released. This is called multiphoton ionization photoelectron spectroscopy. The observed spectra contain information about the final states of the ions produced by multiphoton excitation when the resonant intermediate states are well identified or about the resonant states when the final ionic states are well defined. A schematic model is shown in Fig. 11 for observing the multiphoton ionization photoelectron spectra. The photoelectron energy is denoted by K. The photoelectrons from a sample irradiated by a pulsed tunable dye laser are detected by a time-of-flight energy analyzer. The photoelectron kinetic energy K(eV) is calculated by the formula $v = (5.93 \times 10^7)K^{1/2}$, where v(cm/sec) is the velocity of the observed photoelectron.

F. Miscellaneous Detection Methods

Some other methods have been proven to be sensitive enough for detection of a two-photon process.

One of these methods is thermal blooming, which is also based on the energy-loss mechanisms in the excited states. The transfer of energy from the excited states into heat causes a change of refractive index of the material under investigation. This is then detected by changes in the optical behavior of the material; that is, by a weak focusing or defocusing of the exciting laser beam. The thermal blooming technique has been applied to the test molecule benzene. Two-photon spectra of liquid benzene were measured point by point in the range between 360 and 530 nm. More recently a spectrum generated by this technique yielded additional information about the position of the $^1E_{1g}$ state of benzene. Another detection method for the two-photon absorption is based on changes of the susceptibility of the material under investigation in the presence of the light field. In a strong light field there

are higher order contributions $\chi^{(n)}$ to the susceptibility: The real part of the nonlinear susceptibility Re $\chi^{(3)}_{ijkl}$ produces an intensity-dependent index of refraction that may turn the polarization vector of the incoming light wave, and the imaginary part of the third-order susceptibility Im $\chi^{(3)}_{ijkl}$ produces an intensity-dependent absorption coefficient, which increases if there is a resonance at the two-photon energy. These two-photon resonances in the third-order susceptibility $\chi^{(3)}$ may be detected in several ways.

Two-photon resonances of gases such as SO_2 and NO can be detected in a four-wave mixing experiment. Since this method creates no real population of the resonant state, the detection of the resonances does not directly depend on the dynamical pathway followed by the excited state. The dynamics of the resonant states enters only through a damping parameter, thereby limiting the magnitude of the resonance term. It has been shown that this method is suitable for obtaining absolute two-photon cross section by comparison of the two-photon resonances with coherent antistokes Raman resonances of $\chi^{(3)}_{ijkl}$. Two-photon cross sections are then given in terms of the accurately known Raman cross sections. As demonstrated for Na vapors, the change of polarization produced by the imaginary part of the third-order susceptibility $\chi^{(3)}_{ijkl}$ can be detected for observation of the two-photon spectrum. An extension of this method of the case of molecules seems possible even though the sensitivity is not expected to be better than that of the other methods discussed above. The general virtue of two-photon absorption detection via $\chi^{(3)}_{ijkl}$ is the calibration of two-photon cross sections on the basis of Raman cross sections.

G. Doppler-Free Multiphoton Spectroscopy

One of the advantages of multiphoton spectroscopy is elimination of the inhomogeneous Doppler broadening in the spectra. For an ambient atomic or molecular gas, there is an isotropic velocity distribution that brings about different shifts for the atoms or molecules with different velocity components in the direction of light propagation. The average of these shifts results in a Doppler broadening in the optical transition. For a Maxwell–Boltzmann velocity distribution, a Gaussian line profile in the spectra is characterized by a full-width-at-half-maximum (FWHM) as follows:

$$\begin{aligned}\Delta\omega_{\mathrm{D}} &= (2\omega_{\mathrm{o}}/c)[2\ln(2kT/m)]^{1/2} \\ &= (2.163\times 10^{-7})\omega_{\mathrm{o}}(T/M)^{1/2}, \end{aligned} \qquad (52)$$

where ω_{o} is the optical transition angular frequency, T the temperature (K), m the mass in kilograms, and M the molecular weight of the particles in atomic mass units. Typically, the Doppler width in angular frequency $\Delta\omega_{\mathrm{D}}$ for a polyatomic molecule such as benzene C_6H_6 with $M = 78$ at room temperature is $\Delta\omega_{\mathrm{D}} = 1.67$ GHz for $\omega_{\mathrm{o}} = 40{,}000\ \mathrm{cm}^{-1}$. This value of the Doppler width is several times larger than the average spacing of rovibronic transitions in polyatomic molecules. Therefore, in this case it is not possible to observe single Doppler-broadened lines, but the envelope of the line produces a typical rotational contour of the vibronic band.

The frequency shifts of Doppler-limited and Doppler-free two-photon absorption are shown in Fig. 12, in which the interaction of the particle with two monochromatic light beams with frequencies ω_1 and ω_2 is presented. In general, the Doppler broadening in angular frequency is given by $\Delta\omega = \Delta\mathbf{k}\times\mathbf{v}$, where $\Delta\mathbf{k}$ is the change in momentum of the laser light and $\mathbf{v}$ is the atomic or molecular velocity. For each particle with velocity $\mathbf{v}$ whose propagation component in the propagation direction of the laser beams is v_x, the optical frequency ω_{o} is shifted by $(\omega_1+\omega_2)v_x/c$. When an ensemble of the particles in thermal equilibrium is investigated, this yields a broadening of the transition line according to Eq. (47). If two light beams with frequencies ω_1 and ω_2 propagate in opposite direction, as shown in the lower part of Fig. 12, and the particle absorbs one photon from each beam, then the corresponding Doppler shifts have opposite signs and the residual Doppler shift is given as $(\omega_1-\omega_2)v_x/c$. The shift cancels exactly to zero if the frequencies of both laser beams are equal to each other.

A typical experimental setup for observing the Doppler-free two-photon absorption is shown in Fig. 13. There are optics for polarization of photon and focusing laser beam, and there is a sample cell between the laser and detection system. Applications of the Doppler-free two-photon absorption of atoms and molecules are presented in Section V.

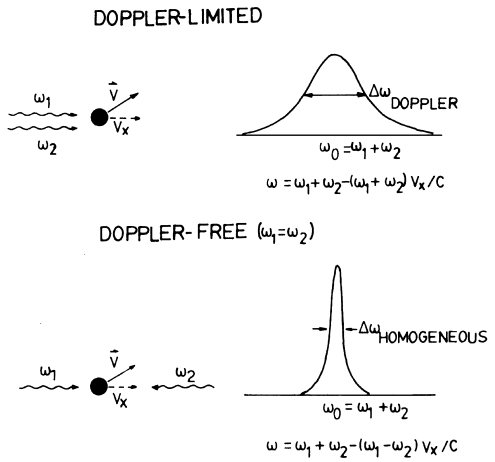


FIGURE 12 Principle of Doppler-limited and Doppler-free two-photon absorptions.

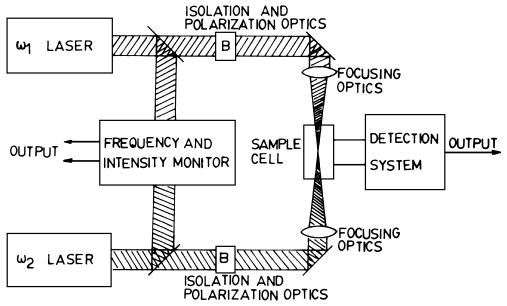


FIGURE 13 Experimental setup for observing the Doppler-free two-photon absorption.

H. Ionization-Dip Spectroscopy

This is a high-resolution multiphoton spectroscopy based on competition between ionization and stimulated emission (or stimulated absorption). Figure 14 shows the principle for the two-photon ionization case. Two photons at ω_1 induce efficient ionization because resonance enhancement occurs through intermediate electronic state m. A high-intensity laser probe at frequency ω_2 is introduced. When $\omega_1 - \omega_2$ matches a suitable vibrational frequency ω_{vi} in the electronic ground state, stimulated emission competes with ionization; this decreases the effective population in state m and reduces the ionization signal (i.e., ion dip). Ion dips may also take place if state ω is at a higher energy than state m.

IV. CHARACTERISTICS AND SPECTRAL PROPERTIES

Laser-intensity dependence and polarization behavior in multiphoton transitions, which are the main characteristics of multiphoton spectroscopy, are presented first. An application of the polarization behavior to a two-photon transition of a molecular system is described. Spectral properties of atomic and molecular multiphoton transitions are finally described, focusing on their difference between nonresonant and resonant multiphoton transitions.

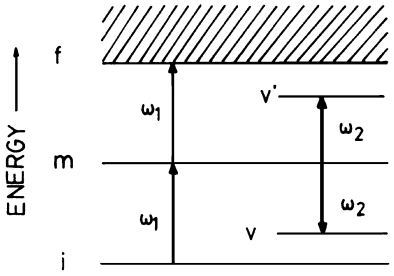


FIGURE 14 Simplified energy level diagram for ionization-dip spectroscopy.

A. Laser-Intensity Dependence

1. Perturbative Regime—Formal Intensity Law

Measurement of the laser-intensity dependence is very important to understand the mechanism of multiphoton processes. The laser-intensity dependence observed in the multiphoton transitions can be classified into two types. One type originates from an intrinsic laser intensity dependence, called the formal intensity law, and the other type from geometrical effects of the focused laser beam applied in the region of material–photon interaction. The latter is sometimes called the $\frac{3}{2}$-power law in multiphoton ionization dissociation experiments because product yields are proportional to $I^{3/2}$, the $\frac{3}{2}$-power of laser intensity, or in some cases to noninteger powers of the laser intensity irrelevant to the intrinsic intensity dependence. The $I^{3/2}$ dependence has been explained by the laser intensity change due to conical focusing in the material–photon interaction region.

The formal intensity law means that the n-photon transition rate constant is proportional to the nth order of laser intensity, I^n. This can be understood from the nonresonant multiphoton rate expression derived in the ordinary time-dependent perturbation theory; for an n-photon absorption experiment in which a single laser beam with frequency ω_R is irradiated, the transition probability from i to f states, $W_{i\to f}^{(n)}$, is written as follows:

$$W_{i\to f}^{(n)} = I^n \sigma_{i\to f}^{(n)} \tag{53a}$$

or

$$W_{i\to f}^{(n)} = \frac{I^n \sigma_{i\to f}^{(n)}}{(\hbar\omega_R)^n}, \tag{53b}$$

where I is the photon flux in units of photons per area per time and the photon intensity in units of energy per area per time in Eq. (53a) and (53b), respectively, and $\sigma^{(n)}$ is the cross section in units of (area)^n (time)$^{n-1}$, written as follows:

$$\begin{aligned}\sigma_{i\to f}^{(n)} = 2\pi(2\pi\alpha)^n \omega_R^n \Bigg| \sum_{m_1}\sum_{m_2}\cdots\sum_{m_{n-1}} & \\ \times \langle f|\mathbf{r}\cdot\mathbf{e}|m_{n-1}\rangle \cdots \langle m_1|\mathbf{r}\cdot\mathbf{e}|i\rangle / & \\ \left\{\left[\omega_{m_{n-1}} - \omega_i - (n-1)\omega_R\right]\right. & \\ \left.\cdots(\omega_{m_1} - \omega_i - \omega_R)\right\}\Bigg|^2 & \\ \times \delta(\omega_f - \omega_i - n\omega_R). & \end{aligned} \tag{54}$$

The formal intensity law has been utilized to determine orders of multiphoton processes. It should be noted that this law holds for nonresonant multiphoton transitions in low-intensity-laser experiments. The geometric effect of

the focused laser beam can be eliminated by setting up the laser in the crossed atomic or molecular beam.

2. High-Intensity Regime—Rate Equation Approach

Use of a high-intensity laser for resonant multiphoton transitions may result in a deviation from I^n dependence even after elimination of the geometric effect. A qualitative interpretation of the deviation can be made by using the rate equation approach. For simplicity, let us consider a resonant two-photon ionization shown in Fig. 15.

In the case of the weak laser field in which the effect of stimulated emission is negligible, and, furthermore, in the time regions of $W_{mi}^{(1)}t \leq 1$, the ionization rate is approximately proportional to $W_{fm}^{(1)}W_{mi}^{(1)}t$, which indicates a quadratic intensity dependence. The ion number is also given by the quadratic intensity dependence. On the other hand, in the case in which a strong laser intensity is applied, the stimulated emission process makes an important contribution, and the deviation of the ionization rate from the quadratic intensity dependence; that is, linear intensity dependence takes place. Two cases can be seen depending on the relative magnitudes between $W_{mi}^{(1)}(W_{im}^{(1)})$ and $W_{fm}^{(1)}$.

When $W_{mi}^{(1)} > W_{fm}^{(1)}$, an equilibrium between the initial and resonant states is achieved and the excitation process from the resonant to the final state can be observed as the apparent transition. When $W_{mi}^{(1)} < W_{fm}^{(1)}$, in which the resonant and ionized states are strongly coupled in terms of the dipole transition, the initial excitation process $i \rightarrow m$ can be observed as the apparent transition.

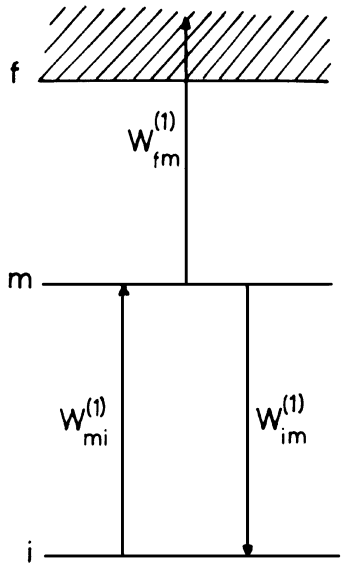


FIGURE 15 A simple model for a resonant two-photon ionization.

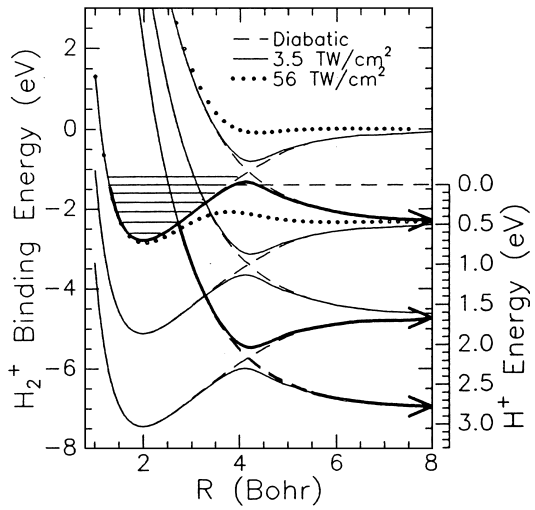


FIGURE 16 Bond softening of molecular hydrogen ion H_2^+ using a 532-nm laser light. The potential curves are calculated by the Floquet method. Solid curves indicate adiabatic potential at the intensity of 3.5×10^{13} W/cm^2. Dashed curves indicate diabatic (free-field) $1s\sigma_g$ and $2p\sigma_u$ states shifted by $n\hbar\omega$ for an n-photon process. [From Backsbaum, P. H., Zavriyev, A., Muller, H. G., and Schumacher, D. W. (1990). *Phys. Rev. Lett.* **64,** 1883.]

3. High-Intensity Regime—Bond Softening and Above-Threshold Dissociation

At high-intensity regimes, molecules exhibit dissociation that is different from multiphoton dissociation. This is called above-threshold dissociation. Above-threshold dissociation takes place at perturbative regimes. Interaction between molecules and intense laser fields brings about potential curves that are softened or flattened in the vicinity of a multiphoton resonance, as shown in Fig. 16. At the intensity of 3.5×10^{13} W/cm^2, the vibrational quantum number $v = 6$ of hydrogen molecular ion H_2^+ is no longer bound. Three channels of the above-threshold dissociation of H_2^+ are indicated in Fig. 16. Both multiphoton absorption and stimulated emission are involved in the bond softening and hardening processes. Bond softening causes the molecule to dissociate through possible channels corresponding to the absorption of one, two, or more photons. The resultant molecular dissociation fragments appear with kinetic energy equivalent to less than one photon. By analyzing the kinetic energy spectra of dissociative species (protons and deuterons), the magnitudes of bond softening can be estimated.

B. Polarization Behavior

Cross sections of multiphoton transitions of atoms or molecules in solids, liquids, and gases depend on whether linearly or circularly polarized laser light is applied. This is called the polarization dependence, one of the important

characteristics of multiphoton spectroscopy. Measurement of the polarization dependence makes it possible to assign the excited-state symmetry of the material and to obtain information about the mechanism of the transition. In order to see the origin of the polarization behaviors, the concept of photon angular momentum are briefly mentioned first. Polarization behaviors in multiphoton ionization of atoms and those in two-photon absorptions of both nonrotating and rotating molecules are then described.

1. Photon Angular Momentum and Polarization Effects

A photon state is characterized not only by its linear momentum but also by the polarization vector $\mathbf{e}_\lambda$, which transforms like a vector and is considered to be the intrinsic angular momentum (spin). Circular polarization vectors constructed by the linear polarization vectors $\mathbf{e}_x$ and $\mathbf{e}_y$ as follows:

$$\mathbf{e}_\pm = \mp(1/\sqrt{2})(\mathbf{e}_x \pm i\mathbf{e}_y) \tag{55}$$

are associated with the spin component $m = \pm 1$ of spin 1 where the quantization axis has been chosen in the photon propagation (z) direction. On the other hand, linear polarization vectors $\mathbf{e}_x$ and $\mathbf{e}_y$ are eigenstates of the intrinsic angular momentum but are not eigenstates of the projection of the spin on the polarization direction. Therefore, in an electronic transition induced by one-photon absorption or emission, one unit of angular momentum is transferred between the electromagnetic field and the electron in the dipole approximation. Because of a different selection rule for the spin component between linearly and circularly polarized photons—that is, the selection rule of the magnetic quantum number $\Delta m = 0$ for linearly polarized photons along the z axis and $\Delta m = \pm 1$ for photons circularly polarized and propagating along the z axis, in which $(+)$ and $(-)$ correspond to right and left polarization, respectively—transition rates of multiphoton processes in which more than one unit of angular momentum are transferred depend on the polarization applied. A simple example (Fig. 17) in a hydrogenic, one-electron model shows the angular-momentum channels available for four-photon transition from an S initial state. As a result of the difference in the number of available channels with different cross sections, the total transition rates of the multiphoton process depend on the polarization.

One of the polarization effects can be seen in multiphoton ionization of atoms and molecules, which is analyzed by using a method of multiphoton ionization photoelectron spectroscopy. The final ionized state is in the continuum and can adequately be expressed as a superposition

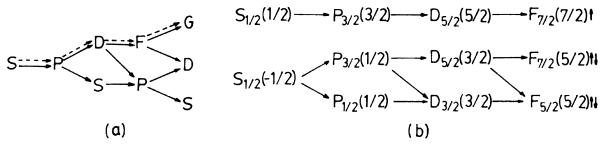


FIGURE 17 Angular momentum channels available for four-photon ionization of an atomic S state. (a) The linearly (——) and circularly (– – –) polarized cases are illustrated in the case of a negligibly small spin–orbit coupling. (b) The case for atoms with spin–orbit split levels and right circular polarization. The numbers in parentheses are the spin projection quantum numbers, while the arrows illustrate the possible orientations of the free electron spin. [From Lambropoulos, P. (1976). *Adv. Atomic Mol. Phys.* **12,** 87.]

of partial waves with well-defined angular momentum l; its electronic part is written as follows:

$$|f(r)\rangle = 4\pi \sum_{l=0}^{\infty} i^l \exp(-i\delta_l) G_l(\mathbf{k}, \mathbf{r}) \times \sum_{m=-l}^{l} Y^*_{lm}(\Theta, \Phi) Y_{lm}(\theta, \phi), \tag{56}$$

where δ_l is the phase shift, G_l the radial part of the partial wave, and Y^*_{lm} and Y_{lm} are spherical harmonics. The spherical coordinates of the wave vector/radius vector ($\mathbf{k}$) and ($\mathbf{r}$) are denoted by (k, Θ, Φ) and (r, θ, ϕ), respectively. All angular momenta in the final state are available for the multiphoton transition; that is, a photon of any arbitrary polarization leads to ionization, in contrast to a bound–bound transition. The total ionization rate again depends on the photon polarization. For example, a three-photon ionization with a circularly polarized photon leads to a photoelectron of orbital angular momentum $l = 3$ (F wave); on the other hand, in the case of a linearly polarized photon, it leads to a photoelectron whose state is a superposition of $l = 1$ (P wave) and $l = 3$, as shown in Fig. 17.

Another interesting example of polarization behavior in multiphoton processes can be seen in bound–bound transitions of molecules.

2. Polarization Behavior of Nonrotating Molecules

Let us consider a nonresonant two-photon absorption of randomly oriented nonrotating molecules excited by two lasers with polarization vectors $\mathbf{e}_1, \mathbf{e}_2$, that is, $(\mathbf{e}_x, \mathbf{e}_y)$ or $(\mathbf{e}_+, \mathbf{e}_-)$, and angular frequencies ω_1 and ω_2. The two-photon transition probability from i to f states, $W^{(2)}_{i\to f}$, is given by the following:

$$W^{(2)}_{i\to f} = 2\pi(2\pi\alpha)^2 I_1 I_2 \omega_1 \omega_2 |S_{fi}(\omega_1 \mathbf{e}_1, \omega_2 \mathbf{e}_2)|^2 \times \delta(\omega_{fi} - \omega_1 - \omega_2), \tag{57}$$

where $S_{fi}(\omega_1\mathbf{e}_1,\ \omega_2\mathbf{e}_2)$, the two-photon transition amplitude, takes the following form:

$$S_{fi}(\omega_1\mathbf{e}_1,\ \omega_2\mathbf{e}_2) = \sum_m \left[\frac{\mathbf{e}_2 \cdot \mathbf{R}_{fm}\mathbf{e}_1 \cdot \mathbf{R}_{mi}}{\omega_{mi} - \omega_1} + \frac{\mathbf{e}_1 \cdot \mathbf{R}_{fm}\mathbf{e}_2 \cdot \mathbf{R}_{mi}}{\omega_{mi} - \omega_2} \right]. \quad (58)$$

The polarization vectors are usually expressed in a laboratory coordinate system, while the electronic transition moment operator $\mathbf{r}$ is expressed in the molecular system, and therefore a coordinate transformation has to be taken into account in order to evaluate the two-photon transition amplitude. This can be carried out by introducing the Euler angles, which specify the molecular coordinate system with respect to the laboratory one. After the transformation is accomplished, the two-photon transition probability average over all the molecular orientations, $\langle W^{(2)}_{i\to f}\rangle$, can be written as follows:

$$\langle W^{(2)}_{i\to f}\rangle = \delta_F F + \delta_G G + \delta_H H, \quad (59)$$

where F, G, and H, which are experimentally controllable polarization variables, are given by the following:

$$F = 4|\mathbf{e}_1 \cdot \mathbf{e}_2|^2 - 1 - |\mathbf{e}_1 \cdot \mathbf{e}_2^*| \quad (60a)$$

$$G = -|\mathbf{e}_1 \cdot \mathbf{e}_2|^2 + 4 - |\mathbf{e}_1 \cdot \mathbf{e}_2^*| \quad (60b)$$

$$H = -|\mathbf{e}_1 \cdot \mathbf{e}_2|^2 - 1 + 4|\mathbf{e}_1 \cdot \mathbf{e}_2^*|^2 \quad (60c)$$

In Eq. (59), δ_F, δ_G, and δ_H, which are characterized by the molecular quantities, laser flux, and laser detuning, take the following forms:

$$\delta_F = \frac{1}{15}(2\pi\alpha)^2 I_1 I_2 \omega_1 \omega_2 \times \sum_a \sum_b S^{aa}_{fi} S^{bb^*}_{fi} \delta(\omega_{fi} - \omega_1 - \omega_2) \quad (61a)$$

$$\delta_G = \frac{1}{15}(2\pi\alpha)^2 I_1 I_2 \omega_1 \omega_2 \times \sum_a \sum_b S^{ba}_{fi} S^{ba^*}_{fi} \delta(\omega_{fi} - \omega_1 - \omega_2) \quad (61b)$$

$$\delta_H = \frac{1}{15}(2\pi\alpha)^2 I_1 I_2 \omega_1 \omega_2 \times \sum_a \sum_b S^{ba}_{fi} S^{ab^*}_{fi} \delta(\omega_{fi} - \omega_1 - \omega_2) \quad (61c)$$

where S^{ba}_{fi}, the component of the two-photon transition tensor in the molecular coordinate system, is defined by the following:

$$S^{ba}_{fi} = \sum_m \left[\frac{\langle f|r_b|m\rangle\langle m|r_a|i\rangle}{\omega_{mi} - \omega_2} + \frac{\langle f|r_a|m\rangle\langle m|r_b|i\rangle}{\omega_{mi} - \omega_1} \right]. \quad (62)$$

It should be noted that since Eq. (61a) is expressed in terms of the absolute square of the trace of the two-photon transition tensor, $\delta_F \neq 0$ is satisfied only for the case of a transition to a totally symmetric state when the initial state is of a totally symmetric.

Depending on the combination of polarizations used, some experimental cases can be considered: (1) two linearly polarized photons with parallel polarization; (2) two linearly polarized photons with perpendicular polarization; (3) one linear and one circular with linear polarization perpendicular to the plane of the circular polarization; (4) both circular, in either the same or opposite sense, with perpendicular propagation; (5) both linear, with $\theta = 45^\circ$ between the two polarization vectors; (6) one linear and one circular, with linear polarization in the plane of the circular polarization; (7) both circular in the same sense, with parallel propagation; and (8) both circular in the opposite sense, with parallel propagation.

In Table I, the values of F, G, and H that correspond to these cases are presented.

As a simple example of the polarization dependence, let us consider two cases of a two-photon transition from a totally symmetric ground state to a nontotally symmetric state: in one case the transition is excited by the laser beam with two linearly polarized photons [case (1), in which the transition probability is denoted by $\langle W^{(2)}_{i\to f}\rangle^{\uparrow\uparrow}$] and in the other case the transition is induced by lasers of two circularly polarized photons with parallel propagation [case (7), in which the transition probability is denoted by $\langle W^{(2)}_{i\to f}\rangle^{\circlearrowright\circlearrowright}$]. From Table I in this case, the ratio is given by the following:

$$\langle W^{(2)}_{i\to f}\rangle^{\circlearrowright\circlearrowright} / \langle W^{(2)}_{i\to f}\rangle^{\uparrow\uparrow} = \tfrac{3}{2}.$$

When one assigns a two-photon absorption of molecules, it is important to know the tensor patterns, which depend only on the symmetry of the molecular states relevant to the transition. The tensor patterns are tabulated in Table II. Here the initial state is assumed to belong to the totally symmetric representation A. The tabulated quantity is as follows:

TABLE I Values of the Polarization Variables *F*, *G*, and *H* for Eight Two-Photon Transitions[a]

	Case							
Polarization variable	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
F	2	−1	−1	$-\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{2}$	−2	3
G	2	4	4	$\frac{7}{2}$	3	3	3	3
H	2	−1	−1	$-\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{2}$	3	−2

[a] From Monson, P. R., and McClain, W. M. (1970). *J. Chem. Phys.* **53,** 29.

TABLE II Cartesian Tensor Patterns for Two-Photon Processes[a]

1. Groups C_1 and C_i

$$A = \begin{bmatrix} s_1 & s_2 & s_3 \\ s_4 & s_5 & s_6 \\ s_7 & s_8 & s_9 \end{bmatrix}$$

2. Groups C_2, C_3, and C_{2h}

$$A = \begin{bmatrix} s_1 & s_4 & 0 \\ s_5 & s_2 & 0 \\ 0 & 0 & s_3 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 & s_6 \\ 0 & 0 & s_7 \\ s_8 & s_9 & 0 \end{bmatrix}$$

3. Groups C_{2v}, D_2, and D_{2h}

$$A_1 = A = \begin{bmatrix} s_1 & 0 & 0 \\ 0 & s_2 & 0 \\ 0 & 0 & s_3 \end{bmatrix}, \quad A_2 = B_1 = \begin{bmatrix} 0 & s_4 & 0 \\ s_5 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$B_1(C_{2v}) = B_2 = \begin{bmatrix} 0 & 0 & s_6 \\ 0 & 0 & 0 \\ s_7 & 0 & 0 \end{bmatrix}, \quad B_2(C_{2v}) = B_3 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & s_8 \\ 0 & s_9 & 0 \end{bmatrix}$$

4. Groups C_4, C_{4h}, and S_4

$$A = \begin{bmatrix} s_1 & s_3 & 0 \\ -s_3 & s_1 & 0 \\ 0 & 0 & s_2 \end{bmatrix}, \quad B = \begin{bmatrix} s_4 & s_3 & 0 \\ s_5 & -s_4 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad E = \begin{bmatrix} 0 & 0 & s_6 \\ 0 & 0 & -is_6 \\ s_7 & -is_7 & 0 \end{bmatrix}, \quad \text{and} \quad \begin{bmatrix} 0 & 0 & s_6^* \\ 0 & 0 & is_6^* \\ s_7^* & is_7^* & 0 \end{bmatrix}$$

5. Groups C_{4v}, D_4, D_{2d}, and D_{4h}

$$A_1 = \begin{bmatrix} s_1 & 0 & 0 \\ 0 & s_1 & 0 \\ 0 & 0 & s_2 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 0 & s_3 & 0 \\ -s_3 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad B_1 = \begin{bmatrix} s_4 & 0 & 0 \\ 0 & -s_4 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad B_2 = \begin{bmatrix} 0 & s_5 & 0 \\ s_5 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

$$E = \begin{bmatrix} 0 & 0 & s_6 \\ 0 & 0 & -is_6 \\ s_7 & -is_7 & 0 \end{bmatrix}, \quad \text{and} \quad \begin{bmatrix} 0 & 0 & s_6^* \\ 0 & 0 & is_6^* \\ s_7^* & is_7^* & 0 \end{bmatrix}$$

6. Groups C_3 and $S_6 = C_{3h}$

$$A = \begin{bmatrix} s_1 & s_3 & 0 \\ -s_3 & s_1 & 0 \\ 0 & 0 & s_2 \end{bmatrix}, \quad E = \begin{bmatrix} s_4 & is_4 & s_5 \\ is_4 & -s_4 & -is_5 \\ s_6 & -is_6 & 0 \end{bmatrix}, \quad \text{and} \quad \begin{bmatrix} s_4^* & -is_4^* & s_5^* \\ -is_4^* & -s_4^* & is_5^* \\ s_6^* & is_6^* & 0 \end{bmatrix}$$

7. Groups C_{3v}, D_3, and D_{3d}

$$A_1 = \begin{bmatrix} s_1 & 0 & 0 \\ 0 & s_1 & 0 \\ 0 & 0 & s_2 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 0 & s_3 & 0 \\ -s_3 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad E = \begin{bmatrix} s_4 & is_4 & s_5 \\ is_4 & -s_4 & -is_3 \\ s_6 & -is_6 & 0 \end{bmatrix}, \quad \text{and} \quad \begin{bmatrix} s_4^* & -is_4^* & s_5^* \\ -is_4^* & -s_4^* & is_5^* \\ s_6^* & is_6^* & 0 \end{bmatrix}$$

8. Groups C_{3h}, C_6, and C_{6h}

$$A = \begin{bmatrix} s_1 & s_3 & 0 \\ -s_3 & s_1 & 0 \\ 0 & 0 & s_2 \end{bmatrix}, \quad E_1 = \begin{bmatrix} 0 & 0 & s_4 \\ 0 & 0 & -is_4 \\ s_5 & -is_5 & 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 0 & 0 & s_4^* \\ 0 & 0 & is_4^* \\ s_5^* & is_5^* & 0 \end{bmatrix}, \quad E_2 = \begin{bmatrix} s_6 & -is_6 & 0 \\ -is_6 & -s_6 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} s_6^* & is_6^* & 0 \\ is_6^* & -s_6^* & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

9. Groups C_{6v}, D_{3h}, D_6, and D_{6h}; groups $C_{\infty v}$ and $D_{\infty v}$

$$A_1 = \Sigma^+ = \begin{bmatrix} s_1 & 0 & 0 \\ 0 & s_1 & 0 \\ 0 & 0 & s_2 \end{bmatrix}, \quad A_2 = \Sigma^- = \begin{bmatrix} 0 & s_3 & 0 \\ -s_3 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad E_1 = \Pi = \begin{bmatrix} 0 & 0 & s_4 \\ 0 & 0 & is_4 \\ s_5 & is_5 & 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 0 & 0 & s_4^* \\ 0 & 0 & -is_4^* \\ s_5^* & -is_5^* & 0 \end{bmatrix}$$

$$E_2 = \Delta = \begin{bmatrix} s_6 & -is_6 & 0 \\ -is_6 & -s_6 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} s_6^* & is_6^* & 0 \\ is_6^* & -s_6^* & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

continues

TABLE II (*Continued*)

10. Groups T and T_h, $\omega = \exp(2\pi i/3)$

$$A = \begin{bmatrix} s_1 & 0 & 0 \\ 0 & s_1 & 0 \\ 0 & 0 & s_1 \end{bmatrix}, \quad E = \begin{bmatrix} s_2 & 0 & 0 \\ 0 & \omega s_2 & 0 \\ 0 & 0 & \omega^* s_2 \end{bmatrix} \text{ and } \begin{bmatrix} s_2^* & 0 & 0 \\ 0 & \omega^* s_2^* & 0 \\ 0 & 0 & \omega s_2^* \end{bmatrix}$$

$$T = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & s_3 \\ 0 & s_4 & 0 \end{bmatrix} \text{ and } \begin{bmatrix} 0 & 0 & s_4 \\ 0 & 0 & 0 \\ s_3 & 0 & 0 \end{bmatrix} \text{ and } \begin{bmatrix} 0 & s_3 & 0 \\ s_4 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

11. Groups O, O_h, and T_d, $\omega = \exp(2\pi i/3)$

$$A = \begin{bmatrix} s_1 & 0 & 0 \\ 0 & s_1 & 0 \\ 0 & 0 & s_1 \end{bmatrix}, \quad E = \begin{bmatrix} s_2 & 0 & 0 \\ 0 & \omega s_2 & 0 \\ 0 & 0 & \omega^* s_2 \end{bmatrix} \text{ and } \begin{bmatrix} s_2^* & 0 & 0 \\ 0 & \omega^* s_2^* & 0 \\ 0 & 0 & \omega s_2^* \end{bmatrix}$$

$$T_1 = \begin{bmatrix} 0 & s_2 & 0 \\ -s_2 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \text{ and } \begin{bmatrix} 0 & 0 & -s_3 \\ 0 & 0 & 0 \\ s_3 & 0 & 0 \end{bmatrix} \text{ and } \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & s_3 \\ 0 & -s_3 & 0 \end{bmatrix}$$

$$T_2 = \begin{bmatrix} 0 & s_4 & 0 \\ s_4 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \text{ and } \begin{bmatrix} 0 & 0 & s_4 \\ 0 & 0 & 0 \\ s_4 & 0 & 0 \end{bmatrix} \text{ and } \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & s_4 \\ 0 & s_4 & 0 \end{bmatrix}$$

[a] The tabulated quantity is $S^{ba}(A \rightarrow J)$. The table is divided into 11 sets of groups. Within each set, the groups either are isomorphic or differ from each other only by the inclusion of the center of inversion element. The tensors are labeled by the symbol J only, and that symbol is often simplified by dropping primes and subscripts, due to the variability of nomenclature among different groups within the same isomorphic set. The word "and" is written between tensors that belong to the different parts of a degenerate transition; such pairs must always be used together, for example, for a twofold degeneracy, according to $W^{(2)} = |\mathbf{e}_1 \cdot \mathbf{S}^1 \cdot \mathbf{e}_2|^2 + |\mathbf{e}_1 \cdot \mathbf{S}^{11} \cdot \mathbf{e}_2|^2$. The basis sets for symmetry species A and B are always unambiguous. When the group has one E species, the basis set is $[x + iy, x - iy]$, except in the tetrahedral and octahedral groups (sets 10 and 11), where the basis is $[u + iv, u - iv]$, with $u = 2z^2 - x^2 - y^2$ and $v = 3^{1/2}(x^2 - y^2)$. When the group has two E species, the basis of E_1 is $[x + iy, x - iy]$ and the basis of E_2 is $[(x + iy)^2, (x - iy)^2]$. In groups T and T_h (set 10), the basis of species T is (x, y, z). In groups O, O_h, and T_d, the basis of T_1 is (x, y, z) and the basis of T_2 is (yz, zx, xy). In sets 10 and 11, note that $1 + \omega + \omega^* = 0$. [From McClain, W. M., and Harris, R. A. (1977). *Excited States* **3,** 2.]

$$S_{fi}^{ba} = \langle f^J |^A S^{ba} + {}^B S^{ba} + \cdots | i^A \rangle,$$

where $A, B, \ldots, J$ represent the names of the symmetry species (A, B, E, T, etc.).

3. Polarization Behavior of Rotating Molecules

Rotational structures in multiphoton spectra of molecules in gases are well resolved by using a narrow-band tunable dye laser. In these experiments, the different rotational branches of the same band of two-photon excitation spectra of gaseous molecules differ in their polarization behavior. In this subsection, the polarization behavior seen in rotational contour in a nonresonant two-photon absorption of a rotating symmetric molecule is treated. The initial and final rovibronic states are respectively specified as $|i\rangle = |i, J_i, K_i, M_i\rangle$ and $|f\rangle = |f, J_f, K_f, M_f\rangle$, where i and f in the right-hand side denote the vibronic states of the initial and final states, respectively; and J, K, and M refer to the total angular momentum, the component of J along the molecular fixed z axis, and that of J along the Z axis of the laboratory coordinates.

In the absence of a magnetic field, each JK level is $(2J + 1)$-fold degenerated. After summations over M_i and M_f, the transition probability is as follows:

$$\begin{aligned} W^{(2)}_{J_i K_i \rightarrow J_f K_f} &= 2\pi (2\pi\alpha)^2 I_1 I_2 \omega_1 \omega_2 \frac{1}{2J_i + 1} \\ &\times \sum_{M_i} \sum_{M_f} |\langle J_f K_f M_f| \\ &\times \hat{S}_{fi}(\omega_1 \mathbf{e}_1, \omega_2 \mathbf{e}_2) | J_i K_i M_i \rangle|^2 \\ &\times \delta(\omega_{fi} - \omega_1 - \omega_2), \end{aligned} \tag{63}$$

where the rotational quantum number dependence of the two-photon transition operator, $\hat{S}_{fi}(\omega_1 \mathbf{e}_1, \omega_2 \mathbf{e}_2)$, has been omitted as follows:

$$\begin{aligned} \hat{S}_{fi}(\omega_1 \mathbf{e}_1, \omega_2 \mathbf{e}_2) = \sum_m \Bigg[&\frac{\mathbf{e}_2 \cdot \mathbf{r} | m \rangle \langle m | \mathbf{e}_1 \cdot \mathbf{r}}{\omega_{mi} - \omega_1} \\ &+ \frac{\mathbf{e}_1 \cdot \mathbf{r} | m \rangle \langle m | \mathbf{e}_2 \cdot \mathbf{r}}{\omega_{mi} - \omega_2} \Bigg], \end{aligned} \tag{64}$$

in which m denotes the vibronic states of the intermediate states. This approximation is valid for nonresonant

transitions. After performing the transformation between the laboratory and molecular coordinates in the two-photon transition amplitude by using the spherical-coordinate basis set and evaluating the matrix element, the two-photon transition probability can be written in the product form of the geometrical factor C, the molecular factor M, and the rotational factor R as follows:

$$W^{(2)}_{J_iK_i\to J_fK_f} = \sum_{J=0}^{2} C_J M_J R_J, \tag{65}$$

where the geometrical factors C_J, which are functions only of the polarization vectors $\mathbf{e}_1$ and $\mathbf{e}_2$, are given by the following:

$$\begin{aligned} C_0 &= |\mathbf{e}_1 \cdot \mathbf{e}_2|^2/3 \\ C_1 &= |\mathbf{e}_1 \times \mathbf{e}_2|^2/6 \\ C_2 &= (1 - C_0 - 3C_1)/5. \end{aligned} \tag{66}$$

The molecular factors M_J are written as follows:

$$\begin{aligned} M_J &= \left|M^{(J)}_{\Delta K}\right|^2 \\ &= (2J+1)\left|\sum_a \sum_b \begin{pmatrix} 1 & 1 & J \\ -a & -b & \Delta K \end{pmatrix} S^{ba}_{fi}\right|^2 \end{aligned} \tag{67}$$

with $|\Delta K| \le J$. The matrix ( ) is called the Wigner $3j$ symbols. Equation (67) can be evaluated with the aid of the symmetry property of the Wigner $3j$ symbols. All the possible values of the molecular factors are tabulated in Table III.

TABLE III Molecular Factors M_J of the General Expression in Eq. (67) for Two-Photon Absorption in Rotating Molecules in Spherical and Cartesian Coordinates[a]

		Molecular factors $M_J = \|M^{(J)}_{\Delta K}\|^2$ with $\|\Delta K\| \leqq J$	
J	ΔK	Spherical coordinates	Cartesian coordinates
2	± 2	$\|M_{\pm\pm}\|^2$	$\frac{1}{4}(M_{xx} - M_{yy})^2 + \frac{1}{4}(M_{xy} + M_{yx})^2$
	± 1	$\frac{1}{2}\|M_{\pm 0} + M_{0\pm}\|^2$	$\frac{1}{4}(M_{xz} + M_{zx})^2 + \frac{1}{4}(M_{yz} + M_{zy})^2$
	0	$\frac{1}{6}\|M_{+-} + M_{-+} + 2M_{00}\|^2$	$\frac{1}{6}(2M_{zz} - M_{xx} - M_{yy})^2$
1	± 1	$\frac{1}{2}\|M_{\pm 0} - M_{0\pm}\|^2$	$\frac{1}{4}(M_{xz} - M_{zx})^2 + \frac{1}{4}(M_{yz} - M_{zy})^2$
	0	$\frac{1}{2}\|M_{+-} - M_{-+}\|^2$	$\frac{1}{2}(M_{xy} - M_{yx})^2$
0	0	$\frac{1}{3}\|M_{+-} + M_{-+} + M_{00}\|^2$	$\frac{1}{3}(M_{xx} + M_{yy} + M_{zz})^2$

[a] Molecular factors are listed for different ΔK and given in the molecular frame system. Spherical coordinates r_+, r_0, and r_- correspond to $-2^{-1/2}(x+iy)$, z, and $2^{-1/2}(x-iy)$, respectively. Matrix elements M_{ba} correspond to S^{ba}_{fi} in Eq. (67). [From Metz, F., Howard, W. E., Wunsch, L., Neusser, H. J., and Schlag, E. W. (1978). *Proc. R. Soc. London Ser. A.* **363,** 381.]

The rotational factors R_j are defined as follows:

$$R_J = (2J_f + 1)(2J_i + 1)\begin{pmatrix} J_i & J_f & J \\ -K_i & K_f & \Delta K \end{pmatrix}^2 \tag{68}$$

From this expression the rotational selection rule can be obtained as follows:

$$\begin{aligned} R_0 \neq 0 \quad &\text{for} \quad \Delta J = 0, \\ &\quad\ \Delta K = 0; \quad \text{Q branch only} \\ R_1 \neq 0 \quad &\text{for} \quad \Delta J = 0, \pm 1, \\ &\quad\ \Delta K = 0, \pm 1; \quad \text{P, Q, R branches} \\ R_2 \neq 0 \quad &\text{for} \quad \Delta J = 0, \pm 1, \pm 2, \\ &\quad\ \Delta K = 0, \pm 1, \pm 2; \quad \text{O, P, Q, R,} \\ &\qquad\qquad\qquad\qquad\ \text{S branches.} \end{aligned}$$

As an application of the theory described above, let us consider the ratio of the linearly to the circularly polarized two-photon absorption probability. Noting that the geometrical factors for the linearly and circularly polarized laser beams are given by $C_0 = \frac{1}{3}$, $C_1 = 0$, and $C_3 = \frac{2}{15}$ and $C_0 = 0$, $C_2 = 0$, and $C_2 = \frac{1}{5}$, respectively; the ratio can be expressed as follows:

$$\frac{W^{(2)\uparrow\uparrow}_{J_iK_i\to J_fK_f}}{W^{(2)\circlearrowright\circlearrowright}_{J_iK_i\to J_fK_f}} = \frac{2}{3} + \frac{5M_0R_0}{3M_2R_2}. \tag{69}$$

This equation indicates that for nontotally symmetric transitions characterized by $M_0 = 0$, the ratio is independent of the rotational quantum numbers J and K, and is given by $\frac{2}{3}$. The same behavior of the ratio is also expected for the rotational lines of branches except Q branch of a totally symmetric transition, because $M_0 \neq 0$ and $R_0 = 0$. The ratio of the Q branch in which $M_0 \neq 0$ and $R_0 \neq 0$ deviate from $\frac{2}{3}$, and the magnitude of the deviation depends strongly on M_0R_0/M_2R_2, in which usually $M_0 > M_2$ and $R_0 > R_2$, and then on the rotational quantum numbers.

C. Spectral Properties of Multiphoton Transitions

Spectral properties of the multiphoton transitions of atoms and molecules depend strongly on whether the resonance condition is satisfied. In this subsection, resonance effects in atomic multiphoton transitions are first presented, focusing on the role of the intermediate states, and then the appearance of the vibronic structures in molecular multiphoton transitions is presented, focusing on the difference

in the spectral intensity distribution between nonresonant and resonant multiphoton transitions.

1. Resonance Effects in Atomic Multiphoton Transitions

A drastic intensity enhancement in the multiphoton transitions can be observed when the energy of the photon is close to that of an intermediate state. An example of the resonance enhancement is shown in Fig. 18, in which two-photon transition rates of ^{23}Na for the 3S (hyperfine level $F = 2$, see Section V.A) $\rightarrow 4D_{5/2}$ and 3S ($F = 2$) $\rightarrow 4D_{3/2}$ transitions are recorded as a function of the wavelength of the second laser λ_2 by detecting the fluorescence from the excited D state. The frequency of the first laser is adjusted to make it close to the 3S–3P doublet. As the frequency approaches to $3S–3P_{3/2}$ absorption line, the cross sections for the two-photon transitions to the $4D_{5/2}$ and $4D_{3/2}$ levels are enhanced by a factor of 10^8. The deep minimum, known as the Fano profile, originates from an interference between $3P_{3/2} \rightarrow 4D_{1/2}$ and $3P_{3/2} \rightarrow 4D_{5/2}$ transitions, and on the other hand, since $3P_{1/2} \rightarrow 4D_{5/2}$, there is no such interference in the two-photon transition via the $3P_{1/2}$ state. Generally, resonance peak positions depend on the laser intensity in a moderate laser intensity range, as shown in Fig. 19. Here variations of the number of ions in four-photon ionization of Cs atoms are drawn as a function of laser frequency in the neighborhood of the resonant three-photon transition 6S $\rightarrow$ 6F. The dashed line shows the resonance shift linear with respect to the laser intensity I. The origin of the resonance shift was explained in Section II.C.

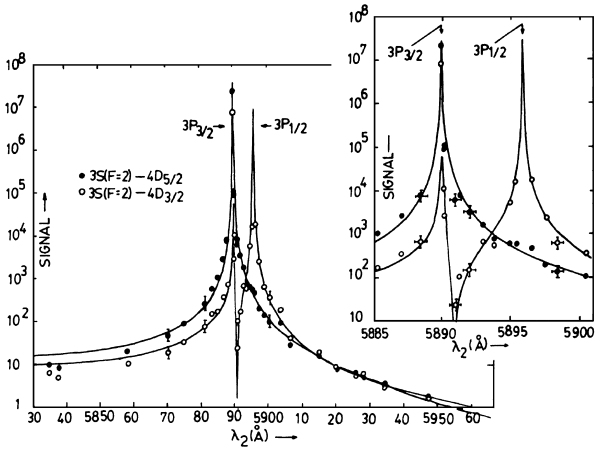


FIGURE 18 Resonant enhancement of the two-photon absorption rate of Na; $\hbar(\omega_1 + \omega_2)$ is fixed at the $3S(F=2)–4D_{3/2}$ or $3S(F=2)–4D_{5/2}$ transition, while $\hbar\omega_1$ is tuned through the (one-photon) yellow doublet. The points are experimental and the curves are theoretical. The insert shows the behavior in the region from 5885 to 5900 Å with an expanded horizontal axis. [From Bjorkholm, J. E., and Liao, P. F. (1974). *Phys. Rev. Lett.* **33,** 128.]

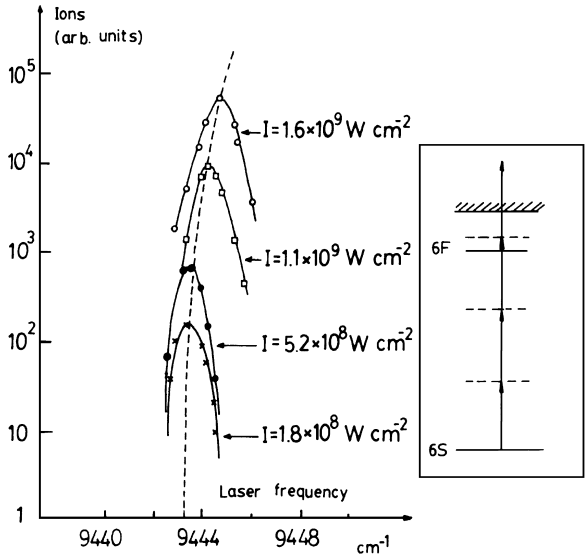


FIGURE 19 Variation of the number of ions in the four-photon ionization of Cs as a function of laser frequency in the neighborhood of the resonant three-photon transition 6S $\rightarrow$ 6F. Dashed line shows the resonance shift for increasing values of laser intensity *I*. [From Mainfray, G., and Manus, C. (1980). *Appl. Opt.* **19,** 3934.]

2. Vibronic Coupling in Molecular Two-Photon Transitions

From the symmetry argument in a previous section it can be understood that two-photon absorptions between the g and g or u and u electronic states of molecules with inversion symmetry take place with high intensity. However, two-photon transitions between electronic states with different inversion symmetry $u \leftrightarrow g$ are also observed, although their intensity is very weak. The $u \leftrightarrow g$ transition, called the forbidden two-photon transition or vibronically induced two-photon transition, can be allowed by coupling of nuclear vibrations of u inversion symmetry with electrons (vibronic coupling). As an example of the vibronically induced two-photon transition, one can see the two-photon absorption from the ground state $^1A_{1g}$ to the first excited singlet state $^1B_{2u}$ in benzene. For the two-photon transition of benzene, using two photons with identical frequency, the tensor pattern belongs to the A_{1g}, E_{1g}, and E_{2g} irreducible representations of the D_{6h} point group. The species of vibrations inducing the two-photon absorption ($^1B_{2u} \leftarrow {}^1A_{1g}$), Γ_j can be specified from the symmetry consideration, $\Gamma_j \times B_{2u} \times A_{1g} = \{A_{1g}, E_{1g}, E_{2g}\}$; that is, the inducing modes belonging to the b_{2u}, e_{2u}, and e_{1u} species. In Fig. 20, the

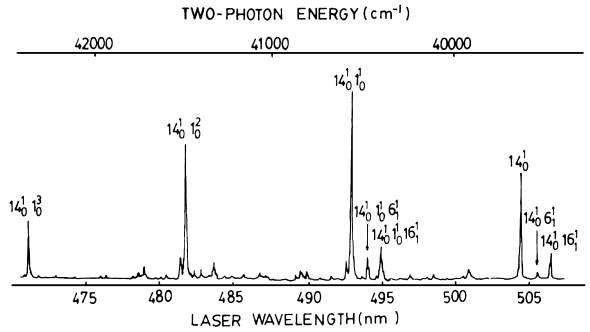


FIGURE 20 Normalized multiphoton ionization spectrum of benzene resonant with S_1. [From Murakami, J., Kaya, K., and Ito, M. (1980). *J. Chem. Phys.* **72,** 3263.]

normalized four-photon ionization spectrum of benzene two-photon resonant with the S_1 state is shown. The vibronic structure consists mainly of the v_1 totally symmetric mode progression that starts from the vibronically induced 14_0^1 band. The inducing mode v_{14} is the C—C bond-alternating vibration belonging to the b_{2u} species. The two-photon fluorescence excitation spectrum of the $^1B_{2u} \leftarrow {}^1A_{1g}$ transition of benzene shows very similar vibronic structure.

In order to predict which mode acts as the inducing mode and to evaluate the two-photon transition probability, magnitudes of the transition amplitude have to be calculated. This can be carried out first by expanding the transition moments to the first order of Q_j in the Born–Oppenheimer approximation, for example, as follows:

$$\begin{aligned}\langle m|\mathbf{r}|i\rangle \simeq\ & \langle \Theta_{mv}(Q)|\Theta_{iv}(Q)\rangle \\ & \times \langle \Phi_m(r, Q)|\Phi_i(r, Q)\rangle|_{Q=Q_0} \\ & + \sum_j \langle \Theta_{mv}(Q)|Q_j|\Theta_{iv}(Q)\rangle \\ & \times \frac{\partial}{\partial Q_j}\langle \Phi_m(r, Q)|\mathbf{r}|\Phi_i(r, Q)\rangle|_{Q=Q_0}, \qquad (70)\end{aligned}$$

where $\Phi(r, Q)$ and $\Theta(Q)$ denote the electronic and vibrational wave functions, respectively, and $\langle \Theta_{mv}(Q)|\Theta_{iv}(Q)\rangle$ is the optical Franck–Condon overlap integral. The vibronically induced transition originates from the second term of Eq. (70). The term Q_j denotes the inducing modes. Summation over the intermediate states in the transition amplitude is then performed neglecting the vibrational quantum number dependence in the energy denominator of the transition amplitude in the case of nonresonant two-photon transitions. In resonant cases, on the other hand, the vibrational quantum number dependence has to be evaluated explicitly. Several methods—the Green's function method, numerical method, path integral method, and so on—can be applied to the evaluation of the resonant two-photon transition probability of molecules.

3. Evolution from Multiphoton Ionization to Tunnel Ionization for Atoms and Molecules

In intense laser fields, there are two types of the ionization processes from atoms: multiphoton and tunnel processes, as schematically shown in Fig. 21. For example, for atoms induced by intense, near-infrared laser pulses, such as Ti:sapphire laser pulses, tunnel ionization competes

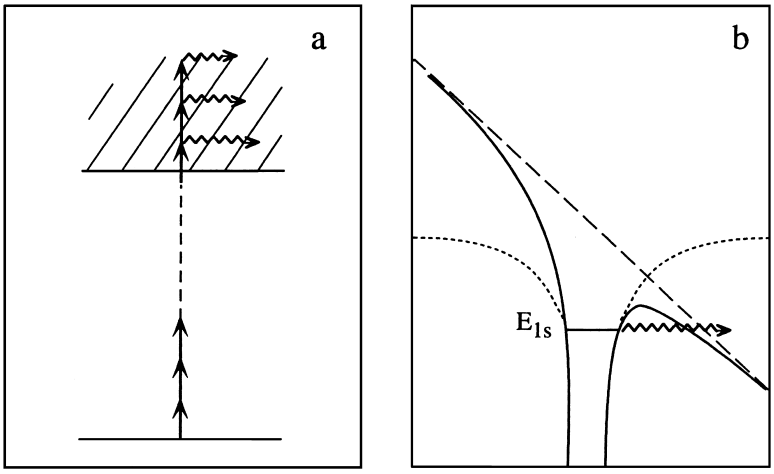


FIGURE 21 Ionization processes induced by intense laser fields: (a) The multiphoton ionization process; (b) the tunnel ionization process is indicated by an arrow. Bold curves show instantaneous, effective, electronic potential for hydrogen atom in an intense laser fields. Dotted curves show the bare electronic potential. A broken line shows the interaction potential between the electron and the laser field. A tunnel ionization from the lowest electronic energy level denoted by E_{1s} is indicated by an arrow.

with multiphoton ionization. In general, these two processes can be identified by observing the band width in photoelectron spectra. Broad bands are due to the tunneling ionization, and sharp bands repeated with the photon energy period are due to the multiphoton process or the above-threshold ionization process, as shown in Fig. 22.

The adiabaticity or Keldysh parameter γ is used to classify these processes. This is expressed as the ratio of the tunneling time to the period of the frequency of the laser as follows:

$$\gamma = \omega_L/\omega_T,$$

where ω_T is the tunneling frequency and ω_L is the frequency of the laser. For $\gamma \gg 1$, the multiphoton mechanism makes the main contribution. This is the case for ionization by visible or UV lasers. On the other hand, for $\gamma \ll 1$, the tunnel mechanism dominates. The adiabaticity parameter is expressed in terms of the ionization potential I_p and pondermotive energy of a free electron in a laser field of amplitude E as follows:

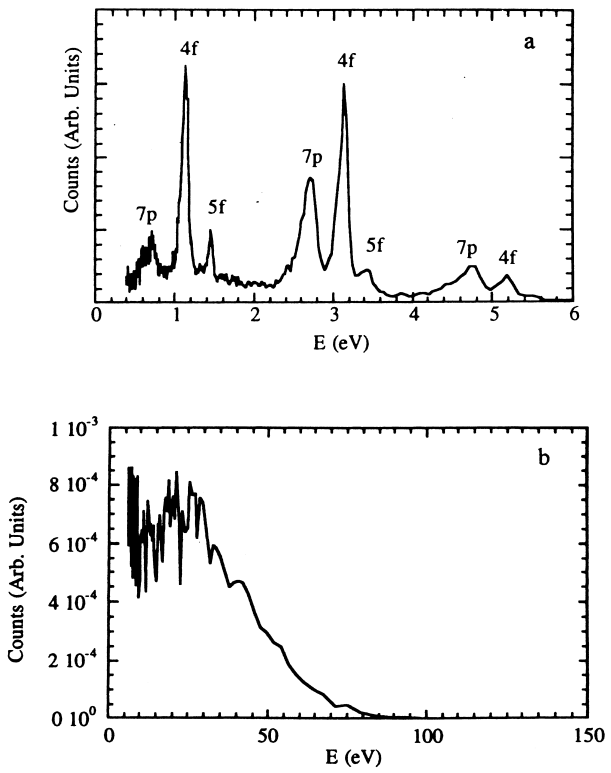


FIGURE 22 (a) Photoelectron spectrum from ionization of xenon by 617-nm, 100-fsec laser pulses with an intensity of 6.2×10^{13} W/cm^2. This is a typical spectrum of atoms induced by multiphoton ionization (above-threshold ionization). (b) Photoelectron spectrum from ionization of helium by the same laser pulses with an intensity of 1.5×10^{15} W/cm^2. This is a typical spectrum from atoms induced by tunnel ionization. [From Mevel, E., Breger, P., Trainham, R., Petite, G., Agostini, P., Migus, A., Chambaret, J-P., and Antonetti, A. (1993). *Phys. Rev. Lett.* **70,** 406.]

$$U_p = e^2E^2/(4m\omega_L) \quad \text{as} \quad \gamma = \sqrt{I_p/(2U_p)}.$$

Figure 22 shows typical examples of photoelectron spectra due to multiphoton ionization and tunnel ionization of atoms. The photoelectron spectrum from multiphoton ionization of xenon by 617-nm, 100-fsec laser pulses of intensity 6.2×10^{13} W/cm^2 is shown in Fig. 22a, in which a structure repeated with the photon energy period can be clearly seen. This is a feature of multiphoton ionization, that is, above-threshold ionization. Tunnel ionization from helium at 1.5×10^{15} W/cm^2 is shown in Fig. 22b, in which there is no structure above 30 eV. The chaotic spiking seen below is not reproducible and shows no relation with the photon energy. This is a typical spectrum from atoms induced by tunnel ionization.

As in the case of atoms, tunnel ionization competes with multiphoton ionization or above-threshold ionization in the case of molecules at high intensities. The photoelectron spectra for benzene, naphthalene, and anthracene ionized by using a 780-nm, 170-fsec laser pulse with an intensity of 3.8×10^{13} W/cm^2 are shown in Figs. 23a–23c, respectively. The spectral features in these figures reveal a decrease in the discrete features associated with above-threshold ionization and dominant role of the tunnel ionization as the molecular size increases. The intensity of laser pulses used in these experiments corresponds to a regime of multiphoton ionization in the Keldysh-type model. In Fig. 23a, there is a series of sharp bands from 0.3 to 0.16 eV, which is attributable to ionization from the e_{1g} molecular orbital. The structure with photon energy of 1.5 eV, observed from 2.5 to 15 eV, reveals the feature of above-threshold ionization. A broad electron distribution ranging from 0 to 15 eV, which is due to a tunnel ionization process, is observed. In Fig. 23b, above-threshold ionization features are observed, but a large contribution of tunnel ionization is overlapped. In Fig. 23c, a broad band structure due to tunnel ionization is observed from 0 to 15 eV. A semiquantitative argument indicates that the evolution from the multiphoton to tunneling ionization from the molecules investigated is due to an increase in the electronic delocalization.

V. APPLICATIONS

In this section, some of the typical applications of multiphoton spectroscopic methods to atoms and molecules are presented.

A. Doppler-Free Multiphoton Transitions in Atoms and Molecules

In Doppler-free multiphoton spectroscopy, one can observe many interesting phenomena that cannot be

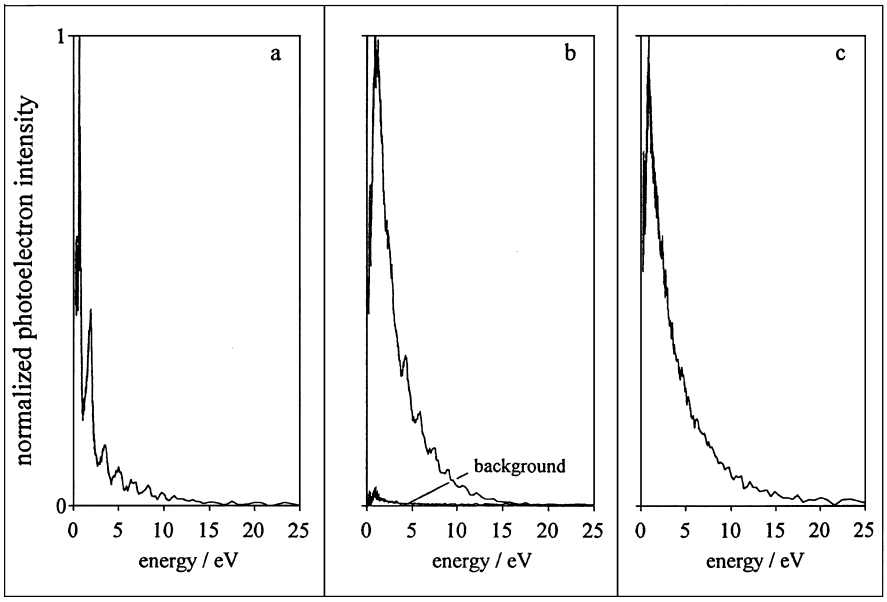


FIGURE 23 Photoelectron spectra measured using 3.8×10^{13} W/cm^2, 780-nm, 170-fsec duration laser pulses for molecules (a) benzene (C_6H_6), (b) naphthalene ($C_{10}H_8$) and (c) anthracene ($C_{14}H_{10}$). The integrated, pressure-corrected photoelectron currents are normalized to benzene. A typical background spectrum is plotted to scale in b. [From DeWitt, M. J., and Levis, R. J. (1998). *Phys. Rev. Lett.* **81,** 5101.]

resolvable in a Doppler-limited spectroscopy, such as fine and hyperfine splittings of excited states, isotope shifts, Zeeman splittings, collision-induced broadenings and shifts, and rovibronic structures. Doppler-free two-photon absorption was first demonstrated in sodium vapor by using the fluorescence detection method. A hyperfine splitting of the 3^2S–5^2S transition at 6022 Å of ^{23}Na is shown in Fig. 24. The ^{23}Na nucleus has spin $\frac{3}{2}$, and it interacts with the spin of the unpaired electron by means of a magnetic hyperfine Hamiltonian $H = A\mathbf{I} \cdot \mathbf{S}$. The splitting of the $F = 2$ and $F = 1$ hyperfine levels of an S state is thus $2A_{nS}/h$ in frequency units. The selection rules for the transition result in two absorption lines separated by $\Delta v = (A_{3S} - A_{nS})/h$; since the ground-state hyperfine splitting is well known for sodium, the splitting of the 5S state can be measured.

In the case of the two-photon absorption of linearly polarized light traveling in the same direction, the hyperfine splitting cannot be resolved, as shown in the bottom of Fig. 24. The suppression of the Doppler background in an even-quantum absorption process can be carried out based on the angular-momentum selection rules when the initial and final state angular momenta are equal. Since the orbital angular momentum vanishes in the initial and final states of sodium $3S \rightarrow 5S$ transition, the selection rule $\Delta L = 0$, $\Delta m = 0$ applies. In the case of using circularly polarized light, the transitions can take place only when the atoms absorb one quantum with angular momentum $+1$ from one laser beam and a quantum with angular momentum -1 from the oppositely propagating beam, as shown in the upper part of Fig. 24. Absorption of two quanta from a single circularly polarized beam requires $\Delta m = \pm 2$; it is impossible to excite the atoms. On the other hand, when linearly polarized laser beams pointing in opposite directions are used, as shown in the middle part of Fig. 24, effects of the Doppler broadening still exist. This originates from the absorption of the two-quanta photon of the linearly polarized beam, $\Delta m = 0$, traveling in the same direction.

The ratio of the intensity of the $F = 2$ to that of the $F = 1$ line is 5:3 from the statistical weights of the F states in the 3S ground level. From the separation of the doublet, a value of $A_{5S}/h = 78 \pm 5$ MHz can be obtained for the hyperfine interaction constant in the 5S state. With the success of the initial experiments with sodium vapor, characteristics of alkali atom states (for example, see Fig. 18) and those of other atomic states have been clarified by using Doppler-free multiphoton spectroscopy.

Applications of Doppler-free multiphoton spectroscopy to measurement of rovibronic states of molecules such as Na_2, NO, and C_6H_6 (benzene) have been reported in detail. Especially, spectral lineshapes for rovibronic transitions in polyatomic molecules like benzene overlap due to Doppler broadening, they cannot be resolved in conventional spectroscopy, and the Doppler-free spectroscopy is necessary. Information about geometrical structures of

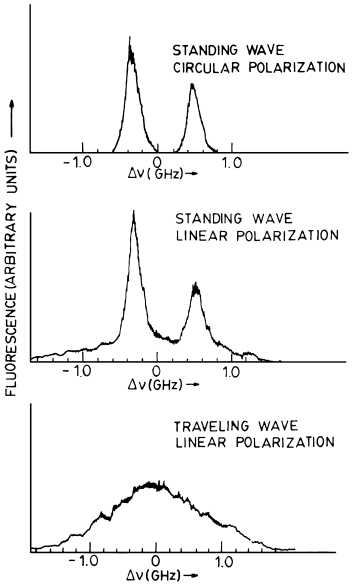


FIGURE 24 Two-photon absorption signal on the 3S → 5S transition of atomic ^{23}Na. The experimental traces record the observed resonance fluorescence intensity at 330 nm (4P–3S transition), following the two-photon absorption. [From Bloembergen, N., and Levenson, M. D. (1976). Doppler-free two-photon absorption spectroscopy. *In* "High Resolution Laser Spectroscopy" (K. Shimoda, ed.), pp. 315–369. Springer-Verlag, Berlin.]

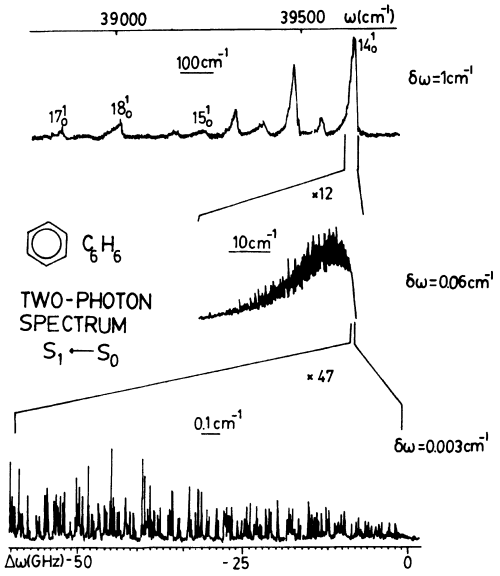


FIGURE 25 Part of the two-photon spectrum of C_6H_6 as measured with different spectral resolution. The middle trace represents the highest resolution possible in Doppler-limited spectroscopy. Only in the Doppler-free spectrum (lower part) are single rotational lines resolved. [From Lin, S. H., Fujimura, Y., Neusser, H. J., and Schlag, E. W. (1984). "Multiphoton Spectroscopy of Molecules," Academic Press, Orlando, Florida.]

electronically excited states and nonradiative processes such as intramolecular vibrational energy redistribution and electronic energy relaxation can be obtained from the line position and width only after removal of the Doppler broadening. In Fig. 25 is a part of the two-photon spectrum ($S_1 \leftarrow S_0$) of benzene as measured with different spectral resolutions shown. Many lines corresponding to the rotational transitions can be seen only in the Doppler-free spectrum shown at the bottom of Fig. 25. This spectrum is the blue edge of the Q branch ($\Delta J = 0$) of the totally symmetric transition 14_0^1 induced by the vibration υ_{14} in the $^1B_{2u}$ electronic state. For this type, two-photon absorption, the Doppler-free spectrum can be observed by using countercircularly polarized light as indicated in atomic S–S two-photon transitions.

B. Multiphoton Ionization of Molecules via Rydberg States

There have been many experimental results on the multiphoton ionization of molecules involving Rydberg states (e.g., iodine, nitric oxide, and aromatic molecules). Mechanisms of the ionization and characteristics of the Rydberg states have been clarified. In Fig. 26, the resonant $(2+1)$ and $(3+1)$ multiphoton ionization spectrum of *trans*-1,3-butadiene is shown together with the energy level diagram. The spectrum is separated into two regions, one to the blue side of 410 nm and one to the red side. In the former region, the structure of the multiphoton ionization spectrum is characteristic of an allowed two-photon resonance with the $\tilde{B}$ state designated by Herzberg: that is, the rate-determining step of the $(2+1)$ multiphoton transition is the initial two-photon transition process. The $\tilde{B}$ state with ^1B_g symmetry is formed by removal of a π-electron to an S-type Rydberg orbital. Many three-photon resonances with Rydberg states have been measured, from the $(3+1)$ four-photon ionization regions to below 410 nm. The structures of the observed spectra reflect those of the initial three-photon absorption process, which is very similar to the vacuum UV spectra because three-photon transitions have the same selection rules as one-photon transitions in a C_{2h} molecule. Quantum defects for the Rydberg series have been identified to clarify the character of the Rydberg orbitals.

C. Low-Lying Electronic Excited States (^1A_g) of Linear Polyenes

Much experimental and theoretical attention has been given to locating low-lying "hidden" electronic excited valence states (^1A_g) of linear polyenes because of their photochemical and biochemical interest. One of the fruitful applications of multiphoton spectroscopy is the direct observation of the excited

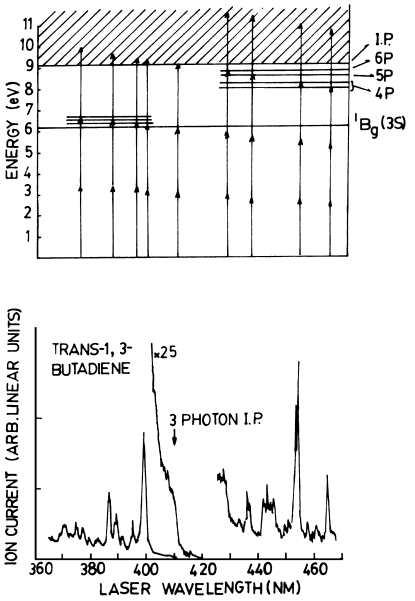


FIGURE 26 Multiphoton ionization spectrum of *trans*-1,3-butadiene, along with an energy-level diagram showing some of the states that appear in the spectrum. [From Johnson, P. M. (1980). *Acc. Chem. Res.* **13,** 20.]

states of linear polyenes, *trans*-1,3-butadiene, *trans*-1,3,5-hexatriene, *trans*, *trans*-1,3,5,7-octatetraene, and so on. The ^1A_g-^1A_g transitions that are forbidden for the one-photon process have been observed in the two-photon excitation spectra. The two-photon excitation spectrum of all-*trans*-diphenylhexatriene in ether-isopentane-ethanol (EPA) solvent at 77 K is shown in Fig. 27. The origin of the lowest excited A_g state is located at 25,050 cm^{-1}, which is at about 900 cm^{-1} below the origin of the first one-photon allowed $^1B_u \leftarrow {}^1A_g$ transition. The ordering of the electronic energy levels of linear polyenes depends,

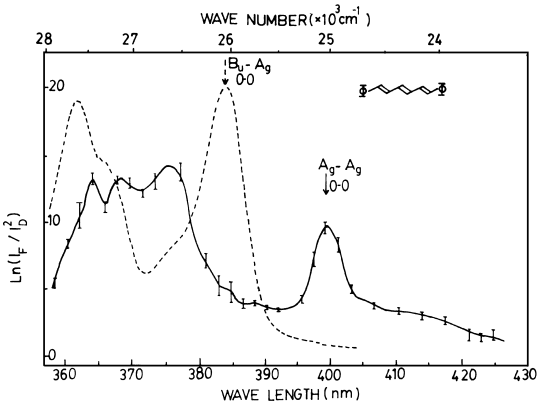


FIGURE 27 Two-photon excitation spectrum of all-*trans*-diphenylhexatriene in EPA at 77 K. The one-photon absorption is shown by the dashed curve. [From Fang, H. L. B., Thrash, R. J., and Leroi, G. E. (1978). *Chem. Phys.* **57,** 59.]

of course, on the substitution groups, as well as on experimental conditions such as the solvent used and the temperature.

D. Above-Threshold Ionization of Atoms

Multiphoton ionizations of atoms by a strong laser field can lead to the production of electrons at energies corresponding to the absorption of extra photons as well as to the absorption of the minimum number of required photons. This process involving the extra photon absorption is called the above-threshold (multiphoton) ionization or continuum–continuum transition. Developments in measuring the energy spectrum of photoelectrons makes it possible to observe the above-threshold ionization phenomenon. Since measurement of the above-threshold ionization of xenon atoms under the irradiation of a frequency-doubled Q-switched Nd:YAG (yttrium aluminum garnet) laser, experimental and theoretical studies on the mechanism of the above-threshold ionization have both become very active because by analyzing the spectra one can obtain information on the magnitude of the continuum–continuum transition probability and can also understand dynamics taking place above the ionization threshold, where the simple perturbation theory breaks down.

Figure 28 shows electron energy spectra arising from multiphoton ionization of Xe by a Nd:YAG laser of 1064 nm wavelength at several pulse energies indicated.

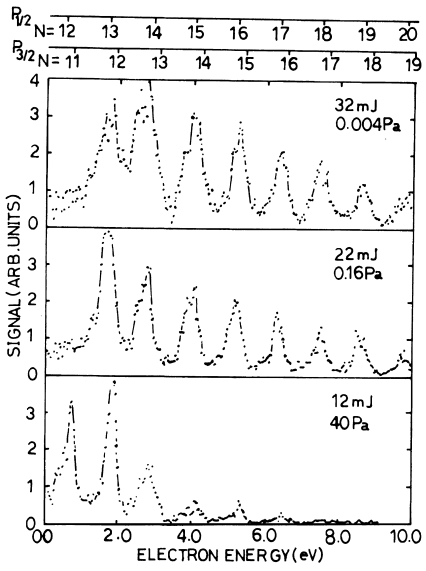


FIGURE 28 Electron spectra from multiphoton ionization of xenon at 1064 nm. The vertical scales are normalized. The pulse energy (in units of millijoules) and pressure at which each spectrum is taken are given. The estimated intensity is pulse energy (mJ) × 2.10^{12} W/cm^2. [From Kruit, P., Kimman, J., Muller, H. G., and van der Wiel, M. J. (1983). *Phys. Rev.* **A28,** 248.]

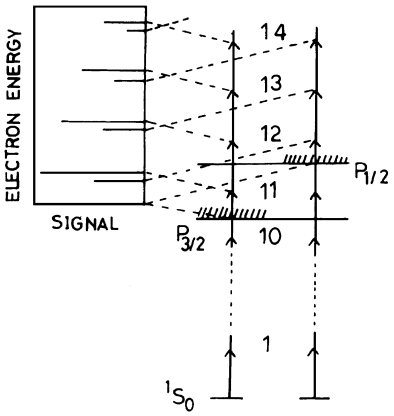


FIGURE 29 Energy-level diagram of xenon. A part of the above-threshold ionization of xenon with 1074-nm photons is schematically shown. [From Kruit, P., Kimman, J., Muller, H. G., and van der Wiel, M. J. (1983). *Phys. Rev.* **A28,** 248.]

Xenon pressure chosen in such a way that the total electron signal in each spectrum is 25–50 electrons per pulse is also indicated in Fig. 28. From this figure it can be recognized that 12- to 19-photon ionization processes with higher relative probabilities take place in addition to 11-photon ionization. A schematic explanation for the origin of the electron energy spectrum of Xe that involves two ionization potentials, at 12.127 eV for the $^2P_{3/2}$ core and at 13.44 eV for the $^2P_{1/2}$ core, is given in Fig. 29. Angular distributions of photoelectrons resulting from the above-threshold ionization are measured to study the mechanism. The above-threshold ionization phenomenon is also observed for other atoms.

E. Above-Threshold Ionization and Dissociation of H_2

Multiphoton excitations of molecules above ionization threshold frequently lead to dissociations of the ions because the dissociation continuum overlaps with the ionization continuum in most cases. The nonstationary state embedded in these continua is sometimes called the superexcited state. Dynamics of such superexcited states is well studied mainly on hydrogen by using the multiphoton spectroscopy. Figure 30, for example, represents comparison of H_2^+ and H^+ ions produced via six-photon ionization of H_2 through four-photon resonance on $E, F\,{}^1\Sigma_g^+ v_E = 0$ (vibrationless state) at laser intensity of 1.5×10^{11} W cm^{-2}. The schematic energy diagram relevant to the multiphoton transition is given in Fig. 31. The fragment ions H^+ are formed by photodissociation of H_2^+ after six-photon absorption as follows:

$$H_2(X, v = 0) + 6E_p \rightarrow H_2^+(v_+) + e,$$
$$H_2^+(v_+) + E_p \rightarrow H^+ + H(1s),$$

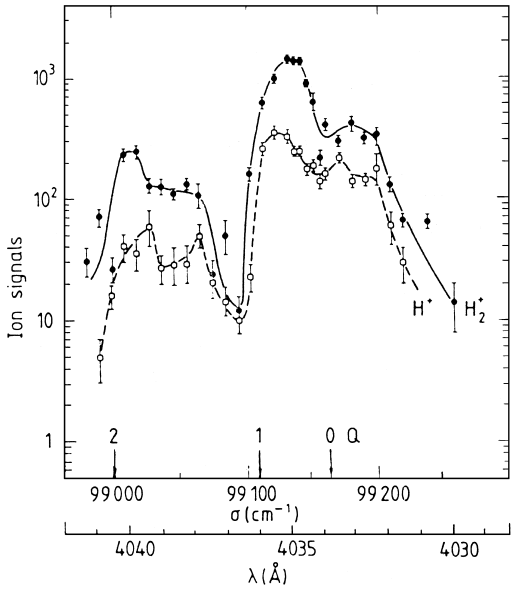


FIGURE 30 H_2^+ and H^+ resonance profiles as a function of the four-photon energy, σ, and of the laser wavelength, λ, at laser intensity of 1.5×10^{11} W cm^{-2}. [From Normand, D., Cornaggia, C., and Morellec, J. (1986). *J. Phys. B: At. Mol. Phys.* **19,** 2881.]

where E_p denotes the one-photon energy of the laser. Such an investigation on the dynamics of the superexcited states of the simplest molecule gives us important information in clarifying mechanisms of resonant multiphoton ionization dissociations of large molecules such as hydrocarbons as well.

F. Multiphoton Ionization Mass Spectroscopy

This method consists of the multiphoton ionization combined with mass detection and allows us to identify atoms and molecules by the optical spectrum related to a resonant intermediate state as well as by their mass. Isotopic species, for example the ^{13}C molecule, can be preferentially ionized in a natural isotopic mixture by shifting the wavelength from the absorption band of the light species if the intermediate state spectrum shows sharp features. The scheme of the setup for mass-selective ion detection has already been shown in Fig. 10.

The fragmentation patterns in the multiphoton ionization mass spectroscopy depend on the laser intensity, and they are different from those obtained by electron impact excitation, charge-exchange excitation, and other methods. In the multiphoton ionization mass spectroscopy, ions with small mass weights can be produced compared with those obtained by other methods. The fragmentation pattern in the mass spectrum of benzene (C_6H_6) is shown in Fig. 32. One can see that atomic fragment ions C^+ are produced with a high probability. To interpret the

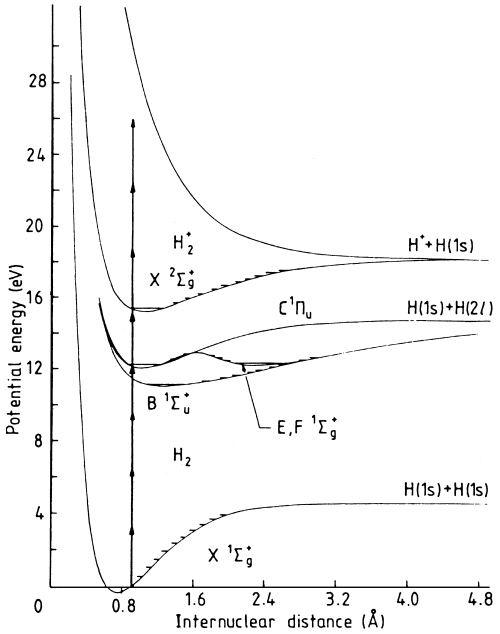


FIGURE 31 Potential energy diagram relevant to the six-photon ionization of H_2 with a four-photon resonance on the vibrationless state of E, $F\,^1\Sigma_g^+$ electronic state. [From Normand, D., Cornaggia, C., and Morellec, J. (1986). *J. Phys. B: At. Mol. Phys.* **19,** 2881.]

fragmentation pattern of polyatomic molecules, several theoretical treatments based on information-theoretical statistical and rate equation approaches have been proposed.

G. Femtosecond Multiphoton Spectroscopy

Femtosecond multiphoton spectroscopy utilizes a resonant multiphoton excitation with femtosecond pulses whose durations are shorter than molecular vibration time. Dynamics in superexcited states above ionization and/or dissociation limits can be studied directly. The final continuum states are analyzed by time-of-flight spectrometers for ionic fragments and electrons. The merit of using such an ultrashort pulse is that the observation is related to excitation processes taking place at restricted internuclear distances but not to further laser-induced excitation or ionization of the fragments. In Fig. 33, the time-of-flight spectra of ions formed and electrons ejected from Na_2 molecular beams are shown after excitation of laser pulses with 120-fsec time duration and 616-nm central frequency. The time needed for the fragments to separate to 10 Å is in the range of 0.1 to 1 psec depending on the recoil energy, W.

The "fast" Na^+ ions in Fig. 33(a), whose ejected electron peaks are around 330 nsec in Fig. 29(b), are due to resonant multiphoton and dissociation processes shown in Fig. 34(a). The slow Na^+ ions are produced by autoionization-induced fragmentation mechanism from bound doubly excited state $^1\Pi_u(4s+3p)$ of Na^{**} indicated in Fig. 34(b),

$$\mathrm{Na}^{**}\left[{}^1\Pi_u(4s+4p)\right] \rightarrow \mathrm{Na}_2^{**}(2\Sigma_u^+) + e$$
$$\rightarrow \mathrm{Na}^+ + \mathrm{Na}(3s) + W$$

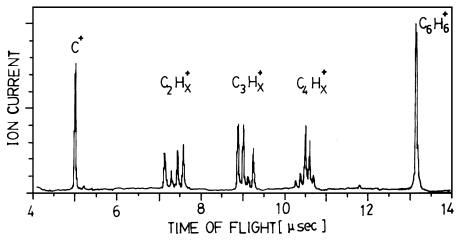


FIGURE 32 Fragmentation pattern in the mass spectrum of benzene (C_6H_6) obtained by two-step photoionization with UV laser light at 2590.1 Å. At high intensities ($>10^7$ W cm^{-2}), smaller molecular ions are observed. The molecular ions are analyzed with the time-of-flight mass spectrometer. (From Boesl, U., Neusser, H. J., and Schlag, E. W. (1978). *Z. Naturforsch., Teil.* **A**(33), 1546.]

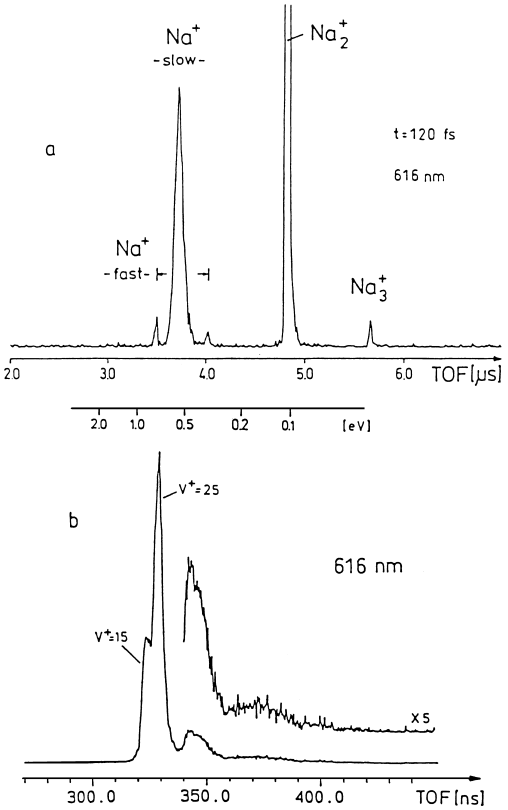


FIGURE 33 (a) Time-of-flight spectrum of ions formed by the interaction of femtosecond laser pulses with Na_2 molecular beam; (b) time-of-flight spectrum of the ejected electrons. [From Baumert, T., Bühler, B., Thalweiser, R., and Gerber, G. (1990). *Phys. Rev. Lett.* **64,** 733.]

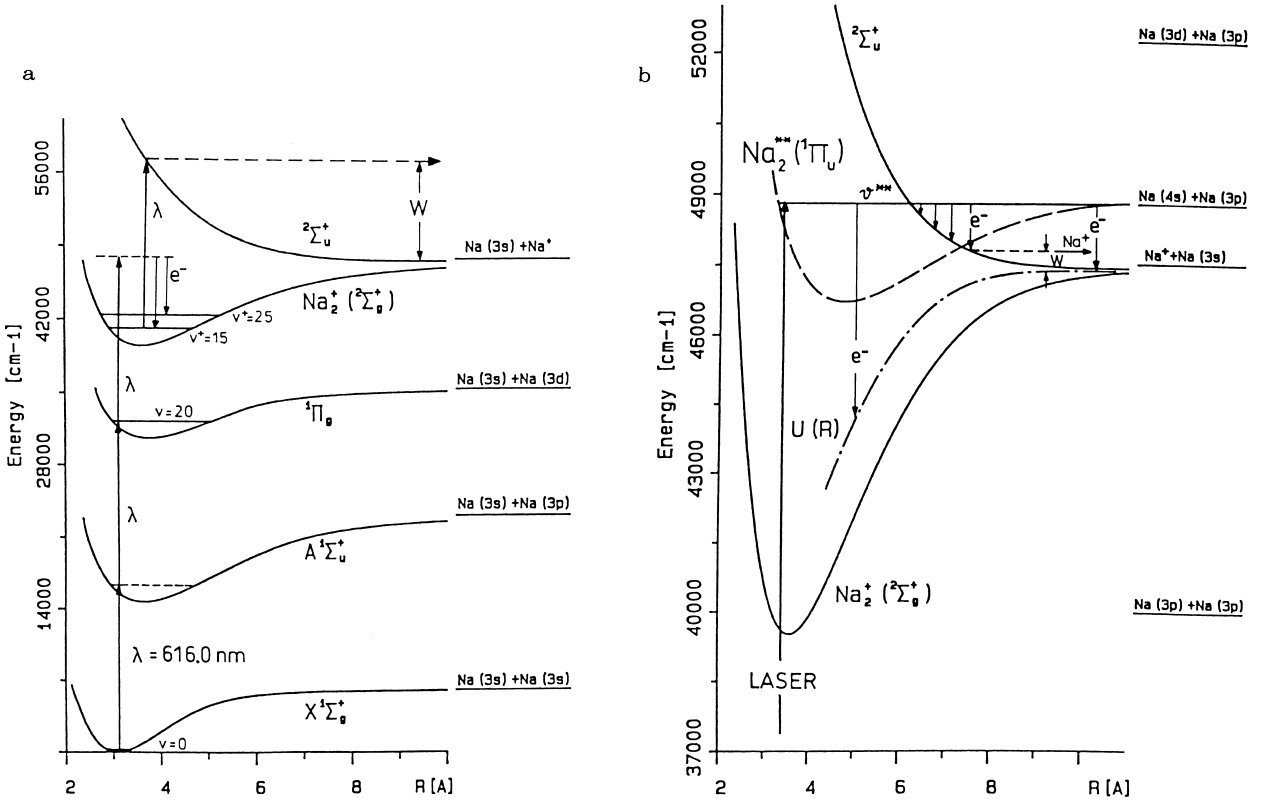


FIGURE 34 (a) Potential energy diagram of Na_2 illustrating the process leading to fast Na^+ ionic fragments; (b) excitation and autoionization processes of the doubly excited $^1\Pi_u(4s + 3p)$ state of Na_2. [From Baumert, T., Bühler, B., Thalweiser, R., and Gerber, G. (1990). *Phys. Rev. Lett.* **64,** 733.]

The spectra shown above are obtained by applying the femtosecond, one color laser apparatus. Dynamics of transition states in chemical reaction as well as the superexcited states can be further studied by using femtosecond time-resolved, multicolor multiphoton spectroscopic methods.

The processes of multiphoton ionization and dissociation of molecules induced by intense femtosecond infrared laser pulses is thought to be different from that induced by nanosecond UV laser light. This is because, first, the interaction times are comparable to or shorter than the time scale of excited state dynamics of molecules and, second, bcause the tunnel ionization process plays a dominant role in multiphoton ionization irradiated by infrared laser pulses with a peak intensity of 10^{14} to 10^{15} W/cm^2. Figure 35 shows a mass spectrum of deuterated benzene (C_6H_5D) irradiated by laser pulses with 2×10^{15} W/cm^2 at 50 fsec and at 790 nm. The spectrum consists of parent ion peaks charged at 3^+ and 2^+ and a number of (M - nH) satellites in addition to the strongest peak of the parent ion mass (79). The two expanded areas in Fig. 29 correspond to the multiple-charged parent ions and the satellites. Two half-mass peaks at 37.5 and 38.5 in Fig. 29c correspond to $C_6H_3^{2+}$ and $C_6H_3D^{2+}$, respectively. The appearance of the multiple charged parent ions is very different from that irradiated with nanosecond pulses at 10^9 W/cm^2, in which low-mass fragments are dominant. Application of very intense laser pulses to molecules induces a Coulomb explosion, which creates multiple-charged atoms and fragments with small masses due to the Coulomb repulsion forces of multiple-charged parent ions.

H. High-Order Harmonic Generation

Atoms or molecules with absorbed energy higher than their ionization potential emit a series of high-energy photons at odd harmonic frequencies of the applied laser field, which is similar to third-order harmonic generation, shown in Fig. 6. This process is called high-order harmonic generation (HG). Harmonic generation is an elastic scattering process and preserves the phase relation between the incident and the emitted photons; a high-order HG is generated when an electron is recombined with the core at return time. A high-order HG can be used as a soft X-ray source of a compact tabletop.

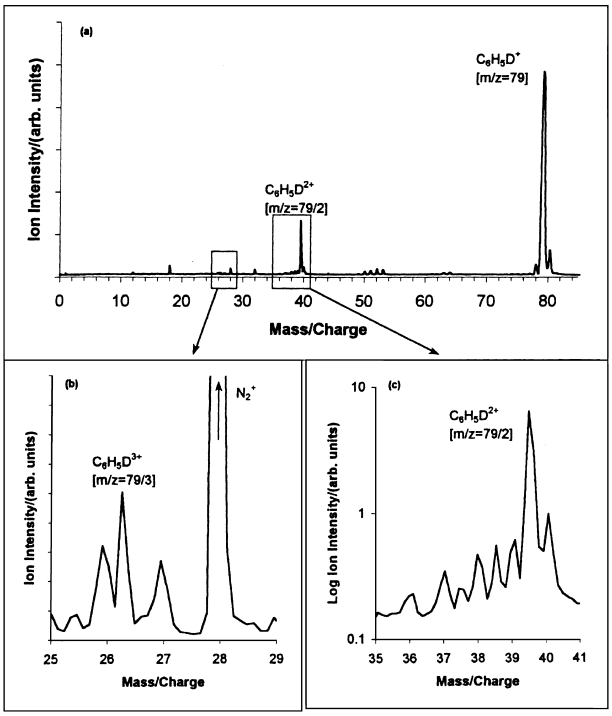


FIGURE 35 Mass spectrum of deuterated benzene with expanded areas around the 2^+ and 3^+ parent ions. The laser intensity is 2×10^{15} W/cm^2 with a pulse width of 50 fsec and a wavelength of 790 nm. [From Ledingham, K. W. D., Singhal, R. P., Smith, D. J., McCanny, T., Graham, P., Kiliv, H. S., Peng, W. X., Wang, S. L., Langley, A. J., Taday, P. F., and Kosmidis, C. (1998). *J. Phys. Chem.* **102,** 3002.]

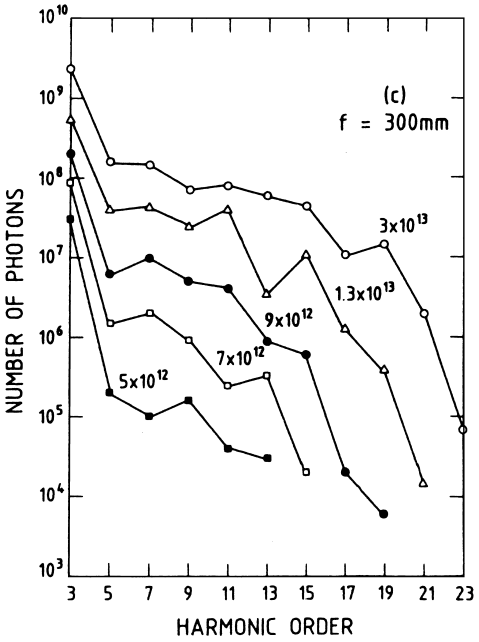


FIGURE 36 Number of harmonic photons in Xe at 15 torr at several laser intensities of a Nd:YAG laser (1064-nm wavelength) with a focal length of 300 mm. [From Lompre, L. A., L'Huillier, A., Ferray, M., Monot, P., Mainfray, G., and Manus, C. (1990). *J. Opt. Soc. Am. B* **7,** 754.]

In Fig. 36 is shown a high-order HG spectrum of Xe, which is characterized by a flat distribution as a function of harmonic order, known as the plateau, followed by a sudden decrease in efficiency, referred to as the cutoff. The cutoff determines the highest harmonic order from atoms or ions. The cutoff is evaluated by $I_p + 3.17\ U_p$ in the semiclassical method, where I_p is the ionization potential and U_p is the ponder motive energy. The maximum photon energy produced by high-order HG, $\hbar\omega_{max}$, is therefore given by the cutoff $\omega_{max} = I_p + 3.17\ U_p$. This cutoff law indicates qualitatively that the shortest wavelength of the high-order HG can be generated as a result of interaction between atoms or ions having a high ionization potential and the laser field with a low frequency.

I. Nonsequential Multiple Ionization of Atoms

The nonsequential multiple ionization of atoms is induced by a breakdown of the single-electron-excitation approximation. In this approximation, multiple ionization takes place sequentially. Evidence of nonsequential multiple ionization is the appearance of a bump in the ion yield-versus-laser intensity curve of atoms as shown in Fig. 37. Here helium (He) ion yields induced by linearly polarized, 100-fsec, 780-nm laser pulses are shown covering 12 orders of magnitude dynamic range. Calculations obtained within a single active electron (SAE) approximation are shown as solid lines and those obtained within an ac-tunnel model are shown as dashed lines. In an SAE approximation, the time-dependent Schrödinger equation is solved under the condition in which the electron moves in response to the laser in the time-independent field of the remaining electron in the ground state. Nonsequential multiple ionization is dependent on a polarization; that is, it is suppressed by circularly polarization. Two models, a rescattering model and a shake-off model, have been proposed for nonsequential multiple ionization. In the rescattering model, nonsequential multiple ionization is explained by an inelastic collision between a previously ionized electron and an electron that can be further ionized located near the parent ion. In the shake-off model, the removal of the first electron is so rapid that the other electrons cannot follow adiabatically to the energy states, and there is a possibility that some of the electrons are excited to an ionized state (shake-off process) as well as to higher bound excited states (shake-up process).

VI. SUMMARY

In the early 2000s, multiphoton spectroscopy is a widely used tool for clarifying their dynamical behaviors as well as for determining electronic structures and vibrational

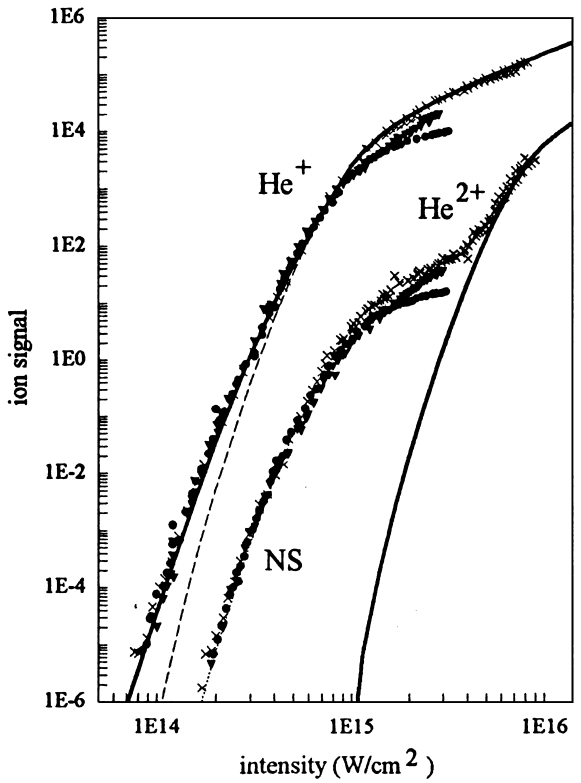


FIGURE 37 Measured He ion yields for linearly polarized, 100-fsec, 780-nm laser pulses. Calculations are shown as solid (SAE) and dashed (ac-tunnel) lines. The solid curve on the right is the calculated sequential He^{2+} yields. [From Walker, B., Sheehy, B., Dimauro, L. F., Abostini, P., Schafer, K. J., and Kulander, K. C. (1994). *Phys. Rev. Lett.* **73,** 1227.]

structures. The use of multiphoton spectroscopy has expanded to various research fields due to developments of laser technology. For example, intense ultrashort pulses generated by a Ti:sapphire laser with wavelength 800 nm have opened up new fields of research on multiphoton phenomena, such as tunnel ionization and generation of multiply charged ions, which can only be explained by nonperturbation theory, not by simple perturbation theory. Multiphoton spectroscopy already has further potential applications for clarification of new phenomena induced by nonlinear matter–radiation interactions.

In this article, the theoretical basis and experimental principles of multiphoton spectroscopy have been presented. Perturbative and nonperturbative treatments based on both the time-dependent and time-dependent theories, have been described from the theoretical view point. Experimental methods for detecting photon and photoelectrons have been described in detail. Characteristic features and spectral properties of multiphoton spectroscopy have been presented. Differences in laser-intensity dependence between perturbative, weak field regime and nonperturbative, high-intensity regimes have also been described. Photon polarization behaviors in atomic and molecules have been described. Finally, typical examples of multiphoton spectroscopy applied to atomic and molecular physics and chemistry have been given.

SEE ALSO THE FOLLOWING ARTICLES

COLLISION-INDUCED SPECTROSCOPY • INFRARED SPECTROSCOPY • LASERS, DYE • MICROWAVE SPECTROSCOPY, MOLECULAR • PHOTOACOUSTIC SPECTROSCOPY • PHOTOCHEMISTRY BY VUV PHOTONS • RAMAN SPECTROSCOPY

BIBLIOGRAPHY

Bandrauk, A. D., ed. (1994). "Molecules in Laser Fields," Marcel Dekker, New York.

Bederson, B., and Walther, H. (1995). "Advances in Atomic, Molecular, and Optical Physics," Vol. 35, Academic Press, New York.

Evans, D. K., and Chin, S. L. F., eds. (1994). "Multiphoton Processes," World Scientific, Singapore.

Lambropoulos, P., and Walther, H. J. (1997). "Multiphoton Processes 1996," Institute of Physics, Bristol, UK.

Lin, S. H., Fujimura, Y., Neusser, H. J., and Schlag, E. W. (1984). "Multiphoton Spectroscopy of Molecules," Academic Press, Orlando.

Lin, S. H., Villaeys, A. A., and Fujimura, F., eds. (1999). "Advances in Multi-photon Processes and Spectroscopy," Vol. 12, World Scientific, Singapore.



Photochemistry by VUV Photons

Michael N. R. Ashfold
Phillip A. Cook
University of Bristol

I. Sources of VUV Radiation
II. Nature of Excited Electronic States Populated by Absorption at VUV Photon Energies
III. Possible Fates of a Molecule Following Excitation with a VUV Photon
IV. Methods for Studying VUV Photochemistry
V. Conclusions and Outlook

GLOSSARY

Absorption spectrum Plot of the wavelength (or frequency) dependence of the propensity for a given molecule to absorb electromagnetic radiation.

Photodissociation The fragmentation of a molecule following absorption of one or more photons.

Potential energy surface A function that, for a particular electronic state, describes how the molecular energy varies with nuclear configuration. This is normally calculated for a mesh of discrete molecular geometries using *ab initio* quantum mechanical methods.

Radiationless transitions Processes by which population transfers from one electronic state to another without emission of a photon. Examples include internal conversion (population transfer between two states of the same spin multiplicity), intersystem crossing (population transfer between states of different spin multiplicity), and predissociation (population transfer from a bound to a dissociative state).

Radiative transitions Processes by which population in an excited state decays to levels of lower energy by emission of a photon. Examples include fluorescence and phosphorescence.

Rydberg state An excited electronic state of an atom or molecule in which at least one electron is considered to be in a spatially diffuse, "hydrogen-like" orbital with a principal quantum number greater than that of any of the valence electrons.

Vacuum ultraviolet (VUV) The region of the electromagnetic spectrum lying within the wavelength range 100–200 nm.

THE VACUUM ULTRAVIOLET (VUV) region of the electromagnetic spectrum is generally taken to span the wavelength range $100 \leq \lambda \leq 200$ nm (corresponding to the energy range 6.2–12.4 eV or, in wavenumber units, 50,000–100,000 cm^{-1}). The long wavelength limit is determined by the onset of significant atmospheric absorption, associated with the Schumann-Runge absorption system of molecular oxygen, while the high energy

limit [where the VUV merges into the extreme ultraviolet (XUV) region] corresponds to the wavelength below which there are no readily available transparent crystalline window materials (even lithium fluoride is opaque at $\lambda < 105$ nm). The VUV region is of great interest to photochemists, since the photon absorption provides sufficient energy to break any chemical bond (of all molecules, carbon monoxide has the highest bond dissociation energy, D_0(C—O) = 11.19 eV). Such fragmentation can occur directly, or only after the excited molecule has undergone substantial nuclear and, in many cases, electronic rearrangement. For most molecules, however, the alternative process of photoionization is only possible, energetically, at the end of the VUV range. This article reviews some of the various means of studying photochemical processes induced by VUV photon absorption and unravelling details of the interplay and competition between the various possible decay mechanisms.

I. SOURCES OF VUV RADIATION

Available VUV light sources divide into three broad classes. The oldest, simplest, and cheapest are ***resonance lamps*** based on atomic and molecular discharges excited, for example, by passing microwaves through a flowing (or sealed) sample of an appropriate gas. Low pressure discharges provide emission lines at specific discrete wavelengths that are characteristic of the atomic carrier; these lines broaden and are supplemented by molecular emissions at higher pressures to yield continuous emission spectra. These higher pressure sources can be used in conjunction with a monochromator to provide narrower bandwidth VUV excitation (albeit of relatively low intensity) at any VUV wavelength of choice. Table I provides an overview of some of the more common discharge sources and the wavelengths they provide.

TABLE I Traditional Sources of VUV Radiation: Discrete Line and Continuum Sources

Atomic species	Emission wavelengths (nm)
Line sources	
H	121.6
Ar	104.8, 106.6
Kr	116.5, 123.6
O	130.2, 130.5, 130.6
Xe	129.6, 147.0
Hg	184.9
Continuum sources	
Ar	105–155
Kr	125–180
Xe	148–200

Synchrotrons are the present-day successors to the continuum lamps. Synchrotron radiation is produced whenever fast-moving, charged particles are deflected in a magnetic field. Circular electron accelerators and storage rings are useful sources of synchrotron radiation. Such radiation has a number of characteristic and attractive properties. It spans a very wide spectral range, from the infrared (IR) to the deep X-ray region (including the VUV region), with an intensity distribution that is dependent upon the energy of the electron beam and is readily calculable. Synchrotron sources provide short (subnanosecond) pulses of intense, highly collimated radiation with well-defined polarization properties (linear or elliptical). Modern synchrotrons often have additional magnetic structures inserted within the storage ring, e.g., wigglers and undulators, to shift the spectral output to shorter wavelength and to give higher spectral brightness, respectively. As with the early continuum lamps, it is necessary to disperse the broadband synchrotron output through a monochromator in order to obtain linewidths comparable to that from the low pressure atomic resonance lamps or from narrow bandwidth lasers (see below), but synchrotron radiation offers the huge advantage of being continuously tunable. To give an idea of the current state of the art, the Chemical Dynamics Beamline at the Advanced Light Source (ALS) offers a quoted flux of $\sim 10^{16}$ photons per square millimeter per second at a resolution ($E/\Delta E$) of 40. However, after passage through a 6.65-m monochromator, this drops to $\sim 10^{12}$ photons per square millimeter per second for photon wavenumbers $\sim$100,000 cm^{-1} and $E/\Delta E = 3000$, and to $< 10^9$ photons per square millimeter per second at $E/\Delta E = 10^5$ which would correspond to 1 cm^{-1} resolution at these energies.

Lasers provide the third route to accessing the VUV spectral region. The output of a laser is derived from the stimulated emission of radiation from an inverted population distribution. In general, excited state population can decay both by *spontaneous* and by *stimulated* emission. The relative probability of these two processes scales with the third power of the transition frequency. Spontaneous emission from highly excited electronic states thus occurs rapidly, and, as a result, there are few lasers whose fundamental output falls in the VUV spectral region. Two notable exceptions are the ArF exciplex laser (operating at 193.1 nm) and the molecular F_2 laser (at 157.5 and 157.6 nm). The lasing transition in the former is between an upper state in which the atoms are bound as an excited diatomic molecule and a lower level that is dissociative (i.e., unstable with respect to the constituent atoms), thus ensuring the population inversion needed for laser action. Both are pulsed lasers, and available commercial systems provide photon intensities—at their respective operating wavelengths—that are orders of magnitude higher than any synchrotron. The caveat is important, since the use

of lasers to generate high intensity coherent radiation at *other* VUV wavelengths is considerably more involved.

Anti-Stokes Raman shifting, in which stimulated Raman scattering is used to "shift" the frequency of intense fundamental laser radiation as illustrated in Fig. 1a, is one route. For example, using the fourth harmonic of an Nd-YAG laser ($\lambda = 266$ nm, $\overline{\omega} \sim 37{,}590$ cm^{-1}) and a high pressure of H_2 gas as the Raman shifting medium ($\overline{\omega}_{vib} \sim 4155$ cm^{-1}) allows generation of coherent radiation at ~199.78 nm ($\overline{\omega} \sim 50{,}050$ cm^{-1}), ~184.45 nm ($\overline{\omega} \sim 54{,}210$ cm^{-1}), and 171.34 nm ($\overline{\omega} \sim 58{,}640$ cm^{-1})—the third (3AS), fourth (4AS), and fifth (5AS) anti-Stokes components, respectively. Conversion efficiencies depend on both the pressure and the temperature of the gas used in the Raman shifter, but higher order shifts are progressively less efficient. This hinders access to the shorter VUV wavelengths. This limitation, and the fact that the technique only provides continuously tunable VUV radiation if the incident laser source is itself tunable, ensures that a variety of third order nonlinear optical techniques are now the methods of choice for generating tunable VUV radiation in the laboratory. The simplest such technique is third harmonic generation, also known as frequency tripling. In this technique, illustrated in Fig. 1b, the polarization induced in a suitable gaseous medium by a laser field of frequency $\overline{\omega}_1$ acts as the source for a new wave at the third harmonic frequency $\overline{\omega}_3 = 3\overline{\omega}_1$, which can attain useful intensities if the necessary phase matching conditions are satisfied. The efficiencies of such third order processes are greatly enhanced if it can be arranged that a real excited state of the tripling gas is resonant at, say, the energy of two $\overline{\omega}_1$ photons. The combined requirements of efficiency and tunability can then be satisfied by the use of two input radiation fields, with frequencies $\overline{\omega}_1$ and $\overline{\omega}_2$, and generation of resonance-enhanced VUV radiation at the sum ($\overline{\omega}_{sum} = 2\overline{\omega}_1 + \overline{\omega}_2$) and difference ($\overline{\omega}_{diff} = 2\overline{\omega}_1 - \overline{\omega}_2$) frequencies as illustrated in Figs. 1c and 1d. The rare gases (either individually, or as suitable phase matched mixtures) are the nonlinear media used most commonly. They are photochemically inert and allow complete coverage of the wavelength range $100 \leq \lambda \leq 200$ nm; however, much of the longer part of this range can be accessed more efficiently by using an appropriate hot metal (e.g., Mg) vapor as the nonlinear medium. Such methods can provide much higher VUV intensities than current synchrotron sources, but at the expense of broad tunability. For example, several laboratories now use third harmonic generation (in phase matched Kr/Ar mixtures) or four wave difference frequency mixing in Kr to generate $>10^{12}$ photons at the Lyman-α wavelength (121.6 nm, 82,259 cm^{-1}) with ~1 cm^{-1} bandwidth and pulse duration <10 ns.

Finally, for completeness, we note that multiphoton excitations involving the coherent absorption of two or three photons (of the same or different frequencies) from an intense pulse of UV or visible laser radiation provide an alternative means of reaching excited electronic states lying at VUV equivalent energies. The implementation of such multiphoton excitation methods normally relies on the fact that a further one photon absorption by the excited state of interest will result in ion formation, thereby providing a sensitive means of detecting that multiphoton absorption has occurred. The selection rules governing one and multiphoton excitations are often different and therefore complementary.

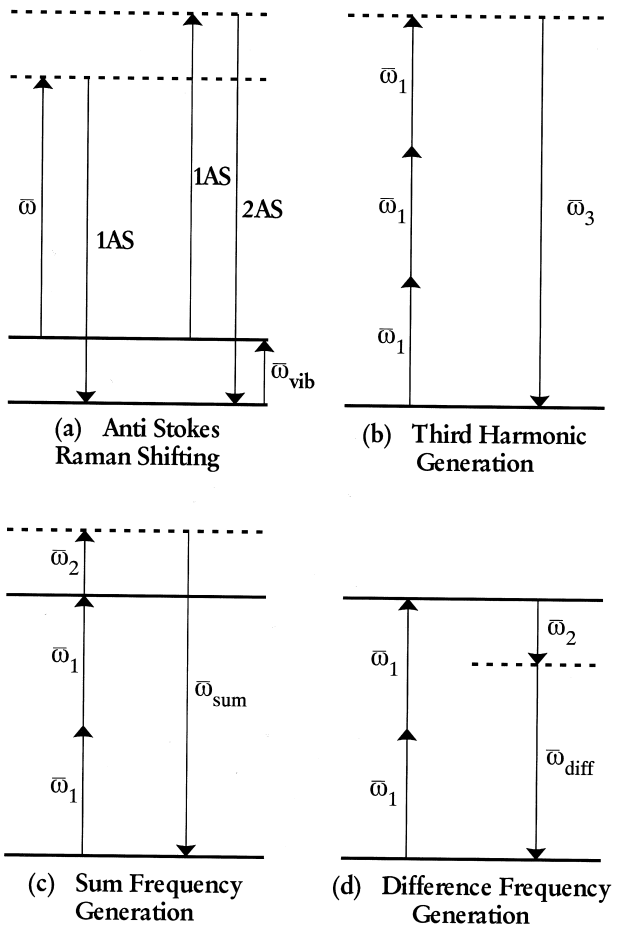


FIGURE 1 Energy level diagrams illustrating VUV generation by (a) anti-Stokes Raman shifting, (b) third harmonic generation, (c) four wave sum frequency mixing, and (d) four wave difference frequency mixing. Solid and dashed horizontal lines represent real and virtual levels, respectively.

II. NATURE OF EXCITED ELECTRONIC STATES POPULATED BY ABSORPTION AT VUV PHOTON ENERGIES

Most of the excited electronic states lying at VUV equivalent energies are Rydberg states, i.e., states that arise as a

result of electron promotion from the highest (or one of the highest) occupied molecular orbitals in the ground state configuration to nonbonding, spatially diffuse, atomic-like orbitals. As in atoms, molecular Rydberg states form series that converge on the various ionisation limits. As a result, the density of such states increases rapidly as these energetic limits are approached. These Rydberg states supplement, and can interact with, the manifold of valence excited states that arise from electronic excitations to the nonbonding and/or anti-bonding valence orbitals that lie above the highest occupied orbitals of the ground state configuration. Since these states lie at energies above the bond dissociation energy, they will generally dissociate, either directly or by predissociation (see below), to neutral fragments in their ground or excited electronic states. Ion pair states—states formed by the union of an anion and a cation—are a third class of excited state that may contribute to the overall manifold of electronic states in the VUV spectral region. The various selection rules governing allowed optical transitions will ensure that only a fraction of the complete manifold of excited electronic states actually contributes to the one photon VUV absorption spectrum; these restrictions can be circumvented to some extent by complementary multiphoton excitation studies of the same (VUV equivalent) energy regions. Similarly, Franck-Condon considerations dictate that both one and multiphoton excitation spectra will only reveal a small fraction of the wealth of vibrational and/or continuum energy levels associated with these excited electronic states. Finally, we must recognize that the high state density, often compounded by spectral line broadening (in the case that the excited state lifetimes are short—see below), will mean that not all excited states that contribute to molecular absorption spectra can actually be recognized and assigned. Theory is playing an ever more critical role in the interpretation of such spectra, reflecting the continual improvements in computing power and the accessibility of *ab initio* quantum chemistry codes for calculation of reliable potential energy functions for *excited* electronic states and of methods for following nuclear motions on such potentials.

III. POSSIBLE FATES OF A MOLECULE FOLLOWING EXCITATION WITH A VUV PHOTON

Figure 2 shows some of the likely fates of an isolated gas phase molecule following photon absorption, in the form of a so-called Jablonski diagram. The molecule in this example is presumed to have singlet spin multiplicity in its ground state (labelled S_0). Allowed electronic transitions involve promotion of one electron from the ground state configuration and, unless spin-orbit coupling is large,

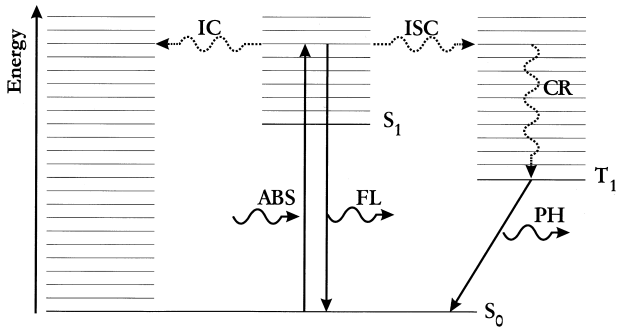


FIGURE 2 Jablonski diagram illustrating some of the various possible fates of a molecule following photoexcitation to excited rovibrational levels within the S_1 electronic state. Key: ABS, absorption; FL, fluorescence; PH, phosphorescence; IC, internal conversion; ISC, intersystem crossing; CR, collisional relaxation. Bold and thin horizontal lines indicate, respectively, electronic origins and excited rovibrational levels built on these origins.

conservation of the overall electron spin. Thus, the excited states appearing in the electronic absorption spectrum will have singlet spin multiplicity also; these are traditionally labelled S_1, S_2, etc. in order of increasing energy. For every excited singlet state there must also be a corresponding triplet state (here labelled T_1, T_2, etc.) in which the two unpaired electrons have parallel spins.

We start by considering a molecule excited to its S_1 state at an energy below that required for dissociation. This molecule may decay by emission of a photon, to a singlet state of lower energy (the S_0 state in this case), or nonradiatively. Radiative decay between two states with the same spin multiplicity is termed fluorescence (FL). In a nonradiative process, however, the energy introduced by photon absorption must be conserved within the molecule. Thus, if the S_1 population decays by transfer to an electronic state whose origin lies at lower energy, overall energy conservation requires that there must be a concomitant increase in the nuclear (vibrational) energy content of the molecule. In some cases, this enhanced nuclear motion may be sufficient to allow isomerization. Both the T_1 and S_0 states are possible acceptors for the decaying S_1 population. Nonradiative decay between states of the same multiplicity is termed internal conversion (IC), while the competing spin-changing $S_1 \rightsquigarrow T_1$ transfer is called intersystem crossing (ISC). Any T_1 population formed in this way will itself ultimately decay. Relaxation within the manifold of vibrational levels associated with the T_1 state must be followed, eventually, by further ISC (to high levels of the ground state) or radiative decay. This latter $T_1 \rightarrow S_0$ emission is called phosphorescence (PH) and, being spin forbidden, is generally slow. The relative efficiencies of these various processes from any particular initially populated level are expressed as the respective quantum yields, ϕ_I, where

$$\phi_{FL} + \phi_{PH} + \phi_{IC} + \phi_{ISC} = 1. \qquad (1)$$

This discussion has tended to focus on decay of the excited state population in the absence of collisions. The relative magnitudes of the various ϕ_I are, however, sensitive to the external environment; collisions generally enhance the rates of both ISC and IC (often bracketed simply as quenching processes) and open up another possible loss mechanism—bimolecular chemical reaction.

The preceding discussion will be valid for any electronically excited molecule, not just one that has specifically absorbed a VUV photon. VUV excitation, however, provides sufficient energy to break at least one bond in any polyatomic molecule, so the picture presented thus far has to be modified further to include dissociation of the excited state molecule, with quantum yield ϕ_{DISS}. Photochemists have spent considerable time and effort unravelling details of the way in which molecules fragment. This is not just because of its intrinsic interest as a route to forming chemical lasers and its relevance to, for example, atmospheric, interstellar, and plasma scientists, but also because of the perceived analogy between a full bimolecular reactive collision, proceeding via a transition state, and the (simpler) "half-collision" in which a photoexcited molecule (the mimic of the reaction transition state) breaks up into products.

Molecular dissociations are normally considered in a number of subcategories. Figure 3 illustrates two generic types of fragmentation mechanism for diatomic molecules. In Fig. 3a, photoexcitation promotes a molecule from a bound ground state to an excited electronic state that has no net bonding and thus dissociates directly into its constituent atoms. The ground and excited electronic states are represented by so-called potential energy (PE) curves. These are theoretical constructs obtained by invoking the Born-Oppenheimer approximation and solving the Schrödinger equation for the two nuclei and the ensemble

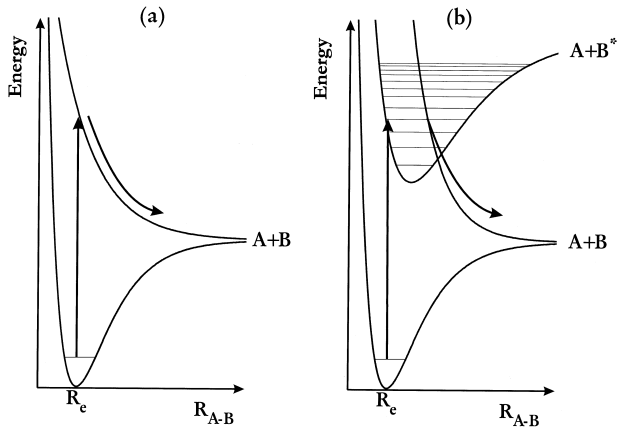


FIGURE 3 Potential energy diagrams for a representative diatomic molecule, AB, illustrating (a) direct dissociation and (b) predissociation.

of electrons that, together, make up the molecule. In this picture, each electronic state has its associated PE curve (so called because it is regarded as an effective potential that governs the motion of the nuclei), which varies as a function of the internuclear separation, R. The equilibrium bond length, R_e, of the ground state molecule is the internuclear separation at which the potential curve for this state has a minimum. In Fig. 3a, the potential curve for the excited state has its minimum at $R = \infty$ and is purely repulsive. The act of photon absorption occurs so rapidly that the molecule has no opportunity to adjust its bond length during the transition. Electronic excitations are thus represented as vertical transitions between such PE curves. The nuclei adjust to their new electronic environment *after* promotion to the excited state potential, in this case by separating into two free atoms on a timescale of the order of one half of a vibrational period ($\sim 10^{-13}$ to 10^{-14} sec). Fragmentations of this kind are termed *direct* dissociations. Transitions to such short-lived, purely dissociative excited states reveal themselves as broad, continuous features in absorption spectroscopy. HF and HCl are examples of molecules that fragment in this way following excitation in the VUV spectral region.

Such behavior should be contrasted with that which would occur in the case of a molecule with PE curves arranged such as in Fig. 3b. The key point to note in this case is that there is a region of configuration space (here corresponding to an internuclear separation larger than the ground state equilibrium bond length) at which the PE of a bound excited state matches that of a second, dissociative electronic state. Initial photoexcitation from the ground state, at or above this energy, will result in population of the "bound" vibrational level, but as the bond extends the excited electronic configuration will evolve, acquiring more of the character of the dissociative state. Each time the nuclei sample the extended geometry they can choose whether to follow the bound or repulsive potential; those that follow the former execute another molecular vibration, but those taking the latter route continue to separate from one another, forming atomic fragments. Such dissociation is termed *pre*dissociation, because the dissociation occurs at an energy below that of the natural dissociation limit of the initially populated bound state. Clearly, predissociation rates can span many orders of magnitude, depending upon the efficiency of the coupling between the bound and repulsive state potentials. The energy-time form of the Uncertainty Principle

$$\delta E \approx \frac{\hbar}{\tau} \tag{2}$$

requires that the energy of an ensemble of molecules in an excited state with average lifetime τ will be blurred to an extent of order δE. Transitions to levels that predissociate very rapidly (i.e., have small τ) show substantial

"lifetime broadening"; e.g., an excited state with $\tau = 10^{-13}$ sec will exhibit associated lifetime broadened linewidths of ~53 cm^{-1} which may prove hard to distinguish from a direct dissociation simply by inspection of the (apparently continuous) absorption spectrum. In contrast, transitions to only weakly predissociated levels can appear as resolved rovibrational structures in absorption spectroscopy and, in extreme cases, the weakly predissociated excited states may exhibit measurable fluorescence quantum yields. The VUV absorption spectra of virtually all diatomic molecules show some fine structure attributable to predissociated excited states when viewed at sufficiently high spectral resolution.

The electronic state density changes little upon progressing from diatomic to polyatomic molecules, but the complexity of the nuclear motions, particularly the number of vibrational degrees of freedom and thus the vibrational state density, increases dramatically. Inevitably, this leads to further spectral congestion, but, more importantly from the photochemical viewpoint, also means that examples of simple direct dissociations (as described above for a diatomic molecule) are rare in the case of electronically excited polyatomic molecules. Much more commonly, the excited molecule will fragment *after* intramolecular vibrational redistribution (IVR) either within the initially prepared excited state (often termed *vibrational* predissociation) or after radiationless transfer (e.g., internal conversion) to an unbound potential energy surface (PES). This latter mechanism is the polyatomic analogue of the electronic predissociation process described above for the case of a diatomic fragmentation.

Figure 4 shows illustrative PESs for the excited states of two hypothetical triatomic molecules which serve to illustrate the distinction between direct dissociation and vibrational predissociation. In the former case, Fig. 4a, a vertical transition from the ground state equilibrium geometry prepares the excited molecule in the exit channel for A—BC bond fission; the accompanying trajectory shows the A and BC fragments separating, directly and irreversibly, under the influence of such a potential. Contrast this behavior with that illustrated by the second example, Fig. 4b, in which the excited state PES has a potential well, with a minimum at configurations corresponding to concerted elongation of both the A—B and B—C bonds. In this case, the initial nuclear motion following vertical excitation involves extension of both bonds; this vibrational motion has to evolve (i.e., undergo IVR on the excited state PES) until there is sufficient kinetic energy directed into the A—BC local stretching mode to surmount the exit channel barrier; only then will the final bond fission occur. Clearly, fragmentations of polyatomic molecules can exhibit a range of timescales. As with a diatomic molecule, "direct" dissociation on a single PES can occur in $<10^{-13}$ sec, but most polyatomic dissociations will occur over a longer timescale, reflecting either the time required for the necessary IVR on the excited PES populated initially or, in the case of electronic predissociations, the time for IC and/or ISC to a lower PES and, quite probably, subsequent IVR on this latter surface.

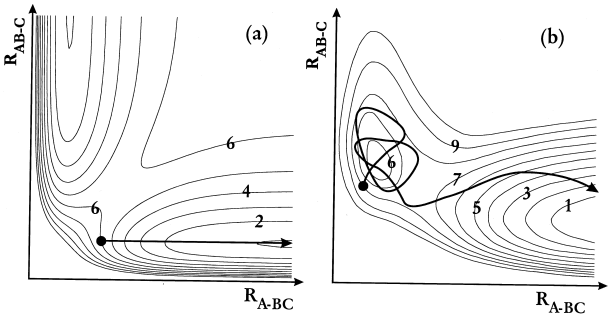


FIGURE 4 Potential energy surfaces for two electronically excited triatomic molecules, ABC (showing, in the form of contour plots, how the PE varies as a function of the two bond lengths, with the interbond angle held fixed). The individual contours are labelled 1, . . . , *n*, with low numbers corresponding to low potential energy. ● identifies the center of the vertical Franck-Condon region; the trajectories (→) illustrate the motion of a configuration point undergoing (a) direct dissociation and (b) indirect dissociation (vibrational predissociation).

IV. METHODS FOR STUDYING VUV PHOTOCHEMISTRY

All but the very simplest molecules (e.g., H_2) exhibit strong absorption throughout most, if not all, of the VUV spectral region. The penetration depth of VUV radiation into most samples is thus very short, and, as a result, the vast majority of photochemical studies involving VUV photons involve low pressure gas samples (especially, recently, jet-cooled molecular beams). This, together with the fact that a VUV photon generally provides more than sufficient energy to break at least one bond in the target molecule, ensures that most studies of photochemistry induced by VUV photon absorption rely on careful investigation of the resulting fragments in order to gain insight into the various dissociation pathways, their quantum yields, and the details of the forces acting during the breakup of the nuclear framework. Table II provides an overview of some of the more commonly employed techniques used in studying photochemistry. Note that none are the exclusive province of photochemists working specifically with VUV photons. The rest of this article is devoted to some brief consideration of the merits of the various techniques, supplemented by a few illustrative examples of molecular photochemistry induced by VUV photon absorption.

TABLE II Observables in VUV Photochemical Studies, and How They May Be Measured

Observable	Experimental technique
Frequency of absorption	Absorption; resonance-enhanced multiphoton ionization (REMPI); photofragment excitation (PHOFEX) spectroscopy
Excited state lifetimes	Absorption linewidths; analysis of photofragment recoil anisotropy; fluorescence decay
Product yields and/or branching ratios	Fluorescence; laser-induced fluorescence (LIF); REMPI detection of products; photofragment translational spectroscopy (PTS); ion imaging
Product vibration, rotation	As for product yields (above)
Product translational energies	PTS; Doppler lineshape measurements
Recoil anisotropy	Ion imaging; PTS; Doppler profiles

A. Absorption Spectroscopy

Absorption spectroscopy was for many years *the* traditional means of gaining some information about the energies, structure, and relative stability of excited electronic states of molecules, and the room temperature VUV absorption spectra of most stable small- and medium-sized gas phase molecules have been known for several decades. In many cases, one or more series of transitions to (predissociated) Rydberg states are observable as relatively sharp features superimposed on a background of continuous absorption; molecular ionization potentials have been deduced by extrapolating to the energy of such series limits. The energies of individual series members (relative to the ionization potential) provide a measure of the extent to which the Rydberg electron penetrates into the localized region in which the remaining valence and core electrons are concentrated (characterized by the so-called quantum defect), and thus some insight into the dominant nature (s, p, d, etc.) of the Rydberg orbital populated by the electronic excitation. Figure 5, which shows the electronic absorption spectrum of a room temperature sample of H_2O vapor, provides an illustrative example. If the Rydberg states predissociate sufficiently slowly, the transitions to them should show resolvable fine structure. Analysis of such structure should enable detailed characterization of the excited state structure and symmetry. In practice, only in diatomics and some of the smaller hydride polyatomic molecules (like H_2O and NH_3) do any of the transitions evident in the room temperature absorption spectra show sufficient resolved rotational structure to allow such spectroscopic analysis. Transitions to the $\tilde{C}^1B_1$ Rydberg state of H_2O (the origin band at ~124.1 nm is clearly evident in Fig. 5) provide one example, which we return to later.

Many of the Rydberg states of larger polyatomic molecules will be just as stable with respect to predissociation. However, the room temperature absorption spectra of these molecules will appear continuous because of the density of overlapping spectral lines. Seeding the molecule of interest in a supersonic molecular beam of inert carrier gas (e.g., argon or helium) provides a means of "cooling" the rovibrational state population distribution into just the few, lowest energy quantum states, thereby relieving much of the complexity (both rotational congestion and hot band absorptions) evident in the room temperature spectrum. To date, however, such jet-cooling techniques have found relatively little use in direct VUV absorption studies. This is a reflection of two practical limitations: (1) the short optical path lengths achievable in a molecular beam and (2) the scarcity of high intensity, high resolution tunable VUV light sources. Synchrotrons are generally the light source of choice for such measurements, but, as with any broadband source that has to be used in conjunction with a monochromator, there is the unavoidable trade-off between resolution and spectral brightness.

B. Resonance-Enhanced Multiphoton Ionization (REMPI)

REMPI, in contrast, is very well suited for use with molecular beam samples. The probability of coherent n-photon absorption typically scales with the nth power of the incident light intensity. Such multiphoton absorption thus tends to be localized in the most intense regions of a

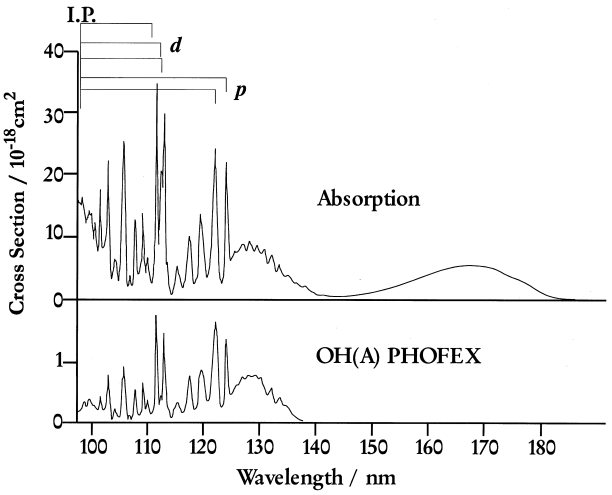


FIGURE 5 VUV absorption spectrum of a room temperature sample of H_2O vapor (upper trace) together with (below) the PHOFEX spectrum for forming electronically excited OH(A) fragments. The first members of Rydberg series associated with excitation to p and d Rydberg orbitals are indicated, as is the wavelength corresponding to the first ionization potential (I.P.). [Adapted from Lee, L. C., and Suto, M. (1986) "Quantitative photoabsorption and fluorescence study of H_2O and D_2O at 50–190 nm," *Chem. Phys.* **110,** 161–169. With permission from Elsevier Science, New York.]

pulsed focused laser beam, which can easily be arranged to overlap with the densest region of a pulsed molecular beam. Figure 6 shows the $3+1$ REMPI spectrum of a 300 K sample of H_2O vapor recorded using excitation wavelengths ~372 nm so as to be resonant, at the three photon energy, with the zero-point level of the $\tilde{C}$ state. This example serves to highlight some of the strengths and limitations of using REMPI to study photochemistry at VUV energies. One obvious strength is the resolution which, if the excited molecules are sufficiently long lived, can be limited by the bandwidth of the available laser radiation, thus allowing observation of a wealth of rotational fine structure that is hard (though, in this particular case, not impossible) to observe in direct one photon absorption. Analysis of the line positions provides information on the symmetry and geometry of the excited state, while, as Fig. 6 shows, modelling the relative intensities of the lines

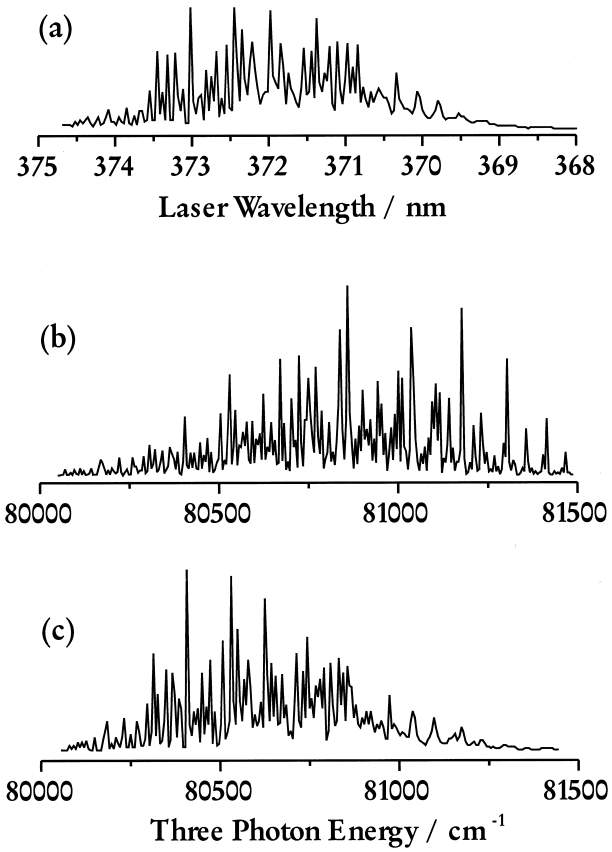


FIGURE 6 (a) $3+1$ REMPI spectrum of a 300 K sample of H_2O vapor following excitation at wavelengths ~372 nm, showing predissociation broadened rotational fine structure. (b) The predicted band contor for the $H_2O(\tilde{C}^1B_1 \leftarrow X^1A_1)$ origin transition calculated using the appropriate three rotational line strength factors. (c) The good match with the experimental spectrum that results after proper allowance for the rotational level dependent predissociation of the $\tilde{C}^1B_1$ Rydberg state. [Adapted from Bayley, J. M. (1985). Ph.D. Thesis, University of Bristol.]

can also provide detailed insight into the predissociation mechanism(s) affecting the excited state molecules. Comparing Figs. 5 and 6, however, also reveals a limitation of the REMPI method, namely, the fact that it discriminates in favor of the more long-lived excited states at the expense of short-lived continuous absorptions. Thus, the technique has found considerable application in unravelling details of excited states of small polyatomics, lying at VUV energies, that predissociate on timescales of the order of a few picoseconds, but it should always be borne in mind that the REMPI spectrum is unlikely to give a representative view of the complete VUV absorption spectrum.

C. Photofragment Excitation (PHOFEX) Spectroscopy

PHOFEX spectra are obtained by monitoring the yield of a chosen photofragment as a function of the wavelength used to excite the parent molecule. Therefore, such spectra provide a view of the wavelength-dependent partial cross section for forming a particular product, P, normally in a particular quantum state (or states). Such spectra may be obtained straightforwardly if the product of interest is formed in an excited electronic state which decays by spontaneous fluorescence. H_2O is such a molecule. The PHOFEX spectrum for forming electronically excited OH(A) fragments is shown below the absorption spectrum in Fig. 5. Several important parameters can be derived from such spectra. For example, the ratio of the PHOFEX signal divided by the total absorption provides a measure of the quantum yield for this excited product channel (~10% at $\lambda \sim 130$ nm, falling to ~2% at $\lambda \sim 100$ nm, and with local minima at wavelengths corresponding to the sharp Rydberg absorption features). Additionally, establishing the long wavelength onset ($\lambda \sim 137$ nm) provides an upper limit value for the dissociation energy for forming these electronically excited products. The analysis of such spectra does require caution, however. As already hinted, the threshold measurement will overestimate the bond dissociation energy if there is an energy barrier in the coordinate leading to the excited products. The quantum yield analysis will be complicated *if* the emitting fragment is formed in a range of rovibrational quantum states, the distribution of which varies with parent excitation wavelength. The Einstein A coefficients and detection efficiencies for the various emitting levels then need to be considered carefully; further complications may arise if, as in the case of OH(A), certain of the product rovibrational levels predissociate sufficiently rapidly that they do not contribute to the measured PHOFEX spectrum.

The low quantum yield of the $H + OH(A)$ product channel following excitation of H_2O vapor with VUV photons serves to emphasize the fact that any complete picture

of a molecular fragmentation requires knowledge of the ground state and/or nonfluorescent products also. Both laser-induced fluorescence (LIF) and REMPI detection methods are used to monitor such photofragments; however, neither is trivial to implement. Use of the former requires the fragment P to have a suitable fluorescing excited state, which generally restricts LIF detection to atomic, diatomic, and a few triatomic products. Both probe techniques involve use of narrow bandwidth excitation lasers. This can be beneficial if the spectroscopy of the electronic transition used to probe P is reasonably well understood, since it allows determination of detailed quantum state population distributions in the product P (see below). However, probing at just one wavelength yields a PHOFEX spectrum for forming products P in just those particular quantum state(s) that are excited at that particular wavelength; this will not correspond to the integral (i.e., nonquantum state resolved) PHOFEX spectrum for forming products P if (as is usually the case) the distribution of population among its quantum states varies as the parent excitation wavelength (and thus the energy supplied to the system) is varied. Mass spectrometry and direct photoionization (using a single VUV photon), with subsequent time-of-flight mass separation, have also been used to measure wavelength-dependent product quantum yields.

D. Photofragment Product State Distributions

Section IV.C described the way in which the partial cross section for forming a particular quantum state of a selected photoproduct can be obtained by monitoring the spontaneous fluorescence, or an LIF or REMPI signal, associated with the product of interest, as the photolysis wavelength is varied. The complementary experiment, in which the parent excitation wavelength is fixed and the probe frequency is scanned, can also provide information about the fragmentation dynamics of the excited parent molecule. Specifically, analysis of such spectra can, in favorable cases, reveal the relative efficiencies with which quantum states (vibration, rotation, spin-orbit, Λ-doublet, etc.) of the product are formed. In the case of jet-cooled H_2O molecules, for example, LIF probing of the $OH(X^2\Pi)$ fragments that result following excitation within the $\tilde{A}^1B_1 \leftarrow \tilde{X}^1A_1$ absorption band ($190 \leq \lambda \leq 140$ nm, see Fig. 5) reveals a preference for their being formed in the upper (more antisymmetric) Λ-doublet states and with only modest rotational excitation, the extent of which scales with the parent rotational temperature; most of the "excess energy" provided by the photolysis photon (i.e., the energy over and above that needed to break the H—OH bond) is released in the form of product translation. The Λ-doublet propensity can be rationalized in terms of conserving the symmetry of the electronic charge distribution about the central O atom (in the plane defined by the three nuclei) as the molecule breaks up. The minimal rotational excitation of the OH(X) fragments reflects the fact that the $\tilde{A}$ state potential is relatively isotropic in the $\angle$HOH bending coordinate, i.e., little or no torque acts on the fragments as they separate. Vibrational excitation of the OH(X) fragments is also modest, but increases with increasing photon energy. This observation and the energy dependence of the balance between OH vibration and product recoil can also be reproduced by detailed calculations of the nuclear motion on a reliable *ab initio* PES for the $H_2O(\tilde{A})$ state.

The photofragmentation of $\tilde{A}$ state H_2O molecules is an unusually favorable example. The dissociation occurs on a single PES, the detailed topology of which can be established via accurate quantum mechanical calculation. The OH(X) fragment can be detected by LIF with high sensitivity. The resulting spectrum is well characterized and can be used to derive quantitative product state population distributions. However, for most excited states of most polyatomic molecules, detailed *ab initio* PESs are not available. Dissociations that involve the transfer of flux between PESs are far less amenable to detailed theoretical treatment. Many fragments are not detectable by LIF (or REMPI) and, of those that are, in many cases their spectroscopy is insufficiently well characterized to allow extraction of reliable quantum state population distributions. Section IV.E describes various forms of an alternative, more general, experimental strategy which goes some way to overcoming these limitations.

E. Photofragment Translational Spectroscopy (PTS) and Ion Imaging

Consider the case of a photoexcited molecule (AB) dissociating into two fragments (A and B, which can be atoms or molecular entities). PTS relies on measurement of the velocity, $\underline{v}$, (and thus kinetic energy, KE) of one known fragment (e.g., A). Energy and linear momentum must be conserved during the dissociation, thus knowledge of $\underline{v}_A$ automatically gives the velocity of the partner fragment B and thus the total recoil energy, E_{trans}. This, in turn, can provide clues as to the internal energy partitioning within A and B via the energy conservation requirement

$$E_{photon} + E_{parent} = \boldsymbol{D}_0(\text{A—B}) + E_{trans} + E_{int}(\text{A}) + E_{int}(\text{B}). \quad (3)$$

E_{photon} and E_{parent} are, respectively, the photon energy and the internal energy present in the parent molecule prior to photoexcitation. The latter is minimized by working with jet-cooled AB molecules in a supersonic molecular beam. $\boldsymbol{D}_0(\text{A}-\text{B})$ is the dissociation energy of the breaking bond, and $E_{int}(\text{A})$ and $E_{int}(\text{B})$ are the internal energies of the

fragments. Three methods of measuring fragment recoil velocities find common usage.

The first, ***Doppler spectroscopy***, is again restricted to the subset of possible fragments that can be monitored by LIF or REMPI. It relies on the fact that the frequency, ν, of light absorbed by a fragment moving with a velocity $\underline{v}_z$ relative to a probing light source differs from that absorbed by the same fragment if stationary (ν_0) by an amount

$$\Delta\nu = \nu - \nu_0 = \pm\nu_0 \frac{v_z}{c}, \tag{4}$$

where c is the speed of light in vacuum. To excite a transition in a fragment moving toward the probe laser source, it is necessary to tune to a slightly lower frequency than if the same fragment was moving in the opposite direction. An ensemble of photofragments will exhibit a spread of absorption frequencies, reflecting their recoil velocity distribution. How well this distribution can be determined depends on the magnitude of the associated Doppler shift and how these compare with other factors contributing to the measured lineshapes (e.g., the spread of initial parent velocities, the probe laser bandwidth, etc.). The characterization also depends on the form of the recoil velocity distribution and, in particular, its spatial anisotropy. This anisotropy arises because the probability of an electric dipole allowed excitation is proportional to the square of the scalar product $\underline{\mu}.\underline{\varepsilon}$. Thus, even though the original sample of parent molecules will normally be randomly oriented in space, the incident radiation will interact selectively with the subset that happens to be lying such that their transition dipole $\underline{\mu}$ is parallel to the electric vector $\underline{\varepsilon}$ of the photolysis radiation. As a result, the photoexcited molecules will be aligned in the laboratory frame with a distribution proportional to the square of the cosine of the angle, θ, between $\underline{\mu}$ and $\underline{\varepsilon}$. If these excited molecules dissociate rapidly (i.e., on a timescale that is fast relative to their period of rotation), then this alignment will normally reveal itself as an anisotropic distribution of recoiling fragments. The angular distribution function will have the general form

$$I(\theta) \propto 1 + \frac{\beta}{2}(3\cos^2\theta - 1), \tag{5}$$

where β is the anisotropy parameter. β takes the limiting values of $+2$ (i.e., $I(\theta) \propto \cos^2\theta$) in the case that $\underline{v}_A$ lies parallel to $\underline{\mu}$ and -1 when $\underline{v}_A$ and $\underline{\mu}$ are perpendicular (i.e., $I(\theta) \propto \sin^2\theta$). An isotropic distribution of recoiling fragments is characterized by $\beta = 0$. Figure 7 illustrates the way in which photofragment Doppler lineshapes can vary with β for the case that the photolysis and probe laser beams are propagating at 90° to each other, that the $\underline{\varepsilon}$ vector of the photolysis laser is in the plane defined by the two laser beams, and that the fragment has a single speed,

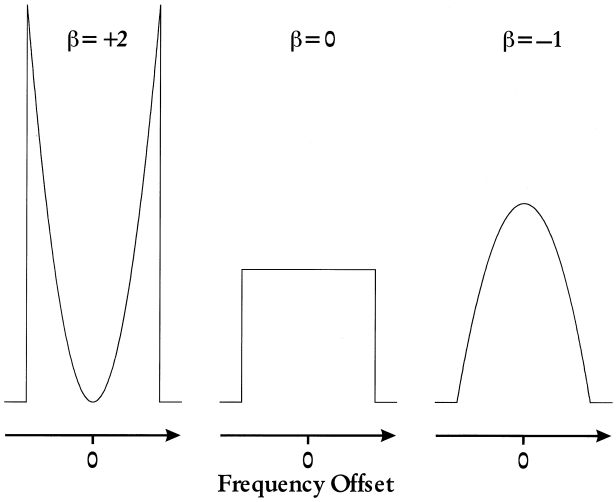


FIGURE 7 Doppler profiles predicted for a photofragment recoiling with a single speed, v, but different degrees of spatial anisotropy characterized by, respectively, $\beta = 2, 0,$ and -1 (see Eq. 5), given that the photolysis and probe lasers propagate along orthogonal axes and that the $\underline{\varepsilon}$ vector of the photolysis laser lies in the plane in which these two beams propagate.

v_A, and is probed by LIF. In most instances, however, fragment A will be formed with a spread of velocities (reflecting the range of possible E_{int} values of the partner B fragment, and any spread of initial velocities and internal energies of the parent molecule). The measured lineshape may be further complicated if there is any correlation between $\underline{v}_A$ and $\underline{J}_A$ [the fragment rotational (or electronic, in the case of an atom) angular momentum]; determining the fragment recoil velocity distribution, unambiguously, from Doppler lineshape measurements can be a nontrivial exercise.

Other ***photofragment translational spectroscopy*** techniques rely on more direct temporally or spatially resolved measurements of the velocities of the recoiling products. Consider a jet-cooled molecular beam of AB molecules excited with a pulse of monochromatic radiation at a well-defined point in space and time. If the energy absorbed exceeds the bond strength D_0(A−B), A and B fragments can be formed. The velocities with which these recoil from the interaction region will be determined by the combined requirements that energy and momentum be conserved; these can be determined by measuring the times taken for the fragment of interest to recoil a well-defined distance, without suffering any collisions, to a detector. In the traditional PTS experiment, shown in Fig. 8a, the photolysis beam is arranged to intercept the molecular beam and induce photodissociation. The small fraction of the resulting products that recoil in the appropriate direction will reach a detector [e.g., the source region of a quadrupole mass spectrometer (QMS)] located a known distance, L, from

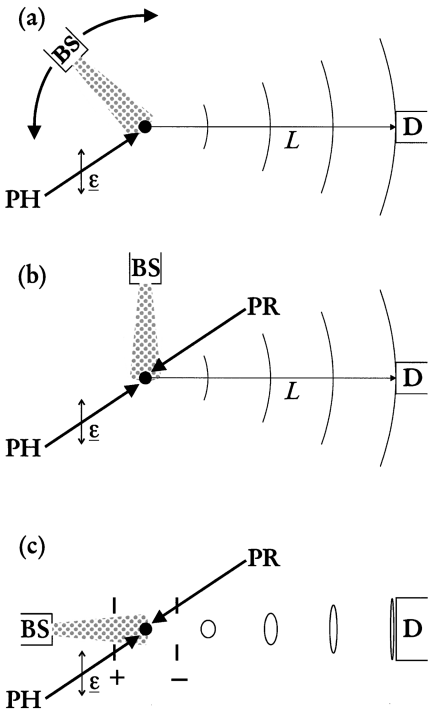


FIGURE 8 Schematics of the various forms of photofragment translational spectroscopy: (a) traditional set-up involving a rotatable molecular beam source (BS) and a detector (D)—in this case a quadrupole mass spectrometer (QMS); (b) Rydberg tagging method in which the fragment of interest (usually an H or D atom) is excited to a high n Rydberg state at source and field ionized after recoiling to the detector (a particle multiplier); (c) ion imaging in which the fragment of interest is ionized, at source, by REMPI and the resulting ion cloud is extracted with appropriate ion optics so as to impact on a time and position sensitive detector (microchannel plate/phosphor screen assembly). PH and PR in this figure indicate photolysis and probe laser beams, respectively. The former is shown with its polarization vector $\underline{\varepsilon}$ aligned perpendicular to the detection axis. The thin lines indicate the way in which the fragments of interest expand from the interaction region, ●, with time after the photolysis pulse.

the interaction volume. Electron impact will ionize some fraction of these products, allowing time-of-arrival measurements of fragments of any chosen mass. Such an apparatus is often referred to as a "universal" detector, reflecting the generality of the detection method. However, this implementation of the technique offers only limited velocity (and thus KE) resolution—limited by the ratio $\Delta L/L$, where ΔL is the length of the ionization region—and is species, but not quantum state, specific. Photofragment translational energy spectra measured in this way can, in favorable cases, and after the appropriate laboratory to center-of-mass frame transformation, provide a measure of the relevant bond dissociation energy and some information about the vibrational (though not the detailed rotational) energy disposal within the recoiling fragments. The angular distribution of the recoiling fragments can also be deduced by measuring the yield of the chosen product as a function of the angle between the polarization vector $\underline{\varepsilon}$ of the photolysis light and the detection axis. These kinds of information will usually suffice to distinguish between direct dissociations and slower predissociations and to reveal the presence (or otherwise) of any significant energy barrier in the dissociation coordinate.

Figure 8b shows a schematic of a high resolution version of the PTS experiment, which has been used in many studies of hydride molecule photolysis. The resulting H(D) atom products are excited to a high n Rydberg state (by two color, double resonant excitation), at source, before escaping the interaction volume. Once "tagged," the H/D photofragments continue to recoil with their nascent velocity distribution; those that happen to fly along the detection axis are field ionized immediately prior to detection by a particle multiplier. Again, angular distributions may be investigated simply by rotating the $\underline{\varepsilon}$ vector of the photolysis radiation relative to the detection axis. The method offers much enhanced resolution [since the flight path from the interaction volume to the detector is well defined) and signal to noise ratio [since each photolysis laser pulse gives rise to a complete time-of-flight (TOF) spectrum of the H(D) atom fragment]. This is illustrated by Fig. 9, which shows a TOF spectrum of H atoms resulting from photolysis of jet-cooled H_2O molecules at 121.6 nm and of the total kinetic energy release (TKER) spectrum that results if it is assumed that the partner fragment is

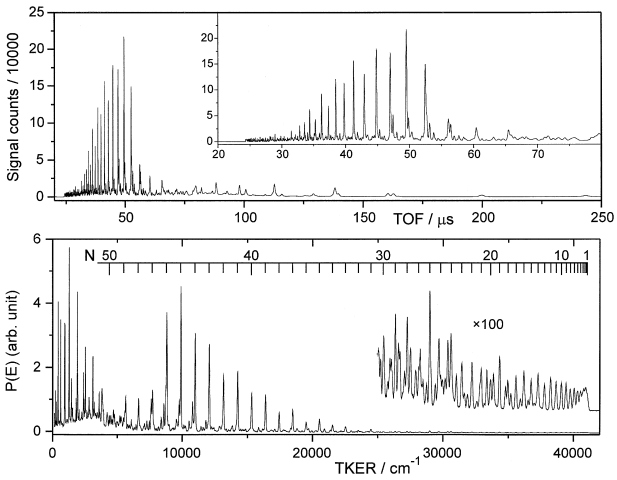


FIGURE 9 (a) H Rydberg atom TOF spectrum resulting from photolysis of jet-cooled H_2O molecules at 121.6 nm recorded with $\underline{\varepsilon}$ aligned parallel to the detection axis. (b) The corresponding TKER spectrum of the H + OH fragments, with the energies of the various rovibrational levels of the $OH(X^2\Pi)_{v=0}$ partner indicated; peaks at low TKER are associated with fragmentations leading to formation of rotationally excited $OH(A^2\Sigma^+)$ products. [Adapted with permission from Harich, S., *et al.* (2000). "Photodissociation of H_2O at 121.6 nm; a state-to-state picture," *J. Chem. Phys.* **113,** 10073–10090. Copyright, American Institute of Physics, 2000.]

the OH radical (i.e., has mass 17 amu). This is confirmed by the observation of clearly resolved features associated with individual rovibrational levels of OH in both spectra. Apart from providing a precise measure of the dissociation energy D_0(H—OH), analysis reveals that the detected H atoms are formed in conjunction with OH partners in both their ground ($X^2\Pi$) and first excited ($A^2\Sigma^+$) electronic states. In both cases these fragments are formed in a wide range of high rotational quantum states. Such behavior has been rationalized in the light of available *ab initio* PESs. The $\tilde{B}^1A_1$ state of H_2O—reached by photoabsorption at 121.6 nm—has a minimum energy configuration at linear geometries. Thus, the initial motion of H_2O molecules prepared on the $\tilde{B}$ state potential following Franck-Condon excitation from the (bent) ground state involves simultaneous H—OH bond extension and ∠HOH angle opening. The PESs for the $\tilde{B}^1A_1$ and ground $\tilde{X}^1A_1$ states are degenerate (a so-called conical intersection) at linear geometries and $R_{H-OH} \sim 1.67$ Å. Molecules that attain linearity at, or before, this amount of H—OH bond stretching can couple to one of two lower PESs and dissociate to ground state products. This coupling explains the observation that both electronically excited *and* ground state OH products are formed in this dissociation, while the strong angular anisotropy of the $\tilde{B}$ state potential accounts for the observed high levels of OH product rotation.

Ion imaging, illustrated in Fig. 8c, is the third PTS technique to be described. Again, a jet-cooled molecular beam of the parent under investigation is photolyzed by pulsed laser radiation. The nascent photofragments of interest are then ionized by REMPI, at source, using the output of a second pulsed laser focused into the interaction volume. The resulting ion cloud continues to expand with the speed and angular distributions characteristic of the initial photolysis event, but is simultaneously accelerated using carefully designed ion optics into a time-of-flight mass spectrometer (TOFMS), at the end of which it impinges on a microchannel plate plus phosphor screen detector assembly. The spot of phosphorescence associated with each ion impact is viewed with a charge coupled device (CCD) camera and accumulated (while scanning backwards and forewards across the full Doppler linewidth of the REMPI probe transition) to build up a complete ion image. This image is a squashed two-dimensional (2-D) projection of the full three-dimensional (3-D) velocity distribution of the nascent fragments; the latter distribution can be recovered using a mathematical (inverse Abel) transform.

Ion imaging offers many potential benefits. It is inherently sensitive. The method ensures species (by virtue of the measured ion TOF) and quantum state (by appropriate choice of probe wavelength) specificity and can be applied to any fragment amenable to REMPI probing. This

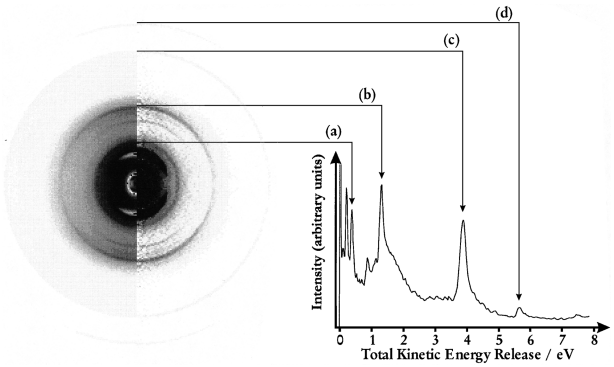


FIGURE 10 Raw (left half) and Abel inverted (right half) images of the $O(^3P_2)$ fragments arising in the 193-nm photoexcitation of jet-cooled O_2 molecules. Fragmentation channels giving rise to the various features in the image, evident in the TKER spectrum shown on the right, are identified as follows: (a) $O_2 \xrightarrow{1\times 226\,nm} O(^3P_2) + O(^3P)$, i.e., photolysis resulting from absorption of one REMPI probe photon; (b) $O_2 \xrightarrow{1\times 193\,nm} O(^3P_2) + O(^3P)$, the process of primary interest in this work; (c) $O_2 \xrightarrow{2\times 226\,nm} O(^3P_2) + O(^1D)$, two photon dissociation induced by the REMPI probe laser; (d) $O_2 \xrightarrow{2\times 193\,nm} O(^3P_2) + O(^1D)$. [Adapted with permission from Bakker, B. L. G., and Parker, D. H. (2000). "Photophysics of O_2 excited by tunable laser radiation around 193 nm," *J. Chem. Phys.* **112,** 4037–4044. Copyright, American Institute of Physics, 2000.]

is illustrated in Fig. 10, which shows both the raw image and a slice through the transformed 3-D distribution of the $O(^3P_2)$ fragments formed following 193 nm photodissociation of O_2 (i.e., within the Schumann-Runge absorption system) and the deduced TKER spectrum. The $O(^3P_2)$ products are probed by $2+1$ REMPI using photons of wavelength $\lambda \sim 226$ nm. The two smallest rings [i.e., the slowest moving $O(^3P_2)$ products] are attributable to predissociation of O_2 molecules to two ground state atoms, following absorption of one photon at, respectively, 226 and 193 nm. All of the larger diameter rings [i.e., faster $O(^3P_2)$ products] arise following multiphoton excitation of the O_2 molecules. Clearly, the image provides a particularly direct visualization of the various active dissociation channels and their relative importances; different velocity subgroups appear in the image with different radii, while the angular variation of the signal intensity in any one ring reveals the recoil anisotropy associated with that particular fragmentation channel. However, these images also serve to illustrate a limitation of the ion imaging method—its limited velocity (and KE) resolution. Our chosen example involves a diatomic molecule fragmenting into two atoms; the various product channels look well resolved but, even in this case, and for the slowest features, the image resolution is insufficient to provide information about the relative branching into the $J = 0$, 1, and 2 spin-orbit levels of the $O(^3P_J)$ partner to the monitored $O(^3P_2)$

atom. For polyatomic systems, where one or (in most cases) both of the resulting fragments are formed in a range of rovibrational states, the rings associated with the various product channels will rapidly merge together and the measured image will appear as an unresolved "blob."

V. CONCLUSIONS AND OUTLOOK

This article summarizes the various sources of VUV radiation available for photochemical studies and then focuses on some of the more probable fates that can befall an isolated gas phase molecule following absorption of a VUV photon and on how these can be investigated experimentally. It has also tried to emphasize the ever-growing importance of companion high level theory (*ab initio* electronic structure calculation of excited state PESs and the couplings between such surfaces, and studies of the nuclear dynamics thereon) in guiding the interpretation of such experimental observations.

The illustrative examples all involve small (diatomic and triatomic) molecules. Each additional atom adds another three vibrational degrees of freedom to the problem; the concomitant increase in the associated density of states (both in the parent molecule and in the products) soon precludes quantum state resolution in the experimental measurements and renders detailed theory impotent. The size of molecule by which experiment and theory are starting to struggle is still depressingly small. CH_4, for example, shows a featureless VUV absorption spectrum for which no definitive analysis yet exists. The H atoms resulting from Lyman-α photolysis of CH_4 have been studied by Doppler spectroscopy, by H (Rydberg) PTS methods, and by ion imaging. REMPI and ion imaging methods have also been used to gain some insight into the H_2 products that are eliminated from CH_4 following photo-excitation at the high energy end of the VUV region. Apart from the H_2 REMPI studies, these experiments provide little in the way of detailed quantum state resolved product distributions, but the deduced KE distributions have been interpreted (with the aid of *ab initio* calculations of various sections through the S_1 and T_1 PESs) in terms of (at least) three competing fragmentation pathways, viz. H_2 elimination and formation of singlet CH_2 radicals on the S_1 surface, and H atom elimination (leaving CH_3 radicals as the partner fragment) following both ISC to the T_1 PES and IC to high levels of the ground electronic state. Some of the resulting CH_3 radicals are formed with such high levels of internal excitation that they must be unstable with respect to further fragmentation; statistical considerations suggest that the major products of this secondary decay will be CH radicals and H_2 molecules. The observation of H atoms as a major dissociation product of the Lyman-α photolysis of CH_4 served to contradict the previous consensus view and has implications for the hydrocarbon balance in the atmospheres of some of the outer planets and their moons (notably Titan). Not surprisingly, our detailed understanding of the primary photochemistry of yet larger, heavier polyatomic molecules is even less complete.

Ab initio theory, and each of the experimental methods outlined above, can all be expected to continue to add to the body of available knowledge relating to molecular photochemistry induced by absorption of VUV photons. Ion imaging studies involving VUV photons are at this time still rare, but they can be expected to increase rapidly in number—so, too, can absorption and PHOFEX type studies involving the new generation of more intense synchrotrons. We are still awaiting the extension of ultrafast pump-probe type studies into the VUV spectral region. Whether such studies can shed much new light on the generally fast processes occurring in isolated gas phase molecules remains unclear, but they could gain in importance as scientists begin to focus on condensed phase problems, e.g., the photophysics of VUV excited molecules trapped in matrices.

ACKNOWLEDGMENT

The authors are grateful to the Engineering and Physical Sciences Research Council for consistent funding of much of the photochemistry work carried out in Bristol and for the award of a Senior Research Fellowship (MNRA) and a postgraduate studentship (PAC), and to Professor R.N. Dixon for preparing Fig. 9.

SEE ALSO THE FOLLOWING ARTICLES

MULTIPHOTON SPECTROSCOPY • NUCLEAR CHEMISTRY • PHOTOCHEMISTRY, MOLECULAR • POTENTIAL ENERGY SURFACES

BIBLIOGRAPHY

Ashfold, M. N. R., and Baggott, J. E., eds. (1987). "Molecular Photodissociation Dynamics," Royal Society of Chemistry, London.

Ashfold, M. N. R., Mordaunt, D. H., and Wilson, S. H. S. (1996). "Photodissociation dynamics of hydride molecules: H atom photofragment translational spectroscopy," *Adv. Photochem.* **21,** 217–295.

Dixon, R. N. (1994). "The dynamics of photodissociation," *Chem. Soc. Rev.* **23,** 375–385.

Heck, A. J. R., and Chandler, D. W. (1995). "Imaging techniques for the study of chemical reaction dynamics," *Annu. Rev. Phys. Chem.* **46,** 335–372.

Herzberg, G. (1966). "Molecular Spectra and Molecular Structure, III: Electronic Spectra and Electronic Structure of Polyatomic Molecules," Van Nostrand, Princeton, NJ.

Okabe, H. (1978). "Photochemistry of Small Molecules," Wiley-Interscience, New York.

Robin, M. B. (1974), "Higher Excited States of Polyatomic Molecules," Vols. 1 and 2; (1985). "Higher Excited States of Polyatomic Molecules," Vol. 3, Academic Press, New York.

Sandorfy, C., ed. (1999). "The Role of Rydberg States in Spectroscopy and Photochemistry," Kluwer Academic, Dordrecht/Norwell, MA.

Schinke, R. (1992). "Photodissociation Dynamics," Cambridge Univ. Press, Cambridge, UK.



Photochemistry, Molecular

Curt Wittig
University of Southern California

I. Electromagnetic Radiation
II. Interaction of Radiation with Matter
III. Photodissociation
IV. Radiationless Decay
V. Photoinitiated Unimolecular Decomposition
VI. Complex Photochemical Pathways

GLOSSARY

Absorption cross section Equates the capacity of a molecule to absorb photons to that of a perfectly absorbing disk whose surface normal is parallel to the **k** vector of the incident radiation; it is usually denoted σ and satisfies Beer's law: $I = I_0 \exp(-\sigma N x)$, where N is the concentration of the absorbing species.

Adiabatic potential Diabatic potentials of the same symmetry are coupled when the energies are close to one another (curve crossing), resulting in adiabatic potentials; the dominant electronic configurations are different on either side of the crossing region.

Born–Oppenheimer approximation Electronic wave functions and energies are calculated at a given geometry by assuming that the nuclei are stationary.

Curve crossing Potential curves intersect (cross) at locations in configuration space. Perturbation terms in the Hamiltonian cause avoided crossings for states of the same symmetry; there is no interaction for states of different symmetries.

Diabatic potential Obtained for a given dominant electronic configuration.

Double resonance The resonant excitation of one transition is required in order that a signal is obtained upon the resonant excitation of another transition.

Electric dipole interaction The interaction energy of an electric dipole moment in the presence of an external electric field.

Franck–Condon factor The square of the overlap integral of nuclear wave functions that belong to different electronic states.

Internal conversion Adiabatic potential surfaces of the same multiplicity are coupled via the nonadiabatic interaction that is due to the coupling of the nuclear and electron motions; this coupling is ignored in the Born–Oppenheimer approximation.

Intersystem crossing Coupling between potential energy surfaces of different multiplicity brought about by spin–orbit interaction.

Intramolecular vibrational redistribution Vibrational energy redistributes within a polyatomic molecule; there are no classical constants of motion or quantum mechanical good quantum numbers for the vibrational degrees of freedom.

Maxwell's equations Equations describing classical, as

opposed to quantum mechanical, electromagnetic radiation.

Photodissociation A molecule dissociates as a consequence of having absorbed one or more photons.

Potential energy surface The electronic energy of a molecule, for a given electronic state, as a function of the molecule's $3N-6$ internal (nonrotational) nuclear coordinates, where N is the number of atoms.

Predissociation A zeroth-order bound electronically excited state undergoes dissociation because it interacts with a repulsive potential.

Quantum fluctuations The quantum chaotic nature of the nuclear dynamics of highly vibrationally excited polyatomics causes the energy levels to behave erratically. This is true in the region of bound states as well as the resonances that lie above the reaction threshold.

Radiationless decay The intramolecular energy transfer process whereby a photoexcited electronic state can be said to transfer its excitation to nonradiative states, i.e., vibrational levels of lower electronic manifolds.

Repulsive potential A repulsive potential energy surface decreases monotonically along a coordinate that leads to fragments, ensuring dissociation; the fragments repel one another along this coordinate.

Resonance A quasi-bound level that has acquired a decay rate and associated linewidth which is usually Lorentzian.

Selection rules Rules that determine the allowed optical transitions for electric dipole, quadrupole, etc., transitions for different polarizations of the electromagnetic field.

Unimolecular decomposition Dissociation is treated statistically by invoking transition states that separate the molecular and product regions. Transition state levels constitute independent open channels, so the rate coefficient is equal to the number of open channels times the rate per open channel. By using statistical mechanics, this is calculated to be $1/h\rho$.

Wave packet A coherent superposition, usually of eigenstates or a convenient basis; the individual phases are such that the term packet is justified somewhere in the temporal evolution.

PHOTOCHEMISTRY, as its name implies, is the chemical change that results from the absorption of electromagnetic radiation, in quantized units of energy called photons, by material substances made of protons, neutrons, and electrons. It is ubiquitous in the world around us—always has been, always will be. It has played a central role in our planet's evolution, from the earliest emergence of primordial plants and animals to the present day, and it will continue to play a central role as long as life as we know it exists on earth. Its importance to the environment, our quality of life, numerous modern technologies, the health sciences, a broad range of scientific research areas, etc. cannot be overstated. All of this is due to the interplay that takes place between matter and electromagnetic waves that were once believed to travel in an ether.

The word photochemistry means different things to different people and has fuzzy boundaries, and applications so varied that practitioners often find little of a scientific nature to discuss. Graduate courses, even within the same academic department but with different instructors, often have no obvious relationship to one another, and conferences on organic photochemistry have little in common with their counterparts in physical chemistry. This is just the way things are in so broad a field.

This overview discusses fundamental aspects of the photochemistry of polyatomic molecules. This is only restrictive in that some phenomena are omitted which are *exclusive* to liquid and/or heterogeneous environments. The material covered ranges from the nature of the electromagnaetic radiation that drives photochemistry to the myriad processes that ensue following radiationless decay. To reinforce the principles, a few examples will be given of benchmark systems that have advanced our understanding of the detailed mechanisms.

I. ELECTROMAGNETIC RADIATION

Coulomb's law is the basis of the classical theory of electricity and magnetism. In the simplest case of a stationary point charge in vacuum, it states that an electric field $\mathbf{E}$ is directed radially outward from a charge Q in the direction $\hat{\mathbf{r}}$, where the caret denotes a unit vector, as shown in Fig. 1a. This is what a stationary observer perceives, for example, when measuring the force experienced by a test charge q. Indeed, this force is used to define the electric field: $\mathbf{F}=q\mathbf{E}$. The electric field due to Q is given in mks units by

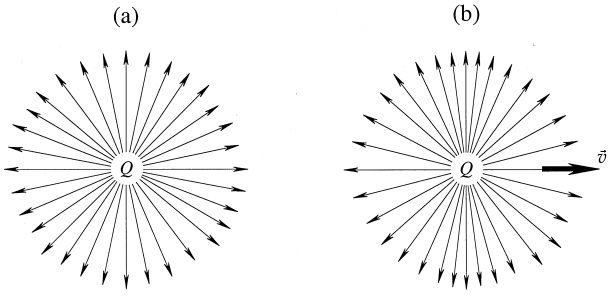


FIGURE 1 (a) The charge Q is stationary. The electric field lines, which point away from the origin when Q is positive, are equally spaced, denoting an isotropic distribution. (b) The charge moves with constant velocity $\mathbf{v}$; note that $\mathbf{v}$ is horizontal. The field lines are more dense in the vertical than the horizontal direction.

$$\mathbf{E} = \frac{Q}{4\pi\varepsilon_0 r^2}\hat{\mathbf{r}}, \tag{1}$$

where ε_0 is a constant called the vacuum permittivity. From Eq. (1) it is clear that $\mathbf{E}$ can also be derived from a potential Φ which is a function of coordinates:

$$\mathbf{E} = -\nabla\Phi, \tag{2}$$

where Φ is given by

$$\Phi = \frac{Q}{4\pi\varepsilon_0 r}. \tag{3}$$

This summarizes the situation for a charge Q at rest in the frame of the observer. In the above equations, no explicit account is taken of the fact that information travels from the charge Q to the observer at the speed of light. Namely, an observer located a distance d from the charge can only sense an effect that influences the charge with a delay time d/c. For a stationary charge, this temporal retardation does not enter.

If Q is moving with a constant velocity $\mathbf{v}$, the electric field seen by the observer depends on the observer's location relative to the charge as well as to the velocity $\mathbf{v}$. Figure 1b indicates the intensity of $\mathbf{E}$ by the density of the field lines. Note that $\mathbf{v}$ is directed horizontally. The field is most intense in the direction perpendicular to $\mathbf{v}$. This is a relativistic effect that depends on velocity; if v/c is small, Figs.1a and 1b look alike. Figure 1b is obtained by using the relativistic Lorentz transformation to obtain the field of the moving charge that is observed in the laboratory reference frame. No static charge distribution can give the field distribution shown in Fig. 1b.

If the charge Q responsible for $\mathbf{E}$ undergoes acceleration, the observer perceives that $\mathbf{E}$ is no longer directed radially outward from Q. It has acquired a component that is transverse to $\hat{\mathbf{r}}$ and is traveling outward at the speed of light.

In Fig. 2, the charge Q has been accelerated abruptly by an impulsive force. This impulsive force causes the field lines to change. At a time d/c after the disturbance, where d is the distance from the origin to the observer, the electric field seen by the observer bends, having acquired a component that is transverse to the radial direction. This transverse field propagates radially outward at the speed of light. In addition, the transverse field varies as r^{-1}, as opposed to the r^{-2} variation of the radially directed electric field given by Eq. (1). A common situation is sinusoidal acceleration of the radiating charges, in which case the transverse field oscillates sinusoidally.

In addition to the above effect on $\mathbf{E}$, any moving charge, whether it is accelerated or not, creates a magnetic field, $\mathbf{H}$. Magnetism is also relativistic. The magnetic field is an electric field that arises when the Lorentz transformation

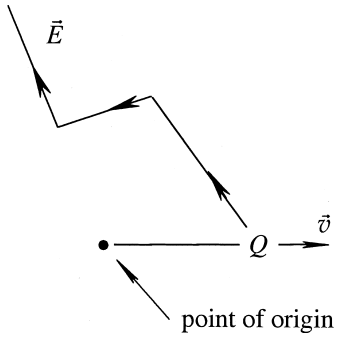


FIGURE 2 The charge was initially at rest at its point of origin when an impulsive force caused its velocity to increase to $\mathbf{v}$. At a given distance from the origin, the electric field was initally directed radially outward from the origin. After the charge's acceleration, its field lines are not isotropic in the laboratory reference system, as was also the case in Fig. 1b. In following the field lines, it is seen that the transverse component travels outward at the speed of light.

is used to calculate the effect of a moving charge density observed in the laboratory reference frame. As in the case of $\mathbf{E}$, when Q is accelerated, the magnetic field $\mathbf{H}$ also acquires a transverse component that varies as r^{-1}. The transverse $\mathbf{E}$ and $\mathbf{H}$ fields are perpendicular to one another.

Thus, the time-varying $\mathbf{E}$ and $\mathbf{H}$ fields are related to one another and together they contain, in an electromagnetic wave, the energy that has been spent in accelerating the charge. The outgoing energy flux (i.e., power per unit area) is given by the Poynting vector, $\mathbf{P}$:

$$\mathbf{P} = \mathbf{E} \times \mathbf{H} \tag{4}$$

whose propagation direction $\hat{\mathbf{k}}$ is perpendicular to both $\mathbf{E}$ and $\mathbf{H}$. Thus, in classical electromagnetic theory, all radiation originates from charged matter that has undergone acceleration. The result is a set of time-varying electric and magnetic fields that satisfy Maxwell's equations for a source-free region:

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}, \tag{5}$$

$$\nabla \times \mathbf{H} = \frac{\partial \mathbf{D}}{\partial t}, \tag{6}$$

where $\mathbf{B} = \mu_0\mathbf{H}$ is the magnetic flux density, the constant μ_0 is called the vacuum magnetic permeability, and $\mathbf{D} = \varepsilon_0\mathbf{E}$ is the electric flux density. Equations (5) and (6) are easily converted into standard forms for the wave equations for the propagating vector fields:

$$\nabla^2\mathbf{E} - \frac{1}{c^2}\frac{\partial^2\mathbf{E}}{\partial t^2} = 0, \tag{7}$$

$$\nabla^2\mathbf{H} - \frac{\partial^2\mathbf{H}}{\partial t^2} = 0, \tag{8}$$

where ∇^2 is the Laplacian operator and $c^2 = 1/\mu_0\varepsilon_0$, where c is the speed of light in vacuum. It is interesting that Maxwell's equations were obtained originally by combining empirical observations for electric and magnetic forces, whose common origin was not appreciated at the time of Maxwell's work. The resulting equations, which are relativistically correct, preceded the theory of special relativity by several decades. Equations (7) and (8) describe all classical electromagnetic radiation, ranging from that which emanates from a blackbody source to the output from the highest resolution laser.

In photochemical applications, the fields come in all varieties. They can be unpolarized, linearly polarized, circularly polarized, or elliptically polarized. For example, a common type of linearly polarized electric field propagating in the $\hat{\mathbf{k}}$ direction is given by

$$\mathbf{E} = E_0\hat{\mathbf{z}} \exp i(\mathbf{k}\cdot\mathbf{r} - \omega t), \tag{9}$$

where ω is the angular frequency of the oscillating field, $k = 2\pi/\lambda$ is the magnitude of the wave vector $\mathbf{k}$, which is perpendicular to $\hat{\mathbf{z}}$, and λ is the wavelength of the radiation. This is called a uniform plane wave; it has no spatial variation in the plane perpendicular to $\mathbf{k}$. It is widely used in descriptions of photochemical phenomena.

In the classical theory summarized above, the radiation is not organized into quanta of energy called photons. When this is done, the energy per unit volume is $h\nu(n + \frac{1}{2})$, where $\frac{1}{2}h\nu$ is the zero-point energy of the vacuum electromagnetic field whose quantum is $h\nu$, h is Planck's constant, and ν is the frequency.

In addition, it is found that the photons have an intrinsic angular momentum or spin, which arises because the photon wave function is a vector field. This intrinsic property is analogous to the intrinsic spins of electrons, protons, and neutrons, which arise because of the relativistic requirement that space and time be treated on an equal footing. The angular momentum of the photon plays a key role in determining the selection rules for photoexcited transitions, as discussed in the next section.

II. INTERACTION OF RADIATION WITH MATTER

Electromagnetic radiation of an appropriate frequency ν that is incident on matter can be partially absorbed, thereby setting the stage for photochemistry. Figure 3 describes this at the most elementary level: a photon of energy $h\nu$ is annihilated by an absorbing molecule, creating an excited state in the process. The simple picture given in Fig. 3 does not take into account the fate of the excited state thus created. This fate is the essence of photochemistry.

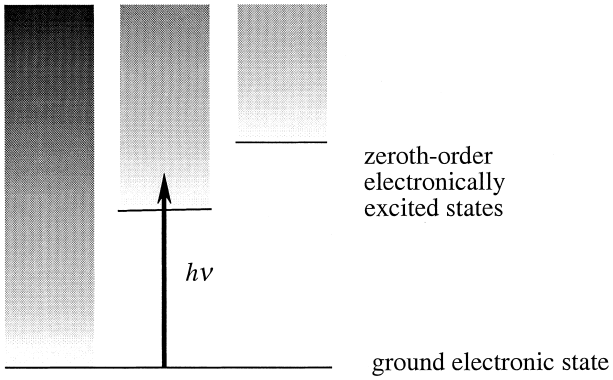


FIGURE 3 Photoexcitation is used to prepare zeroth-order electronically excited states, which subsequently react via one or more mechanisms. Shading indicates variation of the vibrational level density.

There are many different kinds of excited states associated with liquids, solids, and molecules, and photochemistry can occur via a number of different mechanisms and with strikingly different results. For example, a molecule that absorbs an ultraviolet photon of sufficiently high frequency ν has acquired enough internal energy to dissociate. Indeed, some do so rapidly, whereas others do so slowly, and yet others not at all.

For photochemistry to occur, the photon energy $h\nu$ must find its way into a form of energy that promotes chemical change. This may be direct, for example, rapid photodissociation, or indirect, requiring collective nuclear motions and much longer time scales. The former is the most conceptually straightforward, while the latter accounts for most of the photochemistry one is likely to encounter. The corresponding mechanisms are the scientific bases for the numerous phenomena that lie within the domain of photochemistry. They will be identified and discussed below.

The radiation–matter interaction that is responsible for the absorption of a portion of the incident electromagnetic radiation nearly always involves the oscillating electric field component of the radiation. The total energy density of the electromagnetic radiation that propagates in vacuum is made up of equal amounts of energy stored in the electric and magnetic fields, i.e., their time-averaged energy densities, $\varepsilon_0 E^2/2$ and $\mu_0 H^2/2$, are equal. However, the electric field transfers energy to the molecule's charges, i.e., essentially the electrons, much more efficiently than does the magnetic field. This will be the case (i) as long as there is no symmetry restriction that prevents the electric field from inducing a transition except via a very weak interaction and (ii) as long as v/c is small, where v is the characteristic electron speed and c is the speed of light. If there is a symmetry restriction, the magnetic field may win by default.

The ratio v/c is equal to the ratio of the magnitudes of the magnetic and electric forces ($q\mathbf{v} \times \mathbf{B}$ and $q\mathbf{E}$, respectively) experienced by a particle of charge q in the presence of electromagnetic radiation. The small ratio of the magnitudes of the magnetic and electric forces carries over to electrons that occupy molecular orbitals. This is why the electrons available for a transition are excited more efficiently by the electric field part of the incident radiation than by the accompanying magnetic field. The condition $v \ll c$ is satisfied for all cases of interest in the field of photochemistry.

The most common interaction of a molecule with an external oscillating electric field that induces optical transitions is referred to as an *electric dipole interaction*. The term dipole comes from an expansion, in this case of the charge distribution, in terms of what are called multipole moments: monopole, dipole, quadrupole, octupole, etc., as shown in Fig. 4. The simplest dipole moment is that of two point charges, $\pm Q$, separated by $\mathbf{r}$; it is equal to $Q\mathbf{r}$.

The expression for the energy of an electric dipole moment in an external electric field is $-\boldsymbol{\mu} \cdot \mathbf{E}$, where $\boldsymbol{\mu}$ is the dipole moment. It is common practice to introduce this $-\boldsymbol{\mu} \cdot \mathbf{E}$ term into the Schrödinger equation *ad hoc*. This is intuitive and just as accurate as the more formal approach, in which the Hamiltonian is constructed by using the canonically conjugate momentum, $\mathbf{p} - e\mathbf{A}/c$, where $\mathbf{p} = -i\hbar\boldsymbol{\nabla}$ and $\mathbf{A}$ is the magnetic vector potential.

The appropriate matrix element between an initial state $|i\rangle$ and a final state $|f\rangle$ of the molecule is

$$\langle f|\boldsymbol{\mu} \cdot \mathbf{E}|i\rangle, \tag{10}$$

where the minus sign is understood. In the event that the radiation field is treated quantum mechanically, this matrix element does not vanish when the radiation field is in its lowest state and the energy of the state $|i\rangle$ is higher than that of $|f\rangle$. The quantized radiation field is analogous to a mechanical harmonic oscillator. Its lowest level has a zero-point energy, which stymied theoretical physics for years because it appears to lead to an infinite energy when summing over the electromagnetic modes with $0 \leq \nu \leq \infty$. This zero-point level stimulates a transition from the excited state $|i\rangle$ to the lower state $|f\rangle$, thereby creating a photon, a process called spontaneous emission. This is an important *physical* process, but from the perspective of photochemistry, it is unimportant because if the system gives up its energy by emitting a photon, nothing much has happened. Thus, our considerations will be directed toward *nonradiative* processes. In this regard, note that though the photon is used to create an electronic excitation, its ultimate purpose is to produce mechanical and chemical work by moving the nuclei.

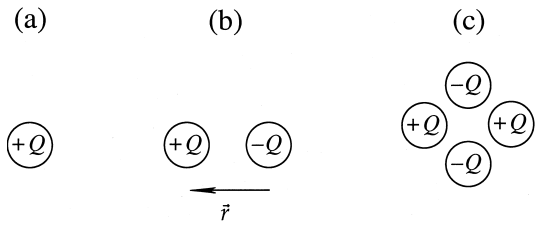


FIGURE 4 (a) The monopole moment is a scalar, i.e., the net charge Q. (b) The dipole moment is a vector, i.e., Q times $\mathbf{r}$; there is no net charge. (c) The quadrupole moment is a tensor; there is no net dipole moment and no unique geometry, e.g., in the CO_2 molecule, the quadrupole moment is due to oppositely directed dipoles that lie along the molecular axis.

In dealing with photochemical phenomena, it is sensible to treat the electric field classically and external to the collection of particles, and rewrite the matrix element given in Eq. (10) as

$$\langle f|\boldsymbol{\mu}|i\rangle \cdot \mathbf{E} \tag{11}$$

where $\langle f|\boldsymbol{\mu}|i\rangle$ is the transition dipole moment matrix element of the molecule. Note that this differs from the permanent dipole moment of a molecule in a state $|j\rangle$, which is given by $\langle j|\boldsymbol{\mu}|j\rangle$. The transition dipole moment is a measure of the charge redistribution that takes place when a transition occurs, whereas the permanent dipole moment reflects the charge distribution in a given state.

In order to evaluate Eq. (11), $\boldsymbol{\mu}$ and $\mathbf{E}$, which are most easily visualized in the molecular and laboratory frames, respectively, should be referenced to a common set of axes. To achieve this, one usually transforms the electric field to a molecule-fixed set of axes. After doing this, it is straightforward to derive a set of selection rules that dictate the excited states that can be accessed from a molecule in a given initial state, which in the present context is usually its ground electronic state. The selection rules also depend on the polarization of the radiation, for example, linear versus circular, but this affects only the spatial alignment and orientation of the excited state relative to the laboratory reference frame, which has nothing to do *per se* with the system's photochemistry.

The selection rules for changes in the total angular momentum and its projection on a space-fixed axis are rigorous for an isolated species in field-free space. These are obtained by using quantum mechanical angular momentum addition, i.e., 3-j symbols, etc. The photon angular momentum is added to the initial-state angular momentum to obtain the possible final states. Nuclear spins can be neglected because the time scales for dynamical processes that involve them is usually much longer than that of photochemistry. Thus, the angular momentum selection rules are $\Delta J = 0, \pm 1$, and $\Delta M = 0, \pm 1$, where J is the total angular momentum not counting the nuclear spins, and M is its projection on a space-fixed axis. Selection

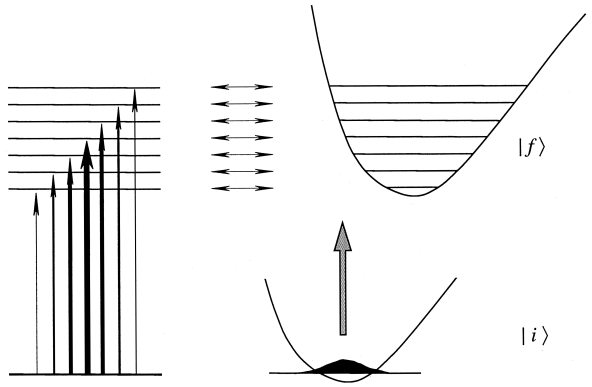


FIGURE 5 On the right, the ground and excited PESs are displaced from one another along one of the molecule's coordinates. The Franck–Condon factors weight the transition strengths, favoring vertical excitation, as denoted by the thicknesses of the lines on the left.

rules that govern the changes of the occupied molecular orbitals of the electrons are obtained straightforwardly by using group-theoretic methods.

In electronic transitions, there are no selection rules that govern the possible changes of the vibrational levels. Instead, there are propensities which can be understood by noting that the nuclei prefer to not move during an electronic transition, as indicated in Fig. 5. In evaluating Eq. (11), the fact that $\boldsymbol{\mu}$ depends only on electron coordinates enables a useful approximation to be introduced. With $|i\rangle$ and $|j\rangle$ written as products of electronic and vibrational wave functions, the $\langle f|\boldsymbol{\mu}|i\rangle$ matrix element in Eq. (11) becomes:

$$\left\langle \chi_f^n \middle| \left\langle \psi_f^e \middle| \boldsymbol{\mu}_e \middle| \psi_i^e \right\rangle \middle| \chi_i^m \right\rangle, \tag{12}$$

where $|\chi_f^n\rangle$ is the nth vibrational level of the final electronic state, $|\psi_f^e\rangle$ is the final electronic state, f is replaced by i for the initial electronic state, and $\boldsymbol{\mu}_e$ operates only on electron coordinates. Thus, Eq. (12) can be written

$$\left\langle \chi_f^n \mid \chi_i^m \right\rangle \left\langle \psi_f^e \middle| \boldsymbol{\mu}_e \middle| \psi_i^e \right\rangle. \tag{13}$$

Squaring this yields the Franck–Condon factors:

$$\left| \left\langle \chi_f^n \mid \chi_i^m \right\rangle \right|^2. \tag{14}$$

The Franck–Condon factors are observed as progressions in electronic spectra when the zeroth-order electronically excited state is bound and there is a geometry change between the ground and electronically excited states. The transfer of an electron from one molecular orbital to another that accompanies an electronic transition usually changes the character and strength of certain bonds. Therefore their equilibrium lengths and angles on the excited potential energy surfaces (PESs) are expected to differ significantly from those of their ground-state counterparts. Indeed, these are pronounced effects that enable a broad range of excited-state energies to be prepared via photoexcitation. Were the electronically excited-state geometry the same as that of the ground electronic state, the electronic absorption spectrum would have a narrow spectral width. This is rarely the case. The broad spectral widths of electronic absorption spectra are due to the geometry changes and the corresponding Franck–Condon factors.

The principles given above also apply to repulsive potential curves. In this case, the Franck–Condon overlap integrals are between the ground electronic state vibrational wave functions $|\chi_i^m\rangle$ and the continuum wave functions, which contribute to the overlap integrals near the classical turning point of the repulsive potential. This results in the ground electronic state vibrational wave function being projected onto the turning point region. In a polyatomic molecule in which the potential along one of the coordinates is repulsive, the Franck–Condon factors take *all* nuclear degrees of freedom into account, i.e., bound and continuum. Thus, upon dissociation, the electronically excited state can yield vibrationally excited fragments.

In addition to electric dipole transitions, there also exist transitions that are due to the higher multipole moments, i.e., quadrupole, octupole, etc. Relative to the electric dipole transitions, the quadrupole transitions are typically 10^4 times weaker. Likewise, magnetic dipole transitions are also typically 10^4 times weaker. Thus, in photochemical systems, we need only consider electric dipole transitions.

In so mature a field, it is inevitable that a number of theoretical formalisms are available for describing photoexcitation. One of the simplest uses time-dependent perturbation theory to calculate the probability per unit time that a final state appears in the total wave function, Ψ. The $-\boldsymbol{\mu} \cdot \mathbf{E}$ perturbation causes the initial and final states $|i\rangle$ and $|f\rangle$ to both appear in Ψ.

This approach can be used to calculate the frequency-dependent absorption cross section $\sigma_{\text{abs}}(\nu)$. This is the effective cross-sectional area that a molecule presents to the incident radiation. As far as the incident radiation is concerned, the molecule is like a black area that absorbs all of the radiation incident on it. Thus, the transmission of radiation through a gas consisting of these absorbers is given by Beer's law:

$$I(\nu) = I_0(\nu) \exp(-\sigma_{\text{abs}} N x), \tag{15}$$

where $I(\nu)$ and $I_0(\nu)$ are the transmitted and incident radiation intensities, respectively, N is the number density of the absorbers, and x is the length of the absorbing medium. This expression is valid as long as the intensity $I_0(\nu)$ is low enough to avoid a phenomenon called saturation. This occurs when the concentration of absorbers N is depleted. The expression for $\sigma_{\text{abs}}(\nu)$ is

$$\sigma_{\text{abs}}(\nu) = \frac{\pi}{\hbar \varepsilon_0 c} \nu_{fi}\, \delta(\nu_{fi} - \nu) |\langle f|\boldsymbol{\mu}|i\rangle \cdot \hat{\mathbf{e}}|^2, \qquad (16)$$

where δ denotes the Dirac delta function and $\hat{\mathbf{e}}$ is the unit vector along the direction of $\mathbf{E}$.

Absorption cross sections are widely used in photochemistry for calculations of photoexcitation rates. Note that σ_{abs} is not limited to the size of the molecule. For example, transitions in atoms frequently have σ_{abs} values exceeding 10^5 Å^2 when $\sigma_{\text{abs}}(\nu)$ is integrated over the absorption linewidth. The atom acts as a small receiving antenna. Its effective dimension for capturing radiation can be the order of the wavelength of the radiation.

Molecular σ_{abs} values are smaller than their atomic counterparts when integrated over the same spectral width as for an atom; 1–10 Å^2 is considered respectable. The charge does not move as far and the absorption strength is spread over the many levels of the nuclear degrees of freedom. Absorption cross sections are often reported incorrectly because proper account is not taken of the spectral distribution of the radiation source. If this is broader than the molecule's absorption line shape, measuring I/I_0 and using Eq. (15) will underestimate the cross section.

When using laser excitation, it is possible to overcome rather small σ_{abs} values because of the high radiation intensities that are currently available. It is commonplace to transport a large fraction of the population in a given level, despite σ_{abs} values as low as $\sim 10^{-7}$ Å^2.

In addition to the simple picture presented above there are many coherent phenomena that arise when considering cases in which $\mathbf{E}$ is carefully controlled. However, in photochemical applications, such phenomena are of minor importance.

Spectral Regions

Photochemistry that occurs in nature at or near the surface of the earth is due to the solar flux that reaches there. The spectral distribution of this radiation extends from about 350 nm to longer wavelengths. Shorter wavelengths are present high in the atmosphere, for example, causing O_2 photodissociation, which results in the protective ozone layer that is present in the stratosphere. Most photochemistry that occurs in nature at or near the surface of the earth is the result of visible or ultraviolet radiation; only in a minority of cases is near-infrared radiation involved. Infrared radiation at wavelengths that are characteristic of fundamental or overtone vibrations plays essentially no role in the photochemistry of natural systems, though the importance of such wavelengths in the research laboratory is another matter, as discussed in a later section.

In the research laboratory, photochemistry is assured at sufficiently short wavelengths. However, for small polyatomic molecules consisting of light nuclei, *i.e.*, exactly the systems that are most amenable to high-level theoretical modeling, the required wavelengths are often in the vacuum ultraviolet, where tunable radiation is not easily obtained. Thus, laboratory studies of benchmark polyatomic molecules have been guided in large part by the need to match absorption spectra with available laser sources. Figure 6 indicates a number of molecules whose study over the last 20 years have contributed greatly to our understanding of small-molecule photochemistry. It is no coincidence that tunable laser frequencies were available for all of these systems.

III. PHOTODISSOCIATION

Direct photodissociation via a repulsive potential curve is the most conceptually straightforward of all photochemical events. Figure 7 indicates schematically photoexcitation in which $-\boldsymbol{\mu} \cdot \mathbf{E}$ operating on one or more low-lying levels creates amplitude on a dissociative excited potential. This excitation is said to be vertical in the sense that the nuclei prefer to remain stationary during the transition. In the case of a diatom, the coordinate r is

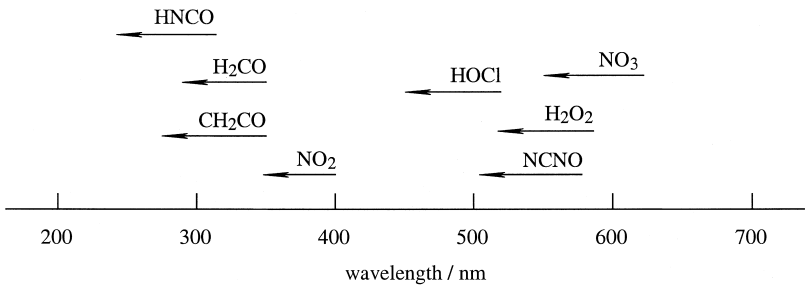


FIGURE 6 Some small polyatomics whose complex photochemistries have been studied in detail. The origins of the arrows indicate reaction thresholds. Except for HOCl and H_2O_2, which are excited by the sequential pumping of overtone transitions, these systems all undergo photochemistry via radiationless decay. Tunable laser frequencies were available for all of these systems.

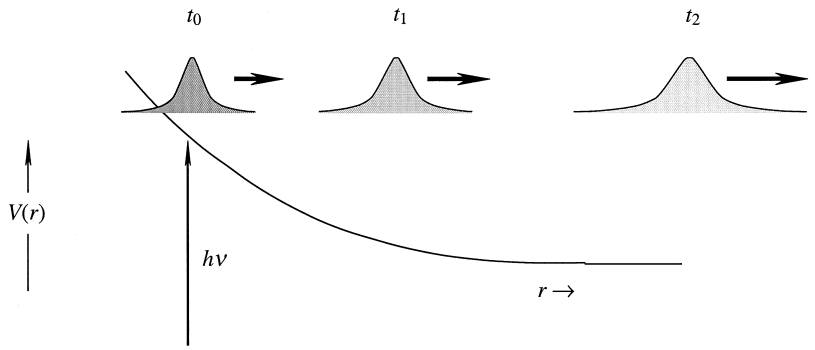


FIGURE 7 Photoexcitation to a repulsive potential. A pulse of radiation of center frequency ν has a short temporal duration and therefore a large linewidth. It creates a wave packet at t_0 near the turning point on the potential. The wave packet evolves rapidly toward large r, spreading as it travels.

the interatomic separation. For polyatomic species, at short distances, the reaction coordinate r might involve simultaneously the positions of several nuclei, while at large distances, it is the interfragment separation.

Dissociation on the excited potential is easily expressed from a time-domain perspective. For example, referring to Fig. 7, consider a photoexcitation process of sufficiently short temporal duration that a spatially localized wave packet is created near the turning point on the repulsive curve. This can be achieved by using ultrafast lasers whose pulse durations are tens of femtoseconds. In this case, the wave packet moves out of the initial excitation region quickly, on time scales of $\sim 10^{-13}$ sec when one of the fragments is small, and on somewhat longer time scales when neither of the fragments is small. Because of the brevity of the overall process, there is insufficient time for the energy implanted in the molecule by the photon to be randomized, i.e., distributed statistically among the molecule's internal degrees of freedom. The product excitations are said to be dynamically biased. Thus, analyses can provide information concerning the photoexcitation and fragmentation processes.

Alternatively, one can imagine a high-resolution experiment carried out by using narrow-linewidth laser radiation. In this case, the excited-state wave function created by photoexcitation is not localized near the turning point on the excited potential. To the contrary, it extends well into the large-r continuum region. As the laser linewidth is made smaller and smaller, the excited-state wave function extends further and further into the continuum. The spectral width of the radiation is too small to create a wave packet localized in the molecular region.

In this high-resolution experiment, the absorption spectrum can be obtained by tuning the laser frequency ν. In so doing, the occupied vibrational levels of the ground PES are mapped onto the repulsive potential curve. For the ground vibrational level, $v = 0$, where v is the ground PES vibrational quantum number for the coordinate shown in Fig. 7, a broad Gaussian-like absorption spectrum is observed: the steeper the repulsive curve, the broader the spectrum. For absorption from the $v = 1$ level, the spectrum has two broad peaks, corresponding to the peaks in the probability density, and so forth for the higher v levels.

Direct photodissociation, for example, as depicted in Fig. 7, is available to all molecules via their repulsive potentials, albeit with the inclusion of some additional complexities discussed below. However, in order to access a molecule's repulsive potentials, the wavelength region must be chosen appropriately, meaning that in many cases the vacuum ultraviolet region is unavoidable.

Prompt dissociation propels fragments into various directions in the laboratory with selectivity, i.e., the fragments are distributed anisotropically in the laboratory reference frame. When the electric field $\mathbf{E}$ of the photolyzing radiation is linearly polarized, its direction in the laboratory is fixed, providing a particularly convenient reference direction.

Figure 8 depicts alignments that derive from the $-\boldsymbol{\mu} \cdot \mathbf{E}$ interaction in the classical limit. In the case of a linear molecule, the transition dipole moment is either parallel or perpendicular to the molecule's axis. Thus, photoexcitation yields distributions of axes that are proportional to $\cos^2 \theta$ and $\sin^2 \theta$ for the parallel and perpendicular transitions, respectively. Axial forces result in recoil velocities being imparted along the directions of the axes. This simple picture is compromised if the molecule rotates during dissociation, thereby lessening the fragment anisotropy. In addition, the anisotropy is lessened by parent bending excitation, even zero-point, which produces velocities that have components which are transverse to the recoil direction.

With diatomic molecules, the atomic fragments are produced in electronic states, the lowest lying of which are the fine structure states due to spin–orbit interaction. For

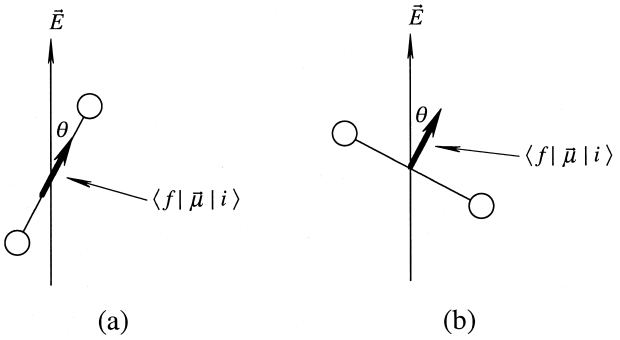


FIGURE 8 Classical picture of parent alignment in direct photodissociation. (a, b) The transition dipole moments are parallel and perpendicular to the molecular axis, respectively. The electric dipole interaction aligns the transition dipole moment to the electric field, thereby creating an anisotropic distribution of excited-state molecules which dissociate via a repulsive potential. The fragments depart axially, producing anisotropic distributions of fragments in the laboratory. (a) Fragments are peaked in the direction of the field; (b) Fragments peak in directions perpendicular to the field.

example, the area of *hot-atom chemistry* exploits direct photodissociation as a means of producing translationally hot atoms, mainly by photodissociating hydrogen halides to prepare fast hydrogen atoms and slow halogen atoms in their two spin–orbit levels. Atomic states that involve excited electron orbitals require shorter photodissociation wavelengths, usually in the vacuum ultraviolet. Their photolytic production can constitute a clean, selective means of preparing excited atoms which, if they are sufficiently metastable, can be used in other studies, for example, of collisional phenomena.

For small polyatomic molecules, modern spectroscopic methods in many cases enable *all* of the product degrees of freedom to be probed, for example, vibrational, rotational, and electronic, including fine structure (spin–orbit). The resulting distributions have proven to be very informative, enabling theoretical modeling of benchmark systems to achieve high levels of quantitative accuracy and improved predictability. Even at a qualitative level, it is clear that a high percentage of the available energy appearing as fragment center-of-mass translational energy indicates repulsion between the separating fragments. At the same time, such exit channel forces also promote rotational excitation, though not to the same extent as relative translational motion.

Exit channel forces are inefficient at imparting vibrational excitation to the products. These excitations are due primarily to geometry changes between the ground and electronically excited PESs, as indicated in Fig. 5. However, the correlation between parent vibrational excitation and product excitations is subtle because the parent vibrational degrees of freedom do not in general evolve in an obvious way to the product degrees of freedom. For example, parent bending and torsional excitations evolve to fragment rotational excitations, orbital angular momentum, and relative translational motion, as shown in Fig. 9. To quantify this requires accurate dynamics calculations.

For polyatomic molecules, anisotropic laboratory frame product distributions arise in a manner that is analogous to the case of linear molecules whose alignments are depicted in Fig. 8. The transition dipole moment is distributed around **E** according to a $\cos^2\theta$ distribution. However, the location of the transition dipole moment in a body-fixed reference frame of the polyatomic molecule is not in general as easily determined as with linear molecules. In addition, the bond that breaks, and there may be a number of equivalent bonds (e.g., C_2H_4, C_2H_6, etc.), is in general neither parallel nor perpendicular to the transition dipole moment. These complicating features notwithstanding, it is straightforward to model anisotropies for the photodissociation of polyatomic molecules, and this has proven to be a useful tool for understanding the photoexcitation and dissociation processes.

In addition to the possibility of photoexciting more than one repulsive potential independently, the different repulsive curves can *cross*, making it possible for the system to evolve into the product region along different repulsive potentials that lead to different product electronic levels. Referring to Fig. 10, one can see that if the zeroth-order diabatic potentials are of the same symmetry, they can interact, yielding the adiabatic potentials indicated in the figure. If the interaction at the crossing region is sufficiently strong, the energy gap between the resulting adiabatic potentials will be large enough to discourage nonadiabatic transitions between the adiabatic potentials during dissociation. In this case, photoexcitation of either adiabatic potential proceeds to products without appreciable population transfer between the adiabatic potentials.

Alternatively, when the energy gap is small, the system can efficiently make a transition from one adiabatic

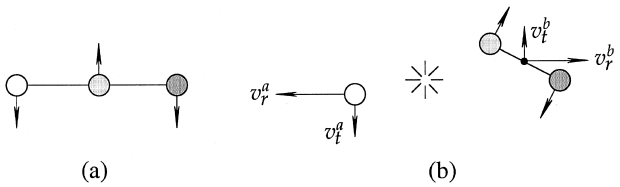


FIGURE 9 (a) The molecule is bending. (b) Following direct photodissociation, the bending motion is transformed, together with axial recoil, into product rotation and translation. Product translation is represented as transverse (orbital angular momentum) and axial recoil components (t and r subscripts, respectively). If the molecule is initially nonrotating ($\mathbf{J}_0 = 0$), the orbital angular momentum **L** is equal in magnitude and directed oppositely to the diatomic fragment's angular momentum, $\mathbf{J}_b$.

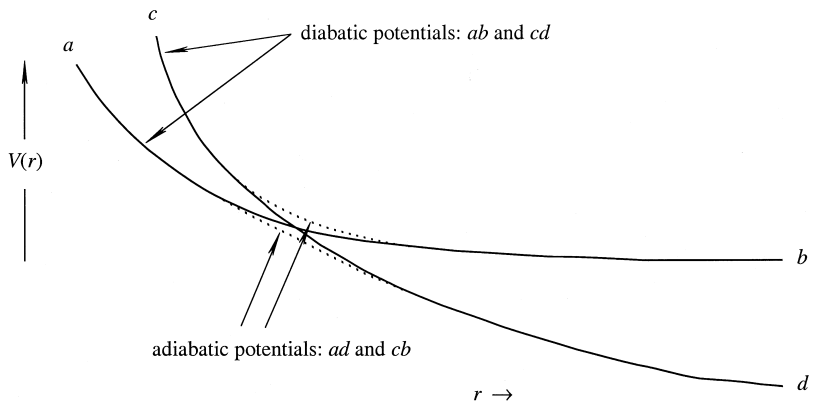


FIGURE 10 The diabatic repulsive potentials *ab* and *cd* interact most strongly in the crossing region. The adiabatic potential curves *ad* and *cb* follow the dashed lines.

potential to the other. In this case, the nuclear motion is sufficiently rapid that the electrons do not reconfigure in the crossing region, as they must if they are to follow the adiabatic potentials. Thus, the system has a high probability of following the diabatic potentials into the product region.

In the intermediate region, where neither the diabatic nor adiabatic representations are good approximations, quantitative analyses are is required. Electronic structure theory is now able to provide high-quality potentials, and exact quantum mechanical dynamics calculations are possible for small polyatomics. From the perspective of photochemistry, it is possible to understand branching ratios for different product electronic states as well as the product state distributions within these channels.

Predissociation

A dissociation mechanism that involves bound *and* repulsive potentials is depicted in Fig. 11 for the case of a diatom. A zeroth-order bound electronic state is crossed by a repulsive potential, raising the possibility that the bound-state vibrational levels acquire linewidths by virtue of their interactions with the repulsive potential. The bound-state vibrational levels are said to undergo predissociation. This picure applies as long as coupling is sufficiently weak to justify using the bound vibrational states of the *ab* diabatic potential in Fig. 11 as a good descriptor. They can be said to acquire a small amount of repulsive state character; they are broadened but otherwise not altered markedly.

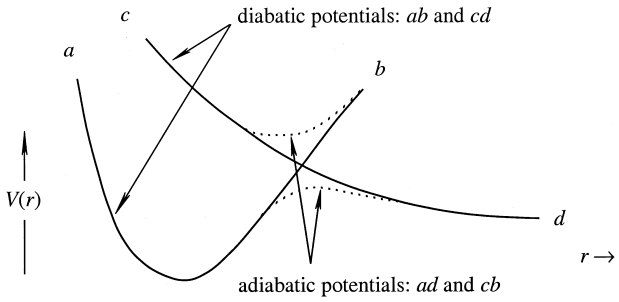


FIGURE 11 The diabatic potentials *ab* and *cd* interact most strongly in the crossing region. Diagonalizing the Hamiltonian with the coupling terms included yields the adiabatic potentials, which are separated by an energy gap whose size depends on the strength of the coupling. When the coupling is weak, some of the *ab* vibrational levels are said to predissociate; they have acquired a small amount of *cd* character.

From the perspective of photochemistry, it is important to know if the system dissociates. This will happen as long as the predissociation rate exceeds both the fluorescence rate and the rate of any collisional process that deactivates the system without causing fragmentation.

If coupling is strong, the diabatic potentials do not provide a good description. Adiabatic potentials result from diagonalization of the electronic Hamiltonian with the coupling term included, and transitions between the adiabatic potentials are caused by nonadiabatic transitions, as discussed in the previous section. In this regime, the term predissociation does not apply to the *ab* potential; one deals with the adiabatic potentials.

Polyatomic molecules, with their additional degrees of freedom, do not have simple one-dimensional curve crossings. The PESs intersect along seams in the multidimensional hyperspace of the molecule's many internal coordinates. However, the same principles apply: levels are said to be predissociative by virtue of having acquired some dissociative character from a repulsive PES.

IV. RADIATIONLESS DECAY

The direct photodissociation processes presented above have been examined thoroughly in numerous research laboratories throughout the world and important mechanisms have been revealed and analyzed theoretically. Most of the

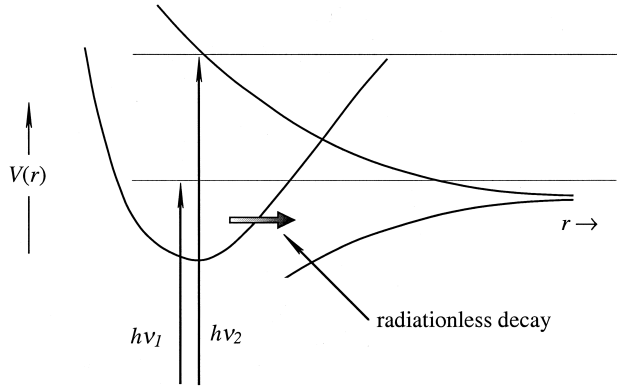


FIGURE 12 Photoexcitation of the excited PESs occurs from near the equilibrium geometry of the ground PES. At this geometry, the energy of the repulsive potential lies well above that of the large-r asymptote. The origin of the bound PES lies below the asymptote, which is a common situation. Thus, the photoexcitation frequencies ν_1 and ν_2 differ significantly.

fundamental physics has been summarized and made accessible to nonspecialists in the monograph by Schinke (1993).

Referring to Fig. 12, one can see that the direct photoexcitation of repulsive potential curves generally occurs at shorter wavelengths than those used to excite zeroth-order bound PESs which undergo radiationless decay. Accessing the repulsive curve in the Franck–Condon region (i.e., vertically from a low-lying vibrational level of the lower PES) occurs at energies well above the reaction threshold shown as the large-r asymptote in the figure. Consequently, relatively short wavelengths, often in the vacuum ultraviolet, are required to photoexcite the repulsive curves, whereas the absorptions of the bound PESs are accessible in more convenient spectral regions, i.e., longer wavelengths.

Mechanisms of radiationless decay in polyatomic molecules have been studied thoroughly, mostly in the 1960s and 1970s. This has yielded a clear understanding of the relevant physics, which can be viewed from complementary time-dependent and time-independent perspectives. The former is indicated in Fig. 13a, where it is assumed that an S_1 level decays via internal conversion to a manifold of S_0 vibrational levels. It treats radiationless decay as a dynamical process, in which an initially excited S_1 level decays with a rate Γ, populating the molecule's S_0 vibrational degrees of freedom, which constitute a "bath."

For simplicity, the S_0 levels are assumed to be bound, i.e., dissociation does not occur, and a single zeroth-order S_1 level is considered. The S_1 level is assumed to be coupled to all S_0 levels that have the same symmetry as the S_1 level. Though the matrix elements V_{S1-S0} differ from one S_0 level to the next, this does not affect the result because of the large number of S_0 levels that are efficiently coupled to the S_1 level, i.e., effects due to the different V_{S1-S0} matrix elements average out.

A. Time-Dependent Perspective

As mentioned above, there are two complementary ways to view this situation. In Fig. 13a, the S_1 level can be said to decay to a dense manifold of S_0 levels. In this case, an approximation called Fermi's golden rule can be used to describe the decay:

$$\Gamma = \frac{2\pi}{\hbar} V^2 \rho, \tag{17}$$

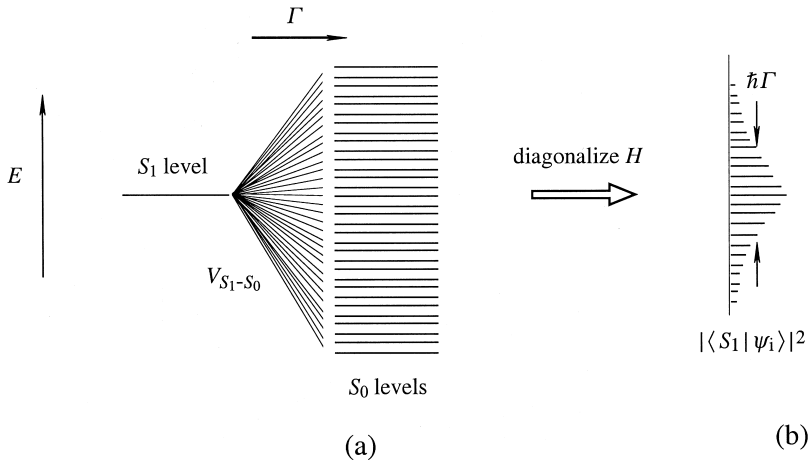


FIGURE 13 (a) An S_1 level is coupled to a manifold of S_0 vibrational levels. These levels are eigenfunctions of H_0. The S_1 level decays nonradiatively with a rate Γ. (b) Diagonalization of $H = H_0 + V$ yields eigenstates ψ_i, which display the fact that S_1 has been dissolved into the S_0 quasi-continuum. The distribution of S_1 character is Lorentzian with a full-width of $\hbar\Gamma$.

where V^2 is the average squared V_{S1-S0} matrix element and ρ is the density of states on the S_0 PES at the energy of the S_1 level. In applying Eq. (17), internal conversion is assumed to be irreversible, i.e., the S_1 level simply disappears:

$$|\psi_{S1}|^2 = e^{-\Gamma t}. \tag{18}$$

This expression would be exact (for constant V_{S1-S0}) if the S_0 levels were truly continuous rather than closely spaced as shown in Fig. 13, i.e., constituting a quasi-continuum. If the quasi-continuum is sufficiently dense, for example, unresolvable in any reasonable experiment, Eqs. (17) and (18) are excellent approximations, and decay can be taken to be irreversible for all practical purposes. However, ρ is sometimes not large enough to justify the use of Eqs. (17) and (18). In this case, the coupling of one or more S_1 levels to the S_0 manifold is best described by calculating the resulting molecular eigenstates, which are mixtures of the wavefunctions of the two manifolds. Large molecules satisfy Eqs. (17) and (18) to a high degree of accuracy, whereas small molecules (triatomics and some tetratomics) are often better served by the latter description.

The time-dependent picture can be illustrated as follows. If the system is excited by using pulsed radiation of sufficiently short duration, the linewidth of the radiation, for example, as given by the Fourier transform of the temporal signal, will exceed $\hbar\Gamma$. In this case, the eigenstates ψ_i, whose percentage S_1 character is indicated in Fig. 13b, are excited simultaneously in proportion to the percentage S_1 character. The phases of the coherently excited ψ_i are such that at $t = 0$ the S_1 level is created as a coherent superposition of the ψ_i. This constitutes the photoexcitation of a resonance which subsequently decays exponentially with the rate Γ. Said differently, on a sufficiently short time scale, the S_1 level cannot know that it is destined to relax, so the short pulse must create the S_1 level.

B. Time-Independent Perspective

With the time-independent perspective indicated in Fig. 13b, the Hamiltonian matrix, whose basis vectors are the S_1 level and the S_0 vibrational levels, is diagonalized, yielding eigenstates ψ_i. Each of the ψ_i contains a small fraction of S_1 character. The S_1 level can be said to be dissolved into the manifold of S_0 vibrational levels, with its character distributed according to a Lorentzian line shape, i.e., the distribution is given by $|\langle S_1 \mid \psi_i\rangle|^2$ versus E. The full-width at half-maximum of the Lorentzian curve is $\hbar\Gamma$.

To illustrate the time-independent picture, consider an experiment in which the spectral resolution is sufficient to resolve the ψ_i. The spectrum might resemble the one shown in Fig. 13b. When an individual ψ_i is excited, its fluorescence decay rate $\Gamma_{\mathrm{fl},\psi i}$ is much less than the S_1 fluorescence decay rate Γ_{fl}:

$$\Gamma_{\mathrm{fl},\psi i} = \Gamma_{\mathrm{fl}}|\langle S_1 \mid \psi_i\rangle|^2. \tag{19}$$

The S_1 fluorescence decay rate Γ_{fl} has been diluted by the fractional S_1 character of the ψ_i eigenstates. This is a large effect. A Γ_{fl} value of 10^7 sec^{-1} can easily be diluted to $\Gamma_{\mathrm{fl},\psi i}$ values $\sim 10^4$ sec^{-1} for a small molecule, and to a much greater extent for larger molecules, thus the term *radiationless* decay. Note that in the high-resolution experiment, the radiationless decay rate Γ of the S_1 level is not observed directly. It is inferred from the Lorentzian width of the spectrum.

C. Hierarchical Coupling

In some cases, the coupling of the S_1 level to the S_0 vibrational manifold occurs in a hierarchical manner. By this is meant that a small subset of the S_0 vibrational degrees of freedom facilitates coupling much more than the other S_0 vibrations. These are called promoting modes. These modes in general do not have good quantum numbers at the energies appropriate to S_1–S_0 coupling, i.e., intramolecular vibrational redistribution (IVR) mixes the S_0 vibrational levels to the extent that there are no good vibrational quantum numbers, and the promoting modes are no exception.

The coupling hierarchy is that S_1 couples to the promoting modes before the latter are degraded by IVR, as illustrated in Fig. 14. The broad width reflects the S_1–S_0 coupling via the promoting modes, not unlike Fig. 13b. The narrower widths are attributed to IVR. In the example shown, the time scales differ by an order of magnitude.

D. Internal Conversion and Intersystem Crossing

These fundamental coupling mechanisms are responsible for essentially all cases in which photon absorption can be said to create a zeroth-order electronically excited state which undergoes radiationless decay. Usually one or the other is responsible for radiationless decay but on occasion they act in competition and/or sequentially. Because the initially excited state can be said to decay, the ensuing photochemistry is that of a lower electronic state.

In the case of a closed-shell molecule whose lowest excited state S_1 undergoes radiationless decay via internal conversion, ultraviolet excitation of the zeroth-order S_1 level leads to the ground electronic state S_0. Higher electronic states (S_2, S_3, etc.) are usually less accessible

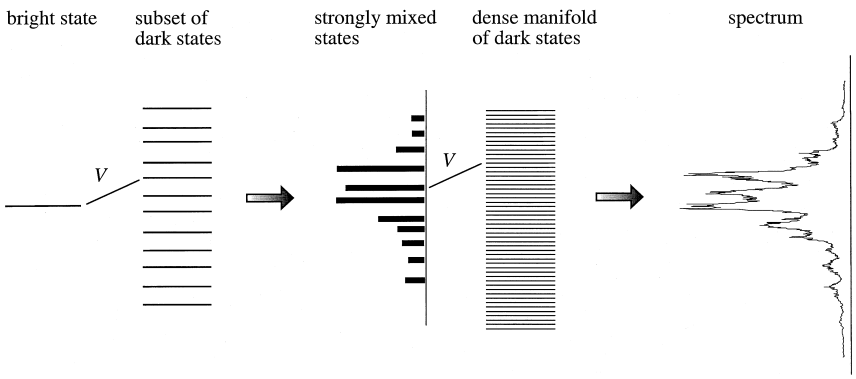


FIGURE 14 The bright state is coupled more strongly to a subset of the dark states (i.e., the promoting modes) than to the rest of the dark states. This yields a broad spectral width. The strongly mixed states are coupled by IVR to the dense manifold of dark states. The spectrum on the right was obtained at the NO_3 $^2E' \leftarrow {}^2A_2'$ origin with a 6 K sample.

because shorter wavelengths are required. The same ideas apply to radicals, where the ground state and optically accessible excited states are usually doublets.

The term in the molecular Hamiltonian that is responsible for internal conversion is the nuclear kinetic energy operator T_N. In the Born–Oppenheimer (BO) approximation, electronic wavefunctions are calculated by assuming that the nuclei are stationary, in which case there is no contribution from T_N when calculating electronic wavefunctions at a given geometry. Though the BO approximation is the cornerstone of molecular science, when dealing with electronically excited states, particularly the issue of whether they decay via radiative or nonradiative processes, it often fails. However, this should not be judged as a serious shortcoming of the BO approximation. Spontaneous emission lifetimes of the zeroth-order electronically excited states are typically 10^{-8}–10^{-6} sec, which is very long compared to molecular time scales (e.g., compare this to vibrational periods, which are typically 10^{-13} sec). Thus, for an isolated molecule, internal conversion need only be faster than the spontaneous emission rate for it to play the dominant role in determining the fate of the excited state.

As mentioned in earlier sections, when the PESs of the electronic states involved in internal conversion are nearly degenerate, breakdown of the BO approximation can be severe. This near proximity of PESs occurs in specific regions of configuration space called conical intersections. These are seams where the diabatic PESs come together and the adiabatic surfaces are separated by a small gap. The motions of the nuclei on the upper adiabatic surface can induce a strong coupling of the electronic surfaces because the energy separation between electronic states is small. Indeed, when the adiabatic surfaces are separated by a small gap, nonadiabatic transitions induced by T_N can occur with high probability. Examples of this abound. Common atmospheric molecules that display this effect are NO_2 and NO_3. Their zeroth-order electronically excited states are very strongly coupled to their ground electronic states.

Radiationless decay via intersystem crossing from S_1 leads to the excitation of triplet levels. The lowest triplet T_1 is favored, though others are known to result in intriguing pathways. The operator in the Hamiltonian responsible for intersystem crossing is that of spin–orbit interaction. It is possible to write this in terms of single-electron operators and introduce these into electronic structure programs. An excellent discussion of intersystem crossing in the context of radiationless transitions in polyatomic molecules is presented in the book by Medvedev and Osherov (1995).

The issue of competition between internal conversion and intersystem crossing is subtle because of the different level densities within the T_1 and S_0 vibrational manifolds. At a given S_1 energy, the S_0 manifold is much more dense than the T_1 manifold because of the large energy difference between S_0 and the T_1 origin. When the T_1 manifold is sparse, i.e., the average separation between adjacent levels is comparable to or larger than the S_1–T_1 coupling matrix elements, coupling is erratic or infrequent. An example of this is given in Section V.

Since about 1980, radiationless decay via internal conversion and intersystem crossing has become the subject of renewed interest for the purpose of carrying out detailed studies of unimolecular decomposition, which is known to play a major role in photochemistry. Unimolecular decomposition refers to the dissociation of a polyatomic molecule by a statistical mechanism in which, in classical terms, the nuclei move about without regularity, occasionally reaching critical configurations from which reaction occurs.

V. PHOTOINITIATED UNIMOLECULAR DECOMPOSITION

The radiationless decay of a zeroth-order electronically excited PES to one or more lower PESs is often accompanied by unimolecular decomposition. For example, ultraviolet photons having wavelengths shorter than 300 nm are sufficiently energetic to bring about dissociation in a large variety of molecules. For wavelengths shorter than 200 nm, any polyatomic molecule that absorbs a photon will have acquired enough energy to dissociate.

Laboratory studies of such processes have, for the past two decades, been dominated by the use of laser techniques applied to gaseous samples that have been cooled to 1–10 K by using supersonic expansion into vacuum. The data of the highest quality in this research area are the result of detailed experimental studies that have used sophisticated pump-probe laser methods. Within the last decade, double resonance methods, though difficult to implement, have become increasingly popular, adding yet another level of sophistication.

The state-to-state era of photoinitiated unimolecular decomposition began with studies in which parent excitation was well characterized (nozzle cooling, collision-free conditions, high-resolution lasers, etc.) and reaction products were probed state-selectively by using tunable radiation, thus enabling all of the populated product rovibronic levels to be detected. The systems that have led to the greatest increase in our understanding have involved the ground PES, though triplet states have also been implicated in some systems.

In principle, many wavelength regions can be reached with continuous tunability. However, there are severe technical limitations due to the VUV cutoff around 190 nm and the limited availability of the nonlinear optical materials that are needed to convert laser output radiations to wavelengths below around 200 nm. Thus, the most thorough studies have been carried out with wavelengths greater than 200 nm, as noted in Fig. 6. Fortunately, this has not proven to be restrictive. The benchmark systems that have emerged have proven to be adequate for unraveling the fundamental processes that are germane to essentially all systems that undergo unimolecular decomposition following radiationless decay.

Consider first the small-molecule limit, in which the radiationless decay mechanism is internal conversion and the density of states on the ground PES is sufficiently sparse that all of the molecular eigenstates in the bound region can be resolved spectroscopically. Putting aside for the moment dissociation, the eigenstates are mixtures of the ground and electronically excited states:

$$\psi_i = \sum_m a_m \phi_m^* + \sum_n c_n \phi_n, \tag{20}$$

where ψ_i is an eigenstate, ϕ_m^* is the mth vibrational level of the zeroth-order electronically excited PES, ϕ_n is the nth vibrational level of the zeroth-order ground electronic state, and a_m and c_n are expansion coefficients. Here we consider just two PESs ϕ^* and ϕ and their respective vibrational levels. Rotations are of secondary importance and are therefore ignored at this level of description. The ψ_i given by Eq. (20) are themselves zeroth order in the sense that we have set aside the fact that they are coupled to dissociation continua. In the simplest cases, when the ψ_i couple to dissociation continua, they acquire linewidths $\hbar\Gamma$ which can be spectrally resolved and reflect their dissociation rates Γ, as noted in Fig. 15. Such quasi-bound levels are called resonances. However, this simple picture only applies to the threshold region. At higher energies, resonances exist but their widths are overlapping and the spectra do not display easily interpreted line shapes.

The vibrational density of states of the ground PES is always much larger than that of the electronically excited state at the same energy (Fig. 3). Consequently, the ψ_i indicated in Eq. (20) are dominated by the ϕ_n. The electronically excited state provides the needed photoabsorption cross section, but the molecular dynamics is that of the ground PES.

With few exceptions, the nuclear motions of polyatomic molecules on their ground PESs are highly irregular at energies slightly below the lowest reaction threshold. In this regime, the dynamics can be said to be quantum chaotic. There are no good quantum numbers that describe the participating degrees of freedom, and the eigenfunctions are composed of random mixtures of the basis vectors of a separable Hamiltonian. Such quantum chaos is expected to extend to energies above the lowest reaction threshold. This is the basis for the statistical theories of unimolecular decomposition, which assume that energy is randomized among the participating degrees of freedom (e.g., vibrational) on a time scale that is short relative to that of decomposition.

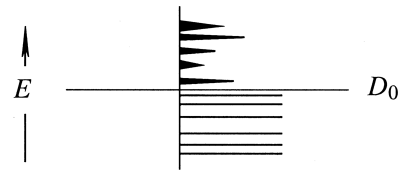


FIGURE 15 Below the reaction threshold D_0, the nearest neighbor level distribution indicates quantum chaos. Above D_0, this character is carried over to quasi-bound levels (resonances) that can only be resolved just above D_0.

The most important randomization process is IVR. Characteristic IVR times are typically a few hundred femtoseconds, which is short relative to nearly all reaction time scales. Without taking rotations into consideration, vibronic symmetry species are preserved and must be dealt with separately, whereas when rotations are taken into consideration, the vibronic symmetry species can be mixed by Coriolis interaction. It is interesting that calculated rates depend very little on vibronic symmetry because its preservation or breaking affects the reactive flux and molecular vibrations to the same extent, leaving the rate coefficient unaffected.

A. Standard Statistical Models

The quantum chaotic dynamics of the bound molecular region persist above the reaction threshold. This gives rise to quantum mechanical fluctuation phenomena, for example, in rates and product state distributions. Fluctuations in the rates are most prominent just above threshold; with product state distributions, the fluctuations extend to higher energies. When there is sufficient averaging, the fluctuations of the rates become too small to be of concern, and statistical mechanics can be used to calculate the rate coefficient. This is achieved by introducing a transition state that separates the molecular and product spaces. The phase volumes of the transition state and molecular regions (ΔV_{TS} and ΔV, respectively) are identified, the ratio of the former to the latter is taken, and this is divided by the transit time through the transition state region, τ_{TS}. This yields the microcanonical unimolecular decomposition rate coefficient in terms of the phase volumes:

$$k(E) = \frac{\Delta V_{TS}}{\Delta V}\frac{1}{\tau_{TS}}. \qquad (21)$$

The phase volumes are obtained by using statistical mechanics, yielding the simple formula

$$k(E) = \frac{K}{h\rho}, \qquad (22)$$

where K is the number of independent transition-state levels at energy E and $1/h\rho$ is the rate per transition-state level. Thus, $k(E)$ is a monotonically increasing function of E. Equation (22) is the fundamental result of the statistical model. Figure 16 shows the steplike behavior predicted by Eq. (22). In practice these steps are difficult to observe experimentally because they are rounded and the data are usually of insufficient quality to justify a positive identification. It is also straightforward to predict the product internal and translational excitations for microcanonical systems by using statistical mechanics.

The first molecule whose photoinitiated unimolecular decomposition was studied thoroughly at the state-to-state

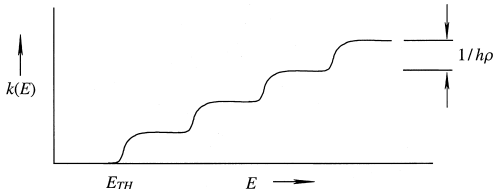


FIGURE 16 Equation (22) predicts that the unimolecular decomposition rate coefficient $k(E)$ increases in steps in units of $1/h\rho$. Tunneling and other nuances round the edges.

level is NCNO → CN + NO. On the negative side, NCNO is explosive and light-sensitive, while on the positive side, its absorption spectrum goes from the near-ultraviolet to the near-infrared, making photoexcitation easy. Prior to these studies, it was believed that photoexcited NCNO dissociated via a nonstatistical mechanism such as coupling to a repulsive PES. This turned out to not be the case: radiationless decay is dominated by internal conversion and reaction takes place on S_0. Because of the similar masses, kinematic bias is minimal. In addition, the simplicity of the product species makes it possible to probe all possible product excitations. The fact that there is no barrier to the production of two radicals enables the products to be formed in their lowest quantum states, as shown in Fig. 17.

The dense electronic absorption spectrum of expansion-cooled NCNO enabled energies in excess of reaction threshold to be varied continuously. Product state distributions for the CN and NO products were recorded for a broad range of energies, and it was found that the data could be fitted by using statistical theory. Likewise, the microcanonical rate constants $k(E)$ were also fitted by using statistical theory. Even correlations between the CN and NO levels were found to be statistical. Thus, NCNO is a benchmark system for the regime where quantum fluctuations are averaged out and a monotonic variation of $k(E)$ versus E is anticipated. Because there is no barrier

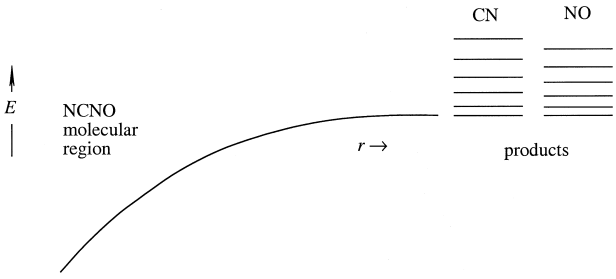


FIGURE 17 Barrierless unimolecular reactions can yield products in their lowest quantum states. Just above threshold, the CN and NO radicals can be formed with no vibrational or rotational energies whatsoever and minimal translational energy (Baer and Hase, 1999).

along the reaction coordinate other than the endoergicity, the transition state is loose (i.e., it resembles the products) and $k(E)$ increases rapidly over a modest energy range.

In a number of cases, molecules that were at first believed to dissociate via a nonstatistical mechanism were later discovered to react via a unimolecular decomposition mechanism on the ground PES. There is no sharp division in time, but one should look carefully at studies reported prior to around 1980 insofar as mechanisms are concerned. Even NO_2, whose absorption spectrum had been studied for a century, was believed to decompose nonstatistically until the 1980s, when it was shown that it decomposes via a unimolecular decomposition mechanism.

Another molecule that belongs to this regime is ketene, CH_2CO. Its ultraviolet absorption spectrum is continuous, enabling $k(E)$ versus E to be determined. Here, steps in $k(E)$ versus E have been observed, in agreement with and confirming the prediction of Eq. (22).

B. Quantum Fluctuation Phenomena

When the distribution of energy levels is sufficiently sparse, it is sometimes possible to observe the quantum fluctuation phenomena described above. Such behavior may at first appear to contradict Eq. (22), which predicts that the rate increases in units of $1/h\rho$ as the different transition-state levels are accessed. However, there is no inconsistency. Equation (22) is correct *on average,* but the chaotic nature of the dynamics in the bound region is manifest as erratic changes in the rate as the energy is varied. For example, the rates of individual resonances can differ by orders of magnitude, while their average rates follow Eq. (22) as indicated in Fig. 18. These fluctuations are most prominent just above threshold where the resonances are most sparse, i.e., for the first few values of K. As E increases, the fluctuations of the rates average out.

The first experimental work in this area was carried out with deuterated formaldehyde (see Baer and Hase, 1996). It was found that the decomposition rates for the lowest energy pathway, i.e., $D_2CO \rightarrow D_2 + CO$, changed erratically

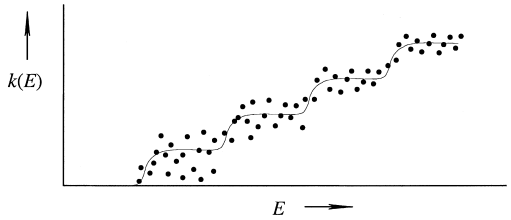


FIGURE 18 Though quantum fluctuations of the resonance decay rates cause $k(E)$ to change erratically from one resonance to the next, on average $k(E) = K/h\rho$ still applies. The fluctuations decrease with increasing K.

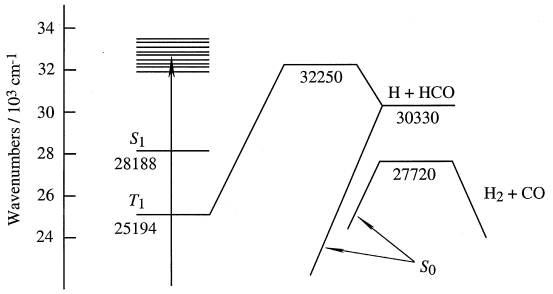


FIGURE 19 Formaldehyde energies (cm^{-1}). S_1 rovibronic levels are coupled to both T_1 and S_0. The $H_2 + CO$ channel is accessed solely via a tight transition state on S_0. The $H + HCO$ channel is accessed via both a loose transition state on S_0 and a tight transition state on T_1. For D_2CO, the S_1 origin lies below the barrier on S_0 to $D_2 + CO$.

from one quasi-bound level to the next. The relevant part of the potential is shown in Fig. 19, which shows that the lowest energy path to $D_2 + CO$ products involves a barrier. Tunneling through this barrier is significant because of the light deuterium atoms. The observed fluctuations confirm the chaotic dynamics of the bound region, in which the eigenfunctions differ markedly from one level to the next.

C. Vibrational Photochemistry

The first era of vibrational photochemistry began in the 1970s when it was discovered that with sufficiently intense CO_2 laser radiation, polyatomic molecules could absorb many photons and decompose via a unimolecular decomposition mechanism. Popularity waned in the 1980s because it was not possible in such experiments to control the amount of energy implanted in the molecule.

This shortcoming can be overcome by using high-frequency overtones, as shown in Fig. 20. In this scheme, the first step excites an overtone transition of a hydrogen stretch vibration. With currently available laser technology, it is possible to transport a significant fraction of the molecules from a given vibrational state to a higher vibrational state. Even $\Delta v = 4$ transitions can be pumped to near-saturation conditions.

Major successes have been achieved with H_2O_2 and HOCl (yielding $OH + OH$ and $OH + Cl$, respectively), which are listed in Fig. 6. In both cases, sequential overtone transitions (e.g., $\Delta v = 2$ followed by $\Delta v = 4$) were used to promote molecules from the ground vibrational level to energies above the bond dissociation energy D_0. Because this is a double-resonance method, state selectivity is superb.

The results obtained to date augur well for the future. The strategy indicated in Fig. 20 is rather general, though

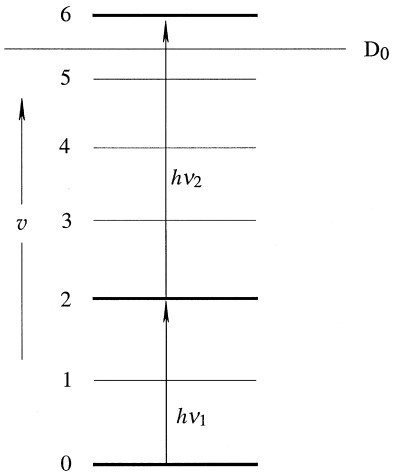


FIGURE 20 Unimolecular decomposition on S_0 is photoinitiated via vibrational overtone transitions by using high-energy pulsed lasers. The diagram indicates hydrogen stretch vibrations: $\Delta v = 2$ followed by $\Delta v = 4$. Because anharmonicity increases upon vibrational excitation, the second step, $h\nu_2$, can cause more quanta to change than the first step, $h\nu_1$, while maintaining a comparable pumping efficiency.

currently restricted to high-frequency vibrations, and assures reaction on the ground PES. Moreover, its attractiveness is dependent on laser technology, which can be expected to advance steadily, as it has for three decades.

VI. COMPLEX PHOTOCHEMICAL PATHWAYS

In general, multiple mechanisms and pathways may participate in photoinitiated chemical change. The paths may be competitive, sequential, independent, etc. Such studies currently account for a large amount of research in this field.

It is known that ultraviolet photochemistry often yields more than a single set of chemical products. In addition, some of the products are formed in electronically excited states and with intriguing rovibrational product state distributions. Because of the complexity, these observations have been mainly of archival value. Suggested mechanisms have been speculative because the experiments were able to probe only a fraction of the possibilities.

Photoexcitation of more than one PES may proceed without quantum mechanical interference, in which case the total frequency-dependent cross section $\sigma_{\text{abs}}(\nu)$ is the sum of the individual cross sections $\sigma_i(\nu)$ for photoexciting the different PESs:

$$\sigma_{\text{abs}}(\nu) = \sum_i \sigma_i(\nu), \tag{23}$$

where the branching ratios for the different $\sigma_i(\nu)$ are given by $\sigma_i(\nu)/\sigma_{\text{abs}}(\nu)$. Note that the rate coefficients for the different pathways do not affect the branching ratios because the paths are independent. For example, suppose $\sigma_{\text{abs}} = \sigma_1 + \sigma_2$, where σ_1 is to a repulsive state and σ_2 is to a zeroth-order bound PES that undergoes radiationless decay and unimolecular decomposition. The branching ratios are simply $\sigma_1/(\sigma_1 + \sigma_2)$ and $\sigma_2/(\sigma_1 + \sigma_2)$, even though the decomposition rates for the different paths may differ by orders of magnitude.

A. Loose versus Tight Transition States

The situation is more interesting when the photoexcited state decays via two or more mechanisms. An example that illustrates such competition is NO_3, whose radiationless decay below the lowest reaction threshold is shown in Fig. 14. This radical is the main oxidizing agent in the atmosphere during the nighttime. It does not survive in sunlight because it absorbs strongly at visible wavelengths and dissociates.

Figure 21 shows relevant energies. The barrier on the $^2A_2'$ ground PES to $NO + O_2$ products lies slightly

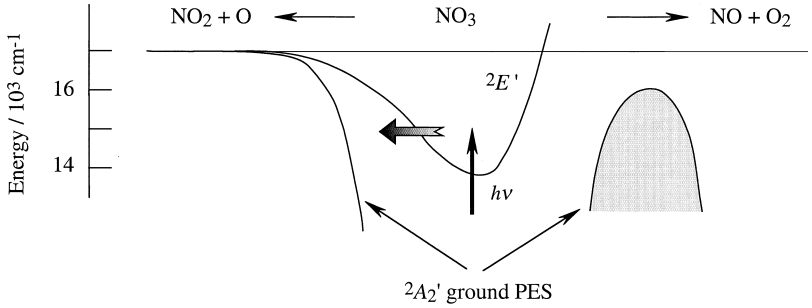


FIGURE 21 The $^2E'$ PES decays via internal conversion (shaded horizontal arrow) to the $^2A_2'$ ground PES, which undergoes unimolecular decomposition to $NO + O_2$ and $O + NO_2$ products via tight and loose transition states, respectively. Unimolecular decomposition via a loose transition state can also occur on the $^2E'$ PES. Energies are relative to the ground state.

below the $O + NO_2$ threshold. Below this threshold, the only possible products are $NO + O_2$. Rate coefficient measurements carried out in this energy region have confirmed a unimolecular decomposition mechanism, including heavy-particle tunneling below the barrier. The $O + NO_2$ channel has no barrier; it opens as soon as these products are energetically accessible.

Just above the $O + NO_2$ threshold, the two chemically distinct product channels compete, with $NO + O_2$ winning. The $O + NO_2$ rate increases rapidly with E and dominates several hundred wavenumbers above the threshold. This competition is between reaction pathways that proceed via a tight transition state to $NO + O_2$ products versus loose transition states to $O + NO_2$ products.

There are two pathways that have loose transition states, the $^2A_2'$ ground state and the $^2E'$ excited state. The latter also has a large internal conversion rate, as shown in Fig. 14. Thus, internal conversion competes with unimolecular decomposition on the $^2E'$ PES. Just above threshold, internal conversion and unimolecular decomposition on the $^2A_2'$ ground state dominates, whereas at higher energies, reaction via $^2E'$ becomes important.

In this example, $^2E'$ internal conversion yields $^2A_2'$, which reacts via loose and tight transition states, yielding $O + NO_2$ and $NO + O_2$ products, respectively. Unimolecular decomposition also occurs directly on $^2E'$ via a loose transition state, in competition with internal conversion.

B. Internal Conversion versus Intersystem Crossing

Referring to Fig. 19, we see that formaldehyde can decompose to two sets of chemically distinct products: $H_2 + CO$ and $HCO + H$. The former is accessed solely via S_0, while the latter is accessed via two paths, a barrierless one on S_0 and a T_1 path having a modest barrier. Thus, S_1 levels decay via competitive pathways that involve internal conversion and intersystem crossing.

Just above the $H + HCO$ threshold at 30,330 cm^{-1}, the S_0 path dominates because the T_1 barrier is prohibitively large. Alternatively, well above the T_1 barrier, the rate coeficient for T_1 is significantly larger than its S_0 counterpart because at a given energy the density of states on T_1 is much smaller than the density of states on S_0; see Fig. 3 and Eq. (22). Because T_1–S_0 coupling is insignificant, the main issue is the relative decay rates: $S_1 \rightarrow T_1$ versus $S_1 \rightarrow S_0$.

This competition is most interesting in the region near and just below the T_1 barrier, as shown in Fig. 22. Above the barrier, the T_1 resonances overlap, assuring a reasonably smooth variation of S_1–T_1 coupling strength. Below the barrier, the S_1–T_1 coupling is very strong when the S_1 level lies close to a T_1 level, and weak when it is separated from the T_1 level by an energy that exceeds the magnitude of the coupling matrix element.

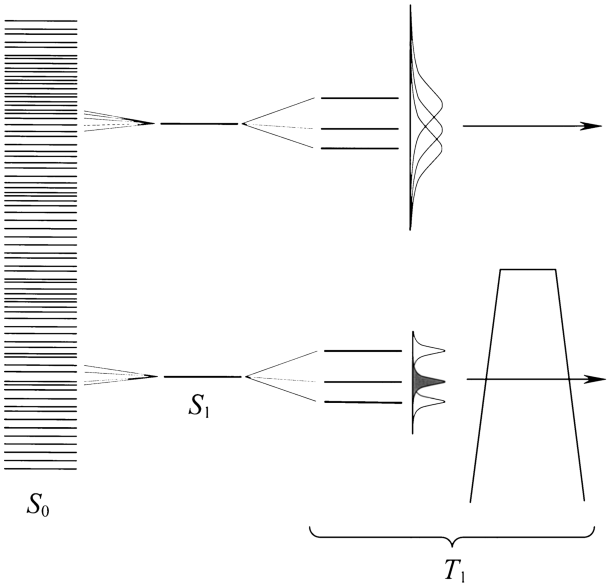


FIGURE 22 The S_1 levels shown are coupled to S_0 on the left and T_1 on the right, below and above the T_1 barrier. The T_1 resonances broaden rapidly with increasing energy.

SEE ALSO THE FOLLOWING ARTICLES

CHAOS • ELECTRODYNAMICS, QUANTUM • ELECTRON SPIN RESONANCE • KINETICS (CHEMISTRY) • PHOTOCHEMISTRY BY VUV PHOTONS • LUMINESCENCE • ORGANIC CHEMICAL SYSTEMS, THEORY • POTENTIAL ENERGY SURFACES • QUANTUM THEORY

BIBLIOGRAPHY

Baer, T., and Hase, W. L. (1996). "Unimolecular Reaction Dynamics," Oxford University Press, New York.

Berge, P., Pomeau, Y., and Vidal, C. (1984). "Order within Chaos," Wiley, New York.

Cohen-Tannoudji, C., Dupont-Roc, J., and Grynberg, G. (1992). "Atom–Photon Interactions," Wiley, New York.

Finlayson-Pitts, B. J., and Pitts, J. N., Jr. (2000). "Chemistry of the Upper and Lower Atmosphere," Academic Press, New York.

Jackson, J. D. (1975). "Classical Electrodynamics," 2nd ed., Wiley, New York.

Lefebvre-Brion, H., and Field, R. W. (1986). "Perturbations in the Spectra of Diatomic Molecules," Academic Press, New York.

Louden, R. (1983). "The Quantum Theory of Light," 2nd ed., Oxford University Press, New York.

Medvedev, E. S., and Osherov, V. I. (1995). "Radiationless Transitions in Polyatomic Molecules," Springer-Verlag, New York.

Mullin, T. (1993). "The Nature of Chaos," Oxford University Press, New York.
Okabe, H. (1978). "Photochemistry of Small Molecules," Wiley, New York.
Porter, C. E. (1965). "Statistical Theories of Spectra: Fluctuations," Academic Press, New York.
Purcell, E. M. (1985). "Electricity and Magnetism," 2nd ed., McGraw-Hill, New York.
Sargent, M., Scully, M. O., and Lamb, W. E. (1974). "Laser Physics," Addison-Wesley, Reading, MA.
Schinke, R. (1993). "Photodissociation Dynamics," Cambridge University Press, Cambridge.



Radiation Effects in Electronic Materials and Devices

Andrew Holmes-Siedle
Brunel University

Victor A. J. van Lint
VvL, Inc.

I. Description of Radiation Physics
II. Making Radiation
III. Quantifying Radiation
IV. Interaction of Radiation with Matter
V. Transient Ionization Effects
VI. Long-Term Ionization Effects
VII. Displacement Effects
VIII. Single-Event Phenomena
IX. Radiation Effects in Systems and Technology
X. Special Uses of Radiation

GLOSSARY

Absorber Matter intervening in the path of a radiation beam that absorbs energy from that beam.

Absorption coefficient (μ) Coefficient in Lambert's law, $I = I_o \exp(-\mu d)$, which describes the attenuation of the radiation penetrating an absorber.

Annealing The rearrangement of atoms or charges in a material with time after irradiation. High temperature may be needed to produce these effects but some annealing occurs even at room temperature.

Bremsstrahlung X ray produced when particles (electrons mainly, but also heavier particles) are accelerated by interaction with matter. Bremsstrahlung (German for braking radiation) is sometimes called white radiation to distinguish it from the characteristic line spectra produced by different chemical elements.

Charge carrier Electrons or holes in a solid; electrons or ions in a gas.

Defect cluster Local concentration of lattice defects produced by a single primary or secondary recoil atom.

Displacement effects Effects on material properties of collisions resulting in the removal of atoms from their normal lattice sites.

Dose (D) Energy deposition per unit target mass, usually by ionizing radiation. Used broadly for energy from radiation accumulated in matter. Used in dosimetry for the energy absorbed per unit mass of material, usually by ionization processes. Units are the rad and Gray,

which are equivalent, respectively, to 100 ergs/g and 1 J/kg. Therefore, 1 rad = 0.01 Gray or 1 cGy.

Dose rate Rate at which energy is transferred to a material by a radiation beam, for example, in units of rad per seconds (or Gray per seconds).

Electron volt (eV) Unit of energy possessed by a particle or photon. The kinetic energy acquired by an electron when accelerated through an electrical potential drop of 1 V.

Energy flux Passage of energy per unit time and area in the form of penetrating particles, not necessarily stopped. Typical units are J/cm^2-sec and W/m^2.

Exposure Important term in dosimetry that expresses a fluence in terms of its effect on a dosimetric medium, nearly always air at STP, and the number of air ions produced per unit mass. An exposure of 1 Roentgen is exposure to the fluence of a given radiation that, in air at STP, generates 2.58×10^{-4} Coulomb (C) of ionic charge per kilogram. Used in contrast to dose, which, for a given fluence, varies from absorber to absorber.

Fluence Time-integrated flux of particles or photons. Unit: cm^{-2}. It is useful to add the symbol for the particle, for example e/cm^2, or even 1 MeV e/cm^2.

Flux Number of particles 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 cm^2. In both cases, the unit is cm^{-2} sec^{-1}. It should be noted that this definition, commonly used in the nuclear community, differs from that for luminescent flux used in the optics community, in which flux is the total number of energy per unit time of radiation, irrespective of the area.

Gray Radiation absorbed dose unit of the Systeme Internationale (SI), of value 1 J kg^{-1} and equal to 100 rad.

Hardening, hardened Used to describe improvement in the tolerance of a device or a system to a radiation environment. Originates from the military term for a site that is invulnerable to attack. Second, denotes increase of average energy of a beam of radiation due to the selective removal of a lower-energy component of the beam. Third, denotes the changes in mechanical properties of some metals, induced by high fluences of particle radiation.

High-energy physics (HEP) Study of subatomic, elementary particles. Experiments involve colliding particle beams of extraordinarily high energy. Apparatus of great size and complexity is designed to track the products of the collision and decay of particles. From elementary particle experiments, a theory has evolved for the detailed structure and origin of matter.

Holes Effective positive charge carriers in a solid due to empty electron states in an almost full band.

Integrated circuit Semiconductor chip on which a large number of interconnecting device functions have been formed.

Interstitial Atom inserted in between lattice sites in a crystal.

Ionization effects Large class of radiation effects that involve the removal of an electron from an atom or the excitation of an electron from a filled band in a solid.

Logic upset Change in logic state resulting when pulses of radiation of high dose rate generate photocurrents of significant magnitude in semiconductor devices. These can so alter the voltages at circuit nodes that the device misinterprets the disturbance as a logic signal and changes logic state.

Long-term ionization effects Effects on material properties due to trapped charge carriers or molecular rearrangement following ionizing events.

Mobility Ratio of charge carrier drift velocity to electric field.

Nonionizing energy loss The deposition of the energy of an energetic particle by means other than ionization, e.g., to producing displaced atoms. Often abbreviated to NIEL.

Positron Positive electron; a particle with the same mass as the electron but the opposite charge.

Range Distance into an absorber to which a particle is likely to penetrate. Since stopping is a statistical process, several ranges (practical, maximum, extrapolated, etc.) are defined.

Roentgen Unit of exposure equivalent to the generation of 2.58×10^{-4} C of ions per kilogram of air.

Shield Absorber structure giving protection from radiation.

Single-event upset Logic upset produced by a single energetic ion.

Surface effect In metal-oxide semiconductor (MOS) devices, diodes and bipolar transistors, the various effects that occur at the surface of the active semiconductor material.

Transient ionization effects Effects on material properties of separation of electrons from atoms, due to mobile charge carriers.

Vacancy Unoccupied lattice site in a crystal.

RADIATION PHYSICS deals with the wide range of effects observed when high-energy radiation, in the form of particles or photons, interacts with matter. "High energy" covers the range from hundreds of electron volts (eV) of energy to many millions of electron volts (MeV). The field of radiation physics is distinct from nuclear physics, covering the wide range of chemical and physical effects that follow the initial interaction of a high-energy particle or photon with resting atoms of matter. In many cases,

only the electron clouds around them are affected. Nuclear physics covers the internal structure of the atomic nucleus and the processes of fission and fusion. Many details of the interaction of radiation with living organisms fall under a separate subject heading, namely, radiobiology.

Many themes of this article have developed from the field of Radiation Effects, a branch of applied radiation physics concerned with achieving tolerance to radiation in electronic devices and systems. Some machines, particularly electronic systems, must survive in high radiation environments (e.g., in space). Some of the important effects in this field are transient in nature; that is, they are present only while the material is being exposed to radiation. Most of these are manifestations of electrons and ions, produced by "ionization." Other important effects are long-lived, increasing in degree with the accumulation of radiation exposure, or "radiation fluence." These include displacement effects in solids: those due to the displacement of atoms from their normal lattice sites by collision with energetic particles.

Radiation types of interest includes X-rays, alpha-, beta-, and gamma-rays, cosmic rays, and many forms of artificially accelerated particles, such as beams of electrons, protons, or ions. Important sources of radiation are electrostatic generators including X-ray machines, geomagnetically trapped charged particles, and the neutron and gamma radiation given off during the decay, fission or fusion of radioisotopes, such as in nuclear reactor materials. We of course live in a "background environment" of high-energy radiation deriving from natural sources (minerals, space) and some man-made sources (X-rays, nuclear reactors). The use of radiation sources in medicine also requires the methods of radiation physics in the planning and control of treatment. Such direct use for human benefit is only one of the ways in which a knowledge of radiation physics unlocks information about the universe and the structure of matter. Some of the specialized applications of this field of knowledge are described at the end of this article.

I. DESCRIPTION OF RADIATION PHYSICS

Radiation physics covers the effects that occur when high-energy radiation interacts with matter. This field is distinct from nuclear and high-energy physics, even though many of the particles involved are common to both. The radiation types of interest include high-energy photons, such as X-rays and gamma-rays (γ), and a multitude of particles, charged and uncharged. The energy of individual radiation particles is expressed in kilo- or mega-electron volts (keV or MeV). The radiation in question arises from both natural and artificial sources. Natural sources include space radiation (i.e., cosmic rays) and radium and uranium in rocks. Artificial sources include concentrated radioisotopes such as cobalt-60 irradiators, accelerators such as X-ray and electron-beam machines, nuclear chain reactions in reactors or weapons, and (not commonly realized) dilute radioisotopes in coal fly ash, food, and cigarette smoke.

The subject of radiation effects includes relatively little in the way of new physical processes. For example, the interaction of radiation with matter is normally studied as part of nuclear physics, the behavior of electrons and ions in gases is part of gaseous electronics, and the properties of defects in solids are treated in solid-state physics. Therefore, this article does not treat any of these subjects comprehensively, but summarizes those parts that are most important for radiation effects on the materials used in electronic systems.

The techniques 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 effect of radiation on living tissue falls 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 part of radiation physics. Radiation deposits energy in matter. This energy is then distributed among the atoms and molecules in a great variety of ways. The energy excites atoms to higher energy states or moves them about in the material. Figure 1 shows a flow chart containing the sequence of events that occur when radiation interacts with a solid. A serious disruption of the existing order in the solid can be produced, and this disruption is often called radiation damage. In liquids and gases, the final effects of radiation are different from those observed in solids, because, in fluids atoms are normally mobile while excited atoms and molecules can move long distances to interact chemically with other species. (The field of radiation chemistry and the effects of radiation on living tissues are not discussed specifically here.)

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. As Fig. 1 shows, *secondary* processes are complex and indeed are only well understood in a few cases, such as the various results of the displacement of atoms in silicon by an electron beam. Certain classes of material, important in technology, show very strong responses to radiation. The disciplines of solid-state and plasma physics are important in understanding these responses.

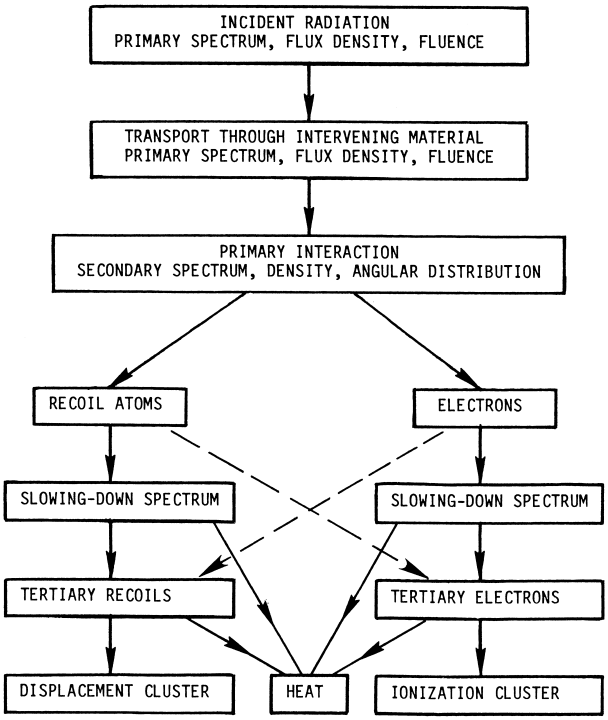


FIGURE 1 Radiation interactions.

The technological uses of radiation physics are very wide and are dealt with in a separate section. 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 the 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. Radioisotopes that are useful in irradiation experiments are listed in Table I.

The relative importance of radiation interactions are determined by the applications, especially the radiation environments in which electronics may be required to operate. The environments of interest include the following:

1. *Nuclear power plants.* Some of the control electronics must be located in areas in which they may be exposed to γ-rays and neutrons generated during fission as well as γ-rays from fission product decay. The rates of exposure are usually low, but the accumulated γ dose and neutron fluence can be very large.
2. *Particle accelerators.* Modern high-energy, high-current accelerators can deliver high exposure rates and accumulated exposures to electronics that must operate in their vicinity, especially the experiments in the target area.
3. *Nuclear explosions.* The effect of intense pulses of γ-rays, X-rays, and neutrons from a nuclear explosion can have severe transient and long-term effects on electronics, such as military systems and diagnostic equipment used during underground nuclear tests.
4. *Spacecraft.* Earth-orbiting satellites are exposed to trapped electrons and protons in radiation belts, cosmic rays, and solar protons. Even more severe environments, including radiation belts containing heavy ions, may be encountered by planetary probe spacecraft.

II. MAKING RADIATION

A. Overview

Figure 2 shows a survey of the broad range of energy values that have to be considered under the term "radiation." Most of the radiation of interest has an energy above 1 keV, 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. Machines which accelerate charged particles to over 1 TeV are now available. The shaded areas in Fig. 2 show that different fundamental effects in matter occur with different threshold radiation energies.

B. Photons

Photons that have energies in the keV and MeV range (X rays and gamma-rays) have the ability to penetrate 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. Energies from 30 keV to 3 MeV are common. Small fluxes of X-rays are also generated naturally in radioisotope samples by the collision of beta rays within the sample itself or with the capsule in which the sample is contained. In electron-beam devices (cathode-ray tubes and electron accelerators), a hazard to humans may be created by the unintentional generation of X-rays when the beam collides with the wall of the chamber.

Gamma-rays are high-energy, often 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 atom. These photons have energies of 1.17 and 1.33 MeV. Gamma rays are commonly created during nuclear chain reactions, such as

TABLE I Half Lives, Principal Photon Energies and Dose Rates from Gamma Emitters

Nuclide	Half-life	Principal γ energies (MeV)	Dose rate at 1 m from 1 Ci (rads/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 γ 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 1 g.

those that occur in a nuclear reactor core or a nuclear explosion. The isotopes contained in nuclear fuels represent concentrated sources of gamma-rays. These can be used for experimental irradiation in spent-fuel ponds. When certain equipment used in reprocessing nuclear fuel is directly exposed to the isotope sample, there is danger that exposed optical and electronic parts may degrade in performance. Thus, the equipment has to be radiation hardened (see later).

C. Electrons and Positrons

Electrostatic electron accelerators range in energy from 0.1 keV to 10 MeV; the beam currents range from microamperes to kiloamperes. To reach energies above 10 MeV, electrons are accelerated by radio-frequency energy in a machine called a linear accelerator. Natural sources of high-energy electrons include β-rays from isotopes and electrons in space which have been accelerated in magnetic fields. There is a particularly high concentration of electrons trapped around those planets that have high magnetic fields, such as Earth and Jupiter. It is inconvenient for space vehicle designers that these trapped radiation belts cover desirable altitudes for unmanned operational satellites. The radiation dose acquired in these regions is a significant source of radiation damage to components. Manned vehicles at present avoid the trapped radiation belts for the sake of the personnel. The antiparticle of the electron, the positron, is accelerated in the same way as the more common negative electron.

D. Protons

One especially well-known form of proton accelerator is the cyclotron, which uses radiofrequency energy. Nuclear reactions also produce protons in a material sample. Megavolt protons are also found trapped in the magnetic fields of planets such as the Earth and Jupiter. These probably

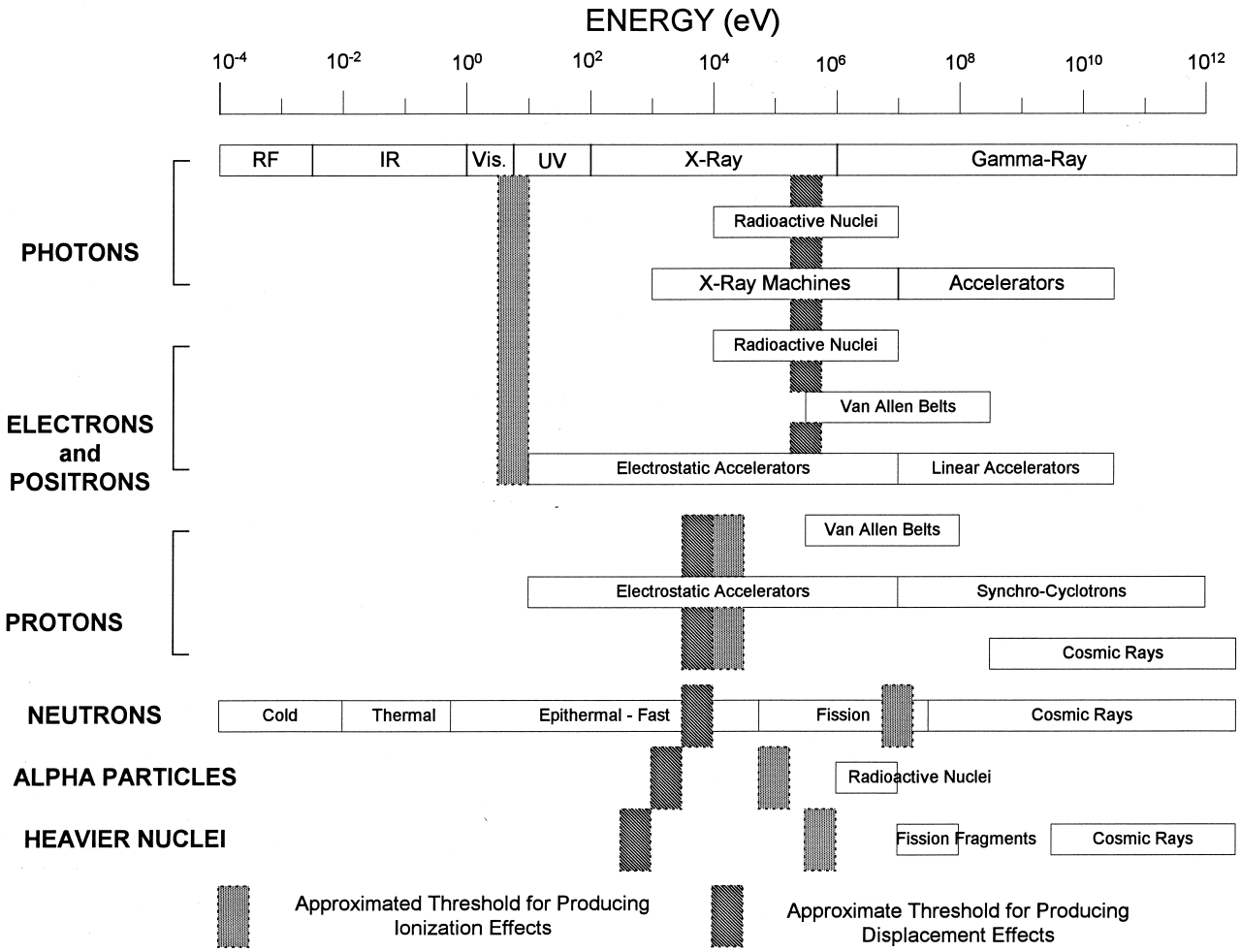


FIGURE 2 Energy ranges for various radiation environments and thresholds for radiation effects.

originate in the particles emitted from the sun in bursts associated with solar flares. A less energetic, steady stream of protons is emitted by the sun and is called the solar wind.

E. Ions

Alpha-particles (which are ions) consist of high-energy helium nuclei; these too are emitted by radioisotopes and stars and thus are found in interplanetary space. 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 produce large amounts of radiation damage in the solid so treated, usually requiring post-implantation high-temperature annealing. Very energetic ions of all known atomic masses are found in space. These are called cosmic rays.

F. 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 materials, especially carbon or hydrogen-containing moderator, reduces neutron energy to produce thermal neutrons. Cold neutrons of much lower energy can also be produced for research purposes. On collision with matter, fast neutrons produce much damage while thermal neutrons produce radioactivation. Atomic 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 at a single energy of 14 MeV.

G. Unstable Particles

When proton beams of GeV energy collide with matter, a complex array of unstable particles are produced. Among the better-known are muons, pions, and neutrinos. Many unstable particles are generated when cosmic rays, particularly very high-energy protons, probably originating from distant stars, strike the atmosphere.

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, fluence, and energy flux 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 a given 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-born energy in solids. For this, 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, define the radiation exposure. These forms of statement of a quantity of radiation do not describe the exact energy absorbed in a given material but are useful if the energy dependence of the exposure effect is similar to the radiation effect of interest.

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 of 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.

Method 2 arises from the fact that air ionization chambers are routinely used for quantifying ionizing radiation. Exposure in air due to a given radiation source is expressed in Roentgen units, or Coulombs per kilogram, based on the number of ions created in a given volume of air. However, for the radiation testing of electronics, it is better to express exposures by method 1, particularly if the effect of interest is not due to ionization (e.g., neutron-induced displacements).

D. Dose and Kerma

The term dose is a useful general description of the energy per unit mass that has been "dumped" into a material by a high-energy particle on its way through. The value for dose with crude specifications about the irradiating particles determines the magnitude of biological and chemical radiation effects. For crystalline solids such as silicon and metals, some further complications arise. It may be necessary to divide energy deposition into two fractions: ionization and atomic displacement. A new word for energy deposition has been introduced to assist with these distinctions. The word is *kerma*, meaning kinetic energy released in a material. To distinguish between the two forms of energy deposition, we qualify this word and speak of ionization kerma when referring to the portion of energy going into ionization, and so on.

Radiation dose and kerma are both measures of energy deposited. The new international unit is the Gray (Gy), which represents energy deposition per unit mass of one joule per kilogram. Many authoritative publications still employ the older practical unit, the rad, representing 100 erg/g. A dose of 1 Gy thus equals 100 rad. An exposure of one Roentgen deposits 86.9 rad in air.

IV. INTERACTION OF RADIATION WITH MATTER

A. General

The laws by which radiation is absorbed by matter are derived from multiple interactions of photons or particles with the atoms of the material. This section describes the primary processes, which are common to all matter. Later sections describe the mechanisms by which these primary processes are linked to the end effects such as atomic displacement and charge build-up.

For purposes of radiation effects, the relevant features of the interaction of radiation with a target material include the following:

1. Cross sections for the primary interactions that lead to significant effects. These include interactions that dominate the loss of energy from the incident radiation and those that produce unique progeny (e.g., a dense cluster of deposited energy).
2. Integration over the secondary effects that follow a primary interaction in terms that are appropriate to the effect being considered. For example, for ionization processes a measure of the total density of

electron-ion pairs produced is the ionization energy density. For some applications it is necessary also to account for the microscopic distribution of this energy.

These considerations lead directly to placing bounds on the effects and to scaling the effects produced by different incident radiation exposures (e.g., the relative effects of different energy spectra). The interactions of primary interest are summarized below.

B. Electrons

The rate of energy loss by electrons moving through silicon is summarized in Fig. 3. Electrons of energy up to 10 MeV lose most of their energy in ionizing collisions: target electrons are removed from their host atoms by Coulomb collisions. Thus, the rate of energy loss is nearly proportional to the density of electrons in the target, which is proportional to the mass density times Z/A. Therefore, the ionization-loss data in Fig. 3 for energies up to a few MeV can be applied directly to other materials, using a Z/A factor for a first-order correction. These primary interactions produce secondary electrons, mostly with a $1/E_s^2$, spectrum (where E_s is the energy of the secondary electron), which in turn lose energy mainly by ionization. The overall result is a slowing-down spectrum of electrons with an approximately $1/E_s^3$ shape and a particle density given by the ionization energy deposited per unit volume divided by an average energy loss per ion pair, E_p. As a result a complex track is produced, with a core of electron–ion pairs along the electron's path and small spurs where secondary electrons have been ejected at various angles.

The net effect of the gradual energy loss illustrated in Fig. 3 is that the energy of an electron passing through matter gradually decreases until it eventually stops at a distance defined as its range. An initially monoenergetic beam of electrons will be spread in energy by the statistics of energy loss, producing range straggling. Figure 4 illustrates a particular measure of penetration, the practical range, for electrons of various energies in aluminum and silicon. Since in most materials the energy loss is approximately proportional to density, it is conventional to represent the range in terms of distance times density, e.g., in g/cm^2.

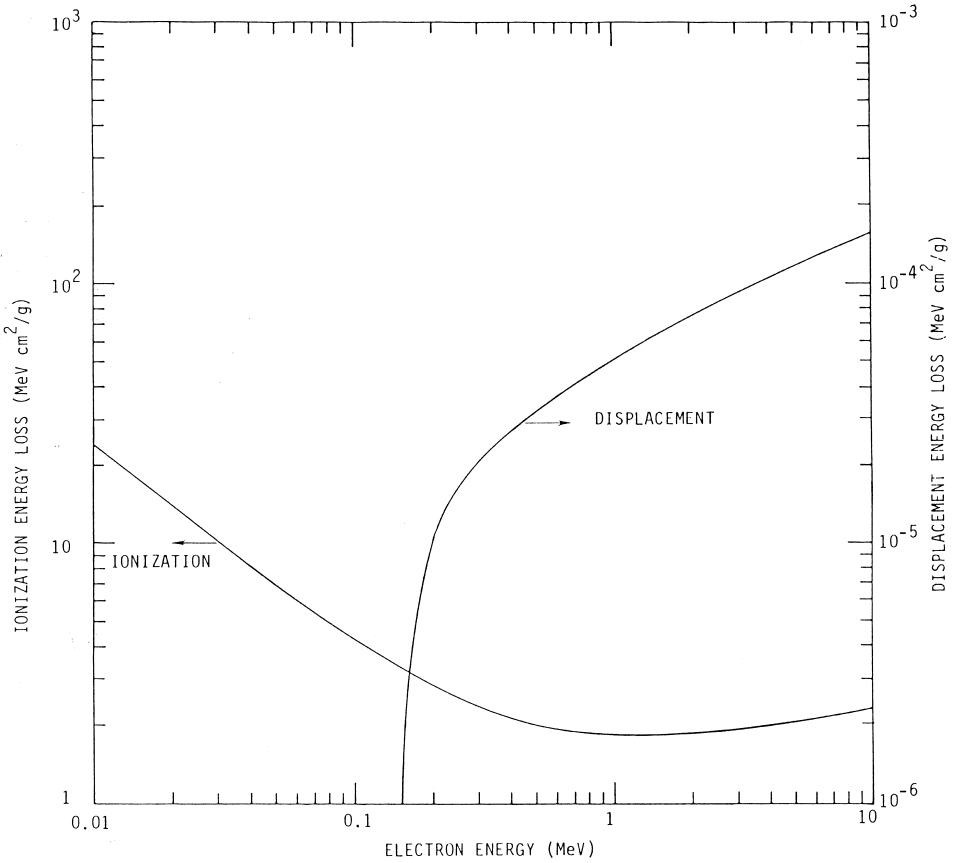


FIGURE 3 Electron energy loss in silicon.

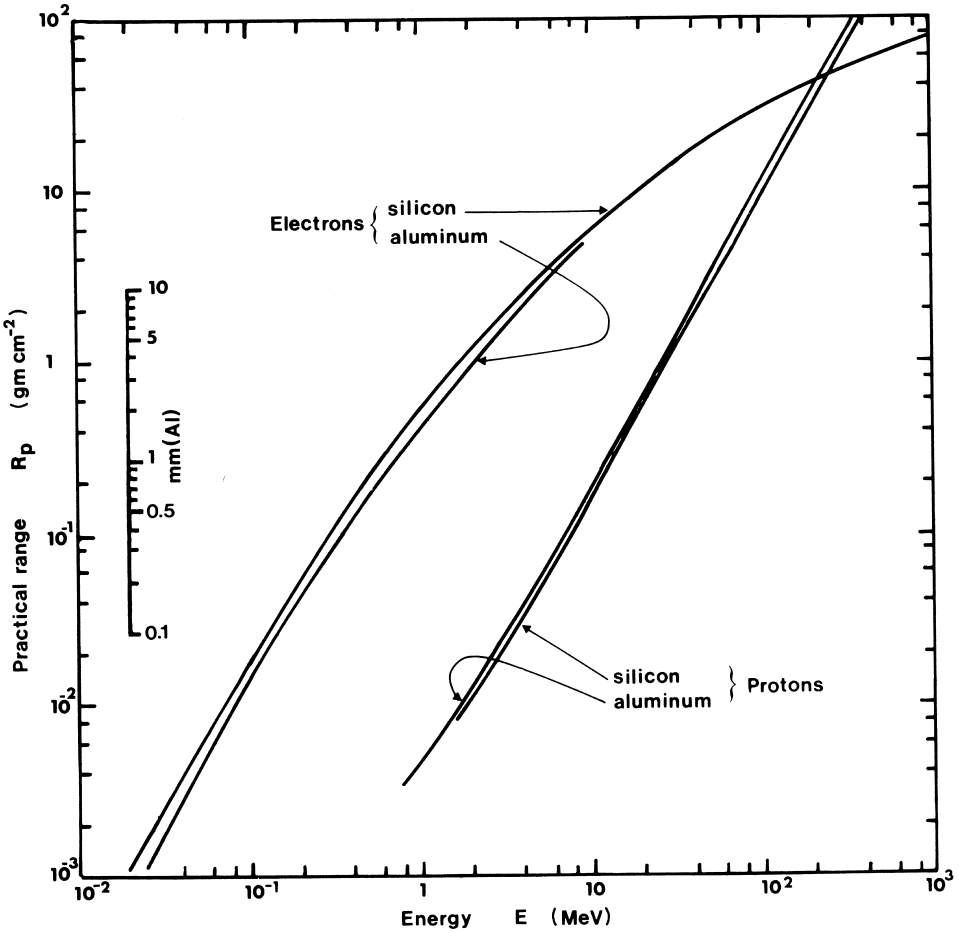


FIGURE 4 Range–energy curves.

Electrons with multi-MeV energies also emit bremsstrahlung radiation during acceleration in the electric field of the nucleus as a significant means of energy loss. Since the acceleration is proportional to nuclear charge, the rate of radiation energy loss is proportional to Z^2; that is, its relative importance increases with increasing atomic number. While bremsstrahlung removes energy from the incident electron, only a small fraction of the lost energy is deposited near the site of the interaction (i.e., the recoil energy imparted to the nucleus). Whether the radiated energy contributes to the radiation effect of interest depends on subsequent interactions of the emitted photons and on the large-scale geometry of the target.

Although most of the electron energy loss produces secondary electrons, which produce further ionization, a small fraction of an electron's energy is lost in close Coulomb collisions with nuclei, which may impart sufficient energy to the target nucleus to displace it from its position in a solid lattice. Such collisions also produce significant deflections in the electron's trajectory manifested as multiple scattering. The rate of energy loss for electrons in silicon by displacing collisions is also shown in Fig. 3. Since these are also Coulomb collisions, the energy spectrum of the recoil atoms is $1/E_s^2$. The electron energy threshold at 145 keV is due to a recoil-energy threshold at 12.9 eV, below which a silicon atom is unable to escape from its lattice site. The displacement process must be considered, even though the rate of energy loss by displacements is small compared with the ionization loss, because the effects are different. For example, ionization energy in silicon produces strictly transient effects, lasting only until the resulting electron–hole pairs undergo recombination; displacement energy produces a permanent change in important material properties, for example, the conductivity and minority carrier lifetime.

C. Photons

Energetic photons (e.g., γ and X-rays) carry no charge and little momentum. Most initial interactions of photons with a solid are with the electrons. The probability of an interaction is low, so γ-rays and X-rays are regarded

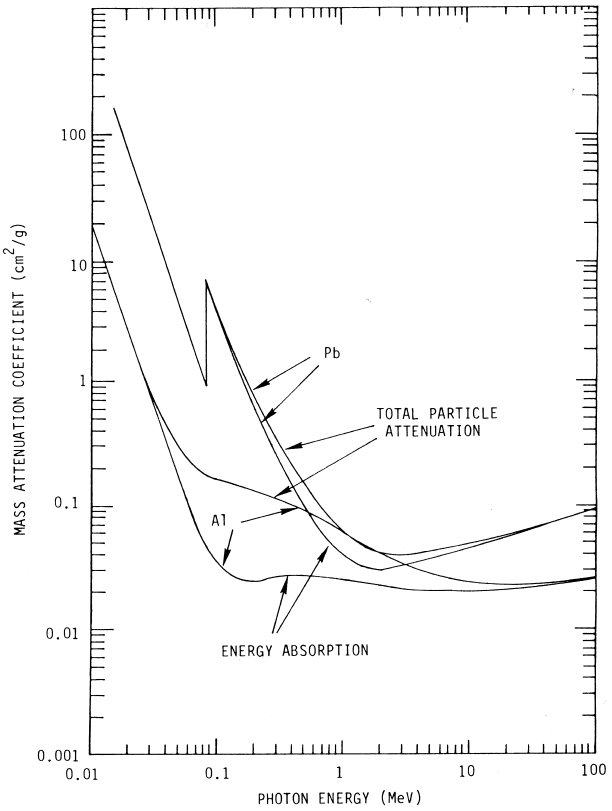


FIGURE 5 Photon absorption.

as penetrating radiation, compared with electrons of few MeV energies. Figure 5 summarizes the interaction length for photons of various energies in a few materials. As distinct from electrons, in which the ionization loss is the result of many small-loss interactions, the photon interactions are discrete. At the lower photon energies the energy loss is by a photoelectric process: the photon is absorbed and a photoelectron is given its energy. At medium energies Compton scattering dominates: the photon is scattered to produce a lower-energy photon, with part of its energy imparted to a target electron. At higher energies electron–positron pair production dominates, almost all of the photon energy is converted to mass and kinetic energy of the pair. Eventually, the positron is annihilated; if at rest it produces a characteristic pair of 0.511 MeV photons moving in opposite directions.

The Compton process depends only on the target electron density: the mass absorption coefficient scales as Z/A. The photoelectric and pair-production processes depend on higher powers of Z. It is useful to note that over a reasonable range of photon energies (e.g., ~100 keV to 3 MeV in aluminum) the energy absorption rate (i.e., the product of the Compton interaction cross section times the fraction of the photon energy imparted to the target electron) is almost independent of photon energy. Thus, the energy imparted to Compton electrons per gram of target material from a photon beam in this energy range can be estimated by dividing the photon energy fluence by 40 g/cm^2. As in the case of bremsstrahlung radiation from electrons, the contribution of the Compton-scattered photons to the radiation effect depends on subsequent interactions and the overall geometry. Since the interaction events are discrete, the transmission of a photon beam through an absorber can be represented by an exponential relationship: $I = I_o \exp(-\mu x)$, where I is the intensity after passing through a thickness, x, of a medium with absorption coefficient, μ. The intensity of surviving incident photons can be calculated by using the total absorption coefficient; the remaining energy fluence can be estimated by using the energy absorption coefficient.

Photons at higher energies can also undergo photonuclear reactions. Examples of such processes are (γ, n), (γ, p), (γ, α) and (γ, fission) reactions. In each case the result is not only the emission of one or more particles, the subsequent reactions of which may have to be accounted for, but also a recoil nucleus that deposits its energy very close to the site of the nuclear reaction. Typical photonuclear reaction cross sections have thresholds above 10 MeV and rise to peak values of 10^{-27} to 10^{-26} cm^2.

D. Neutrons

Since neutrons have no charge, their principal primary interaction is with the nucleus of target atoms. At energies up to a few MeV the most likely process is elastic scattering. The rate of energy loss by elastic scattering is approximately inversely proportional to the target atomic mass number, A. Therefore, hydrogen-containing materials are particularly effective in slowing down neutrons. Most neutron collisions deliver more energy to a target atom than needed to displace it from a solid lattice. Figure 6 presents the effective rate of energy loss by displacing collisions as a function of neutron energy in silicon. At the highest recoil energies the portion of the recoil atom's energy that is dissipated in ionization has been subtracted.

Neutrons also undergo nuclear reactions, producing energetic particles and recoil nuclei. Even low-energy (e.g., thermal) neutrons can undergo nuclear capture, usually accompanied by the emission of one or more photons. Of particular interest is thermal neutron capture in ^{235}U, which produces fission with ~200 MeV of kinetic energy in its fragments.

E. Protons

Proton interactions combine many of the features of a charged particle, like an electron, with a nuclear particle,

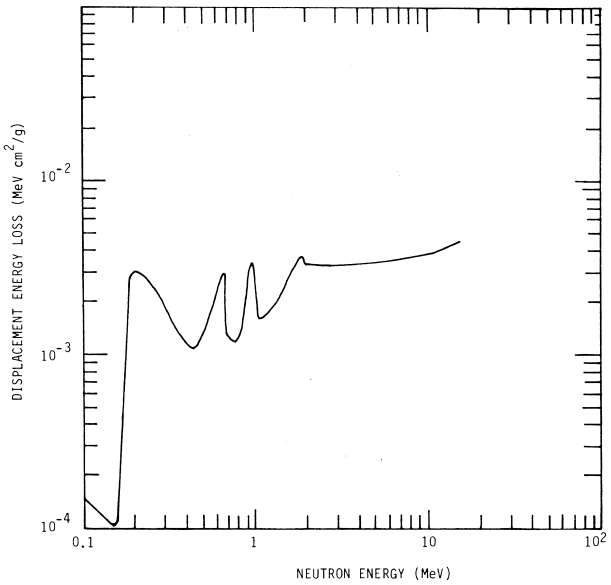


FIGURE 6 Neutron displacement energy loss in silicon. [Reprinted with permission from Smith, E. C., Binder, D., Compton, P. A., and Wilbur, R. I. (1996). "Theoretical and experimental determination of neutron energy deposition in silicon." *IEEE Trans. Nucl. Sci.* **NS-13,** No. 6, 11. Copyright 1966 IEEE.]

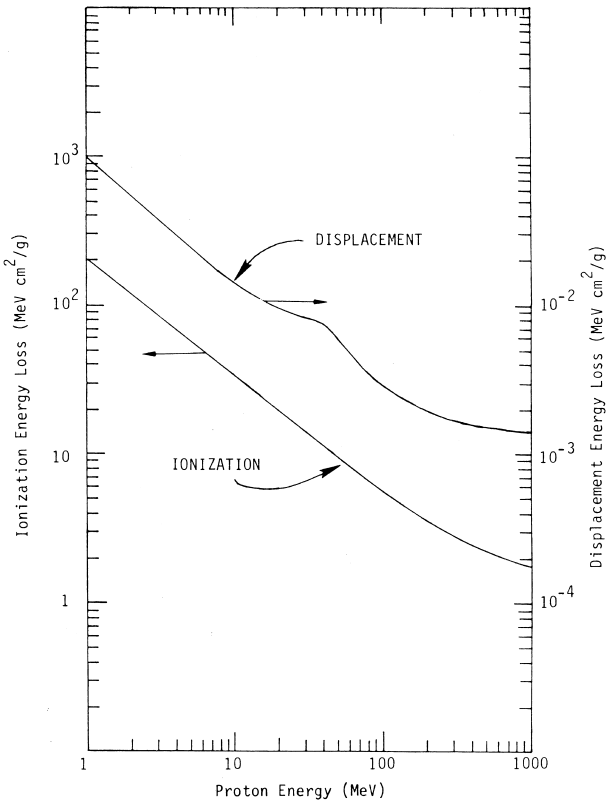


FIGURE 7 Proton energy loss in silicon.

like a neutron. The rate of energy loss by ionization and by displacements for protons as a function of energy is shown in Fig. 7. The portion of the curves below –20 MeV is essentially due to Coulomb interactions (electron emission for the ionization process and nuclear scattering for displacements). At higher energies nuclear reactions become important, especially in producing additional displacement energy by imparting energy to recoil atoms. The net range–energy relation is shown in Fig. 4. Since proton energy loss rarely involves large energy transfers, range straggling of protons is small and a monoenergetic beam of protons retains a small energy spread as it penetrates matter.

F. Heavier Energetic Particles

Heavier charged particles (e.g., α particles, heavier nuclei, fission fragments) interact similarly to protons, with the following attributes: (1) The rate of ionization energy loss is proportional to the square of the particle's net charge and inversely proportional to the square of their velocity. (2) When the velocity is less than the classical orbital electron velocity, the particle's charge becomes partially neutralized by capturing and retaining electrons.

Thus, when a fast heavy ion slows down its rate of ionization increases to a high peak, then decreases as its charge decreases. Once it slows to velocities below the classical velocity of the outermost electron orbits, it moves as a neutral atom, losing energy primarily by collisions between screened nuclear Coulomb fields.

G. Recoil Atoms

Whenever a close encounter occurs with a target nucleus, by Coulomb, nuclear elastic, or nuclear inelastic interactions, a significant kinetic energy is imparted to the target atom. In many cases the velocity of the recoiling atom is less than the classical orbital velocities of its electrons. In this case, the atom will move as a neutral entity through the target material, losing energy by screened Coulomb collisions with other target atoms. As a result many other atoms are displaced near one another, resulting in the initial recoil energy being distributed in a short-range displacement cascade. The effective size of such structures is measured by the range–energy relationship for recoil atoms, as illustrated for silicon in Fig. 8. If the recoil energy is sufficiently high, a portion of its energy is imparted to ionization (Fig. 9).

H. Cosmic Rays

Cosmic rays are mostly protons and heavier particles (i.e., helium nuclei), albeit at very high energies. However, a

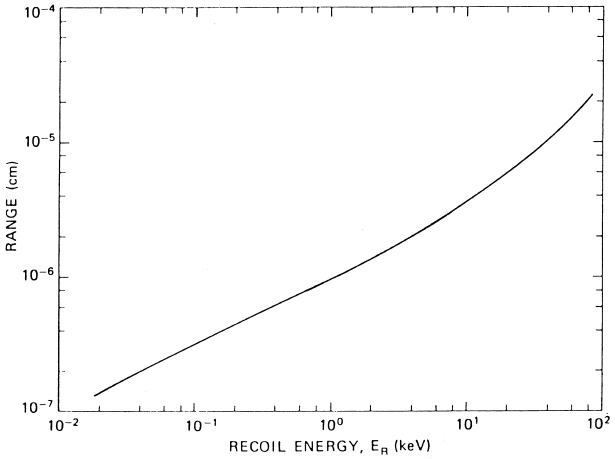


FIGURE 8 Experimental range–energy curve for silicon. [Reprinted with permission from Nichols, D. K., and van Lint, V. A. J. (1996). "Energy loss and range of energetic neutral atoms in solids." *In* "Solid State Physics," Vol. 18 (F. Seitz and D. Turnbull, eds.), Academic Press, New York. Copyright 1996 Academic Press.]

unique attribute of cosmic rays is that there are also significant fluxes of extremely energetic (multi-GeV) heavy ions (e.g., carbon, iron nuclei). Such particles are capable of delivering a very high local density of ionization, which is of particular interest for single-event phenomena. Typical values for the energy-loss spectrum of cosmic ray iron nuclei are shown in Fig. 10.

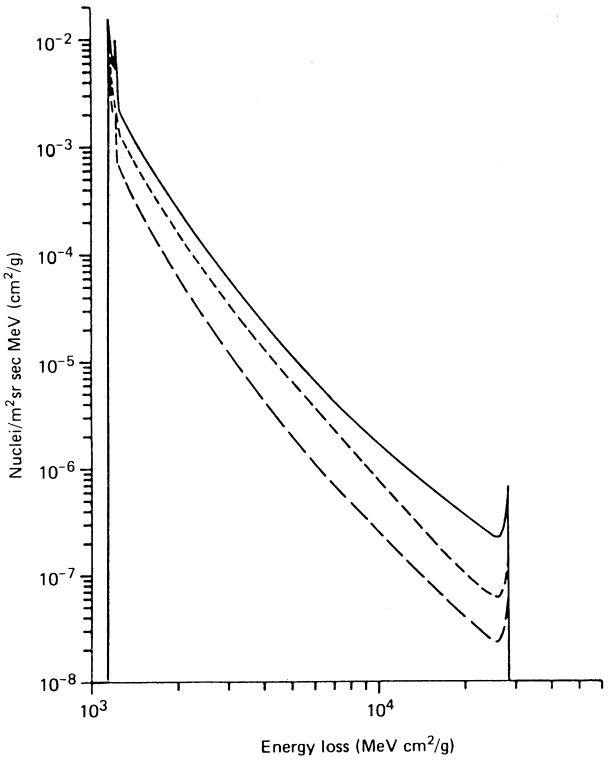


FIGURE 10 Differential energy-loss spectrum for cosmic ray iron nuclei. [Reprinted with permission from Petersen, E. L., Shapiro, P., Adams, J. H., Jr., and Burke, E. A. (1982). "Calculation of cosmic-ray induced soft upsets and scaling in VLSI devices." *IEEE Trans. Nucl. Sci.* **NS-29,** No. 6, 2055. Copyright 1982 IEEE.]

V. TRANSIENT IONIZATION EFFECTS

When ionization is produced in a target material, electrons and ions (or holes), which are temporarily mobile, are produced. The ionization energy loss per electron–ion or electron–hole pair produced, E_p, is independent of the type of ionizing particle and depends primarily on the target material. Table II presents a summary of the E_p values for a variety of materials. As a rule of thumb E_p is equal to two to four times the ionization potential of the target material, the correct value being closer to two times for gases with high ionization potentials and closer to four times for semiconductors with low ionization potentials.

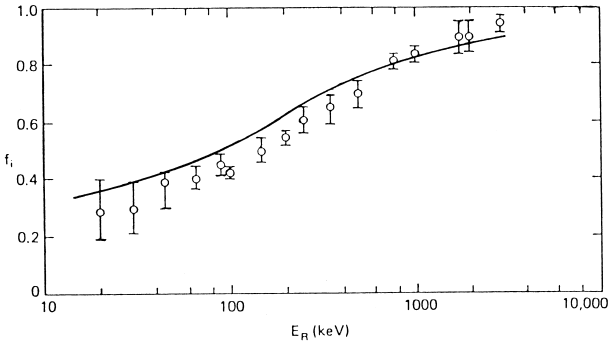


FIGURE 9 Fraction of energy f_i going into ionization as a function of recoil E_R. [Reprinted with permission from Sattler, A. R. (1965). "Ionization produced by energetic silicon atoms within a silicon lattice," *Phys. Rev.* **138,** A1815.]

TABLE II Energy Loss Per Electron–Ion Pair

Material	E_p (eV/pair)	Ionization potential (eV)
Si	3.7 ± 0.1	1.11
Ge	3.0 ± 0.3	0.67
N_2	36.3	15.5
Air	35.0	14.9
He	46.0	24.5
KCl	38	10.7
SiO_2	18	8.5

The effects of these charge carriers may include the following:

1. The conductivity of the target material is increased because these charge carriers will drift under the influence of an electric field.
2. Electric fields will be created within the target material because these charge carriers will tend to flow under the influence of gradients in their density. This is the source of ionization-induced photovoltages at interfaces.
3. Light may be emitted during transitions in which the mobile carriers are captured into lower energy states (e.g., during electron–ion recombination).

A. Gases

Transient ionization effects can be illustrated by considering the effect of a very short uniform ionization pulse on a parallel-plate gas ionization chamber (Fig. 11). Consider first a low, microscopically uniform ionization density produced in a short pulse. The effect of this ionization is to produce a low density of electron–ion pairs distributed throughout the gas volume. The electrons and ions will move in opposite directions under the influence of the applied field; eventually the electrons will be collected by the anode, and at a much later time the ions will be collected by the cathode. The resulting current pulse in the external circuit is illustrated in Fig. 11a.

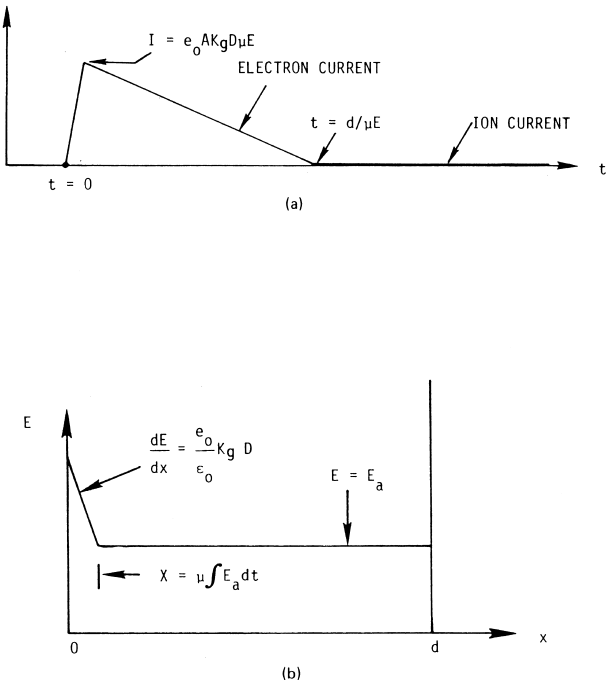


FIGURE 11 Parallel-plate ionization chamber. (a) Low-dose current pulse; (b) high-dose electric field profile.

If an electronegative gas species (e.g., O_2) were added to the gas, the electrons might attach to form negative ions before they drifted to the anode. This would not change the total amount of charge collected, but it would change the time scale. If the duration of the ionization pulse were long compared with the electron and ion collection times, the current in the ion chamber circuit would follow the ionization rate, representing an equilibrium between the generation of electron–ion pairs by the radiation and their collection at the electrodes.

Consider now increasing the intensity of the short-pulse ionization. This increases the local density of electrons and ions, with the result that some pairs may undergo electron–ion recombination before they are removed from the ion chamber. If the electrons are attached, the positive and negative ions are even more likely to undergo ion–ion mutual neutralization, because they will spend more time in the gas before sweepout. Furthermore, as the electrons move toward the anode, they leave behind a positive space charge in the vicinity of the cathode. This increases the electric field near the cathode but decreases it in the remainder of the ion chamber where the electrons are (Fig. 11b). As a result, the fast component of current is decreased in amplitude and increased in time scale. If attachment were occurring. the magnitude of the early-time charge collection would decrease. If recombination is important. the total charge collected will decrease.

If the ionization were not created microscopically uniform but were actually produced along a number of heavily ionizing particle tracks, the space charge effects described near the cathode could occur near each particle track. Moreover, the local proximity between positive ions and electrons would increase the electron–ion recombination rate, which is called preferential recombination. If the electron from each ionization event were slowed to thermal energy very close to its parent ion (e.g., in a high-density medium), the Coulomb attraction between the opposite charges could be sufficient to overcome thermal energy of motion. Then most of the electrons would be recaptured by their parent ion. This process is called geminate recombination.

Even if no electric field were applied to the ion chamber, the electrons formed in the ionization pulse would diffuse out of the gas to the walls at a faster rate than the positive ions. This process would result in a positive space charge, which would tend to inhibit further electron diffusion. If the initial ionization density were large enough, a potential difference of a few times kT would produce an equilibrium between the electron diffusion gradient and electron drift in the opposing electric field. This is the equivalent of a Langmuir sheath in plasma physics.

Ionization-induced conductivity in air illustrates methods of estimating transient ionization. In air at atmospheric pressure electron-ion pairs are formed at the following rate:

$$1\ \text{rad(air)} = 100\ \text{erg/g}$$

$$\text{deposits:}\quad 100\ \text{erg/g} * 1.2 * 10^{-3}\ \text{g/cm}^3$$

$$/(1.6 * 10^{-12}\ \text{erg/eV}) = 7.5 * 10^{10}\ \text{eV/cm}^3$$

$$\text{produces:}\quad 7.5 * 10^{10}\ \text{eV/cm}^3/(35\ \text{eV/ion pair})$$

$$= 2.18 * 10^{9}\ \text{ion pairs/cm}^3$$

The electrons move under the influence of an electric field at a velocity per unit electric field given by their mobility μ_n. They are attached to O_2 molecules by a three-body process (i.e., proportional to oxygen pressure squared) in a time, τ_n, dependent on electric field and water vapor concentration. Representative mobilities and attachment rates are given in Fig. 12.

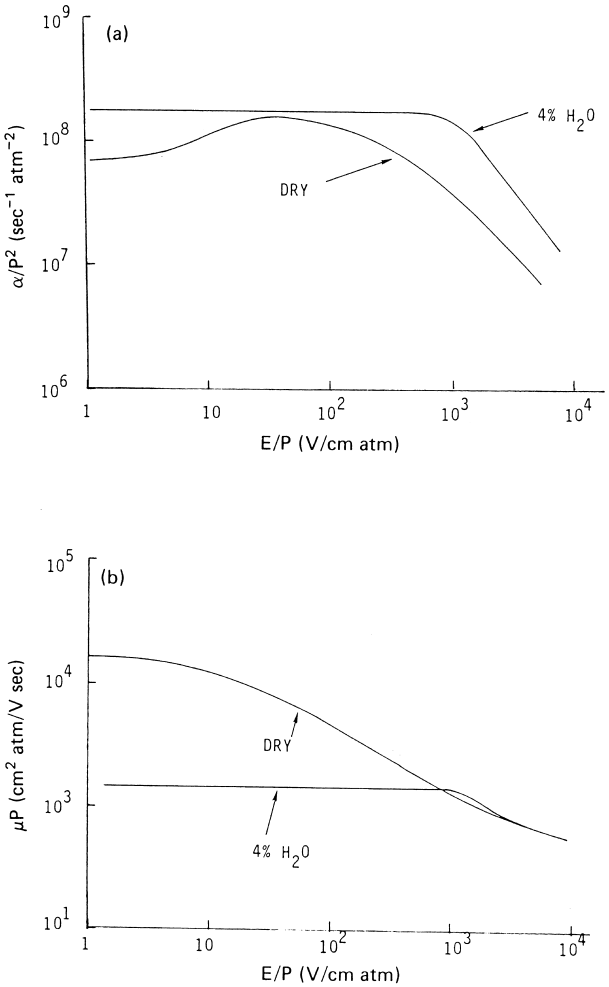


FIGURE 12 Electron reaction rates in air. (a) Attachment rate; (b) mobility.

Under ionization exposure that is long compared with the attachment time, the ionization-induced conductivity is

$$\sigma = n_{e_o}\mu_n = e_o K_g \dot{D}\mu_n\tau_n,$$

where e_o is the electronic charge, K_g is the electron–ion density formed per unit dose ($K_g = 2.18 * 10^9$ cm^{-3}/rad in air at atmospheric pressure) and $\dot{D}$ is the dose rate. Since K_g. is proportional to pressure, μ_n inversely proportional to pressure, and τ_n inversely proportional to pressure squared,

$$\sigma(P) = 3.49 * 10^{-10}\,\mu_{n1}\tau_{n1}\dot{D}/P^2,$$

where P is the pressure in atmospheres and μ_{n1}, τ_{n1} are the values of mobility and attachment time, respectively, at 1 atm. For reasonably large electric fields (100<E/P< 1000 V/cm atm), $\mu_{n1}\tau_{n1}$ in moist air is approximately constant at $\sim 3 \times 10^{-5}$ cm^2/V. This leads to an approximate value of

$$\sigma(P)(\text{Siem/cm}) \approx 10^{-14}\,\dot{D}(\text{rad/sec})/P^2(\text{atm})$$

Eventually the electrons and ions recombine, releasing the energy of the ionization potential. A large fraction of this energy may be radiated as photons at wavelengths extending from the visible into the ultraviolet. The processes and emissions are similar to those produced by any other mechanism for creating electron–ion pairs, e.g., sparks, lightning.

B. Semiconductors

Many of the processes described above for gases have analogs in semiconductors. Ionizing radiation produces electron–hole pairs in silicon at a rate of one pair for every 3.7 eV of ionizing energy deposition, corresponding to a carrier density of 3.9×10^{11} cm^{-3} times the dose in rad(Si). There is no evidence for geminate recombination in silicon, at least at room temperature, indicating that the electron range is sufficiently long to escape the Coulomb field of its parent hole. While they are free, the electrons and holes drift under modest-value electric fields at a velocity proportional to their respective mobilities, which depend somewhat on the doping and defect density. For nominally pure silicon at room temperature, the mobilities for electrons and holes are $\mu_n = 1300$ cm^2/V-sec and $\mu_p = 370$ cm^2/V-sec respectively, leading to an excess conductivity produced by a dose D of

$$\Delta\sigma(\text{S/cm}) = e_o\Delta n(\mu_n + \mu_p) = 0.0105D[\text{rad(Si)}].$$

The electrons and holes will recombine directly, or be trapped separately and eventually recombined, in semiconductors in a time determined by the density of trapping centers. The mean excess carrier lifetime may vary from less than 1 nsec to longer than 1 msec, depending on material purity and doping. In typical silicon devices

it varies from 1 nsec to 1 μsec. Part of the recombination energy may be emitted as photons, especially in semiconductors like gallium arsenide, in which the minimum of the conduction band occurs at the same momentum as the maximum of the valence band.

Though the effect of ionization on the conductivity in semiconductors is noticeable, the effect on reverse-biased *pn* junctions is much more dramatic. The increase in conductivity is compared with the quiescent conductivity due to majority carriers. The reverse current across a *pn* junction is proportional to the minority carrier concentration, which is typically 10 orders of magnitude less than the majority carrier concentration. The charge is collected from a distance of approximately a minority-carrier diffusion length, $L_D = \sqrt{D\tau}$, where D is the minority carrier diffusion coefficient and τ the minority carrier lifetime. The charge is collected during a time of approximately τ. Typically, for $\tau = 3{*}10^{-8}$ sec and $D = 30$ cm^2/sec, $L_D = 10^{-3}$ cm. The charge flow across a reverse-biased junction of area 10^{-4} cm^2 exposed to a dose of 1 rad(Si) collecting carriers from a depth of 10^{-3} cm is:

$$Q_{1\text{rad}}(\text{Coul}) = 1.6{*}10^{-19}\text{Coul/carrier}{*}3.9{*}10^{13}\ \text{carriers/(cm}^3 - \text{rad)}{*}10^{-7}\text{cm}^3 = 0.62\text{pCoul.}$$

Such a charge flowing in the 30-nsec collection time produces a peak current of 21 μA, which is many orders of magnitude greater than the normal reverse leakage current. A similar estimate can be produced for long-pulse exposures. A steady-state dose rate exposing the same junction would produce a steady reverse leakage current of 0.62 pA times the dose rate.

The foregoing examples illustrate the process of estimating the response of a semiconductor device to a pulse of ionizing radiation. The relation between ionizing dose and carriers generated per unit volume is determined by the energy loss per ion pair, which is independent of the particle producing the ionization. The effect of those carriers on the semiconductor device is determined by the device geometry and doping. The effective volume from which a junction can collect minority carriers in normal operation usually is also the volume from which a reverse-biased junction will collect ionization-induced carriers.

C. Insulators

The principal transient manifestations of ionization effects in insulators are ionization-enhanced conductivity and optical emission. Although the amount of energy required to form an electron–hole pair in an insulator is believed to be approximately three times the insulator band gap, geminate recombination is frequently important. This implies that the effective density of mobile carriers that escape from their original partner is a small fraction of the total density of carrier pairs created and that this fraction depends on the electric field E. A theoretical derivation of the escape probability, φ(E), by Onsager yields

$$\phi(E) = \phi(0)\left[1 + e_o^3 E \Big/ \left(8\pi\kappa\varepsilon_o k^2 T^2\right)\right],$$

where $\varphi(0)$ is the escape probability at zero electric field.

For most insulators the effective time for immobilization (trapping) the mobile carriers is very short ($<$1 nsec). Therefore, the measured transient conductivity almost always represents an equilibrium between the generation of free carrier pairs and trapping at defect sites preexisting in the insulator. Nevertheless, apparent conductivity decay components much longer than 1 nsec, even extending to many hours are observed in many insulators. These are believed to be related to shallow traps: A carrier trapped at a defect with a small ionization energy can be released thermally, contributing to a delayed conductivity component until it is eventually captured at a deep trap. Since most technological insulators, like polymers, have a wide range of chemical impurities and structural defects, it is not surprising to observe a wide range of apparent trapping energies.

An alternative description of carrier transport in insulators replaces the concept of mobility, which is based on wave functions effectively extending over many lattice spacings, by a concept of hopping conduction. At any one time a carrier is effectively localized at one lattice site, but its wave function extends slightly to neighboring sites. There exists a small probability per unit time that the carrier can tunnel into the adjacent site. The probability of the carrier appearing at adjacent sites in various directions is influenced (e.g., biased) by the electric field. There is also a wide range of effective barrier heights for the transitions. Obviously, the lower height transitions will occur more rapidly than the remainder. This model, which is called the continuous-time random-walk (CTRW) model, predicts carrier motion and delayed conductivity that depends on the logarithm of time, rather than as a sum of exponentials. Data on carrier transport in thin, thermally grown films of amorphous SiO_2 appear to fit this model well.

In addition to complications with geminate recombination and delayed conductivity, insulator conductivity experiments are chronically affected by contact barriers. As in the gas ion chamber described above, when carriers that are swept away from a contact by the electric field are not replenished by that contact a space-charge layer is established. As a result the applied voltage is distributed between a boundary layer with a high electric field and the bulk of the sample, in which the field is decreased from the preionization value. As a result, the apparent conductance of an insulator sample with electrodes decreases as a function of accumulated charge transfer through the sample. Attempts to circumvent this problem

with noncontact methods of measuring conductivity (e.g., by microwave absorption) have not yet been generally successful with insulators. In addition to the nonlinearities discussed above, carrier trapping mechanisms are responsible for a change in the ionization-induced conductivity with accumulated dose. It was established long ago that photoconductors could exhibit nonlinear responses as a function of exciting intensity. A distribution of trapping sites was used to explain these results, which were generally measured under conditions of long, steady illumination. A simple formula was developed,

$$\sigma = K\dot{D}^{\Delta},$$

where σ is the transient conductivity measured at the dose rate, $\dot{D}$, K is a constant that depends on the trap density and cross sections, and Δ is a numerical constant that depends on the shape of the distribution of trap energy levels, with values typically $0.5 < \Delta < 1.0$. Early workers in transient-ionization-induced conductivity used the same relationship to fit their data. However, it was demonstrated theoretically and experimentally that this relationship is inappropriate to short-pulse nonequilibrium ionization exposures. Instead, the proper relationship for prompt ionization response (i.e., without delayed conductivity) is

$$\sigma = \dot{D}F(D),$$

where $F(D)$ is a function of accumulated ionizing dose that depends on the material trap structure. $F(D)$ must approach a constant, $F(0)$, at low enough doses that individual ionizing particle tracks do not overlap. The transition between this short-pulse relation and the steady-state formulation previously applied to photoconductivity occurs as follows:

1. As dose is accumulated at a particular dose rate, the traps will gradually be filled to a state in which their occupancy no longer changes. In this state the rate of capture of holes and electrons into each trap is equal, because the throughput of carrier pairs is greater than the number of available traps.
2. If measurements are made with ionization pulses whose duration is longer than some of the delayed conductivity relaxation times. the apparent transient conductivity includes such delayed conductivity components. Since the relative amount of delayed conductivity to prompt conductivity may change as a function of dose delivered in the delayed component relaxation time, there can be an apparent dose rate dependence of this transient conductivity.

Table III presents a typical set of coefficients $F(0)$, i.e., the ratio of conductivity to dose rate in a variety of insulators in the low dose limit. The purest materials—those in which the trap density is smallest—exhibit the largest coefficients (e.g., sapphire, crystalline quartz), presumably due to higher mobilities and longer trapping times. The lowest values are associated with polymers. Figure 13 presents examples of the dependence of the ionization-induced conductivity coefficient on accumulated dose.

TABLE III Typical Insulator Conductivity Coefficients

Material	$F(0)$ (Siem-sec/rad-cm)
Single-crystal MgO	$5*10^{-15}$
Single-crystal Al_2O_3	$4*10^{-15}$
Single-crystal SiO_2	$2*10^{-16}$
Pure fused silica	$3*10^{-16}$
Polycrystalline Al_2O_3	$6*10^{-17}$
Teflon	$8*10^{-18}$
Polyethylene	$2*10^{-18}$
Mylar	$5*10^{-19}$
Kapton	$5*10^{-19}$

D. Charge Transfer

Inevitably, when electronic equipment is exposed to nuclear radiation, charge is transferred across interfaces by the motion of energetic charged particles. As a result currents are induced into the electronic circuits. If the incident radiation consists of charged particles that are stopped in the electronic assembly, the cause of the charge transfer is obvious: The particles are depositing charge where they stop. Even if they reach the end of their range in nonconducting material, their image charge is still induced into the nearest conductors and the associated currents may appear in the circuits.

If the incident radiation is composed only of uncharged particles. charge transfer is still likely to occur. For example, a solid surface exposed to a pure photon beam will emit electrons, which have been liberated near the surface by photoelectric, Compton, or pair production interactions. The electron emission coefficient for a variety of materials exposed to various photon energies is shown in Fig. 14. At photon energies in the photoelectric interaction regime the emission coefficient falls off with increasing energy because of the steep energy dependence of the photoelectric interaction cross section. In the Compton interaction regime the emission coefficient increases with energy, because the thickness of the layer from which the electrons can escape increases with increasing range of the higher-energy Compton electrons.

If a neutral beam, like a beam of photons, is moving through a homogeneous solid, the electrons emitted from a particular layer tend to be replaced by those emitted from adjacent layers. Thus, the charge transfer tends to cancel out, except for two effects:

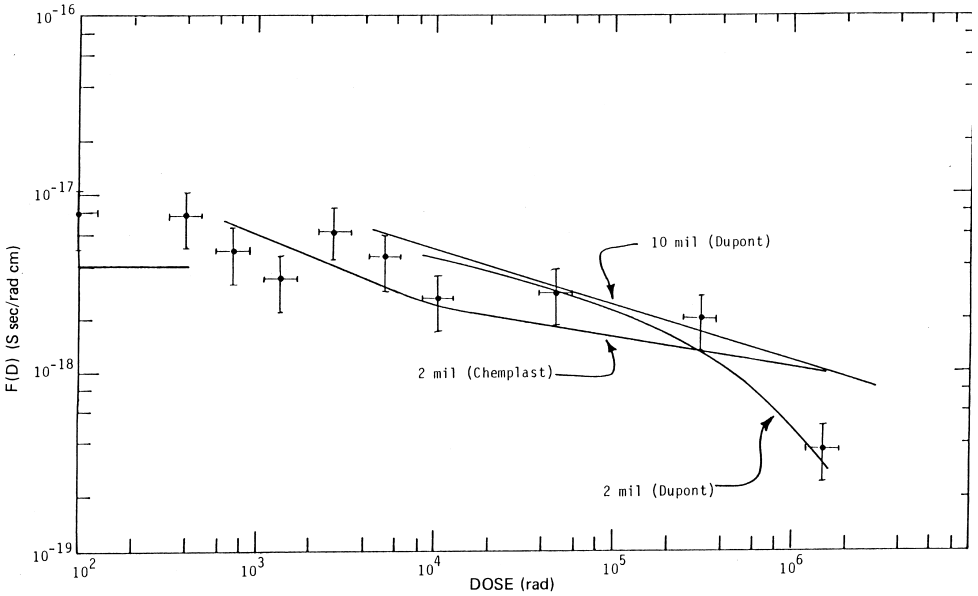


FIGURE 13 Summary of ionization-induced conductivity in Teflon samples.

1. There is still a net current flowing in the direction of the photon beam, and this current produces a magnetic field. If the current and field are sufficiently intense, significant voltages can be induced into circuitry by magnetic coupling.

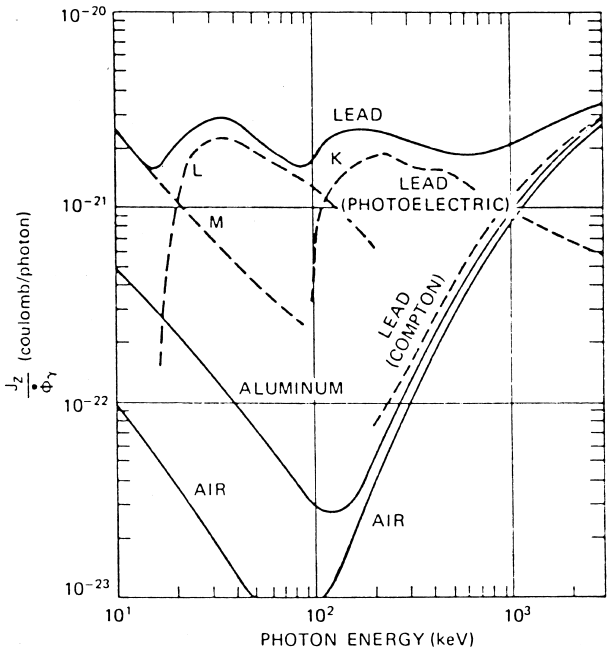


FIGURE 14 Photon-induced electron emission. [Reprinted with permission from van Lint, V. A. J., Flanagan, T. M., Leadon, R. E., Naber, J. A., and Rogers, V. C. (1980). "Mechanisms of Radiation Effects in Electronic Materials," Vol. 1, Wiley, New York. Copyright 1980 John Wiley and Sons.]

2. The cancellation of charge is not quite complete, because the photons are slightly attenuated as they pass through the material. The net deposited charge density in a homogeneous material is equal to the charge moving across any plane divided by the photon absorption length.
3. Near the interface between dissimilar materials there can be net charge transfer due to differences in electron emission coefficients. This imbalanced charge is deposited within one electron range of the interface.

Neutral particles other than photons can also be responsible for charge transfer, because almost all nuclear particles produce charged progeny. For example, neutron irradiation of a hydrogen-containing material will produce recoil protons that carry charge, just as the electron secondaries from photon interactions do. There is a strong dependence of proton emission on neutron energy (because of the steep proton range–energy relationship) and material composition (because heavier atoms receive less energy from neutron collisions and are less likely to be charged). Even thermal neutrons can produce charge transfer via the gamma rays that are emitted when they are captured in nuclei.

E. Effects on Electronic Devices and Circuits

Under the most likely situations, the most important transient ionization effects are the excess currents induced into reverse-biased semiconductor junctions. At low ionization

rates these represent excess leakage currents: at higher rates they can saturate the junction and make it temporarily inactive. In digital circuits these currents can temporarily change the logic state of a gate circuit or permanently change the state of a flip-flop or other memory device. The magnitude of the transient effect of a short pulse of ionizing radiation can be estimated by calculating the charge generated per unit volume of semiconductor material, which is directly proportional to the dose deposited in the device, and multiplying by an effective charge collection volume. The collection volume for most devices is determined by the junction area and an effective collection thickness. Apart from funnel effects, the collection thickness is equal to the sum of the thickness of the depletion layer in the reverse-biased junction and the lesser of a diffusion length or the epitaxial layer thickness, that is, the thickness of the layer from which carriers can diffuse to the junction. The depletion layer component of this current is collected essentially instantaneously (<1 nsec); the other component is collected in the time required for minority carriers to diffuse across the remainder of the collection layer (e.g., the minority carrier lifetime if the diffusion length is the limiting thickness).

For long ionization pulses the estimate proceeds along the same lines, except that the dose rate is used to calculate the rate of carrier generation per unit volume. This rate multiplied by the collection volume yields the junction current directly.

Capacitor leakage is usually the next most important effect on electronics. Although insulator conductivity current densities are usually many orders of magnitude below *pn*-junction photocurrent densities, capacitors usually have large areas and thin dielectrics, and they are frequently used in applications in which a small change in voltage can have serious and long-lasting effects. The simplest way to estimate the effect of a short ionization pulse (short compared with the resistance–capacitance (RC) recharging time of the circuit) depositing a dose D in a charged capacitor dielectric is to estimate the voltage change by

$$V = V_o \exp(-D/D_o),$$

where V is the voltage on the capacitor after the pulse, V_o the voltage before the pulse, and D_o an almost constant depending on the dielectric material: D_o is related to the ionization-induced conductivity coefficient $F(D)$ by

$$D_o = \kappa \varepsilon_o / F(D),$$

where κ is the dielectric constant and ε_o the permittivity of free space. Typical values of D_o range from 10^4 to 10^6 rad.

For an ionization pulse that is long compared with the capacitor recharge time, the current flowing through the capacitor exposed to an ionization pulse at a dose rate of $\dot{D}$ is

$$I = CV\dot{D}/D_o,$$

where C is the capacitance of the capacitor and V the voltage across it.

In some special applications the conduction in ionized air near the circuit conductors can have a significant effect on an electronic circuit. Unfortunately, ionization-induced conductivity in air is a nonlinear function of electric field and depends on other parameters, such as the humidity of the air. An approximate relationship that can be used for estimating effects at atmospheric pressure is

$$\sigma(\text{Siem/cm}) \approx 1^* 10^{-14} \dot{D}(\text{rad/sec}).$$

The actual conductivity can be larger by an order of magnitude, especially at low electric fields in dry air.

In most electronic circuits charge transfer is usually small compared with junction photocurrents. An accurate estimate of charge transfer is difficult to compute, because it is frequently a difference between various components (e.g., forward minus backward emission, different materials) and depends in a sensitive manner on the exact spectrum. For photons with energy between 100 kV and a few MeV impinging on low- to medium-Z materials an estimate of charge transfer can use 5 pC/cm^2-rad(Si). When the materials and geometry are controlled to balance the charge transfer, an upper limit of the net transfer equal to one-tenth of this value is reasonable.

VI. LONG-TERM IONIZATION EFFECTS

A. Mechanisms

As described in Section V, ionization effects covers a broad range of phenomena in which electrons are excited and leave their parent atom. In gases this creates an electron and a positive ion; in solids, an electron–hole pair. The pair may recombine, or, if a field is present, the carriers may drift apart, producing electrical conduction and possibly a local accumulation of charge. While the initial effects of ionization may fade, certain key performance features stay changed for such a long period that we must regard them as a near-permanent degradation. This "thermally stable" disruption can sometimes be reversed, for example, by heating in an "annealing oven," and so the word "long-term" is the best description. In engineering materials, semiconductors, glasses and plastics are particularly susceptible, and some devices using these materials are mentioned in Table IV. In semiconductors, it is the surface regions which are most affected by ionization.

TABLE IV Materials and Devices That Generally Have Poor Tolerance to Radiation

Semiconductor devices
Optical lenses
Optical fibers
Optical windows (e.g., encoder plates)
Elastomers (e.g., plastic bellows)
Plastic bearings
Lubricants
Adhesives
Hydraulic fluids
Paints
Reflective coatings
Thin insulators
Photosensitive materials
Gas sensors
Liquid-ion sensors
Surface-active reagents
Piezoelectric transducers
Micropositioners

With regard to the time sequence of radiation effects, long-term ionization effects take up where the transient ionization effects, discussed in the last section, leave off. We might choose to place the dividing line at 1 sec or many seconds. transient carriers are usually immobilized in much less than a second but some other stabilization processes take longer. In some cases the immobilization process restores the material to its preirradiation condition, meaning that there are no long-term effects. This is generally the case with excess carriers generated in semiconductors: carriers are trapped and then annihilated with no effects other than the production of some low-energy photons and phonons (e.g., heat).

An ionized gas illustrates a simple type of long-term ionization effect. Eventually, the electrons, negative ions, and positive ions, generated in the gas, will recombine. Electron–ion recombination is a particularly violent process. It releases more than 10 eV of excess energy, much of which appears as kinetic energy of the fragments into which the ion is usually disrupted. The result is that the former gas molecule is converted to a pair of free radicals (e.g., nitrogen atoms), which then react chemically to produce a new compound. Such chemical effects of ionizing radiation are common in gases and organic material, and are called radiolysis. In radiation chemistry, a G factor is commonly used to describe the production efficiency of new species, where G is the number of molecules formed per 100 eV of ionization energy deposited. In air, a species for which one molecule is formed per ion pair would have a G value of 3. In polymers a similar relationship can be used to describe the rate of chain breaking or cross-linking. In lists of the radiolytic products of organic materials, many G values may be given for any one starting compound, because a mixture of reactions between free radicals leads to many different products. The radiolysis of water is probably the first step in radiation damage to living cells. Radiolytic effects at the surface of semiconductors are sometimes detectable.

A second class of long-term ionization effects is related to chemical effects but occurs in crystalline solids. In some materials the energy released by an ionization event, probably during subsequent recombination. can induce an atom to move out of its normal place in the solid lattice. Examples are lattice defects created in alkali halides by subthreshold (for displacement effects) ionizing radiation. These effects are distinct from displacement damage produced by dynamically knocking atoms out of their lattice positions, particularly in materials requiring a high degree of structural regularity, which is dealt with in Section VII. For most other technological materials, and, of course, biological tissue, ionization effects have a greater impact on performance.

For solid-state electronics, the local accumulation of charge is an important effect. The trapping of space charge occurs in thin-film oxide dielectrics. This effect has a strong impact on the performance of metal-oxide-semiconductor devices. The unwanted charge in the oxide film can change the function of a logic circuit so that the device becomes unusable. Because the oxide films used are excellent insulators, the space charge may persist for many years (see later in this section). The extensive use of MOS-based microcircuits in modern electronics thus provides a major problem for the engineer addressing the use of a microcircuit in radiation. An integrated circuit is both small and structurally complex; while manufacturing technique is constantly changing. Add to this the complexity of the engineering material and of the long-term effects described above, and it is not surprising that the interpretation and prediction of effects in such components is a field for specialists. Thus the fields of radiation effects engineering and the "radiation hardening" of electronics (see Section IX) has developed as a combination of radiation physics and advanced solid-state engineering.

B. The Deleterious Long-Term Effects of Ionization

1. Semiconductor Devices

In semiconductor devices, there are three important effects of ionization:

1. Photocurrents are produced when electron-hole pairs are swept apart by a field, particularly the built-in

field at a *p-n* junction. Currents of this type were discussed in Section V.
2. The build-up of a net positive charge in oxide films changes electric fields within the device. In oxides used as insulators or passivation in modern integrated circuits, these fields may in turn produce inversion layers in the silicon. Field effect transistors (FETs) are designed to operate in response to electric fields from an electrode. The radiation-induced fields may give rise to severe leakage and the switching to the "off" condition of some devices that are meant to be "on."
3. The disruption of chemical bonds at interfaces between semiconductors and insulators. The resultant "interface states" again interfere with semiconductor action in the surface region.

Long-term effects 2 and 3 are the ones that probably produce the majority of the engineering problems associated with operating electronics in space and in nuclear facilities and will be discussed in detail later.

If some of the carriers generated in an insulator become trapped permanently near the surface of a semiconductor device, its properties are permanently changed. This effect has a particularly strong impact on the performance of metal-oxide-semiconductor (MOS) devices. The effect is shown schematically in Fig. 15. The sensitivity of many metal-oxide- semiconductor (MOS) field effect devices to ionizing radiation is the direct consequence of the mechanism that is seen in Fig. 15, i.e.,

1. Ionizing radiation creates electron–hole pairs in the gate oxide, many of which escape geminate recombination under the influence of the field present in the oxide. Mathematically, a "field-dependent charge yield," $F(E)$, describes this escape process. We find the value of the field, E, from the combination of applied voltages and built-in fields.
2. Given a strong field or near proximity to an interface, an electron is very mobile and will escape rapidly from the oxide, leaving behind a net deficit, a population of "holes," spread throughout the oxide.
3. At room temperature the holes gradually move to an interface. For a positive voltage on the gate electrode, this interface is the one between the silicon and the oxide layer (Si/SiO_2 interface). For a negative gate voltage this is the interface between the oxide and the gate electrode.
4. On arrival at the interface, some of the holes pass across and disappear innocuously, but others are trapped in deep oxide trap levels near the physical interface.
5. In the case of positive bias there is a net negative shift in gate threshold voltage when the electrons are swept out, which gradually increases as the holes approach the silicon interface and decreases as some holes pass across the interface into the semiconductor. Under negative bias the charge motions are reversed but again, except in a few special cases, a net positive charge remains in the oxide.
6. Subsequent to the generation of holes and electrons in the oxide, there can be violent releases of energy similar to that described for gases in section VIA. One of the effects of this release can be changes in chemical bonding, a result which is especially noticeable near the silicon/silicon dioxide interface. Changes in bonding in this somewhat unstable region produces a new interface energy states. Depending on the Fermi level in the adjacent semiconductor, the states may acquire a charge which further changes the effective gate threshold voltage; their effect on conduction and the ease of charging and discharging of these states strongly affects some other characteristics of the MOSFET.
7. An important final step is a back-flow of electrons into the oxide from the silicon. This happens slowly because it is controlled by tunnelling and is limited to a depth in the oxide of 10 nm. That process, however will clearly be of great significance in many modern devices, which have oxides of the same order of thickness.

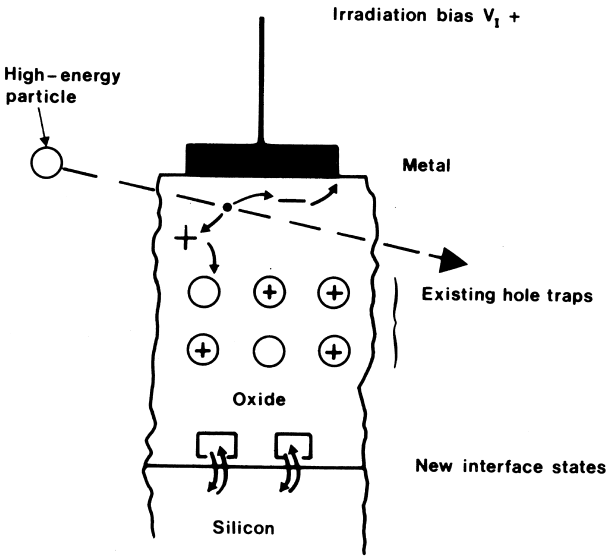


FIGURE 15 Ionization effects in a thin oxide film.

The process by which radiation-generated holes move in the thermally grown amorphous silica layer on a MOS device is particularly important. Figure 16 presents a composite curve of the change in flatband voltage of MOS

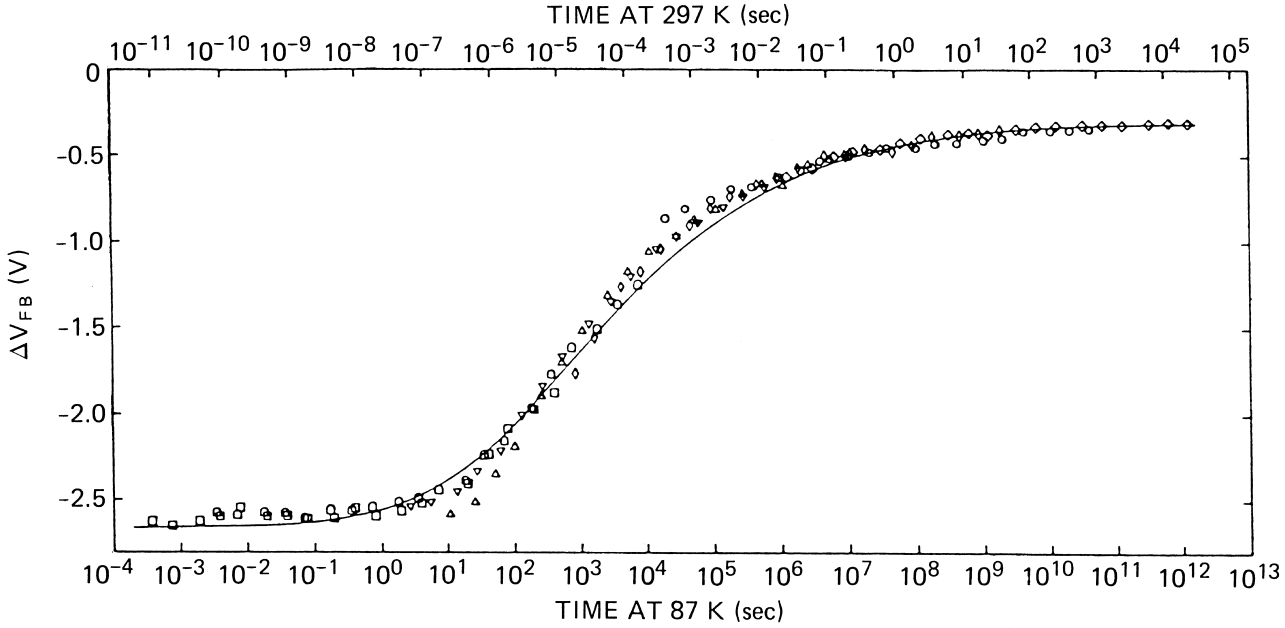


FIGURE 16 Composite universal recovery curve obtained by translating MOS capacitor recovery curves according to the temperature-dependent activation energy. The solid curve is the response calculated from the CTRW model for $\alpha = 0.22$. The time scales appropriate to both 297 and 87 K are indicated. Hughes Aircraft dry SiO_2; oxide thickness is 875 Å; $V_g = 9$ V; dose, 20 krad. ◇, 377 K; ○, 297 K; ⋄, 244 K; △, 194 K; ▽, 166 K; 121 K; □, 87 K. [Reprinted with permission from McLean, F. B., Boesch, H. E., Jr., and McGarrity, J. M. (1976). "Hole transport and recovery characteristics of SiO_2 gate insulators." *IEEE Trans. Nucl. Sci.* **NS-23,** No. 6, 1506. Copyright 1976 IEEE.]

capacitors, which measures the product of charge density in the dielectric times distance from the gate electrode, as a function of time at various times after a rapid pulse of ionizing exposure. The initial value shown, −2.7 V, represents the shift produced by uniformly distributed holes whose density corresponds to an ionization energy loss of 18 eV per electron–hole pair formed. These irradiations were performed at sufficiently high electric bias field for essentially all the pairs to escape geminate recombination. After scaling the individual temperature data by the measured activation energy, the universal curve exhibits a very broad time range for the transition from an initial uniform hole distribution to a final state at which a small fraction of the holes are apparently permanently trapped near the silicon interface: The 10 and 90% points on the transition curve are separated by approximately five orders of magnitude in time. This curve has been explained and fitted to the data by the solid line in Fig. 16 by a continuous-time- random-walk (CTRW) model of hole transport. This model requires polaron effects, in which a significant contribution to the hole trapping energy comes from the lattice distortion produced by the presence of the hole charge.

In summary, there are short-term current and voltage changes which settle down into (a) sheets of trapped charge in the oxide, stable in the long term; the "long term" can vary from seconds to many years, depending on the details of the oxide structure, i.e., the way it was grown and the temperature.; and (b) changes in interface states which are always stable for very long times. Since the effects of trapped charge and interface states can be in opposite directions, and they are likely to change on different time scales, the long-term recovery behavior can be complex.

These detailed descriptions have been arrived at only after extensive research by semiconductor device physicists and radiation physicists. Investigations into charge movement and trapping have been performed for a large range of microelectronic devices, a variety of silicon and oxide preparation methods, and for thicknesses of the silicon dioxide film varying from several micrometres to 10 nm. Research has clarified the role of impurities (such as hydrogen) and especially high-temperature treatments on the efficiency with which the holes are trapped and the rate of production of interface states. Studies of the "bleaching" or annihilation of the holes by a stream of electrons have been useful. Although these investigations have led to an understanding of some of the variables that give rise to undesirable radiation response (e.g., large hole trapping or interface state generation efficiencies), there may still be some variables which are still not fully controlled because not yet fully appreciated. It is clear that most of the deleterious effects occur in the oxide very near the interface

(i.e., within 5 nm of the metallurgical interface with the silicon). This observation suggests that not only the interface states but also the dominant hole traps are the result of strains in the transition region between the silicon crystal lattice and the amorphous oxide.

Given the need to control these two types of defects—hole traps and interface states—the physical nature of the defects which hold the charge has been studied by physical methods such as electron spin resonance and theoretical modelling of chemical bonds. Although the SiO_2 is an amorphous material, the hole traps can be regarded in the same light as "displacement damage" in silicon—missing or substitute atoms in a regular matrix of atoms. The physical investigations indeed suggest that the traps are closely allied to those found in crystalline oxides, including quartz. The favored candidate for the hole trap is the oxygen vacancy, with the hole residing on the deprived silicon next to it.

"Interface states" are energy levels which are in electronic contact with the semiconductor region. That is, they are special in that their electronic occupancy depends on the Fermi level in the semiconductor. The physical investigations in this case suggest that the traps are closely allied to those found in solid–vacuum interfaces. At this solid-solid interface, a foreign atom, hydrogen, is occupying some of the chemical bonds to silicon (hydrogen treatment is part of the industrial "passivation" process to suppress interface states). Chemical energy transmitted from within the oxide when it is excited by radiation, reaches the hydrogen atoms and breaks the chemical link with the silicon. The "dangling bond" produced is the active interface state. Depending on its exact nature, it can exchange charge with the silicon rapidly or slowly, giving "fast" or "slow" states. For some positions of the Fermi level especially when the device is conducting, such interface states will have the same effect as trapped oxide charge; that is, they require a change in gate voltage to compensate for their charge and return the semiconductor to its preirradiation electrical condition.

With the rapid advance in physical knowledge and silicon technology, there is now a generation of "radiation tolerant" silicon integrated circuits based on the MOSFET transistor. A major influence has been the natural evolution to thinner oxides with smaller feature sizes, which produce smaller shifts in threshold voltage. Figures 17 and 18 show the dual effects of thinner oxides and improved radiation-tolerance which have been achieved by means of oxide growth technique. These curves were devised as working diagrams for the engineer. They group MOSFET technology into four categories of tolerance to radiation and

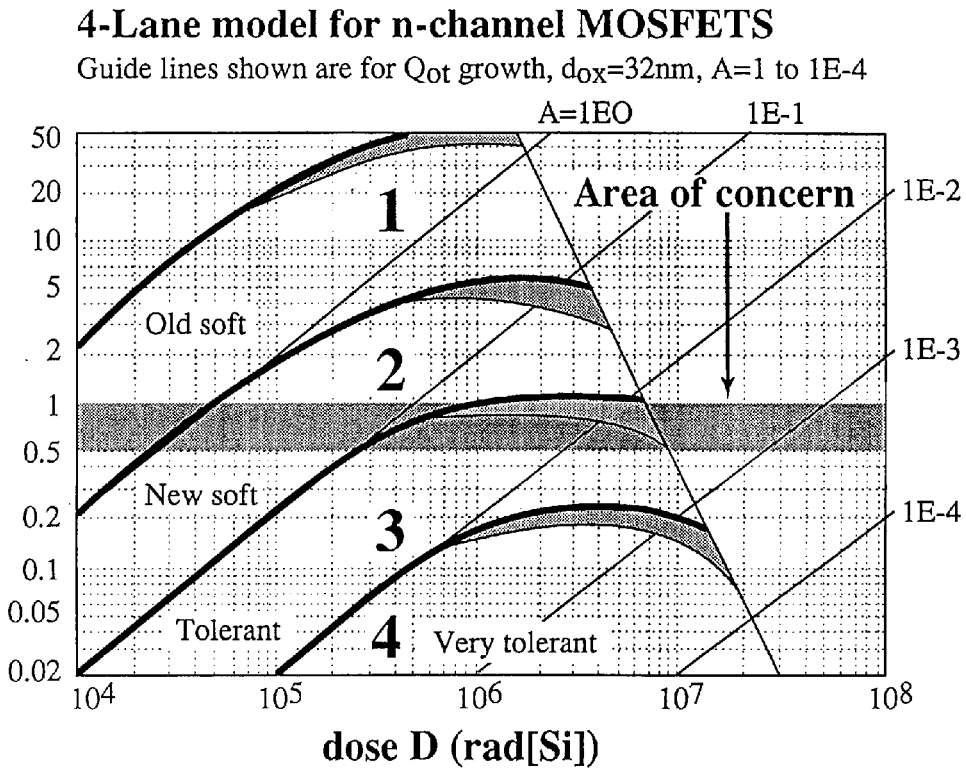


FIGURE 17 Responses of *n*-channel MOS devices to radiation. The four "lanes" classify the devices into categories of total-dose radiation tolerance when exposed under operating gate voltage (positive). [Reprinted with permission from Holmes-Siedle, A., and Adams, L. (1994). *IEEE Trans. Nucl. Sci.* **NS-41,** No. 6, 2613-8. Copyright IEEE 1994].

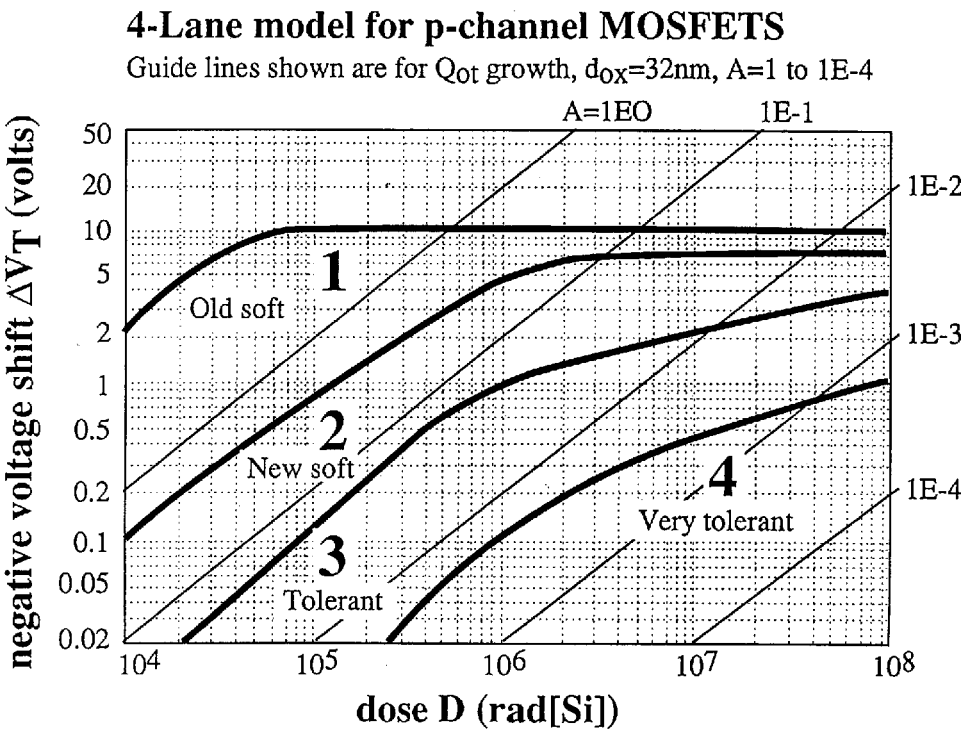


FIGURE 18 Responses of *p*-channel MOS devices to radiation. The plots are as for Fig. 17 but the shapes of the curves are altered by the different polarity of certain charges in the oxide film and the exposure in this case under negative operating voltage. [Reprinted with permission from Holmes-Siedle, A., and Adams, L. (1994). *IEEE Trans. Nucl. Sci.* **NS-41,** No. 6, 2613–2618. Copyright IEEE 1994.]

model or predict the growth of trapped charge in the oxide layers with increasing value of radiation dose. Positive trapped charge is converted here into "negative voltage shift." Experimental plots for most commercial CMOS logic devices devices stay within the four "lanes" in the diagrams bounded by the curved lines. Lane 1 represents early microelectronic technology, here labeled "old soft," in which oxide thickness values of the order of 100 nm were used. Lane 2, labeled "new soft" represents the technology of the 1990s in cases where no attempt was made to "harden" the oxide, i.e., reduce the concentration of charge traps.

The uppermost straight line is the theoretical maximum shift which would be attained for an oxide of thickness 32 nm if every hole generated were captured in the oxide (capture probability, A, of unity). Lanes 3 and 4 represent degrees of hardening of the oxide growth process. As we noted earlier, this is achieved by reducing the number of trapping centres present, which reduces the hole capture probability, A, to values much less than unity. The bold curves fall below the theoretical straight lines because, as the population of charges increases, fields within the oxide reduce the yield of trapped charge per unit dose. Other distortions affect n- and p- channels differently, as explained later.

Figure 17 is the model for *n*-channel MOS devices, often the weakest link in microelectric circuits. The curves are distorted by the effect of negative interface states, discussed earlier. Figure 18 is the model for p-channel MOS devices. In microelectric circuits, these cause less trouble than the *n*-channel type. The curves are distorted by the effect of positive interface states and also fall below the equivalent curves in Figure 17 because the polarities of operating gate voltages, normally used in logic circuitry, are different in the two cases (see the effects of field discussed above).

Figures 17 and 18 describe the case for two different cases of oxide thickness. In fact, as microelectronic technology progressed, oxide thickness values have fallen year by year. The "scaling law" in microelectronics dictates that, as the lateral dimensions of circuit elements are reduced, so the vertical dimensions such as the oxide thickness must also reduce. The continuous evolution of silicon microelectronics requires greater and greater packing density of circuits. This has led to smaller and smaller dimensions for each part of the circuit. Electronics of the "submicron" generation will have some oxide thickness values less than 20 nm. This scaling-down of oxide thickness values will continue to reduce the impact of trapped charge on modern generations of microelectronics.

Some manufacturers of microelectronics advertise an "RH" (radiation-hard) version of their commercial products. Such products would reside in the fourth lane shown in Figs. 17 and 18.

Bipolar transistors are also subject to ionization effects of the same type as the ones described earlier for the MOSFET. For such devices these effects in surface regions are properly called "surface ionization effects," to contrast them with the "bulk displacement effects" described in Section VII. Alongside the large development efforts in silicon microcircuits, described earlier, some semiconductor devices are designed and built with their response to radiation in mind. These include large power-controlling chips made of silicon and also devices made from III to V semiconductors. A new type of detector of radiation, the Radiation-Sensitive Field-Effects Transistor (RADFET dosimeter) operates by maximizing the trapped hole effect and treating the trapped charge as stored dosimetric information. In Fig. 18, the growth of threshold voltage in this device would be a curve near to, or off the top of the figure.

2. Optical Materials and Oxides

Glasses, polymers, and many colorless inorganic salts are colored by exposure to radiation. During exposure, optical materials may emit light as a result of photon emission during carrier-trapping and recombination processes. If the ionization-generated carriers become trapped at defect sites, this may introduce optical absorption at inconvenient wavelengths. Thus, a material that is normally transparent in the visible spectrum may become highly absorbing—dark brown or black.. Examples of this effect are the F centers which occur in all alkali halides. Most glasses contain color centers, produced fortuitously during manufacture. The alkali silicates, alkali borates, and phosphates all contain nonbridging oxygen bonds that can give up an unshared electron. The resulting "trapped hole" absorbs light strongly both in the near UV and blue wavelength ranges. Thus, irradiated multicomponent glasses are often deep red-brown after irradiation. It is difficult to avoid this effect, but partial suppression has been achieved by alterations in glass formulation and the addition of dopants: for example, cerium oxide in silicate glasses. Because of its low impurity content, synthetic fused silica (e.g., Suprasil 1) is relatively immune to coloring under irradiation. However, in the manufacture of optical fibers, pure fused silica is doped to modify the refractive index and melting point. The sensitivity of the fiber to color-center formation under irradiation is thereby increased. Table V gives radiation-induced optical losses for various glasses and optical fibers. Figure 15 shows the

TABLE V Optical Loss Induced by Radiation in Selected Glasses and Optical Fibers[a]

Source	Code	Type	Core material	Form of glass	Response[b] [dB(km rad(Si))$^{-1}$]		
					$\lambda = 0.8\,\mu$m	0.9 μm	1.05 μm
Corning	(CGW)	5010	Pb flint	Fiber	5.4	2.5	0.50
Pilkington	(PBL)	HYTRAN	Pb flint	Fiber	4.5	1.9	0.50
Galileo	(G)	0001AA	Zn crown	Fiber	1.5	0.49	0.25
Schott	(S)	F2	Pb flint	Bulk	1.3	0.69	0.21
		Dividing line for values above and below 1 dB/km at $\lambda = 0.8\,\mu$m					
NRL		GL2382	BaLa crown	Bulk	0.65	0.35	0.16
Galileo	(G)	0001AB	Zn (0.3% Ce) crown	Fiber	0.27	0.0062	0.0026
NRL[c]	—	GL2364	BaLa (1% Ce) crown	Bulk	0.21	<0.18	—
Owens Corning	(OCF)	X-4147A	Zn crown	Bulk	0.040	0.020	<0.016
Schott[d]	(S)	F2G12	Pb (1.2% Ce) flint	Bulk	<0.1	—	—
Schott	(S)	R1	Pb(∼1% Ce)	Fiber	0.0031	0.0015	0.0010
Corning	—	—	SiO_2(Ti)	Fiber	8×10^{-1}	—	—
Corning	—	—	SiO_2(Ge)	Fiber	1.4×10^{-2}	—	—
NRL	—	—	Soda-lime Silicate glass	Bulk	1×10^{-2}	—	—
NRL	—	—	Suprasil I	Bulk	$<1 \times 10^{-5}$	—	—

[a] Reprinted with permission from Evans, B. D., and Sigel, G. E. Jr. (1975). *IEEE Trans. Nucl. Sci.* **NS-22**(6), 2462. Copyright © 1975 IEEE.]
[b] Except where mentioned, readings made 1 hr after γ-irradiation.
[c] 30 min after γ-irradiation.
[d] 9 min after γ-irradiation.

spectrum of the ionization-induced coloration of various glasses.

Another material in which trapped-charge effects are frequently important is crystalline quartz used for precision frequency standards. In this case the effect of the trapped charge is to change the elastic constants of the lattice sufficiently to alter the resonant frequency. This effect is relatively subtle, but it is important because quartz crystal resonators are used for frequency standards with stability in the range of 1 part in 10^7. Research has shown that the magnitude of the frequency shift depends in a nonlinear manner on accumulated dose, presumably because different defect sites become saturated at various dose levels. There are particular crystal-growing processes that will minimize the magnitude of the effect, presumably by minimizing the kind of defects at which charges will be trapped. Synthetic quartz purified by heating in the presence of a high electric field (sweeping) is a favored method.

3. Organic Materials and Polymers

Plastics are used in a variety of ways in devices and electronic systems. The more demanding applications include thin-film insulators in capacitors, stand-off insulators in cables), coatings and encapsulations, membranes, adhesives, optical lenses, pyroelectric sensors, plastic scintillators and electrets. Organic compounds which are not polymeric have only limited uses in devices, such as a few light-sensitive crystalline powders. However, the study of the *G*-factors for the radiolysis of such simple compounds (see Section VI.A) acts as a guide to the radiolysis of engineering plastics. In any organic material, ionization leads to the excitation and breaking of covalent bonds. The bonds may subsequently recombine or rearrange, leading to a large variety of possible products. At an absorbed dose value of 10^7 rad (10^5 Gy), only a few of the more sensitive plastics are altered mechanically or electrically. Even so, many plastics develop a deep coloration at this dose level.

VII. DISPLACEMENT EFFECTS

Displacement effects are the result of displacing atoms from their normal positions in crystal lattices, often known as "Bulk Damage" or "Nonionizing Energy Loss" often abbreviated to NIEL. Although displacement effects have been studied in a wide variety of materials, only displacement effects on semiconductors are considered important in electronics, because the other materials are inherently tolerant to displacement effects at exposures well beyond the maximum tolerance levels of most semiconductor devices. Silicon has been studied most extensively, and it will be used as the example for this discussion. There are similar data for germanium, as well as many compound semiconductors. The particles likely to cause displacement problems are ions, electrons, protons, and neutrons, although any particles possessing high momentum, including unstable ones (e.g., muons and pions), can produce atomic displacements.

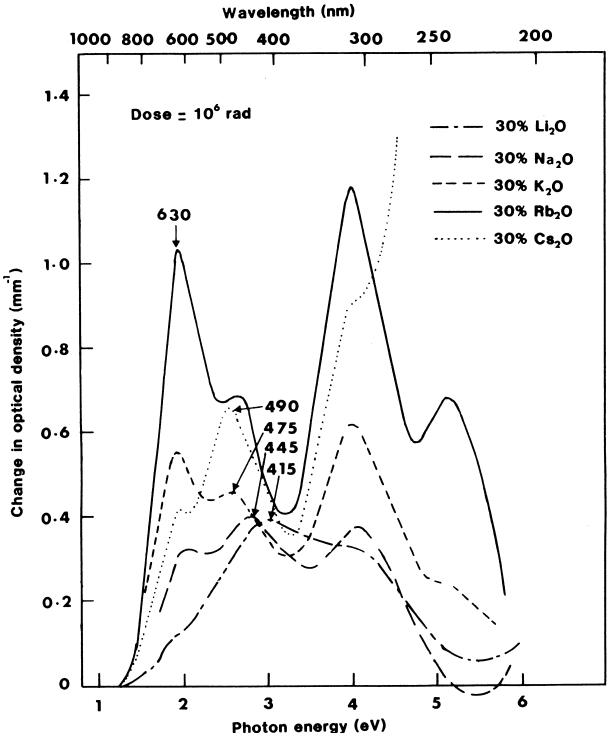


FIGURE 19 Radiation-induced absorption of various alkali silicate glasses. (Courtesy of Phillips Research Laboratories.)

A. Simple Defects

The simplest defects in an otherwise regular crystal lattice are vacancies and interstitials. A vacancy is an unoccupied lattice site; an interstitial is an extra atom inserted between the atoms occupying lattice sites. It might be expected that exposure of an elemental semiconductor (e.g., silicon) to electrons whose energy is slightly above the threshold for displacement effects (i.e., that can displace an atom from its lattice site) would create vacancy–interstitial pairs whose characteristics and effects could then be observed. Unfortunately, thermally activated rearrangements preclude such observations at room temperature and even at liquid-nitrogen temperature. There is even some evidence that the silicon interstitial formed by irradiation at liquid-helium temperature is mobile.

The simplest defect in silicon that is stable at room temperature is the B1 center (also called the A center). It

consists of a substitutional oxygen atom formed by a mobile radiation-induced vacancy migrating to an interstitial oxygen. The interstitial oxygen was electrically inactive, the substitutional oxygen is an electron trap that serves as an effective recombination center. It can also be readily detected by electron spin resonance (ESR). Considering its origin and ease of detection, the B1 center concentration is an excellent measure of the density of mobile isolated vacancies formed in a particular radiation exposure.

The next simplest defect in silicon that is also stable at room temperature is a divacancy: two vacancies located near one another in the silicon lattice. Divacancies are also detectable by ESR as G6 and G7 centers and have been used to measure the onset of more complex defect formation. Figure 20 demonstrates the use of these defects to measure the dependence of radiation effects on electron energy in silicon. The threshold for forming the vacancy by displacing one atom from its lattice position is reflected in the (V + Oxy) curve. The onset of more complex defects is shown by the V^2_{TOTAL} curves. These results are in reasonable agreement with a threshold for displacing a silicon atom at 12.9 eV. It is clear that, as the incident electron energy is raised, ever more complicated defects are formed, as are more of the simpler defects.

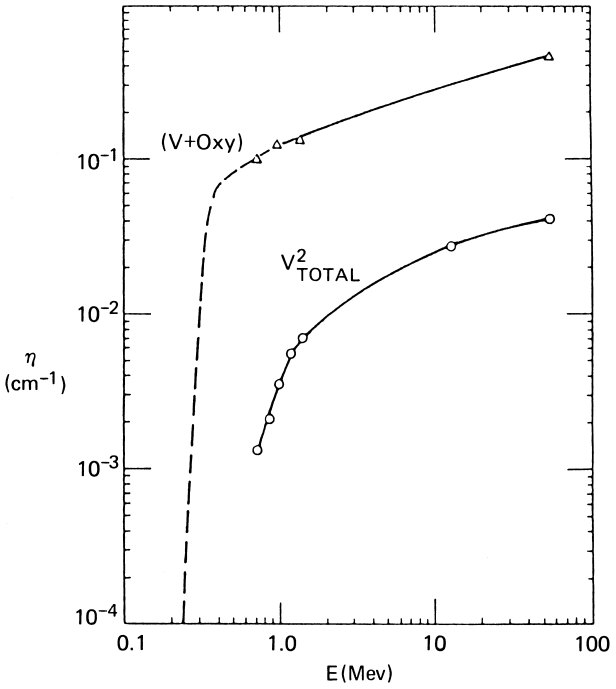


FIGURE 20 Energy dependence of the room-temperature production rate of the divacancy V^2 and of the vacancy–oxygen pair (V + Oxy). [Reprinted with permission from Corbett, J. W., and Watkins, G. D. (1965). "Production of divacancies and vacancies by electron irradiation of silicon," *Phys. Rev.* **138,** A555.]

B. Defect Clusters

The situation for fast neutron irradiation of silicon and other materials is considerably different. A typical silicon recoil from a 1-MeV neutron collision would receive ∼50 keV, producing in turn ∼1000 other displaced atoms within a path length of ∼100 nm, as illustrated in Fig. 21. Such a cluster of displaced atoms cannot be adequately described as a superposition of simple defects: a high concentration of trapping centers will produce a space-charge sphere around it, reflecting the charge removed from the conduction or valence band near the defects. Only a small fraction of the defects in the cluster need to be charged to establish a space-charge barrier around the cluster. This process is illustrated in Fig. 22 which depicts a region of *n*-type semiconductor in which a localized cluster of deep electron traps has just been introduced. Some of the conduction electrons are inclined to be trapped at these sites, and other conduction electrons will then diffuse into this region under the influence of the resulting density gradient. However, this results in a negatively charged region in the defect cluster surrounded by positive charge from the nearby material from which the excess electrons have diffused. This process stops when the space charge shifts the potential of the cluster region sufficiently to place the Fermi level near the trap level. For a large, dense defect cluster only a small fraction of the traps need be occupied to achieve this condition. If such a cluster is inserted into high-resistivity material, the region of positive space charge around the cluster can be much larger than the actual size of the cluster.

The effects of such a cluster and its associated space-charge region on the properties of the semiconductor can

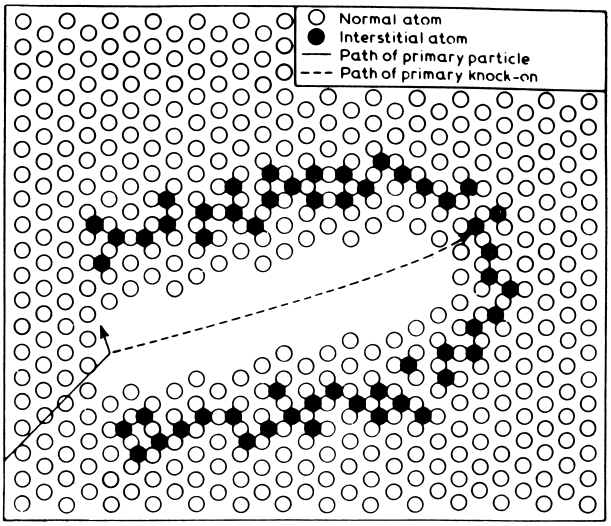


FIGURE 21 Displaced atoms in a crystal lattice, illustrating a cluster produced by ∼1 keV recoil atom.

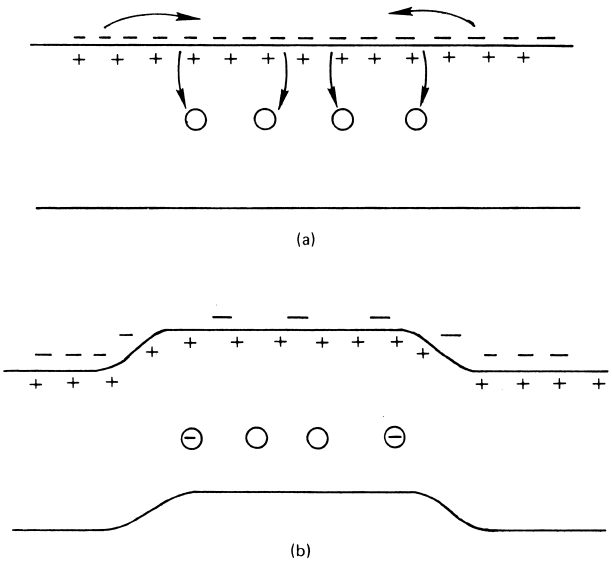


FIGURE 22 Space-charge layer forming around a defect cluster. (a) Initial condition; (b) after charge arrangement.

be very different from the effect of a uniform distribution of point defects. For example, in silicon the effect of the clustered defects on the recombination rate of excess carriers is much stronger than the effect on net carrier density, as compared with the equivalent average density of displaced atoms. This is illustrated in Fig. 23, which presents a calculation of the effects of unannealed clusters of various sizes on the net carrier concentration and on the inverse carrier lifetime in 10 Ω-cm *p*-type silicon. The equivalent number of displaced atoms distributed uniformly would produce an effect approximately proportional to the recoil atom energy, as shown, with a small roll-off at high energy to account for the ionization loss of the recoil atom. Instead, the clusters show an effect on carrier concentration that increases only slightly with recoil energy and an effect on carrier lifetime that is much stronger than linear. Other manifestations of clusters and space-charge layers include the very large factor by which carrier lifetime effects anneal after a short-pulse (<1 msec) displacement exposure, as compared with almost no short-term anneal effect on the conductivity.

C. Short-Term Annealing

Annealing is the rearrangement of atoms or charges in a material with time after irradiation. High temperature may be needed to produce these effects but some annealing effects occurs at room temperature over long periods. This section deals with rearrangements over even shorter periods. We know that stable displacement damage at room temperature is only a small fraction of the calculated initially produced damage; therefore there must be a transition period during which occur the atomic rearrangements responsible for the annealing. Initial measurements on bipolar transistors demonstrated that the effective damage (e.g., change in reciprocal h_{fe}) produced by a short neutron pulse was up to a factor of 6 greater at 1 msec after exposure than the long-term stable damage. The annealing time scale decreased with increasing minority-carrier injection level. Later experiments on solar cells with few-microsecond neutron pulses measured minority-carrier lifetime annealing factors as high as 50 in *p*-type silicon. Actually, the damage measured immediately after exposure in these experiments agreed well with the calculated effect of space-charge regions around damage clusters (see Fig. 23). The annealing process was then modeled as the shrinkage of the space-charge volume during defect rearrangement inside the cluster. The small degree of short-term annealing in majority-carrier density is consistent with this hypothesis. The dependence of annealing rate on minority-carrier injection is also consistent with low-temperature long-term annealing studies, that demonstrate that the vacancy is more mobile in its negative charge state, which is likely to occur in *p*-type silicon only during minority-carrier injection.

D. Scaling Displacement Effects for Different Exposures

Although it is tempting to scale displacement effects produced by different exposures according to the energy imparted in nonionizing collisions by atomic recoils, the strong effects of displacement clusters, particularly in semiconductors, must produce some caution. In addition, the dominant role of pre-existing defects and impurities on the quantity and nature of eventually stable defects, particularly for low energy exposures that produce mostly isolated displaced atoms, increases the complexity of experiments comparing the effects of different irradiations; apparently identical materials may respond very differently. As long as some margin for uncertainty is allowed, it is possible to compare the effects of different exposures if the following requirements are met:

1. One separates the effects of relatively isolated defects (e.g., those produced by <3-MeV electrons) from those produced by energetic recoils (e.g., reactor neutrons).
2. The scaling is done by dividing the recoil energy spectrum into its low-energy (<100 eV) and high-energy parts and associating appropriate damage coefficients with each part).

Figure 24 illustrates how this can be done by presenting the cumulative non-ionizing recoil energy as a function

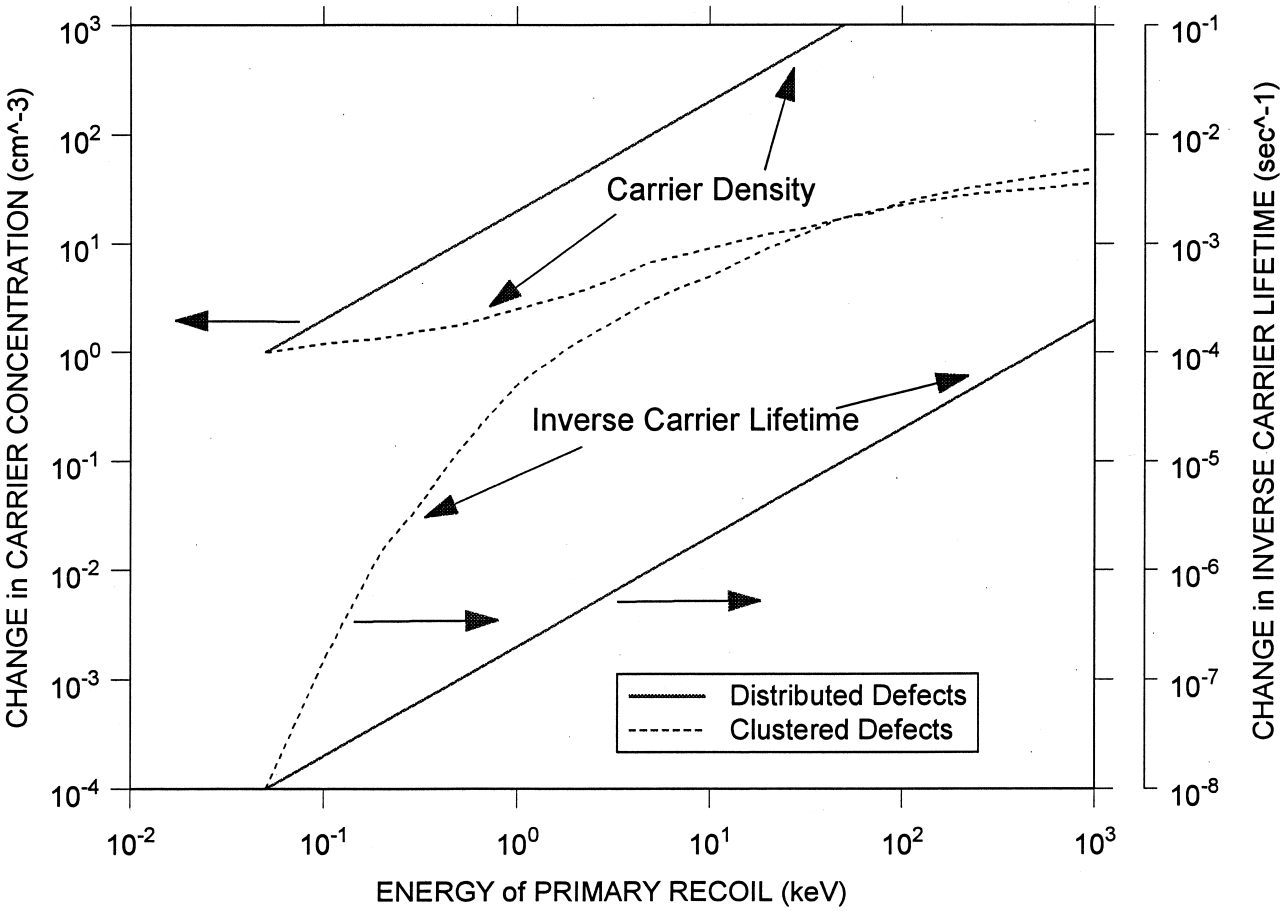


FIGURE 23 Calculated effect of one recoil per cubic centimeter on recombination rate and carrier concentration. [Reprinted with permission from van Lint, V. A. J., Leadon, R. E. and Colwell, J. F. (1972). "Energy dependence of displacement effects in semiconductors." *IEEE Trans. Nucl. Sci.* **NS-19,** No. 6, 181. Copyright IEEE 1972.]

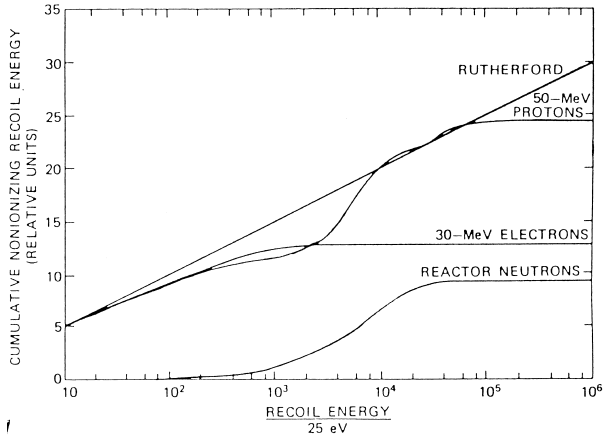


FIGURE 24 Relative amount of recoil energy imparted to recoils with energy $<E_R$ versus $E_R/25$ eV for silicon (renormalized for different particles). [Reprinted with permission from van Lint, V. A. J., Flanagan, T. M., Leadon, R. E., Naber, J. A., and Rogers, V. C. (1980). "Mechanisms of Radiation Effects in Electronic Materials," Vol. 1, Wiley, New York. Copyright 1980 John Wiley and Sons.]

of the energy of the individual recoil atoms for different irradiating particles. For example, it can be seen that the recoil energy spectrum produced by 50-MeV protons is not much different from that produced by a combination of 30-MeV electrons. which produce recoil energies mainly up to 1 keV, and reactor neutrons, which cover the higher-energy recoils. Thus, an appropriate linear combination of electron and neutron data would provide an accurate estimate of the effects of protons on most materials.

In spite of this caution, recent work has shown remarkably good agreement between calculations of nonionizing recoil energy (i.e., displacement curves in Figs. 6 and 7) and experimental data on room-temperature stable damage for various energy neutron and proton irradiations. The calculations are absolute, using measured interaction cross sections. The measurements are normalized to the calculations once only to account for annealing. The relative effectiveness of different neutron spectra is usually expressed by reducing all to a 1-MeV equivalent fluence using the calculated energy dependence. The strong

dependencies on cluster size implied by Fig. 23 do not appear to affect these correlations. It appears that the annealing process distributes the stable damage sufficiently widely to minimize the cluster effects. Some experiments on extra pure silicon used in detectors show the clusters in the process of steady change over long periods (room temperature annealing) and suggest the presence of multivacancies such as V6. New theoretical methods and large computer power have been used to calculate the energy levels of these defects in the silicon bandgap.

E. Effects of Displacements

Displacement radiation effects are manifested most strongly in semiconductors by decrease of the excess-carrier recombination lifetime, τ; second, by changes in the net majority-carrier concentration, n or p; and third, by decreases in the carrier mobilities, μ_n and μ_p. For exposure fluences that do not move the Fermi level too much, the dependence on fluence, Φ, can be represented, in n-type material by

$$1/\tau = 1/\tau_o + K_\tau \Phi$$
$$n = n_o - K_n \Phi$$
$$1/\mu = 1/\mu_o + K_\mu \Phi,$$

where "o" subscripts refer to the property values before exposure.

Semiconductor devices whose properties depend on excess-carrier recombination rates are most sensitive to displacement effects. Solar cells are particularly sensitive. The short-circuit current produced in silicon solar cells by the longer-wavelength photons is proportional to the minority-carrier diffusion length, which varies as the square root of the carrier lifetime. Thus, for this wavelength component,

$$\frac{1}{I_{sc}^2} = \frac{1}{I_{sco}^2} + K_{sc}\Phi.$$

Similarly, for bipolar transistors the common-emitter current gain, h_{fe}, decreases with exposure as

$$\frac{1}{h_{fe}} = \frac{1}{h_{feo}} + K_{hfe}\Phi.$$

In Section VII.D, we mentioned experiments on clustered defects in extra pure silicon. This material is used in "vertex detectors," particle trackers used in high-energy physics. Unusually low levels of dopant (phosphorus, boron) are dictated by the mode of detector action. The result is an unusual degradation problem, caused mainly by "carrier removal" (*see earlier, this section*). Particle trackers in accelerators are exposed to pions and other unstable particles. Tests confirmed our earlier statement that, in crystalline silicon, pions produce roughly the same defects as other charged particles with high momentum such as protons. The power of new theoretical methods and fast computing have been combined to identify the defects causing the problems in detectors while, in accelerator projects, new silicon technology will be used to minimize the electrical problems produced by these defects.

VIII. SINGLE-EVENT PHENOMENA

So far our discussion of transient ionization effects (Section V) has addressed uniform ionization and the effects on conductivity and currents across reverse-biased junctions. Some modern semiconductor devices have very small junction areas and correspondingly small amounts of charge to control the state of critical circuit nodes (e.g., memory nodes). If this critical charge is small enough, a single heavily ionizing particle passing through the junction can induce sufficient charge into the node to change its state, causing a disruption of the data stored at the node. This effect is called a single-event upset (SEU). Whether or not a particular device structure exhibits SEU disturbances depends on a comparison between the critical charge (i.e., charge required to change the information state of a circuit node) and the product of the energy deposition per unit length by the incident particle (i.e., its ionization density) and the length of the path in the device from which charge is collected by the node. The path length is determined mainly by the diffusion profile of the device, but it can be increased somewhat by the electric field created in a densely ionized spike by carrier motion (i.e., the funnel effect). There exist device structures with such a low critical charge that α particles from naturally occurring radioactive materials can produce SEU's. In other devices, heavily ionizing heavy ions in cosmic rays (e.g., iron nuclei) may be required to exceed the SEU threshold. Other processes for producing such intense local ionization include energetic nuclear reactions produced by fast neutrons, protons or mesons, where the local energy is deposited by the fragments of the target nucleus.

The SEU phenomenon can be illustrated using a typical CMOS memory cell, as shown in Fig. 25 and the associated CMOS cross section in Fig. 26. The memory cell is a flip-flop composed of two CMOS inverters. It has two stable states: left or right side output node at high level, the other low. Normally, while not being read or written to, it is disconnected from the bit lines and retains the state into which it was last written. Potential paths of ionizing particles are illustrated in Fig. 26. In normal operation one of the two transistors, n or p channel, will be turned off and have a large voltage between its drain and its local

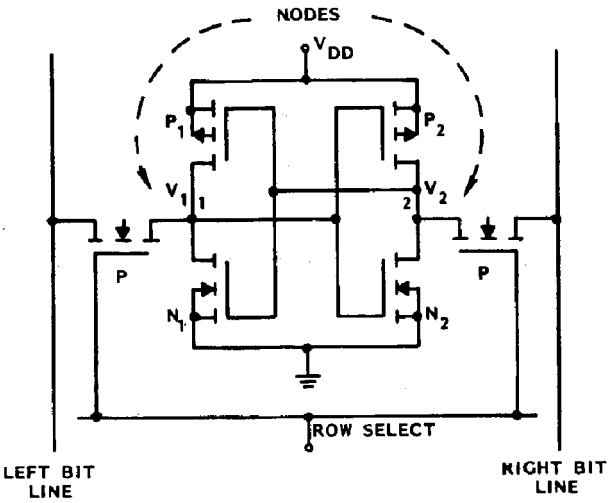


FIGURE 25 Typical CMOS memory circuit. [Reprinted with permission from Sivo, L. L., Peden, J. C., Brettschneider, M., Price, W., and Pentecost, P. (1979). "Cosmic ray-induced soft erros in static MOS memory cells." *IEEE Trans. Nucl. Sci.* **NS-26,** No.6, 5042. Copyright 1979 IEEE.]

substrate. An ionizing particle will cause charge to flow from its substrate to the drain (inverter output line). As a result a voltage pulse appears at the inverter output, which, if sufficiently large, can initiate regenerative action by toggling the other inverter into a change of state. Below the toggling threshold the coupled inverters tend to restore the memory state fairly quickly, so that the rapid charge flow from the depletion layer determines the critical energy deposition.

Estimating the rate of SEU events requires the following information:

1. The critical charge: how much charge must be injected into a node to trigger regenerative action.
2. The device geometry: the area of the sensitive node and the available path length for an ionizing particle to deposit its energy.
3. The energetic-particle exposure: the flux density of particles with sufficient ionizing power to deposit the critical charge within the available path length, either directly or via nuclear interactions.
4. Shielding: the effect of intervening material to attenuate or change the spectrum of the incident particles.

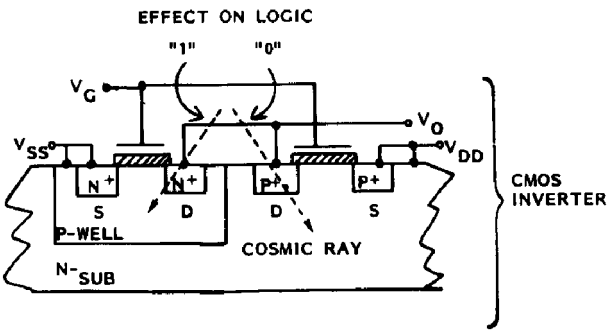


FIGURE 26 Path of cosmic-ray particles through CMOS inverter. [Reprinted with permission from Sivo, L. L., Peden, J. C., Brettschneider, M., Price, W., and Pentecost, P. (1979). "Cosmic ray-induced soft erros in static MOS memory cells." *IEEE Trans. Nucl. Sci.* **NS-26,** No.6, 5042. Copyright 1979 IEEE.]

Larger devices typically are inherently immune to SEU's: their critical charge cannot be generated by the most heavily ionizing events. For most space-based applications energetic iron nuclei in cosmic rays are the most effective at producing SEU's. While the probabilities for upsets due to nuclear interactions of neutrons and protons of modest energy (>10 MeV) are much less, they are still significant for complex electronics exposed in the atmosphere. Some devices must be protected by purified shielding material, because even alpha particles from natural radioactivity can produce upsets.

Experiments exposing specific devices to energetic ions are usually used to measure the effective device cross section as a function of linear energy transfer (LET). An example for a 256-Kbit SRAM is illustrated in Fig. 27. The expected exposure spectrum is then translated into the flux density as a function of LET, as illustrated in Fig. 28. The expected upset rate can then be estimated by convolving the two curves.

Dynamic NMOS memories are particularly susceptible to SEU's, since their state retention depends on retaining the charge on very small capacitors between refresh

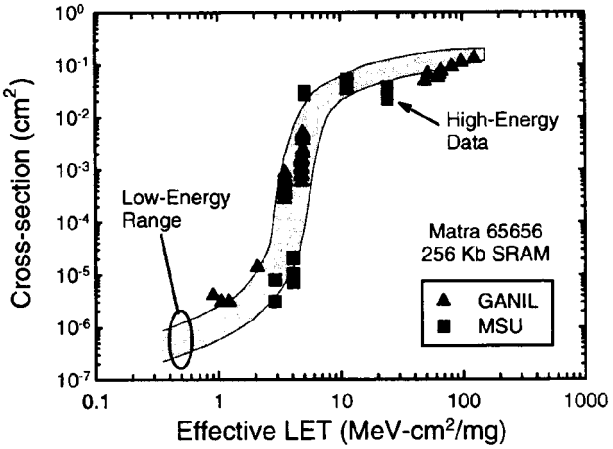


FIGURE 27 SEU cross-section measurements for Matra 256-Kbit SRAM. [Reprinted with permission from Dodd, P. E., *et al.* (1998). "Impact of ion energy on single-event upset." *IEEE Trans. Nucl. Sci.* **NS-45,** No. 6, 2483 Copyright 1998 IEEE.]

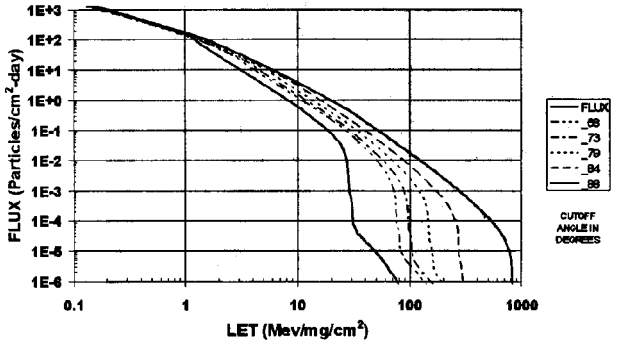


FIGURE 28 Geosynchronous effective flux LET spectrum. [Reprinted with permission from Peterson, E. L. (1995). "SEE rate calculations using the effective flux approach and a generalized figure of merit approximation." *IEEE Trans. Nucl. Sci.,* **NS-42,** No. 6, 1995. Copyright 1995 IEEE.]

operations. A heavily ionizing particle passing through the depletion layer between the capacitor and substrate can discharge a capacitor sufficiently that it appears to be uncharged at the next refresh operation, after which the memory error will be preserved. Since all the charge collected between refresh cycles is effective in discharging the capacitor, both depletion and diffusion-region contributions of the ionization energy deposition contribute to the SEU effectiveness.

Latchup is another, more dramatic, manifestation of single-event phenomena in some devices: the ionizing event triggers the device into a high-current paralyzed state, which may cause permanent thermal damage, or require at least power cycling for recovery. The most common latchup paths are found in CMOS inverters illustrated in Fig. 26. Consider the structure represented, from right to left, by the p^+ source of the p-channel transistor, the n-substrate, the p-well and the n^+ source of the n-channel transistor. It appears to be a $pnpn$ structure similar to a silicon controlled rectifier (SCR). Normally, it does not conduct, because the intermediate n and p layers are biased at the most positive and negative voltages, respectively. Consider a typical state with high output: the n-channel transistor on the left is turned off with bias between the n^+ drain and grounded p-well; the p-channel transistor on the right is turned on, with substrate, source and drain all near V_{DD}. The ionizing event creates minority carriers (electrons) in the p-well, which flow to both the n-transistor drain and into the n- substrate. If there is sufficient impedance between the n-substrate and its bias connection, this current can cause the junction between the n-substrate and p^+ source to become forward biased, i.e., turning on the SCR. This condition will be sustained if the product of bipolar- transistor current gains of the effective pnp and npn transistors is sufficiently greater than unity to supply enough SCR-gate current to maintain the forward bias across the np^+ and n^+p junctions.

Single-event latchup-type damage can occur even in large-scale devices, such as power MOSFET's. Again, a parasitic path is triggered into a conducting state by the high instantaneous energy deposition, allowing excess current to persist and damage the device. Another damage manifestation in such devices is Single-Event Gate Rupture, which produces an electrical punch-through of the gate oxide by a transient electrical overstress. The overstress is produced by the ionizing channel, but without requiring any parasitic regenerative action.

IX. RADIATION EFFECTS IN SYSTEMS AND TECHNOLOGY

A. General

In the foregoing sections, it has been noted repeatedly that practical reasons exist for the study of radiation physics. Advanced electronic and optical systems sometimes have to operate in radiation environments, so that high-technology devices have to be made to survive the radiation effects we describe. Therefore, it might be said that, at the leading edge of technology, the most sophisticated materials are being exposed to the most complicated effects, as they are described above. Radiation physics provides a framework for the science, such as providing units (see Table VI) and makes its contribution to modern technology both in the "radiation hardening" described in this section and the next, which bring together some of the important technological aspects of radiation physics.

B. Radiation Effects in Solid-State Electronics

1. The Problem

In solid-state devices, electronic functions are carried out in microscopic regions of highly structured materials. It is not surprising that, when particles or photons deposit large amounts of energy in such a structure, effects occur that temporarily or permanently interfere with the operation of that device. For example, the npn transistor depends for its amplifying actions on the passage of current through a perfectly ordered crystal lattice. This includes the transport of current across a thin layer known as the base region. If neutrons or high-energy electrons disrupt the crystal lattice, this base transport is disrupted, and the transistor loses its power to amplify electrical signals. A similar effect occurs in single-crystal silicon solar cells. In fact, it was the practical problem of designing solar panels to survive in space that gave rise to one of the first radiation-effects engineering projects.

TABLE VI Radiation Units

1 rad = 100 erg g^{-1}
$= 6.25 \times 10^{13}$ eV $g^{-1} = 10^{-2}$ Gy
1 Mrad $= 6.25 \times 10^{19}$ eV $g^{-1} = 10$ kGy
1 Gy = 1 J $kg^{-1} = 100$ rad
1 kGy = 10^5 rad = 100 krad
1 MGy = 10^8 rad = 100 Mrad
10^{20} eV $g^{-1} = 1.6$ Mrad = 16 kGy

1 roentgen (R) = 86.9 erg g^{-1}(air) = 2.58×10^{-4} C kg^{-1}(air)
1 R of 1–MeV photons = -1.95×10^9 photons cm^{-2}
This fluence deposits
0.869 rad (cGy) in air
0.965 rad (cGy) in water
0.865 rad (cGy) in silicon
0.995 rad (cGy) in polyethylene
0.804 rad (cGy) in LiF
0.862 rad (cGy) in Pyrex glass (80% SiO_2)

1 curie (Ci) of radioactive material produces 3.700×10^{10} disintegrations per second. (activity 37 GBq)
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 ^{6O}Co source gives 1.29 R hr^{-1} at 1 m
Photon flux at 1 m from a 1-Ci point source = 1.059×10^9 cm^{-2} hr^{-1} (assuming one γ-ray photon per disintegration)

Systeme Internationale (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) = 1disintegration /sec = 2.703×10^{-11} Ci
- the old units have been abandoned in some fields but **not in others**
- 1 rad may be called **"1cGy"**

Basic units of biological dose from radiation are the **rem and the sievert (Sv)**. 1 Sv is the absorbed dose to the body of 1 Gy weighted by a quality factor Q.F. that is dependent on the type of radiation involved. This is because the energy absorption from heavily ionizing particles has greater effectiveness for biological damage. **1 Sv = 100 rem.**

These effects were discussed in Section VII, Displacement Effects.

Ionization severely affects electronic circuits that use metal-oxide-semiconductor (MOS) technology. This effect is caused by both particles and photons, since a high mass is not needed. The energy imparted to the solid appears in the form of electrons and holes (positive and negative carriers of electricity). If a voltage is present on the metal electrode, the electrons and holes move apart. This leads to the formation of charge sheets in the oxide layer (see Fig. 15). Positive charge sheets near the semiconductor have a profound effect on the ability of the semiconductor to conduct electrical signals. As a result of this type of effect, MOS integrated circuits are often badly affected during space flights. Special design measures are needed to prolong their survival. This may involve either a major adjustment in the processing used to make the devices or the addition of shadow shields to reduce the space radiation dose reaching the component. These effects were discussed in Section VI, Long-Term Ionization Effects.

For equipment exposed to pulsed radiation, such as nuclear explosions, photocurrents can upset most integrated circuits to act as spurious signals in digital circuits and produce errors in microcomputers, memories, and so forth. These effects were discussed in Section V, Transient Ionization Effects.

A problem that has arisen with the more recent large-scale integrated circuits is the single-event upset. As memory and logic cells on integrated circuits become smaller it is possible for a single heavy ion to deposit enough charge to discharge a logic mode, producing an error. Cosmic rays and radioactive emanations have both been found to produce soft-error upsets in large-scale integrated circuit memory chips. A problem found at one stage in

development was that the minerals used in the ceramic packaging contained radioactivity that caused soft errors. These effects were discussed in Section VIII. Single Event Phenomena

2. Data Banks

A large variety of semiconductor device designs is available on the commercial market. Each type has its own make-up, hence, each may have a differing response to radiation. While the laws of radiation physics help us to make an approximate prediction of this response, a firm prediction is obtained only by testing a reasonable sample of a given commercial type in a laboratory radiation source. Such testing has been proceeding for a number of years. It has been sponsored mainly by space and military projects, which have accumulated extensive banks of engineering data. Since new device designs are constantly being produced, the process of testing and putting the results in data banks is continuous. Efficient use of such data banks by the community of aerospace and nuclear engineers is needed to reduce repetitions in the already costly business of analyzing and testing electronic devices under radiation (for websites containing databases of this kind, see the end of this article).

3. Radiation Hardening of Devices

Research has gone into ways in which the design of semiconductor structures can be altered to reduce the effects of radiation. The improved devices are sometimes known as radiation-hardened devices. We shall illustrate the approach used in device hardening by the case of the very sensitive MOS integrated circuit. The long-term effects of radiation are localized in the silicon dioxide layers. It has been found that these effects can be alleviated by altering the rate of growth of the oxide layers and controlling the composition of the gases used to promote the growth and annealing of the oxide. Finding the best method of process modification has entailed a large amount of research. The mechanisms of degradation of the oxide films first had to be understood. It is now possible to make "megarad-hard" microprocessor circuits. The same circuit made by commercial methods would fail at a dose of 10 krad. Only a few production lines for "hard" device exist in the world.

C. Optical Systems

In the early days of solid-state physics research, it was noticed that crystals of potassium chloride, normally colorless, became purple when exposed to X-rays. Similar color-center effects have since been found in most transparent solids. Most common glasses become dark brown when irradiated, so that the degradation of optical devices under radiation is a serious engineering problem. For example, cameras used to inspect the interiors of nuclear reactors have to be changed frequently because of radiation-induced browning. A particularly vulnerable form of optical device is the optical fiber. If radiation causes an optical medium to darken, then the longer the optical path, the more severe the optical losses produced. In communications, the optical paths in fibers may be several miles long, and the loss of a few decibels per kilometer may cause the failure of the system. Table V shows the very great difference between the losses in doped and pure silica fibers. To make an optical system more tolerant to radiation, materials must be carefully selected; Table V also shows some figures for the great variability in the browning of silicate glasses and also for the beneficial effect of cerium doping on them. In some cases, the shielding of the most sensitive element of a system, such as a lens, may help.

D. Structural Materials

1. Metals

The metal and ceramic parts that go to make up nuclear reactors often receive very high exposure to neutrons and γ rays from the nuclear reactions in the fuels. In the future, similarly intense exposures will occur in the first wall of a fusion power reactor. The radiation damage caused by the neutrons is manifested as the production of voids. The net result is that the metal sample swells significantly. This may lead to the spontaneous rupture of a nuclear fuel, and will certainly lead to the loss of mechanical strength. The selection of the alloys for constructing reactors is thus a vital field of research.

2. Polymers

The mechanical strength of polymers depends on the continuity of the carbon backbones of the polymer molecules. Energy deposition from any form of radiation (including UV light) can cause the rupture, or scission, of the carbon–carbon bonds, leading to loss of strength. However, before this occurs, more complex chemical reactions may actually lead to the toughening of the polymer by cross-linking leading to a three-dimensional network of molecules. Thus, in several commercial processes, polymers are treated with radiation after molding or extrusion. Radiation also "cures" layers of polymers. For example, uncured polymeric coatings on wires or dipped articles can be hardened in place.

E. Shielding

A radiation shield is a mass of material placed between a source of radiation and an object that requires protection,

so as to reduce the radiation dose received. The design of shielding is a combination of techniques, such as the calculation of radiation transport, the selection of a structure, and the choice of materials to suit the situation of the source and object. For large shields, cost analysis is necessary. For special vehicles, such as satellites, weight is, of course of major importance and cost is less important. However, for a large reactor, the aim is to use earth, water, or concrete whenever possible to reduce cost.

The range–energy curves and photon attenuation curves discussed earlier are useful for a rough evaluation of the effectiveness of a given shield, but for the full design of a radiation shield, detailed calculations of radiation transport are required, using computer programs that trace the progress of individual particles through the various layers of a shield. For power reactors the same programs also calculate the heat produced in shields, the cooling necessary, and the complicating effects of the large ducts that penetrate the shields. In power reactors, the shields become radioactive, remaining so after machine shutdown, thus hampering the maintenance procedure. Research is carried out on low-activation shielding for reactors.

In future, magnetic-confinement fusion reactors that are built for power generation are likely to have a very heavy toroidal shield surrounding the reaction vessel. Extensive ducting will be necessary. In the United States the computers used for shield and structure analysis in fusion reactors are some of the most powerful in existence.

In spacecraft, the components of interest are circuits located in electronics boxes. They are surrounded by arrays of other parts, satellite structures, and so on, which act as built-in shielding. Some boxes are thus more protected than others. Computer programs are used to calculate the built-in shielding. The programs used perform ray tracing to determine the directions from which radiation penetrates to the component of interest. A summary of our calculation for a typical spacecraft is presented in Fig. 29.

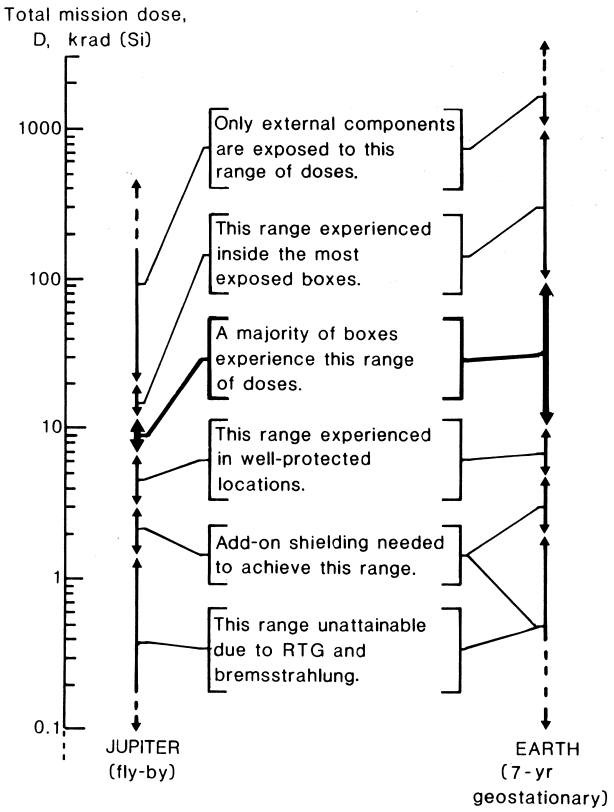


FIGURE 29 Summary of the shielding analysis for a spacecraft.

F. Systems Engineering

On some occasions, electronically controlled machines have to operate in radiation environments. The prime example is unmanned satellites, which often have to operate in the trapped radiation belts that exist around the earth. The electronics in geostationary communications satellites are designed to last for many years, and internal doses are in the region of 10 krad/year. Other examples are military vehicles, which may be exposed to nuclear explosions; remotely handled tools operating in a nuclear reactor facility; and the control and cooling systems in a fusion reactor. The task of designing equipment to survive is a problem in systems engineering. Given the usual constraints on size, weight, cost, and so on, the system has to be optimized. The designer must achieve a balance between the methods available for reducing the effects of radiation. These include (1) changing circuit design to increase tolerance, (2) adding shielding, (3) procuring less sensitive components, and (4) making alternative mechanical layouts. Radiation physics is used in the development of the system approach. The first contribution of radiation physics is to make predictions of the degradation or other responses of the electronic or optical components that incur most of the damage. This involves solid-state physics. Calculations must be made of the individual response of each component of the circuit, and then the combined effect of these must be calculated. This involves electrical engineering. Furthermore, methods to reduce the responses where necessary again may involve solid-state physics, electrical and mechanical engineering, and extensive systems engineering. Finally, a method is developed to compare the merits of the various steps taken.

X. SPECIAL USES OF RADIATION

We have learned to control and use radiation in a large number of ways, most of which are to the benefit of

mankind. First, humans have to be protected from radiation. Second, devices, especially microelectronics and optics, have to be protected. Third, controlled amounts of radiation are a part of life-saving diagnosis and therapy. Fourth, scientists have thought up a large number of other useful techniques which derive from the special powers of high-energy radiation, some of which we will now mention briefly.

A. Radiography—The Familiar One

The most common use of radiation is in radiography, for diagnosing disease. The familiar radiographic X-ray uses photons of energy near 50 keV. This range gives the correct contrast for picturing bones and internal organs or the contents of airline baggage. Chest radiography contributes more than any other man-made source to the overall exposure of the population to radiation. However, the perceived benefits outweigh the perceived risks so that this technique inspires much less anxiety than most other forms of radiation, such as the minute emissions from power stations.

B. Specialized Radiography

The technologies of electron beams, computer control and fast imaging have made advances leading to some unique forms of radiography, described below.

i. **Computed tomography**. A more detailed picture of internal organs is given by a special form of radiography in which X-ray tubes are rotated around the body and a computer reconstructs the images as "slices" through the body.
ii. **High-speed photography**. Electron beams can be switched on and off very rapidly. "Flash X-ray machines" use this method to illuminate high-speed processes such as rotating engines, explosions, and bullets striking a target.
iii. **Aircraft safety**. X-ray machines are used to study clearances and detect cracks in the spars and engines of aircraft. Linear accelerators and cobalt-60 gamma-rays are used to penetrate thick girders, where lower-energy X-rays cannot penetrate. Beams of cold neutrons have also given unique views of the motion of fuel and oil within aeroengines.

C. Life Savers

i. **Radiotherapy**. The careful local irradiation of a cancerous growth can reduce the tumor and greatly prolong the life of the subject. Radiotherapy treatment planning is a highly skilled procedure in which the radiation physicist calculates and then measures with precision during treatment the local radiation dose to a tumor while minimizing the doses that reach healthy tissue. As part of the planning, computers are used to calculate and plot the anticipated radiation doses around an organ for a chosen radiation source. The sources may be implanted radioisotopes or external high-energy beams. The art of medical dosimetry is to control the energy deposited so accurately that neither under- nor overdoses occur. Some tissues other than tumors that undergo dangerous overgrowth are now also being treated with radiation.
ii. **High-energy beams in therapy**. Particle accelerators which operate in the range of billions of electron volts have been used successfully for a dual purpose—discovering new facts about fundamental physics (HEP) and treating deep-seated cancers in the brain and body.
iii. **Personnel protection**. Both man-made radiation and the natural environment are monitored to keep the radiation exposure of humans to the minimum. Certain veins of minerals on the Earth's surface (Brazil and India have outstanding examples) can give dangerous radiation levels. Workers who use radiation must also be protected. Radiation Safety officers use instruments initially developed for nuclear physics but now specially adapted to Personnel protection at work. Miniaturized "pocket dosimeters" will "bleep" to warn of radiation hazards, while miniaturized "dosimeter badges" keep a monthly count of a person's total exposure. Polycarbonate plastic foils which record the tracks of heavy particles are used in the control of dose from Radon, a hazard from the ground in volcanic areas. A new occupational hazard, of long-term exposure to cosmic rays on high-flying airliners, is now being monitored by semiconductor detectors originally designed for spacecraft.
iv. **Detection of explosives**. Neutron beams can detect the high content of nitrogen in an explosive by using detectors which recognize the characteristic backscatter of radiation from an object, even within luggage or a large vehicle. The returns given by this method are more precise than those given by X-ray detection systems.

D. Radiotracers

Radioisotopes can be detected in minute amounts by gamma-photon counting instruments. In this way, small amounts of isotopes can be tracked through the body's chemical processes and provide a "tracer" of the body's functions. Similar methods are used to track and understand industrial chemical or physical processes.

E. Sterilization

A beam of electrons or gamma-rays will kill bacteria and parasites in food, surgical instruments, or sewage by the ionization effect without leaving any radioactivity. At the doses required, no serious degradation occurs in fruit, grain and many products irradiated, although great caution is being exercised before the public is allowed to eat irradiated foods. Industrial radiation treatment machines will, in future, constitute one of the largest uses of electrons and gamma-rays in human service. Radiation physics is used in the design of the processing plant and—as for therapy but on a higher scale—the precise delivery of doses.

F. Curing of Polymers

In industrial processes, polymers are treated with radiation after molding or extrusion to solidify or merely to toughen (cross-link) the existing polymer. Radiation is also useful for "curing" thin layers of polymers such as ink, paint, coatings on wires or dipped articles. The benefit of radiation curing is that it often reduces the pollution caused by older processes as solvents are evaporated off during drying.

G. Level Measurement

The fact that gamma-rays can penetrate thick tubes means that a beam of gamma photons from an isotope, passed across a container can be measured by a counter. The level of a liquid in the container can be measured by a sudden drop in count rate as the surface level passes through the gamma beam. A similar effect can be obtained by the back-scatter of particles.

H. Ion Implantation

Ion beams, generated in accelerators are implanted in solids to modify their properties. High-current ion implanters are available for industrial use. The predominant use of ion beams is in the formation of miniature p-n junctions using kilovolt beams of phosphorus and boron. The theory of displacement effects, described earlier, is now being applied to the better control of junction formation on the submicron scale in microelectronics.

I. Microanalysis

Radiation is used in many ways in the analysis of chemical or crystalline makeup of materials. For example, the identity of a few milligrams of a material can be determined by the diffraction of X-rays from the crystal structure, the measurement of the energy of X-rays given off when an electron-beam probe is played on the sample or the measurement of the radioactivity generated by neutrons.

J. Light from Radition

Luminescence—the conversion of radiation to light in certain "phosphor" compounds—has been a major theme of radiation physics for a 100 years. In the early days, Roentgen and Rutherford both used phosphorescent screens to make radiation visible. The "cathode ray," i.e., the electron is used effectively in modern color monitor screens because phosphor technology has developed greatly from those beginnings. In order to identify individual particles, "scintillation counters" monitor minute light pulses from large, clear volumes of phosphors which are specially grown or cast. The glow in certain irradiated phosphors is also smoothly released during heating. This effect—thermoluminescence—is a leading method for dosimetry and for archaeological dating. New HEP experiments will carry several tons of very dense and highly pure scintillators. These are designed to help uncover new members of the elementary particle family. Thus optics and luminescence will continue to be of importance to microelectronics as one of the skills required in radiation physics.

K. Summary

It can be seen from the above that, while the benefit to humans conferred by radiation when used in health care is of major importance, there are many other ways in which the science of radiation physics is of major benefit. In this last section, we have thus described some of those lesser-known and specialized applications of radiation in the service of mankind.

SEE ALSO THE FOLLOWING ARTICLES

COSMIC RADIATION • DIELECTRIC GASES • DOSIMETRY • INTEGRATED CIRCUIT MANUFACTURE • NUCLEAR PHYSICS • NUCLEAR RADIATION DETECTION DEVICES • NUCLEAR SAFEGUARDS • PARTICLE PHYSICS, ELEMENTARY • RADIATION SHIELDING AND PROTECTION • RADIATION SOURCES • SPACE PLASMA PHYSICS

BIBLIOGRAPHY

Dienes, G. J., and Vineyard, G. H. (1957). "Radiation Effects in Solids," Wiley Interscience, New York. *IEEE Transaction on Nuclear Science*, Proceedings of the Annual Conference on Nuclear and Space Radiation Effects. Vols. NS-19. No. 6. December 1966, through NS-49. No. 6. December 1998.

Johns, H. E., and Cunningham, J. R. (1969). "The Physics of Radiology," 3rd ed., Charles C Thomas, Springfield, Ill.

"Radiation Damage in Semiconductors," 7th Internation Conference on the Physics of Semiconductors, Royaumont (1964), Dunod, Paris (1965).
Vook, F. L., ed. (1968). "Radiation Effects in Semiconductors," Plenum Press, New York.
"Radiation Damage and Defects in Semiconductors" (1972). Institute of Physics, London and Bristol.
"Lattice Defects in Semiconductors" (1974). Institute of Physics, London and Bristol.
"Radiation Effects in Semiconductors" (1976). Institute of Physics, Bristol and London.
Van Lint, V. A. J., Flanagan, T. M., Leadon, R. E., Naber, J. A., and Rogers, V. C. (1980). "Mechanisms of Radiation Effects in Electronic Materials," Wiley, New York.
"Defects and Radiation Effects in Semiconductors" (1981). Institute of Physics Conference Series No. 59. Heyden Publications, London.
Chilton, A. B., Shultis, J. K., and Faw, R. E. (1984). "Principles of Radiation Shielding," Prentice-Hall, Englewood Cliffs, NJ.
Messenger, G. C., and Ash, M. S. (1986). "The Effects of Radiation on Electronic Systems," Van Nostrand Reinhold Co., New York.
Holmes-Siedle, A. G., and Adams, L. (1993). "Handbook of Radiation Effects," Oxford University Press, London and New York.
Zheng, J. F., Tavrola, M. S., and Watkins, G. D. (1995). "22nd International Conference on the Physics of Semiconductors" (D. J. Lockwood, ed.), p. 2363, World Scientific, Singapore.

DATABASE WEBSITES

http://redex.nrl.navy.mil, (U.S. Navy Research Laboratory device data base).
http://radnet.jpl.nasa.gov (device data compiled by NASA Jet Propulsion Laboratory for space projects).
http://flick.gsfc.nasa.gov (device data compiled by NASA Goddard Space Flight center).
http://erric.dasiac.com (latest radiation effects device data for military environments).
http://jpl.nasa.gov (space environments).



Scattering and Recoiling Spectrometry

J. Wayne Rabalais
University of Houston

I. Basic Physics Underlying TOF-SARS
II. TOF-SARS Instrumentation and Experimental Methods
III. Elemental Analysis from TOF-SARS
IV. Applications to Structural Analyses
V. Role of TOF-SARS and Future Directions

GLOSSARY

Blocking cone Excluded volume behind an atom in a crystal into which escaping ions do not penetrate due to the repulsive potential between the atomic cores.

Bulk truncated surface Surface for which the structure is identical to that of the bulk structure.

Channel electron multiplier Device constructed of materials with high secondary electron emissivity which is used to detect and amplify a wide variety of electromagnetic phenomena such as UV and X-ray photons, charged particles (electrons, ions), and high-velocity atoms and molecules.

Coulomb potential Scalar point function equal to the work per unit charge done against the Coulomb force in transferring a particle bearing an infinitesimal charge from infinity to a point in the field of a specific charge distribution.

Impact parameter Minimum perpendicular distance from an ion trajectory to the target atom.

Recoil Atom that is set into motion by an atomic collision or by a process involving the ejection of another particle.

Reconstructed surface Surface for which the symmetry is different from that of the bulk symmetry.

Relaxed surface Surface that has the symmetry of the bulk structure but different interatomic spacings.

Scattering Change in the direction of motion of a particle as a result of a collision or interaction with another particle.

Screening function Function that describes the difference between the atomic number of an element and the apparent atomic number resulting from reduction of the electric field of the nucleus by the space charge of surrounding electrons.

Shadow cone Excluded volume behind a target atom into which projectile ions do not penetrate due to the repulsive potential between the atomic cores.

Sputtering Ejection of atoms or groups of atoms from a surface as a result of energetic ion collisions.

Time-to-amplitude converter Device which measures the time interval between electrical pulses and generates an analog output pulse proportional to the measured time.

ION SCATTERING AND RECOILING SPECTROMETRY consist of directing a collimated beam of monoenergetic ions toward a surface and measuring the flux of scattered and recoiled particles from this surface. When the neutral plus ion flux is velocity selected by measuring the flight times from the sample to the detector, the technique is called time-of-flight scattering and recoiling spectrometry (TOF-SARS). The technique is used for (1) surface elemental analysis by measuring the energies or velocities of the particles, (2) surface structural analysis by monitoring the angular anisotropies in the flux of scattered or recoiled particles, and (3) analysis of ion–surface electron exchange probabilities by determining the ion to neutral ratios in the scattered or recoiled particle flux.

I. BASIC PHYSICS UNDERLYING TOF-SARS

A. Atomic Collisions in the Kiloelectron-Volt Range

The kinematics of atomic collisions in the kiloelectron-volt range is accurately described through classical mechanics by considering the mutual Coulomb repulsion between the colliding atomic cores. The scattered primary atom loses some of its energy to the target atom which, in turn, is recoiled into a forward direction. The energies of the scattered and recoiled atoms and the directions of their trajectories are determined by the masses of the colliding pair and the closeness of the collision.

By applying the laws of conservation of energy and momentum, the TOF of an incident ion of mass M_1 and energy E_0 which is scattered from a target atom of mass M_2 into an angle Θ is given by

$$t_S = L(M_1 + M_2)/(2M_1E_0)^{-1/2}\{\cos\Theta \pm [(M_2/M_1)^2 - \sin^2\Theta]^{1/2}\}^{-1}, \quad (1)$$

where L is the flight distance, that is, the distance from target to detector. For cases where $M_1 > M_2$, there is a critical angle $\Theta_c = \sin^{-1}(M_2/M_1)$ above which particles arriving at the detector must have experienced more than one collision. Recoils that are ejected from single collisions of the projectile into an angle ϕ, that is, direct recoils, have a TOF given by

$$t_R = L(M_1 + M_2)[(8M_1E_0)^{1/2}\cos\phi]^{-1}. \quad (2)$$

Because of the energetic nature of the collisions, molecular fragments are not observed as direct recoils and their energies are independent of the chemical bonding environment.

B. Interatomic Potentials

Scattering in the kiloelectron-volt range is dominated by repulsive potentials of the screened Coulomb type, such as

$$V(r) = [Z_1Z_2e^2/r]\Phi(r), \quad (3)$$

where r is the internuclear separation, the Z_i are the atomic numbers of the collision partners, and Φ is a screening function; there are several good approximations for Φ. Using such a potential, the relationship between the scattering angle Θ, the recoiling angle ϕ, and the impact parameter p can be determined. A small value of p corresponds to a near head-on collision and backscattering. A large value corresponds to a glancing collision and forwardscattering. Fast computer programs to simulate these trajectories have been developed.

C. Shadowing and Blocking Cones

For an ion flux impinging on a crystal surface, the ion trajectories are bent by the repulsive potentials of the atoms such that a *shadow cone* is formed behind each target atom. Ion trajectories are concentrated at the cone edges, much like rain pours off an umbrella, as shown in Fig. 1a. Similarly, if the scattered or recoiled atom is directed toward a neighboring atom, a *blocking cone* is formed behind the neighboring atom as shown in Fig. 1b. The dimensions of the cones can be calculated from classical mechanics or they can be determined experimentally using crystals with known interatomic spacings. Since the radii of these cones are on the order of interatomic spacings, that is, 1–2 Å, the ions penetrate only into the outermost surface layers, making TOF-SARS extremely surface sensitive. This is in contrast to Rutherford backscattering spectrometry (RBS), which uses ions in the million electron volt (MeV) range for which the cone radii are only on the order of 0.1–0.2 Å, allowing the ions to sample the bulk structure.

D. Scattering and Recoiling Anisotropy

When an isotropic ion fluence impinges on a crystal surface at a specific incident angle α, the scattered and recoiled atom flux is anisotropic. This anisotropy is a result

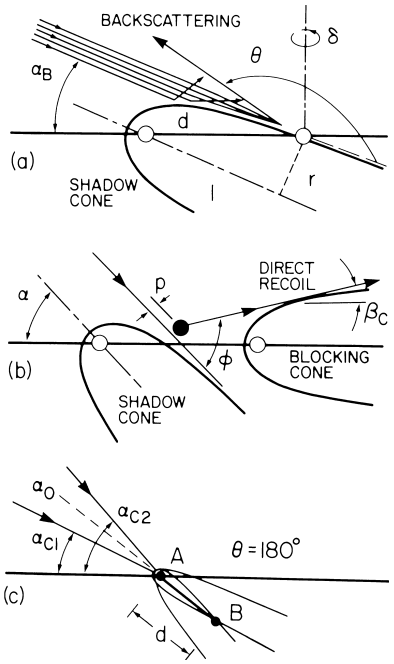


FIGURE 1 Schematic illustrations of (a) backscattering and shadowing, (b) direct recoiling with shadowing and blocking, and (c) second-layer scattering at 180°.

of the incoming ion's eye view of the surface, which depends on the specific arrangement of atoms and the shadowing and blocking cones. The arrangement of atoms controls the atomic density along the azimuths and the ability of ions to channel—that is, to penetrate into empty spaces between atomic rows. The cones determine which nuclei are screened from the impinging ion flux and which exit trajectories are blocked. By measuring the ion and atom flux at specific scattering and recoiling angles as a function of ion beam incident α and azimuthal δ angles to the surface, structures are observed which can be interpreted in terms of the interatomic spacings and shadow cones from the ion's eye view. The anisotropy in the scattered and recoiled flux is best observed by constructing scattering and recoiling structural contour maps, which are plots of the scattered or recoiled intensity in α, δ space.

E. Ion–Surface Electronic Transitions

Electron exchange between ions or atoms and surfaces can occur in two regions, (1) along the incoming and outgoing trajectories where the particle is within angstroms of the surface, and (2) in the close atomic encounter where the core electron orbitals of the collision partners overlap. In region (1), the dominating processes are resonant and Auger electron tunneling transitions, both of which are fast, that is, they occur in $<10^{-15}$ s. Because the work functions of most solids are lower than the ionization potentials of most gaseous atoms, kiloelectron-volt scattered and recoiled species are predominately neutrals as a result of electron capture from the solid. In region (2), as the interatomic distance R decreases, the atomic orbitals (AOs) of the separate atoms of atomic number Z_1 and Z_2 evolve into molecular orbitals (MOs) of a quasi-molecule and finally into the AO of the "united" atom of atomic number $(Z_1 + Z_2)$. As R decreases, a critical distance is reached where electrons are promoted into higher energy MOs because of electronic repulsion and the Pauli exclusion principle. This can result in collisional reionization of neutral species. The fraction of species scattered and recoiled as ions is sensitive to atomic structure through changes in electron density along the trajectories.

II. TOF-SARS INSTRUMENTATION AND EXPERIMENTAL METHODS

A. TOF-SARS System

An instrument for structural studies by ion scattering and recoiling should be capable of continuous variation of the scattering Θ, beam incident α, and crystal azimuthal δ angles (see Fig. 1), generation of a pulsed kiloelectron-volt primary ion beam of low fluence, efficient detection of both ions and neutrals, *in situ* low-energy electron diffraction (LEED), and operation in an ultrahigh vacuum ($<10^{-10}$ Torr) environment. Figures 2 and 3 provide schematics of a TOF-SARS system and the pulsed ion beam with associated electronics. The primary ion beam is a 1 to 5 keV rare gas ion source which has a narrow energy spread, is mass-selected, pulsed at 10 to 40 kHz, with pulse widths of 20 to 50 ns for an average ion current density of <1 nA/cm^2, and has low angular divergence. The detector is a channel electron multiplier or channel plate which is sensitive to both ions and fast neutrals. The sample is mounted on a precision manipulator, and the angles α and δ are computer controlled by means of stepping motors. Scattered and recoiled particles are velocity analyzed by measuring their flight times from the sample to detector, a distance of 1 m. An electrostatic deflector plate near the flight path allows deflection of ions for collection of TOF spectra of neutrals compared to that of ions plus neutrals. Standard timing electronics are used for data collection. The trigger output of a pulse generator, delayed by the time necessary for the pulsed beam to travel from the pulse plate to the sample, starts a time-to-amplitude converter (TAC). The TAC is stopped from the signal output of a particle reaching the detector. The output of the TAC yields a histogram of the distribution of particle flight times. The data are collected into a

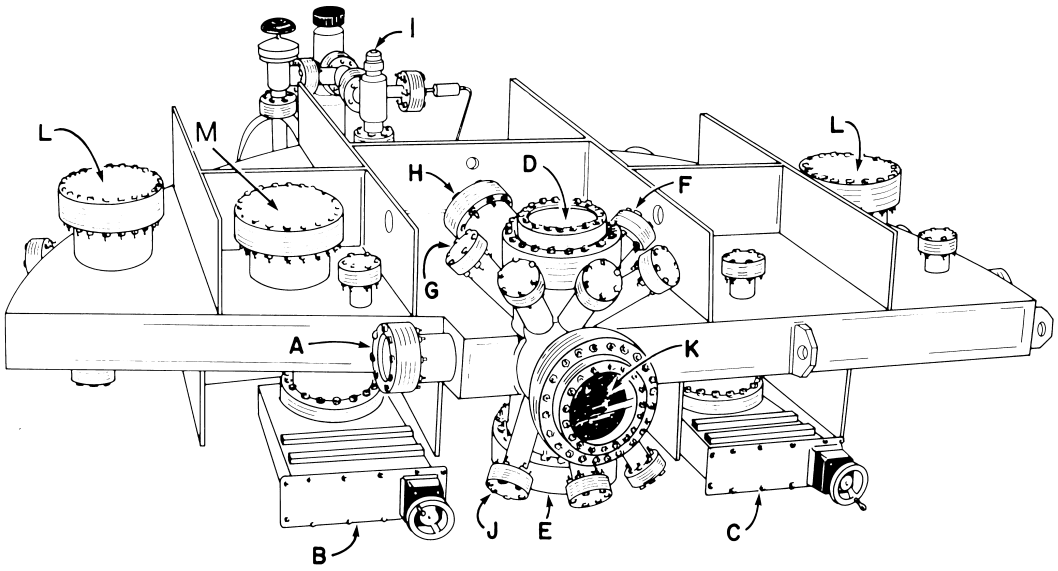


FIGURE 2 Spectrometer system designed for TOF-SARS, electrostatic analysis of scattered and recoiled ions, and conventional surface analysis techniques such as AES, XPS, and LEED. A, pulsed ion beam; B, turbomolecular pump; C, ion pump; D, sample manipulator; E, detector precision rotary motion feedthrough; F, X-ray source; G, electron gun; H, 180° electrostatic hemispherical analyzer; I, sorption pumps; J, sputter ion gun; K, viewport or reverse view LEED optics; L, titanium sublimation pump; M, cryopump.

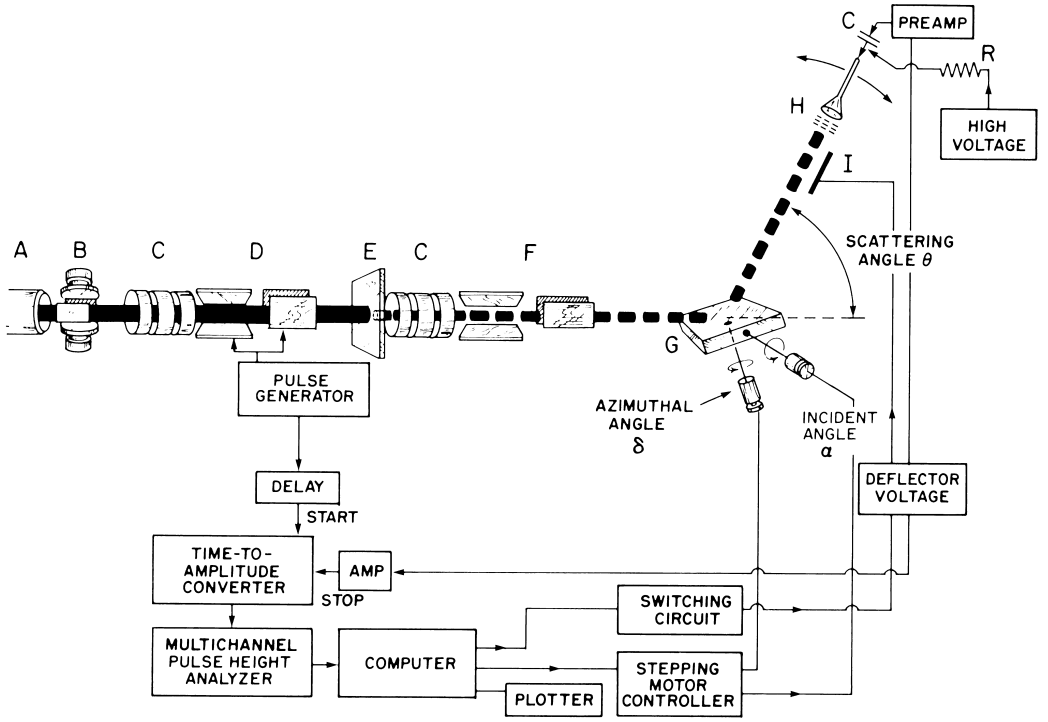


FIGURE 3 Schematic of pulsed ion beam line and associated electronics. A, ion gun; B, Wien filter; C, Einzel lens; D, pulsing plates; E, pulsing aperture; F, deflector plates; G, sample; H, channel electron multiplier with energy prefilter grid; I, electrostatic deflector.

multichannel pulse height analyzer and stored in a computer. TOF spectra can be collected with a dose of $<10^{-3}$ ions per surface atom, making the technique relatively nondestructive. The TOF-SARS instrument also has ports that contain standard surface analysis techniques such as LEED, Auger electron spectroscopy (AES), and X-ray photoelectron spectroscopy (XPS).

B. Sample Preparation

Samples are typically in the form of single crystals with polished surfaces. The surfaces are polished with successively finer grits of alumina down to 0.05 μm. Before insertion into the vacuum chamber, these samples are cleaned ultrasonically using suitable solvents. After the sample is in the chamber and UHV conditions (10^{-10} Torr) have been obtained, the sample is cleaned and annealed by electron bombardment heating from the back of the crystal. Sometimes it is necessary to sputter the surface with rare gas ions to remove impurities that have migrated to the surface. Reannealing is necessary after sputtering to regain a smooth surface. Polycrystalline samples can be used with good results in elemental analysis but with less detailed information in structural analysis. Sample cleanliness is checked by observing recoiled impurities in TOF-SARS or by using standard surface analysis techniques such as AES and XPS. Surface symmetry can be verified by observing the LEED pattern.

III. ELEMENTAL ANALYSIS FROM TOF-SARS

TOF-SARS is capable of detecting all elements either by scattering or by recoiling, or by both techniques. TOF peak identification is straightforward through the use of Eqs. (1) and (2). Collection of neutrals plus ions results in scattering and recoiling intensities that are determined by elemental concentrations, shadowing and blocking effects, and classical cross sections. Spectra from a clean tungsten surface and oxygen and hydrogen chemisorbed on that surface are shown in Fig. 4. For clean W{211}, both the backscattering (BS) and forwardscattering (FS) spectra have sharp peaks at the TOF positions predicted by Eq. (1). The absence of H, C, and O recoils in the FS spectrum indicates that the surface is clean and free of the normal atmospheric contaminants and, specifically, that it has $<1\%$ of a monolayer (ML) of these contaminants. The high background on the long TOF side of the peaks is attributable to ions which have lost energy because of multiple collisions and penetration. When oxygen is chemisorbed on W{211}, an oxygen recoil peak O(R) is observed in the TOF spectrum at the position predicted by Eq. (2). Note that the O(R) peak is on the long TOF side

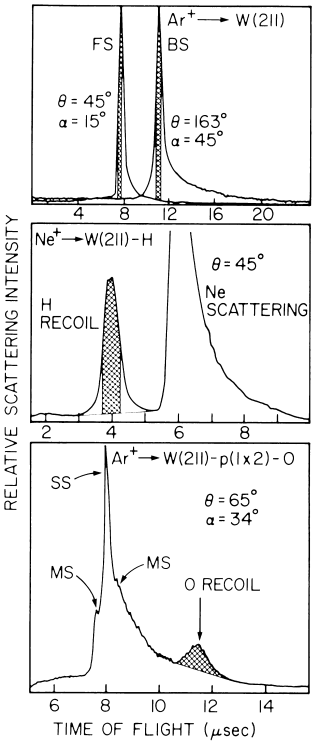


FIGURE 4 TOF spectra for keV Ar^+ scattering from clean W{211} and the oxygen chemisorbed W{211} and 4 keV Ne^+ scattering from the hydrogen chemisorbed W{211} surface. FS, forwardscattering; BS, backscattering; SS, single scattering; MS, multiple scattering.

of the Ar scattering peak. As the scattering angle Θ is reduced to smaller values, the recoil peak shifts toward lower TOF until it appears on the low TOF side of the scattering peak. Such a case is shown for hydrogen chemisorbed on W{211} where the probe was a Ne^+ beam. The intensities necessary for structural analysis are obtained by integrating the areas of fixed-time windows under these peaks.

IV. APPLICATIONS TO STRUCTURAL ANALYSES

Surface structural analyses are carried out by measuring the intensities of TOF peaks corresponding to BS, FS, or recoiled (R) events as a function of the various angles mentioned already. Each of these measurements will now be described along with the type of information obtained.

A. Backscattering

1. BS versus Incident Angle α Scans

When an ion beam is incident on a flat surface at grazing incidence, each surface atom is shadowed by its neighboring atom such that only large p collisions are possible,

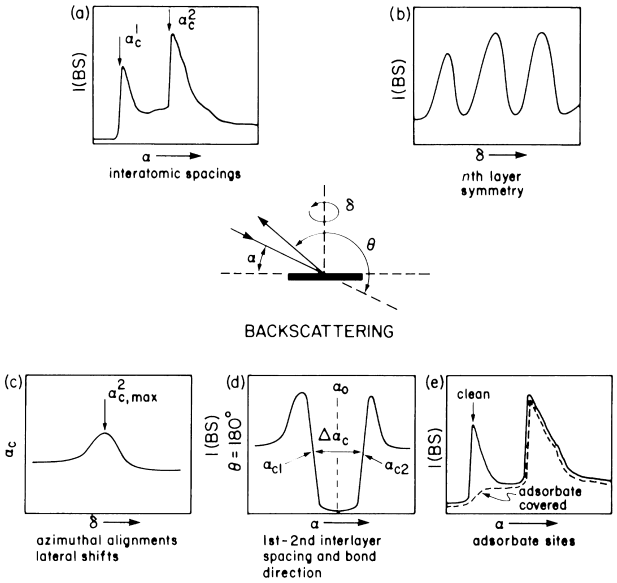


FIGURE 5 Methods of obtaining structural information from backscattering intensities I(BS). (a) I(BS) versus α for interatomic spacings; (b) I(BS) versus δ for nth layer symmetry; (c) α_c versus δ for azimuthal alignments and lateral shifts; (d) I(BS) at 180° versus α for first–second interlayer spacing and bond direction; (e) I(BS) versus α from a clean and adsorbate covered surface for locating adsorbate sites.

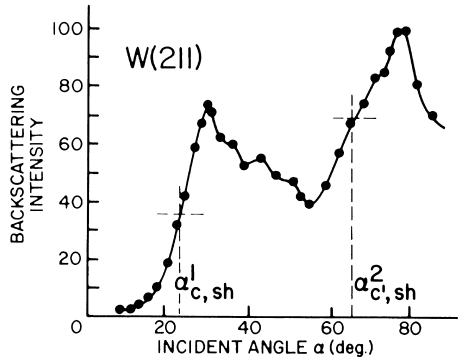


FIGURE 6 Representative plot of I(BS) versus α along the [$\bar{1}$11] azimuth of W{211} using 4 keV Ar^+.

resulting in FS. As α increases, a critical value $\alpha^i_{c,sh}$ is reached each time the ith layer of target atoms moves out of the shadow cone, allowing for small p collisions and, hence, BS as shown in Fig. 1a. If the BS intensity I(BS) is monitored as a function of α, steep rises with well-defined maxima are observed, as shown in Fig. 5a, when the focused trajectories at the edge of the cone pass through the neighboring atom. The first-layer interatomic spacings can be directly determined by measuring $\alpha^1_{c,sh}$ along the directions for which specific crystal azimuths are aligned with the projectile direction and matching the experimentally determined ratio $r/1 = \tan\alpha^1_{c,sh}$ to the (r, 1) coordinates of the shadow cone. The first-layer spacing is then $d = r/\sin\alpha^1_{c,sh}$. The first–second-layer spacing is obtained in a similar manner from $\alpha^2_{c,sh}$ measured along directions for which the first- and second-layer atoms lie in the same scattering plane. An example of such an α scan is shown in Fig. 6.

2. BS versus Azimuthal Angle δ Scans

Rotating the crystal about the surface normal with fixed α provides a scan of the crystal azimuthal angles δ. Plots of I(BS) versus δ, as in Fig. 5b, reveal the surface symmetry. Using low α, scattering occurs only from the first atomic layer, and the plot reveals the symmetry of the outermost layer. Similar information can be obtained about the second atomic layer by using higher α values. Shifts in the first–second-layer registry can be detected by monitoring the $\alpha^2_{c,sh}$ values for second-layer BS along directions near those azimuths for which the second-layer atoms are expected, from the bulk structure, to be directly aligned with the first-layer atoms. The $\alpha^2_{c,sh}$ values are maximum for those δ values where the first- and second-layer atoms are aligned, as shown in Fig. 5c.

3. 180° BS versus Incident Angle α Scans

By focusing the primary ion beam through an aperture in a multichannel plate detector, it is possible to observe BS at $\Theta \approx 180°$. Since the ion incidence and scattering direction are coaxial, variations in I(BS) due to angular variations reflect the local atomic symmetry as shown in Fig. 1c. The I(BS) versus α scans for BS from atom B are expected to exhibit a pattern similar to that shown in Fig. 5d. The low I(BS) at α_0 is due to shadowing of atom B by atom A. The I(BS) peaks on both sides of α_0 are due to focusing by shadowing and blocking cones of atom A as atom B goes in and out of the cones, resulting in two critical α_0 values. The α_0 corresponds to the bond direction and the difference in the critical angles $\Delta\alpha_c$ is related to the bond length d by $d = r/\sin(\Delta\alpha_c/2)$.

4. BS versus α Scans with Adsorbates

Collecting I(BS) data as a function of α for the adsorbate-covered surface and comparing these data to those of the clean surface provide a qualitative location of the adsorption site position. I(BS) versus α structures that are sensitive to the adsorbates are those for which the adsorbate atoms lie along or very near the path of the scattering primary ion. Although the primary ion cannot backscatter from light adsorbates, the presence of the adsorbates near the scattering trajectory results in small-angle deflections of the projectile which are sufficient to produce changes

as shown in Fig. 5e. In this figure, the first-layer scattering peak is perturbed, indicating that the adsorbate lies along this specific azimuth and interferes with the trajectories for first-layer scattering.

B. Forwardscattering

1. FS versus δ Scans

When very small ($<10^\circ$) incident angles α are used, the projectile ion experiences a continuous potential along low index azimuths rather than individual atomic potentials. In scattering from such a continuous potential, the outgoing trajectories are focused at the specular ejection angle β, that is, $\alpha = \beta$. In addition, surface semichanneling, that is, steering of the incident ions by troughs formed from two first- and one second-layer rows, contributes to the I(FS) enhancement along the low index azimuths. If the crystal is rotated about δ with $\alpha = \beta$, plots of I(FS) versus δ exhibit structure as shown in Fig. 7a. Intensity maxima are observed along the principal azimuths due to the specular focusing. If a large exit angle β is used, minima rather than maxima are observed along the high-symmetry axes as shown in Fig. 7b. The surface periodicity can be read directly from these minima in the plots, thereby revealing the crystal structure. An example of such a δ scan is shown in Fig. 8 for a Pt(110) surface in the (1 × 2) reconstruction phase.

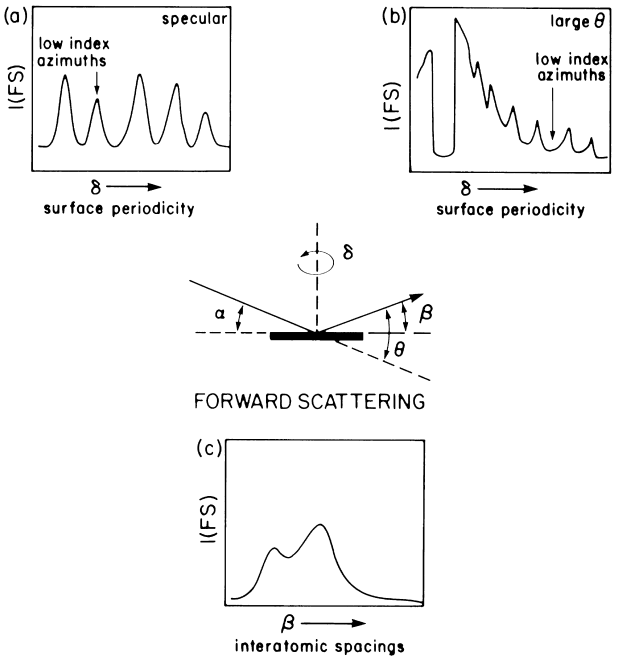


FIGURE 7 Methods of obtaining structural information from forwardscattering intensities I(FS). (a) I(FS) versus δ using specular scattering for surface periodicity; (b) I(FS) versus δ using large Θ for surface periodicity; (c) I(FS) versus β for interatomic spacings.

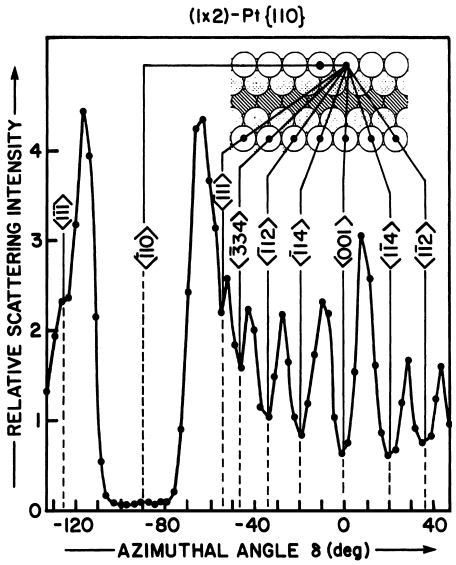


FIGURE 8 Representative plot of I(FS) versus δ for Pt(110) in the (1 × 2) missing-row reconstruction using Ar^+ at $\alpha = 5^\circ$ and $\Theta = 40^\circ$, illustrating surface periodicity.

2. FS versus β Scans

Varying Θ at constant α provides a scan of the exit angle β. At low β the structural features are determined by focusing of scattered trajectories at the edges of the blocking cones of atoms obstructing their escape along that azimuth. For example, if a low α value is used so that only first-layer scattering is observed, the structures will be determined by first-layer atoms blocking their first-layer neighbors. If the crystal consists of a single structure, only one interatomic distance exists along a given azimuth, and hence only one peak is observed. The example in Fig. 7c is typical of two different interatomic spacings along one azimuth, and therefore two structures.

C. Recoiling

1. R versus α Scans

Plots of I(R) versus α provide two critical angles as shown in Fig. 9a. The first critical angle α_c corresponds to the situation where the edge of the shadow cone cast by a neighboring atom focuses the primary ion flux at the appropriate impact parameter for direct recoil into ϕ as shown in Fig. 1b. The steep drop-off at high α corresponds to the critical ejection angle $\beta_c = \Theta - \alpha$. The β_c is determined by focusing of recoil trajectories by the blocking cone of a neighboring atom. This provides a direct measure of

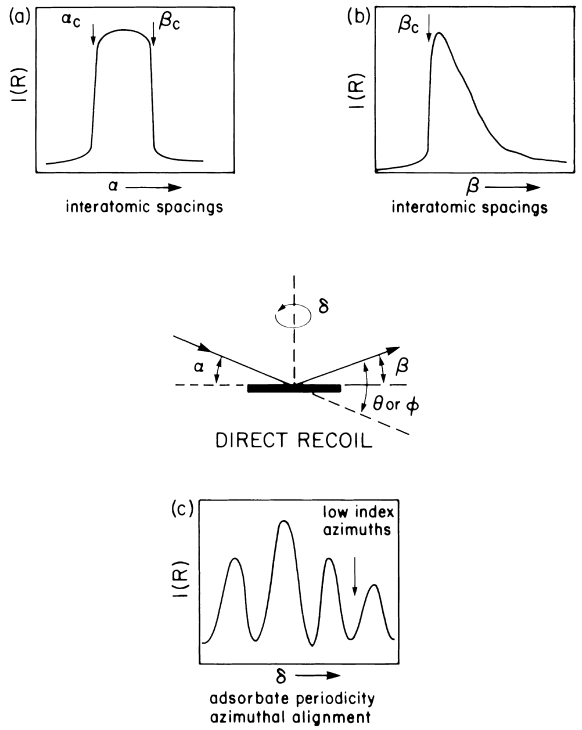


FIGURE 9 Methods of obtaining structural information from recoiling intensities I(R). (a) I(R) versus α for adsorbate interatomic spacings; (b) I(R) versus β for adsorbate interatomic spacings; (c) I(R) versus δ for adsorbate periodicity and azimuthal alignment.

the interatomic spacings between an adsorbate atom and neighboring atoms along the azimuth. An example of such α scans is shown in Fig. 10 for oxygen chemisorbed on W{211}.

2. R versus β Scans

Plots of I(R) versus β as shown in Fig. 9b, obtained by varying Θ, provide a direct measure of the critical blocking angle β_c, which is comparable to the previous measurement.

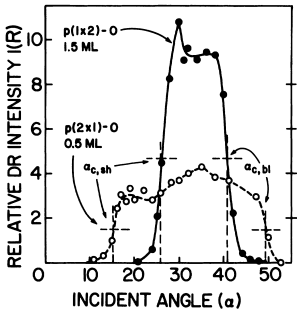


FIGURE 10 Representative plots of I(R) versus α for the p(1 × 2) high (1.5 ML) and p(2 × 1) low (0.5 ML) oxygen coverages on a W{211} surface. ML, monolayer.

3. R versus δ Scans

Rotating the crystal about δ with constant α can reveal the symmetry of the adsorbate sites on the surface. Plots of I(R) versus δ exhibit maxima and minima, as shown in Fig. 9c, resulting from shadowing and blocking of the adsorbate sites because of the different interatomic spacings as δ is changed. This shows the adsorbate symmetry and azimuthal alignment with respect to the substrate crystallographic axes.

4. Scattering and Recoiling Structural Contour Maps (SSCM & RSCM)

TOF-SARS data can be summarized in the form of SSCMs and RSCMs for a given system; these are plots of intensity in α, δ-space. The SSCM for a clean W{211} surface using I(BS) and the RSCM for an oxygen-chemisorbed W{211} surface using I(R) are shown in Fig. 11. These plots reveal the symmetry of the data in α, δ-space and serve as a fingerprint for a specific crystal face and adsorbate structure. Considering the SSCM of clean W, $\alpha^{l}_{c,sh}$ is symmetrical about $\delta = 0°$, $[01\bar{1}]$ as is the first-atomic layer. The row of intense structures increasing diagonally at higher α values is due to scattering from the second and third atomic layers, which are not symmetrical about the $[01\bar{1}]$ azimuth. Considering the RSCM of oxygen on W{211}, the $\alpha^{i}_{c,sh}$ values are symmetrical about the $[01\bar{1}]$ azimuth, as are all of the recoil structures. This shows that the adsorption site is symmetrical with respect to this azimuth. Detailed analysis shows that the clean surface is relaxed relative to the bulk structure and that the oxygen occupies threefold sites in which it is bound to two first- and one second-layer tungsten atoms.

V. ROLE OF TOF-SARS AND FUTURE DIRECTIONS

The TOF-SARS technique contributes to our knowledge of surface science through (1) elemental analysis, (2) structural analysis, and (3) analysis of electron-exchange probabilities. The merits of each of these areas will be considered below.

A. Elemental Analysis

Although TOF-SARS is sensitive to all elements, including hydrogen, the limited resolution of the TOF technique presents difficulties in resolving spectral peaks of high-mass elements with similar masses. The unique feature for elemental analysis is direct monitoring of surface hydrogen. For general qualitative and quantitative surface

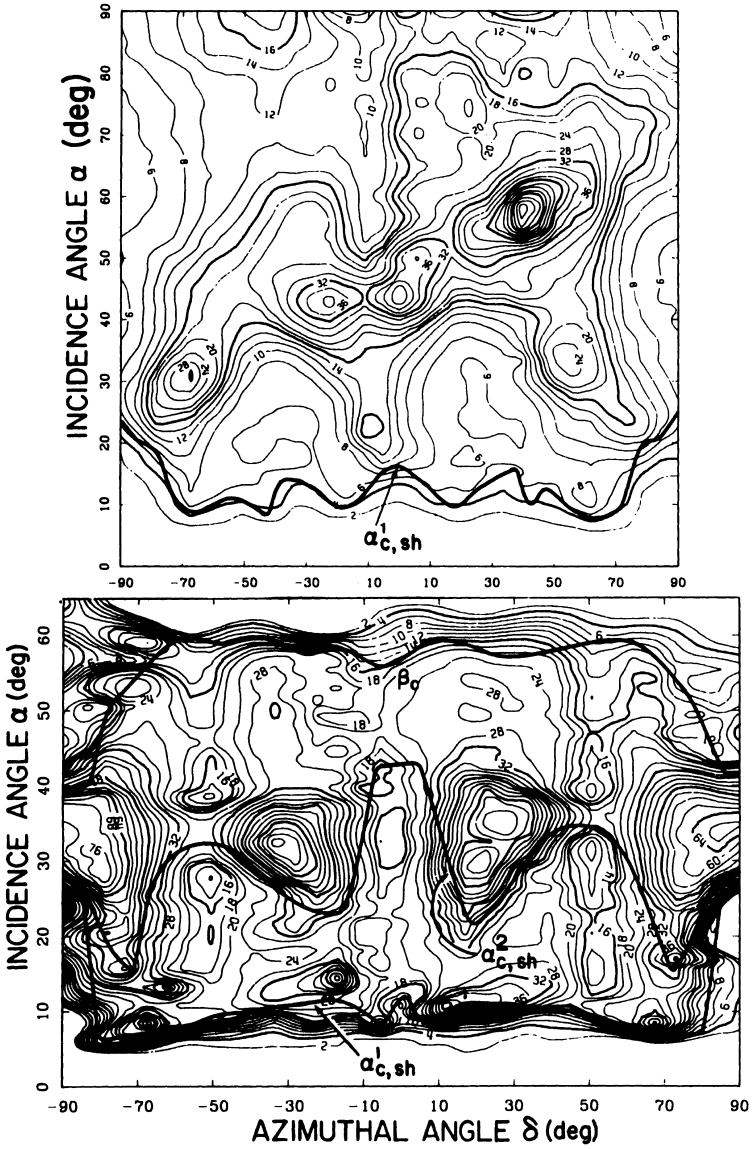


FIGURE 11 (Top) Scattering structural contour map for the clean W{211} surface using 4 keV Ar^+ and $\Theta = 163^\circ$. (Bottom) Recoiling structural contour map for the W{211}-p(1 × 2)-O oxygen-adsorbed surface using 4 keV Ar^+ and $\Theta = 65^\circ$. $\delta = 0^\circ$ is the [01$\bar{1}$] azimuth, $\delta = -90^\circ$ is the [1$\bar{1}\bar{1}$] azimuth, and $\delta = +90^\circ$ is the [$\bar{1}$11] azimuth.

elemental analyses, XPS and AES remain the techniques of choice.

B. Structural Analysis

The major role of TOF-SARS is as a surface structure analysis technique which is capable of probing the positions of all elements with an accuracy of $\leq$0.1 Å. TOF-SARS is sensitive to short-range order—individual interatomic spacings along azimuths. It provides a direct measure of interatomic distances in the first and subsurface layers and a measure of surface periodicity in real space. It is complementary to LEED, which probes long-range order, minimum domain size of 100 to 200 Å, and provides a measure of surface and adsorbate symmetry in reciprocal space. Coupling TOF-SARS and LEED provides a powerful combination for surface structure investigations.

C. Ion–Surface Electron-Exchange Probabilities

One of the unsolved problems in the interaction of low-energy ions with surfaces is the mechanism of charge transfer and prediction of the charge composition of the flux of scattered, recoiled, and sputtered atoms. The ability

to collect spectra of neutrals plus ions and only neutrals provides a direct measure of scattered and recoiled ion fractions. Plots of ion fractions in α, δ-space provide electronic transition probability contour maps which are related to surface electron density along the various azimuths. Consistency between such electron density contours and the SSCMs provides a unique description of the surface electronic and atomic structure.

D. Future Directions

Future developments will include the use of large channel plates or hemispherical grids for spatial resolution of particles ejected through a large solid angle, facilitating rapid and direct collection of an entire scattering and recoiling map. Improvements in optics will provide narrower ion-pulse widths, resulting in enhanced time resolution of the spectra. Development of the simulations will make computer modeling of surface structures routine. TOF-SARS is now well established as a surface structural analysis technique that will have a significant impact in areas as diverse as thin-film growth, catalysis, hydrogen embrittlement and penetration of materials, surface-reaction dynamics, and analysis of interfaces.

SEE ALSO THE FOLLOWING ARTICLES

ATOMIC AND MOLECULAR COLLISIONS • ATOMIC SPECTROMETRY • AUGER ELECTRON SPECTROSCOPY • MASS SPECTROMETRY • SURFACE CHEMISTRY • X-RAY PHOTOELECTRON SPECTROSCOPY • X-RAY, SMALL-ANGLE SCATTERING

BIBLIOGRAPHY

Bu, H., Grizzi, O., Shi, M., and Rabalais, J. W. (1989). Time-of-flight scattering and recoiling. II. The structure of oxygen on the W(211) surface. *Phys. Rev. B* **40,** 10147–10162.

Chang, C. S., Porter, T. L., and Tsong, I. S. T. (1989). In-plane geometry of the Si(111)-($\sqrt{3} \times \sqrt{3}$) Ag surface. *J. Vac. Sci. Tec.* **A7,** 1906–1909.

Fauster, T. (1988). Surface geometry determination by large-angle ion scattering. *Vacuum* **38,** 129–142.

Grizzi, O., Shi, M., Bu, H., Rabalais, J. W., and Hochmann, P. (1989). Time-of-flight scattering and recoiling. I. Structure of the W(211) surface. *Phys. Rev. B* **40,** 10127–10146.

Grizzi, O., Shi, M., Bu, H., and Rabalais, J. W. (1990). Time-of-flight scattering and recoiling spectrometer (TOF-SARS) for surface analysis. *Rev. Sci. Instrum.* **61,** 740–752.

Huang, J. H., and Williams, R. S. (1988). Surface-structure analysis of Au overlayers on Si by impact-collision ion-scattering spectroscopy: $\sqrt{3} \times \sqrt{3}$ and 6×6(111)/Au. *Phys. Rev. B* **38,** 4022–4032.

Mashkova, E. S., and Molchanov, V. A. (1985). "Medium-Energy Ion Reflection from Solids," North-Holland, Amsterdam.

Rabalais, J. W. (1988). Direct recoil spectrometry. *CRC Critical Rev. Sol. St. and Mat. Sci.* **14,** 319–376.

Shi, M., Grizzi, O., Bu, H., Rabalais, J. W., Rye, R. R., and Nordlander, P. (1989). Time-of-flight scattering and recoiling spectrometry. III. The structure of hydrogen on the W(211) surface. *Phys. Rev. B* **40,** 10163–10179.

Williams, R. S., Kato, M., Daley, R. S., and Aono, M. (1990). Scattering cross sections for ions colliding sequentially with two target atoms. *Surface Sci.* **225,** 355–366.



Transition Probabilities and Atomic Lifetimes

Wolfgang L. Wiese
National Institute of Standards and Technology

I. History and Basic Concepts
II. Numerical Determinations
III. The Emission Method
IV. Absorption and "Hook" Techniques
V. Atomic Lifetime Determinations
VI. Atomic Structure Theory
VII. Regularities in Oscillator Strengths: *f*-Sum Rules
VIII. Data Availability

GLOSSARY

Allowed transition Electric dipole (E1) transition between two atomic energy states permitted by quantum mechanical selection rules.

Branching ratio technique Technique for measuring, in emission or absorption, the relative transition probabilities of all lines ("branches") originating from a common atomic level.

Forbidden transition Transition between two quantum states forbidden by ordinary electric dipole selection rules and therefore much weaker than an allowed transition; also known as magnetic dipole (M1), electric quadrupole (E2), and higher-order radiation.

Line strength Square of the quantum mechanical electric dipole matrix element. This quantity is equivalent to the transition probability.

Mean life or radiative lifetime Time during which an assembly of atoms in an excited state decays by spontaneous emission to a fraction $1/e$ of its original number.

Multiconfiguration self-consistent field method Iterative quantum mechanical procedure of finding the wave function for a state of a many-electron atomic system, including electron correlation effects.

Oscillator strength or f value This quantity, equivalent to the transition probability, originated from classical absorption and dispersion theory, is a dimensionless number, and is widely used in astrophysics.

Transition probability Rate at which an atom or ion will make a spontaneous radiative transition from a higher to a lower energy state, i.e., emit a photon; a measure

of the intrinsic strength of a spectral line. Sometimes also called the Einstein A coefficient.

IN ATOMIC spectral line radiation, the transition probability is the quantity that determines the intensity of a spectral line, in addition to lightsource-dependent factors, such as the number of emitters. A characteristic feature of this atomic quantity is that it is generally difficult to determine accurately so that many of the existing data are rather uncertain.

I. HISTORY AND BASIC CONCEPTS

The concept of atomic transition probabilities was introduced by Einstein in 1916 in his quantum theoretical description of the interaction between matter and radiation. He derived Planck's fundamental radiation law by assuming that the following three elementary quantum processes of radiation take place between two excited atomic states of energies, E_k and E_i, involving photons (quanta) of frequency ν_{ik} and of energy $E_k - E_i = h\nu_{ik}$, where h is Planck's constant (see Fig. 1).

1. *Spontaneous emission*. Treating spontaneous emission processes in analogy to radioactive decay processes, Einstein assumed that the number of these events per unit time and volume, as a consequence of electron jumps from a higher atomic energy level k to a lower level i, is given by

$$A_{ki} N_k. \tag{1}$$

Here N_k is the initial number of atoms per unit volume in energy level k, and A_{ki} (the subscript for the initial level is written first) is a constant that is specific to this transition. It is now generally called the atomic transition probability, but it is still sometimes referred to as the Einstein A coefficient (It is actually an atomic transition rate, since it has the dimension of inverse time).

2. *Absorption of radiation*. This quantum process requires the presence of a radiation field with resonant photons, that is, photons of energy $h\nu_{ik}$. The field may originate either from surrounding radiating atoms of the same kind (self-absorption) or from an external radiation source, and thus the rate of these processes is proportional to the energy density $u(\nu)$ of the radiation field. Einstein showed that this resonance process may be positive as well as negative. The usual (positive) absorption of photons by atoms in the lower energy state i is proportional to their number density N_i and occurs at a rate, per second, which raises atoms into state k:

$$B_{ik} N_i u(\nu). \tag{2}$$

3. *Stimulated emission.* The absorption process may also be negative; that is, atoms may be stimulated resonantly by the radiation field to emit photons of the same frequency and in the same direction as the incoming photons. These induced emission processes occur from the energetically higher atomic state k into the lower state i at the rate

$$B_{ki} N_k u(\nu). \tag{3}$$

Einstein, considering a gas in a state of thermodynamic equilibrium, showed that he could derive Planck's law by balancing these three radiation processes and that the three constants A_{ki}, B_{ik}, and B_{ki} are related in a simple manner. However, nowadays only the transition probability for spontaneous emission A_{ki} is in general use.

In the classical electron theory developed earlier, the principal quantity governing the absorption in a spectral line is the number of dispersion electrons $\tilde{N}$. In 1921 Ladenburg established a connection to the quantum concept by equating the classical expression for the total absorption in a spectral line with Einstein's theoretical result and derived a simple relation among $\tilde{N}$, N_i, and B_{ik}. This relation still had a flaw in that it contained the densities of absorbing atoms N_i as well as $\tilde{N}$, which Ladenburg and Reiche (1923) eliminated by splitting $\tilde{N}$ into the product

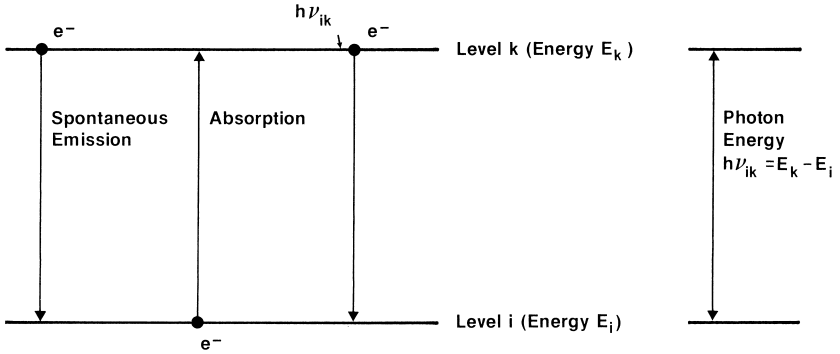


FIGURE 1

$\tilde{N} = N_i f$ (N_i can be interpreted as the number density of resonant electrons occupying state i as well as the number density of atoms in state i). This quantum correction factor f, which is now generally known as the absorption oscillator strength, or f value, can be interpreted as the fraction of optically effective electrons (per atom) for a specific transition and is frequently used in absorption work and astrophysical applications.

In the late 1920s a fully quantum mechanical treatment of the interaction of matter and the electromagnetic field was developed by Born, Dirac, Heisenberg, Pauli, Schroedinger, and others. This yields, in the electric dipole approximation, for the rate of energy loss per atom in a transition $k \rightarrow i$,

$$-\left(\frac{dE}{dt}\right)_{ki} = \frac{16\pi^3 \nu_{ik}^4}{3c^3 \varepsilon_0} \times \langle \psi_i | P_e | \psi_k \rangle^2, \tag{4}$$

where ε_0 is the permittivity of the vacuum; ψ_i and ψ_k are the atomic wave functions for states i and k, respectively, and $\boldsymbol{P}_e = e|\sum_{n=1}^{M} \mathbf{r}_n|$ is the quantum mechanical expression for the electric dipole moment, with e the electron charge and $\mathbf{r}_n (r_n^2 = x_n^2 + y_n^2 + z_n^2)$ the position vector of the nth electron with respect to the nucleus and where the sum is taken over all M electrons of the atom. The square of the matrix element in Eq. (4) has been defined by Condon and Shortley as the line strength S_{ik},

$$S_{ik} = \langle \psi_i | \mathbf{P}_e | \psi_k \rangle^2. \tag{5}$$

This quantity is symmetric in i and k, that is, $S_{ik} = S_{ki}$ (usually written simply as S), and is widely used in theoretical work. Both Eq. (1) and Eq. (4) represent energy loss rates due to spontaneous emission and thus establish the connection between A and S.

The numerical relations among A, f, and S are given in Table I for allowed (i.e., electric dipole or E1) transitions. This type of radiation takes place only between certain quantum states of an atom or ion, and these state combinations are governed by quantum mechanically derived selection rules for optical transitions. Although electric dipole transitions are by far the strongest in any given spectrum, quantum mechanics also shows that higher-order or multipole radiation exists and has provided general expressions for the strengths of these so-called forbidden transitions, which are of drastically weaker strength. This higher-order (multipole) radiation [i.e., magnetic dipole (M1) and electric quadrupole (E2) transitions] follows different selection rules. Also, the relations between S and A are different (f is used only for E1 transitions) and are shown in Table II for M1 and E2 transitions.

Lifetimes of excited atomic states are closely related to transition probabilities. In the absence of external factors, an atomic state k is depopulated by spontaneous electron transitions into all possible lower energy states i at the rate

$$\frac{dN_k}{dt} = -N_k(t) \sum_i A_{ki}. \tag{6}$$

Integration yields

$$N_k(t) = N_{k,0} \exp\left[-\left(\sum_i A_{ki}\right) t\right], \tag{7}$$

where $N_{k,0}$ is the initial population of state k at time $t = 0$. The mean lifetime τ_k of an atomic state k is defined as the time in which $N_k(t)$ decays to the fraction $1/e$ of its original value; thus,

$$\tau_k \equiv \left(\sum_i A_{ki}\right)^{-1}. \tag{8}$$

The upper state of the principal resonance line (labeled R) can decay only to the ground level, so that for this special case, Eq. (8) reduces to $\tau_R = A_R^{-1}$.

II. NUMERICAL DETERMINATIONS

Transition probabilities for electric dipole transitions of neutral atoms typically span the range from about 10^9 s^{-1}

TABLE I Numerical Relations between Transition Probabilities, Oscillator Strengths, and Line Strengths[a]

	A_{ki}	f_{ki}	S
Transition probability $A_{ki} =$	1	$6.670 \times 10^{13} \frac{g_i}{\lambda^2 g_k} f_{ik}$	$2.026 \times 10^{15} \frac{1}{g_k \lambda^3} S$
Oscillator strength $f_{ki} =$	$1.499 \times 10^{-14} \lambda^2 \frac{g_k}{g_i} A_{ki}$	1	$30.38 \frac{1}{g_i \lambda} S$
Line strength $S =$	$4.936 \times 10^{-16} g_k \lambda^3 A_{ki}$	$3.292 \times 10^{-2} g_i \lambda f_{ik}$	1

[a] The transition probability A_{ki} is given in reciprocal seconds, the oscillator strength (or f value) f_{ik} is dimensionless, and the line strength S is given in atomic units, the wavelength λ is in nanometers, and the statistical weight g_n is obtained from the total angular momentum quantum number J_n by $g_n = 2J_n + 1$.

TABLE II Numerical Relations between Transition Probabilities and Line Strength for Magnetic Dipole (M1) and Electric Quadrupole (E2) Lines ("Forbidden" Lines)[a]

	Magnetic dipole (M1)	Electric quadrupole (E2)
Transition probability $A_{ki} =$	$\frac{2.697 \times 10^{10}}{g_k \lambda^3} S$	$\frac{1.120 \times 10^{13}}{g_k \lambda^5} S$
Line strength $S =$	$3.707 \times 10^{-11} g_k \lambda^3 A_{ki}$	$8.929 \times 10^{-14} g_k \lambda^5 A_{ki}$

[a] The units are as in Table I (but the atomic units for these line strengths are different). The statistical weight g is obtained from the total angular momentum quantum number J by $g = 2J + 1$.

for the strongest spectral lines at short wavelengths to 10^3 s^{-1} and less for weaker lines at longer wavelengths. The transition probabilities for given transitions along an isoelectronic sequence, that is, for all atomic systems with the same number of atomic electrons but increasing nuclear charge Z, usually increase strongly with Z, due mainly to the strong increase in the transition energies with Z. However, the transition energies for lines within the same electron shell (where the principal quantum number n does not change; i.e., $\Delta n = 0$) grow much more slowly, and therefore the A_{ki} values increase only gradually. As an example of the growth in the transition probability of a $\Delta n \neq 0$ transition, available data show that for the $2s2p\ ^3P^o$–$2s3d\ ^3D$ transition of the beryllium sequence, the transition probability increases by a factor of about 1.3×10^5 from neutral beryllium (nuclear charge $Z = 4$) to Fe^{22+} ($Z = 26$).

The dimensionless f values vary from ~ 1 for the strongest transitions to $\sim 10^{-4}$ or less for weak lines and generally exhibit no pronounced scaling with Z. For the above-cited example of the $2s2p\ ^3P^o$–$2s3d\ ^3D$ transition in the beryllium sequence, the f value increases slowly, from 0.27 for neutral beryllium to ~ 0.71 for Fe^{22+}.

For the specific case of hydrogenic ions of nuclear charge Z, the following relations exist with respect to neutral hydrogen:

$$A_Z = A_H Z^4 \quad \text{and} \quad f_Z = f_H. \tag{9}$$

[No $\Delta n = 0$ optical transitions are possible because all states of a given n have the same energy (energy degeneracy in hydrogenic species) except for small relativistic and quantum electrodynamic effects.]

The transition probabilities of forbidden lines cover a very wide range, spanning many orders of magnitude. Transition rates higher than 1 s^{-1} are rare for neutral atoms, but for given transitions within an isoelectronic sequence the transition probabilities increase strongly with increasing nuclear charge Z. M1 transitions are usually stronger than E2 transitions.

Atomic transition probabilities are needed for many applications, mainly in astrophysics, fusion energy research, low-temperature plasma technology, laser development, atmospheric science, atomic physics, spectrochemistry, and the lighting industry. The rather divergent demands from these different user communities are the principal driving force for producing large quantities of numerical data. However, the accurate determination of atomic transition probabilities still presents formidable problems, on both the experimental and the theoretical side. Since the 1920s, numerous methods have been developed to measure and calculate these quantities, and more than 7000 papers containing numerical results have appeared in the literature. Nevertheless, the data are far from satisfactory. Results that are essentially exact are available only for the case of hydrogen or hydrogen-like ions. These are atomic systems with a single electron, where quantum mechanics provides exact numerical solutions. In Table III, the transition probability data for some prominent hydrogen lines are listed. For the case of two-electron atomic systems (i.e., helium or helium-like ions), the best calculated results are estimated to be correct to within 0.0001%. For prominent lines of the alkali metals, which possess a single electron outside closed shells, experimental precision measurements have provided data with uncertainties as small

TABLE III Transition Probability Data for Some Prominent Hydrogen Lines

Wavelength (nm)	Common symbol for line	Statistical weights		Transition probability (s^{-1})[a]
		gi	g_k	
121.567	L_α	2	8	4.699 (8)
102.572	L_β	2	18	5.575 (7)
97.254	L_γ	2	32	1.278 (7)
94.974	L_δ	2	50	4.125 (6)
656.280	H_α	8	18	4.410 (7)
486.132	H_β	8	32	8.419 (6)
434.046	H_γ	8	50	2.530 (6)
410.173	H_δ	8	72	9.732 (5)

[a] The number in parentheses indicates the power of 10 by which the transition probability values have to be multiplied.

as 0.15%, and recent calculations for the Li resonance line are estimated to be accurate within 0.002%. For the case of atoms and ions with two valence electrons outside closed electron shells (i.e., the alkaline earths), data for the most thoroughly studied transitions are known to better than 1%, whereas for other transitions and for more complex atoms or ions, such as N I, uncertainties of the order of 5 to 50%, or worse for weak transitions, must be expected. For very complex heavy atoms and ions, such as Cr II or Fe II, and especially for the weaker lines, the best data are in the ±10% range, but uncertainties of factors of 2 are common in calculated data and discrepancies of the order of a factor of 10 have even been found between data from different sources. An understanding of the main features of the experimental and theoretical approaches to determine transition probabilities is therefore of considerable importance for analyzing their strengths and weaknesses. The short discussions that follow are restricted to the main approaches; special techniques and modifications are not covered.

III. THE EMISSION METHOD

The emission method, practiced for more than 60 years, has supplied and is continuing to supply the largest number of experimental transition probabilities. Figure 2a shows the basic experimental setup. Emission spectra are usually generated in plasmas of low or moderate density, produced by some type of electrical discharge, and total line intensities are measured photoelectrically with a grating or Fourier-transform spectrometer and are calibrated on an absolute scale with a spectral radiance standard. Assuming that the plasma column is homogeneous and optically thin (i.e., without self-absorption in the spectral lines), the energy emitted in a transition of atomic electrons from a higher level k to a lower level i of (normally precisely known) frequency ν_{ik}, per unit time and steradian, is related to the atomic transition probability A_{ki} [see Eq. (1)] by

$$I_{ik} = \int_0^\infty I(\nu)\,d\nu = \frac{1}{4\pi} A_{ki} N_k h \nu_{ik} l. \qquad (10)$$

Here $I(\nu)$ is the specific intensity emitted at frequency ν, which must be integrated over the frequency (or wavelength) range of the line profile to obtain the total line intensity I_{ik}.

The determination of A by the emission technique therefore requires the following.

(a) Measurements of the total line intensity A_{ki} (including calibration with a radiance standard) and of the length l of the emitting plasma layer.
(b) Determination of atomic state populations N_k, or number densities.

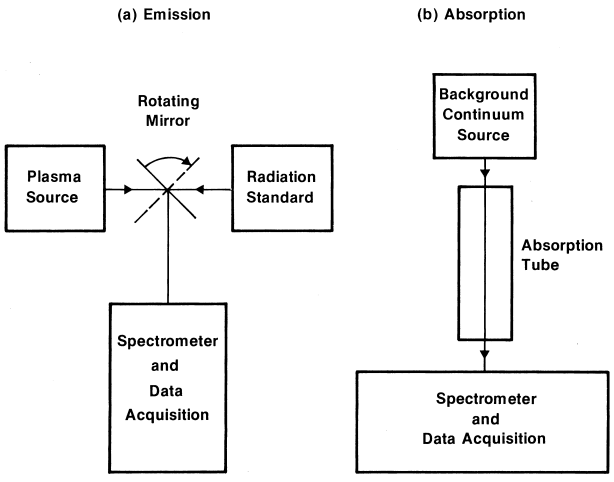


FIGURE 2

A. The "Branching Ratio" Technique

In its most elementary and rather restricted form, the emission technique is applied on a relative scale to a group of lines originating from the same upper atomic level k and ending in various lower levels, 1, 2, 3, ... For such lines, N_k and l are identical so that one obtains, utilizing Eq. (10),

$$\left(\frac{I_{1k}}{\nu_{1k}}\right) : \left(\frac{I_{2k}}{\nu_{2k}}\right) : \left(\frac{I_{3k}}{\nu_{3k}}\right) : \cdots = A_{k1} : A_{k2} : A_{k3} : \cdots . \qquad (11)$$

Since all lines involved originate from the same energy level, there are no special requirements for the plasma source except short-term stability (not required for Fourier transform spectroscopy) and an optically thin emission layer, which are requirements that can be readily tested. The branching ratio technique has been applied especially to heavier elements. For complex spectra the significantly contributing line "branches" from a given level are fairly extensive, comprising often more than 10 lines. Thus, to obtaining a complete set of branches is often a problem. When complete sets of branches can be measured, sums of relative transition probabilities for all lines originating from level k are obtained, which may be converted to absolute data with a known lifetime for this level, since according to Eq. (8), $\tau_k = (\sum_i A_{ki})^{-1}$. Appropriate combinations of branching ratio emission and lifetime measurements are an efficient and accurate approach to determine absolute transition probabilities and have yielded many high-quality results for neutral and singly ionized atoms. For many heavier elements, this approach is the principal source of accurate data.

B. The Complete Emission Technique

The complete emission technique is concerned with the measurement of relative or absolute transition probabilities for numerous lines of an atom or ion. Relative transition probabilities are obtained from measurements of line intensity ratios, with one line arbitrarily chosen as a reference line to obtain a common basis. For this method, the knowledge of the populations of upper states N_k is required, according to Eq. (10). These are not directly measurable, but are related to measurable quantities when conditions of full local thermodynamic equilibrium (LTE), or at least partial LTE for the states under consideration, prevail in the plasma source. In this case, the N_k's can be expressed by Boltzmann factors,

$$N_k = [N_a g_k \mid U_a(T)] \exp(-E_k \mid k_B T), \qquad (12)$$

with T being the plasma temperature, N_a the total number density of species a, $U_a(T)$ the atomic partition function, g_k the statistical weight of state k (related to the total angular momentum number J_k of this state by $g_k = 2J_k + 1$), and k_B the Boltzmann constant. In line intensity ratio measurements within the same species, N_a and $U_a(T)$ cancel, and one obtains from Eqs. (10) and (12), for the relative transition probability of a line x compared with that of the reference line r,

$$A_x = A_r \frac{I_x}{I_r} \frac{\lambda_x}{\lambda_r} \frac{g_{k,r}}{g_{k,x}} \exp\left(\frac{E_{k,x} - E_{k,r}}{k_B T}\right). \qquad (13)$$

This equation is presented on the wavelength (λ) scale, which is generally employed ($\lambda = c\nu^{-1}$, where c is the velocity of light).

Thus, for relative transition probability measurements a temperature determination is required, but since the energy differences $E_{k,x} - E_{k,r}$ are usually not much larger than $k_B T$, the A_x/A_r ratios are insensitive to temperature errors. Often the principal resonance line of a species is chosen as the reference line when lifetime data for its upper state are available and they can be utilized via $\tau_r = A_r^{-1}$ [see Eq. (8)] to put the relative transition probability data on an absolute scale. With such combinations of lifetime and relative emission data, numerous transition probabilities for light elements have been determined efficiently.

Absolute emission measurements are performed in plasmas that are in LTE and thus allow the application of equilibrium relations. Plasmas generated by high-current arcs, plasma torches and shock tubes adhere closely to the LTE criteria. Again, Eq. (10) is applied, with the Boltzmann factor [Eq. (12)] employed for the determination of N_k. Measurements of both the total number density N_a of the investigated species (or the electron density via the equilibrium relations) and the temperature are now required. For the determination of these quantities, a number of well-established spectroscopic and interferometric techniques are available. Nevertheless, accurate absolute emission measurements remain difficult to accomplish. The absolute emission technique has been especially tested on the spectra of Ar I and Ar II, and accuracies of the order of $\pm 10\%$ have been achieved in careful measurements.

IV. ABSORPTION AND "HOOK" TECHNIQUES

The absorption technique is quite similar in concept to the emission method. A schematic setup is shown in Fig. 2b. Radiation from a source emitting a continuum spectrum—a tungsten ribbon lamp, for example—is sent through an absorption tube containing the species to be studied as an atomic vapor. The tube is usually part of an electric furnace and has to be operated at elevated temperatures so that sufficient vapor pressure is generated. In analogy to Eq. (10), one obtains, on the basis of Einstein's quantum concept, for the total energy F_{ik} absorbed in a transition $i \to k$ [see Eq. (2)],

$$F_{ik} = \int_0^\infty F(\nu)\, d\nu = \frac{1}{4\pi} B_{ik} N_i h\nu_{ik} u(\nu) l. \qquad (14)$$

In contrast to the emission method, the energy density of the background radiation field $u(\nu)$ enters, since the frequency of the absorption processes depends on the intensity of the field emitted from the continuum source. [All other symbols are as in Eqs. (2) and (10).] It is customary to work on the wavelength scale λ, where $u(\lambda) = c\lambda^{-2} u(\nu)$, and to measure the line absorption in terms of the incident intensity $= I(\lambda) = (c/4\pi)u(\lambda)$, as

$$W_{ik} = \frac{F_{ik}}{I(\lambda)} = \int_0^\infty \frac{I(\lambda) - I'(\lambda)}{I(\lambda)}\, d\lambda, \qquad (15)$$

where $I'(\lambda)$ is the specific intensity at wavelength λ after passage through the absorbing column, which has to be "optically thin," or without saturation effects (this may be readily tested).

In absorption work, the transition probability for absorption B_{ik} is normally not used; rather, one uses the absorption oscillator strength f_{ik}, which is related to B_{ik} via $B_{ik} = e^2(4\varepsilon_0 m h\nu_{ik})^{-1} f_{ik}$. Substituting this quantity into Eq. (14), one obtains on the wavelength scale,

$$W_{ik} = \left(e^2 / 4\varepsilon_0 m c^2\right) \lambda_{ik}^2 N_i f_{ik} l. \qquad (16)$$

The determination of the oscillator strength thus requires measurement of the total line absorption W_{ik} (sometimes called the equivalent width) and the length l of the absorbing column. Also, N_i must be determined. For an absorption tube in LTE, the Boltzmann factor [Eq. (12)] can

again be applied, so that the determination of N_i reduces to the measurement of the temperature of the absorbing gas and the measurement of the total particle density. Temperature measurements are usually carried out by pyrometric methods, and density measurements make use of known vapor pressure data or are accomplished in an atomic beam setup, with the weighing of the accumulated substance on a microbalance.

Similar to the emission work, many absorption experiments are done on a relative basis (which requires the measurement of a temperature only) and are subsequently converted to an absolute scale by the utilization of available lifetime data. Very precise absorption measurements for several iron group elements have yielded many relative data, with uncertainties in the range from only 0.5 to 1%.

"Hook" measurements are also performed in absorption, but the experimental approach is based on a fundamentally different idea. It utilizes the quantum mechanical formula for the anomalous variation of the index of refraction n in the vicinity of an isolated absorption line, which is linearly dependent on the oscillator strength of the line. In 1912 Rozhdestvenskii developed a method to determine n with a Jamin- or Mach–Zehnder-type interferometer and a thick plane-parallel glass plate placed into the compensating arm of the interferometer. The glass plate produces a very large path difference, which causes a tilting of the interference fringes and the formation of two characteristic hooks symmetric to the center of an absorption line. The distance between the extrema of the hooks, as well as some readily measurable properties of the interferometer, provides the information necessary to determine the product $N_i f_{ik}$. This method, with a limited range of transitions and spectra, was often applied in the past but is rarely utilized nowadays.

V. ATOMIC LIFETIME DETERMINATIONS

The atomic lifetime approach is conceptionally less complex than the other experimental techniques, since the principal measurement—that of a time interval—is accomplished in a direct way. However, transition probabilities may be derived directly from lifetimes only for the special case where the sum in Eq. (8) reduces to a single term, such as for principal resonance lines or for other low-lying excited atomic states where a single transition is the strongly dominating decay channel. Lifetime measurements are, of course, also very valuable for normalizing relative sums of transition probabilities obtained from other methods such as the branching ratio technique.

In the most general case, the rate of depopulation of an excited atomic state k is the result of the following radiative and collisional factors:

$$\frac{dN_k}{dk} = -N_k(t)\sum_i A_{ki} + N_u(t)\sum_u A_{uk} + \text{absorption} + \text{induced emission} + \text{collisions}. \tag{18}$$

The first term represents the spontaneous decay of state k by photon emission, which by itself is the rate required for the interpretation of lifetimes in terms of transition probabilities [this is Eq. (6)]. The other terms distort such measurements and must therefore be reduced to insignificance. The second term represents the repopulation of state k through the spontaneous radiative decay of higher-lying atomic states u, which thus feed electrons into state k as it gets depopulated. This process is known as cascading and applies, of course, only when such states u are initially populated along with state k. The other factors are absorption, induced emission, and inelastic collisions, which may repopulate state k as well as depopulate it. The latter three factors can be rendered negligible by working with atomic vapors at sufficiently low densities. If the ideal condition of spontaneous radiative decay from the selected atomic level is realized, one arrives again at the expression for the radiative lifetime $\tau_k = (\sum_i A_{ki})^{-1}$ [Eq. (8)].

For lifetime measurements, two main approaches are applied.

1. *The laser-induced fluorescence method.* In this method the time interval between the moment of excitation of the atomic valence electron and its subsequent spontaneous radiative decay into a lower atomic state is determined. Atoms are excited in a low-density sample with a pulsed laser which is tuned to the photon energy of a specific transition, so that only a selected atomic energy level gets populated and cascading effects are eliminated. The resulting spontaneous decay, or fluorescence, is observed with a fast photodectector and a transient signal analyzer (see Fig. 3). Large numbers of decay curves may be obtained within minutes with pulsed lasers of moderate repetition rates. Since low-density atomic vapors can be generated from almost any chemical element by sputtering techniques or with electric furnaces, this approach can be widely applied. However, it becomes technically difficult to extend it to multiply ionized atoms.

2. *The beam-laser technique.* Fast ion beams are generated with accelerators of various types, including, for example, Van De Graaffs. To generate ions of various charges, a large range of beam energies, 50 kV to 50 MeV, has been applied. The ion beam is crossed with a tunable cw laser, so that ions or atoms, the latter obtained from charge exchange collisions, are excited into selected atomic states, and the decay fluorescence is observed as a function of distance from the point of laser excitation. The

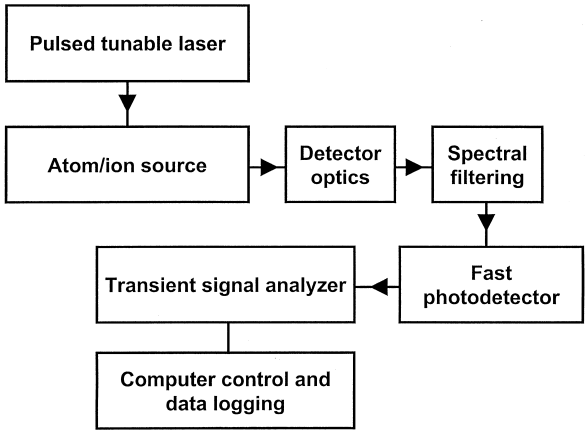


FIGURE 3

atomic beam velocity is accurately determined to convert distance into time. Since this can be readily accomplished and since high-quality decay curves can be generated, lifetimes for resonance lines with uncertainties close to 0.1% have been determined, which constitute the best existing experimental transition probability data. The first technique making use of fast-ion beams is beam-foil spectroscopy. In this approach, the ion beam crosses a thin carbon foil where the ions are stripped of part of their electrons and where some of the remaining electrons are raised into excited states. Thus, a luminous beam emerges from the foil and its decay is recorded. However, the excitation process is indiscriminate, so that radiative cascading from higher levels into the level under study is often pronounced, and this repopulation process lengthens the lifetime. Thus, sophisticated methods have been devised to account for cascading, such as the arbitrarily normalized decay curve (ANDC) technique, which has found many applications.

VI. ATOMIC STRUCTURE THEORY

The principal quantity calculated is the line strength S, as given by Eq. (5). This requires the calculation of the wave functions ψ_i and ψ_k for the lower and upper states of the transition, which must be approximate solutions of the Schroedinger or Dirac equation. For atoms with more than one electron, the Schroedinger equation contains not only a term for the attractive interaction of each electron with the nucleus, but also terms for the repulsive interactions between the electrons, which make an exact solution virtually impossible. These and other many-body aspects represent the major computational challenge of atomic structure calculations.

Various approximate methods of determining transition probabilities, or line strengths, have been developed. Probably the most widely used technique for obtaining approximate wave functions is the self-consistent field method developed by Hartree. This is an iterative variational approach, in which the individual electrons are assumed to move independently in the central field of the nucleus. Thus, one-electron wave functions (orbitals) can be calculated with trial potentials for the fields experienced by each electron due to the assumed charge density distribution representing all other electrons. From the resulting product of one-electron wave functions the overall charge density distribution is compared with the initial one obtained with the trial functions, and this process is repeated until self-consistency is obtained. This basic procedure was improved by Fock, who included terms describing exchange effects between the electrons. Early atomic structure calculations, before the advent of powerful computers, proved to be quite cumbersome for many-electron atoms and thus were limited mostly to single-configuration computations. Also, simplified methods known as Hartree–Fock–Slater, Thomas–Fermi, and Thomas–Fermi–Dirac approximations were then developed. Theorists have also pursued semiempirical approaches utilizing numerous, precisely known experimental energies for the excited atomic states. The coulomb approximation developed by Bates and Damgaard in 1949 is probably the best-known semiempirical method and has produced many useful results, especially for light atoms.

In the single-configuration approach, it is assumed that a discrete state of an atom is well described by a definite electron configuration, with each of the atomic electrons described by a one-electron wave function (orbital). Each electron is assumed to move independently in an average field produced by the combination of the nucleus and the other electrons, and the total wave function for the atomic state is approximated by the product of these one-electron wave functions. This model works best when the jumping electron is well removed from the other (core) electrons and when there is little interaction between them, which is typical for at least moderately or highly excited states, especially for atoms and ions which are alkali-like, with one valence electron outside closed shells.

However, comparisons with experimental results showed early on that the single-configuration approach is inadequate for most atomic transitions, especially those involving states where the electron orbits overlap strongly, such as for $2s$ and $2p$ electrons (shell-equivalent electrons). Therefore, sophisticated multiconfigurational approaches have been developed that treat these mutual interactions between the electrons—the "electron correlation"—by expanding the total many-electron wave function to include contributions from the principal interacting configurations. For example, the ground state of

carbon, which involves the above-cited case of $2s$ and $2p$ electrons, is expected to have the character of a mixture of the $2s^2 2p^2$ and $2p^4$ states plus admixtures of other states of the same parity and angular momentum, such as $2s^2 3p^2$ and $2s^2 3s 3d$. These multiconfiguration approximations, introduced in practical ways with the increasing use of computers in the early 1960s, have cleared up many former discrepancies between experiment and theory. They are routinely applied to transitions between low-lying atomic levels and are known as the multiconfiguration Hartree–Fock (MCHF) [sometimes referred to as superposition-of-configurations (SOC)] and configuration interaction (CI) approaches. For the lighter elements, many comparisons with experimental data indicate that these multiconfigurational approaches—some include as many as several thousand interacting configurations—are generally accurate to within 10%, sometimes 1%, for the stronger transitions. Fine-structure transitions are increasingly calculated in intermediate coupling, with relativistic terms usually included via the Breit–Pauli approximation.

In highly stripped ions, the remaining atomic electrons are strongly bound to the nucleus. Under the influence of this central field, they acquire relativistic speeds, with resulting orbit changes and other effects, so that fully relativistic treatments must be applied. With the increasing interest in highly charged ions due to applications in fusion research, X-ray astronomy, and vacuum ultraviolet and X-ray laser development, relativistic atomic structure treatments have been developed, notably the relativistic MCHF (or Dirac–Fock) and the relativistic random phase (RRPA) approximations.

The assessment of the accuracy of theoretical data is a difficult problem, and usually no uncertainty estimates are given for calculated values. Methods exist, however, for the establishment of rigorous upper and lower error bounds, and these schemes have been successfully applied to some transitions of the helium, lithium, and beryllium sequences.

VII. REGULARITIEXS IN OSCILLATOR STRENGTHS: *f*-SUM RULES

It has been shown by perturbation theory that the oscillator strength or f value for a given transition along an isoelectronic sequence scales approximately with the inverse nuclear charge Z as

$$f = f_0 + f_1/Z + f_2/Z^2 + \cdots. \tag{19}$$

The coefficient f_0 is a hydrogenic f value and specifically becomes zero for transitions where the principal quantum number n remains unchanged ($\Delta n = 0$), because for hydrogenic ions all atomic states with the same n possess the same energy, so that no transitions can take place (hydrogen energy degeneracy). For large values of Z, n tends to approach this hydrogenic value until relativistic effects take over and fundamentally alter the trend.

Therefore, in cases where experimental and theoretical data are available for some key ions along an isoelectronic sequence, numerical data for other ions can be obtained by graphic interpolation or extrapolation. Some of these trends are not smooth but show a pronounced irregularity somewhere along a sequence. Such irregularities are of considerable interest insofar as they always indicate some kind of an interference effect in the atomic structure. Most likely, they are due to the crossing of one of the levels involved in the transition with another level of the same parity and angular momentum. Since such level crossings are often more firmly established than f-value data, they provide a signal that extra care is necessary for the determination of f-value data near such points.

Another regular behavior of f values occurs for the sequence of lines within a spectral series of an atom or ion. A spectral series is a group of lines with a common lower level and a sequence of upper levels for which the principal quantum number n increases successively in steps of 1 while all other quantum numbers remain constant. For the various series of hydrogen, known as the Lyman, Balmer, Paschen, Brackett, and higher series, it has been shown that

$$f = Cn^{-3}, \tag{20}$$

where the constant C is specific to each series. For other atoms a similar behavior is expected since the higher states become increasingly hydrogen-like, but n must be replaced by an "effective" quantum number n^*, derived from the respective excitation energy E by

$$n^* = (Z - N + 1)Rhc/(I - E), \tag{21}$$

where N is the total number of electrons, R the Rydberg constant, and I the ionization potential. The difference $n - n^*$, which gradually decreases for large n and large angular momentum, is called the quantum defect.

With the interpretation of the f value as the fraction of optically effective electrons (per atom) available to participate in a particular transition, it is expected that the sum over all possible transitions from a given state (including transitions to the continuum) should be equal to the number of valence electrons. The Thomas–Reiche–Kuhn sum rule states exactly that.

Of more practical value are partial f-sum rules, especially one derived by Wigner and Kirkwood for various spectral series of one-electron systems. This rule states that the f sum is 1 for an s–np spectral series, 10/9 for a p–nd spectral series, 7/5 for d–nf, and so on.

VIII. DATA AVAILABILITY

The Data Center on Atomic Transition Probabilities at the U.S. National Institute of Standards and Technology (NIST), formerly the National Bureau of Standards (NBS), has critically evaluated and compiled atomic transition probability data since 1962 and has published tables containing data for about 39,000 transitions of the 28 lightest elements, hydrogen through nickel. In these tables, all stages of ionization are covered for which data are available, and forbidden transitions are included. Additional data for many heavy elements are critically compiled in the *Handbook of Chemistry and Physics* (CRC Press). The compilations for elements $Z = 1 - 28$ contain accuracy estimates for each transition probability listed and are generally limited to "reference" data, i.e., data estimated to have uncertainties of less than 50%. This material is also part of a large spectroscopic database on the Internet, the NIST Atomic Spectra Database at the URL address http://physics.nist.gov/asd. This address also contains a link to a comprehensive bibliographic database on atomic transition probabilities from 1914 to the present, maintained by NIST.

SEE ALSO THE FOLLOWING ARTICLES

ATOMIC PHYSICS • ATOMIC SPECTROMETRY • QUANTUM OPTICS • QUANTUM THEORY • TRANSMISSION ELECTRON MICROSCOPY

BIBLIOGRAPHY

Fuhr, J. R., and Wiese, W. L. (h). NIST atomic transition probability tables. In "Handbook of Chemistry and Physics," 71st ed., CRC Press, Boca Raton, FL.

Fuhr, J. R., Martin, G. A., and Wiese, W. L. (1988). "Atomic transition probabilities—Scandium through manganese, and iron through nickel," J. Phys. Chem. Ref. Data **17,** Suppls. 3 and 4.

Huber, M. C. E., and Sandemann, R. J. (1986). "The measurement of oscillator strengths," *Rep. Progr. Phys.* **49,** *397*.

Thorne, A., Litzén, U., and Johansson, S. (1999). "Spectrophysics, Principles and Applications," Springer, Berlin.

Wiese, W. L., Fuhr, J. R., and Deters, T. M. (1996). "Atomic transition probabilities of carbon, nitrogen and oxygen," J. Phys. Chem. Ref. Data, Monogr. 7.

Wiese, W. L., Smith, M. W., and Glennon, B. M. (h). "Atomic Transition Probabilities," Vols. I and II, U.S. Government Printing Office, Washington, DC.



X-Ray Small-Angle Scattering

O. Kratky
P. Laggner
Austrian Academy of Sciences

I. Introduction and Survey
II. Particle Scattering in Monodisperse Solutions of Corpuscular Structures
III. Long-Chain Molecules in Solution
IV. Polymers in the Solid State
V. Inorganic Substances—Materials Science
VI. Instrumentation
VII. Recent Developments

GLOSSARY

Absolute intensity Ratio of scattered intensity and primary intensity.

Average intersection length In two-phase systems, average over all possible intersection (chord) lengths within one phase.

Cross-section factor Scattering intensity times scattering angle; characteristic for size and shape of rodlike particles.

Distance distribution function Number distribution of distances between any two electrons within a particle; obtained by Fourier transformation of the scattering curve.

Invariant Integral small-angle scattering intensity, dependent only on the mean-square electron density fluctuation, that is, the scattering power.

Particle scattering Scattering curve of a single particle averaged over all orientations, experimentally found in dilute solution.

Particle weight per unit length or unit area On rodlike and lamellar particles, respectively, determined through the absolute intensity of the cross-section (or thickness) factor at zero angle.

Persistence length Average length measured along a coiled polymer chain (wormlike chain) in which the director decays to $1/e$.

Porod slope Decay exponent of the scattering intensity curve toward large angles; depending on the dimensionality of the scattering object; directly related to the fractal dimension.

Radius of gyration Root-mean-square distance of electrons from center of gravity; analogous to radius of inertia in mechanics.

Thickness factor Scattering intensity times the square of scattering angles; depending on thickness of lamellar particles and electron density profile vertical to the lamellar plane.

X-RAY SMALL-ANGLE SCATTERING is a method to study the structure of colloid and macromolecular systems. A sharp and well-collimated X-ray beam is made to penetrate the sample under investigation, and the angular dependence of the intensity of the radiation scattered to very small angles (i.e., a few minutes of arc to several degrees) is recorded (scattering curve). A well-developed theory permits the derivation of numerous geometrical and structural parameters from this scattering curve.

1. The most general parameters, obtainable from any sample, are (a) the mean square fluctuation of the electron density; (b) in the case of a two-phase system, (i) the volume fraction, (ii) the inner surface, and (iii) the intersection length; and (c) generalized relationships in the case of multiphase systems; in the case of fractal systems, the mass and surface fractal dimensions, respectively.

2. For corpuscular macromolecular particles in monodisperse solution, X-ray and neutron small-angle scattering yields numerous geometrical and mass parameters; additional information is obtainable by variation of the electron density contrast between solvent and dissolved particles as well as by heavy-atom labeling.

3. The use of high-intensity synchrotron radiation permits time-resolved X-ray measurements, which makes possible the observation of biological processes on a medium time scale.

4. Statistical shape parameters can be obtained from solution of long-chain molecules, such as the average degree of coiling and several mass parameters, for example, mass per unit length or total mass. Moreover, the use of markers frequently extends the obtainable information and allows observations on polymers in the solid state.

5. Additional ways of analysis of natural solid high polymers are discussed for the example of cellulose. For air-swollen, regenerated cellulose, the cross-section analysis leads to information about the elongated supermolecular bundles called micelles. X-ray measurements on stretched and rolled cellulose films yield similar results.

6. Synthetic high polymers in the solid state can mainly be analyzed on the basis of the lamellar stack model.

7. Owing to the sample diversity of inorganic systems, a simple, systematic, and general treatment here is not possible. Successful applications have been obtained in, among others, the field of physical metallurgy, in particular on a special type of cluster, the Guinier–Preston zones. Other applications include the analysis of catalysts, glasses, and ceramics as well as density fluctuations at the critical point in liquids.

8. The problems of the instrumentation start with the selection of the anode material best suited for the type of sample to be studied, the shape of the focus, and the positioning of the camera in front of the X-ray source. In the future, particular emphasis will undoubtedly be placed on high-intensity rotating-anode X-ray tubes and on synchrotron sources. The next step is the selection of the type of camera to be used. Among the three-slit cameras, the Beeman camera has won widespread acceptance. A way to reduce parasitic scattering to a neglectable minimum is the use of the block collimation system. The highest resolution can be obtained with the Bonse–Hart camera, which uses an ideally parallel primary beam by multiple reflection.

Other topics discussed are different techniques to obtain a monochromatic primary beam, methods for detecting the scattered radiation, the measurement of the "absolute intensity" (i.e., the ratio of scattered intensity and primary intensity), and the elimination of errors introduced by intensity fluctuations.

I. INTRODUCTION AND SURVEY

A. The Phenomenon and its Physical Basis

The first observations of diffuse X-ray scattering at small angles, dating back to the early 1930s, were made on solid fibers and colloidal powders. In the meantime small-angle X-ray scattering (SAXS) developed to a powerful analytical method for the investigation of submicroscopic solid or liquid structures. It is bound to the existence of inhomogeneities in electron density that are much larger in size, typically between 10 Å (1 nm) and 10^4 Å (1000 nm), than the wavelength of the X-rays (normally about 1 Å). This makes it generally applicable to systems of colloidal dimensions and dispersity.

The physical basis of the scattering process is identical to that of classical X-ray crystal diffraction. It is the consequence of the elastic interaction between the electric vector of an X-ray wave with electrons. The oscillating field causes the electrons to oscillate with the same phase and frequency, and the accelerated charges become sources of secondary waves, in the purely classical, electrodynamic sense. The scattering intensity I_e of the single electron is given quantitatively by the Thomson theory as follows:

$$I_e(2\theta) = P_0(7.9 \times 10^{-26})\left(\frac{1}{a^2}\right)\frac{1+\cos^2 2\theta}{2}, \qquad (1)$$

where P_0 is the intensity (i.e., the energy passing through 1 cm^2/sec), 2θ is the scattering angle (angle between the primary and diffracted beam), and a is the distance between the electron and the point of observation. The numerical constant is the effective scattering cross section of the electron (in square centimeters). The polarization factor $(1+\cos^2 2\theta)/2$ is effectively equal to unity for small angles 2θ and can therefore be neglected. For a system of many electrons, the amplitudes of the secondary waves

are added with their phase differences in each direction; the scattering intensities are the absolute squares of the amplitude sums.

The plot of the relative intensities versus the scattering angle 2θ is generally called the scattering curve. For the comparison with theoretical scattering curves, it has proven useful to use the following argument:

$$h = (4\pi \sin\theta)/\lambda \quad \text{Å}^{-1}, \tag{2}$$

where θ is half of the scattering angle; by this notation, the wavelength λ is eliminated as a variable. With the same physical meaning as h, the symbols Q or κ can be found frequently in the neutron scattering field. In some parts of the literature the term $s = (2 \sin\theta)/\lambda)$ is used as an angular argument; however, we prefer h in this review as the most widespread one.

B. The Experimental Principle

In most cases, the X-ray source is a sealed X-ray tube with copper anode. The emitted radiation is not immediately usable for the experiment because it is both divergent and polychromatic. A collimator has to be used in order to provide a well-defined, narrow primary beam with an either point- or line-shaped cross section, and a monochromator can be used to select a discrete wavelength: alternatively, the effect or polychromasy can be eliminated numerically in the evaluation of the scattering pattern if the spectral distribution of the primary beam is known.

A partial monochromatization that is sufficient for many purposes can be achieved by the use of filters, such as a 10-μm-thick nickel filter for copper radiation. The specimen is located at the exit of the collimator, and the scattered radiation is monitored at some distance a by a suitable detection device, that is, photographic film or an electronic counter. In the measurement of scattering intensity, both the shape of the scattering curve (i.e., the angular distribution of the scattering intensity on a relative scale) as well as the quantitative comparison to the intensity of the primary beam are of interest. The latter is referred to as absolute intensity measurement.

The aspects of instrumentation are treated in more detail in Section VI.

C. General Types of Scattering Patterns

From the phenomenological point of view, one has to distinguish between isotropic and anisotropic scattering patterns. In the former case the scattering intensities from a primary beam with a point-shaped cross section are evenly distributed over the envelopes of cones with opening angles 4θ; that is, the scattering pattern in the plane of registration is circularly symmetric around the intersection

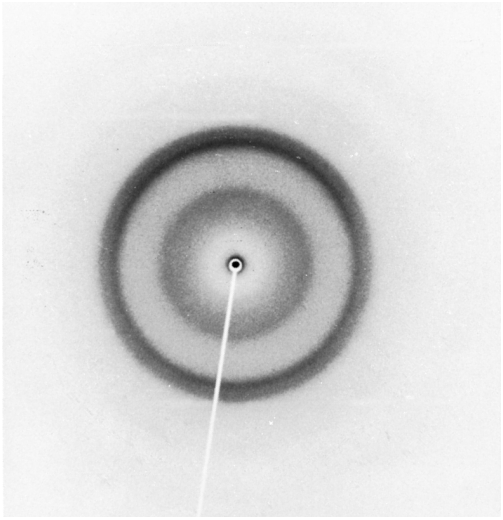


FIGURE 1 X-ray scattering diagram of unoriented rayon fibers.

point of the primary beam with the film; this indicates completely random orientation of the inhomogeneities within the sample. An example is given by the scattering pattern of unoriented regenerated cellulose (Fig. 1). Isotropic scattering is generally also observed with dilute macromolecular solutions. In cases where anisotropy of the small-angle scattering patterns is observed, this indicates some preferred orientation within the specimen is it occurs frequently in natural or synthetic polymers. An example is shown in Fig. 2 by the scattering of oriented rayon fibers; this has been attributed to the existence of microvoids. According to the general law of reciprocity, that the size of an object and its scattering angles are inversely correlated, the pattern in Fig. 2 indicates that these voids are elongated in the direction of the fiber axis.

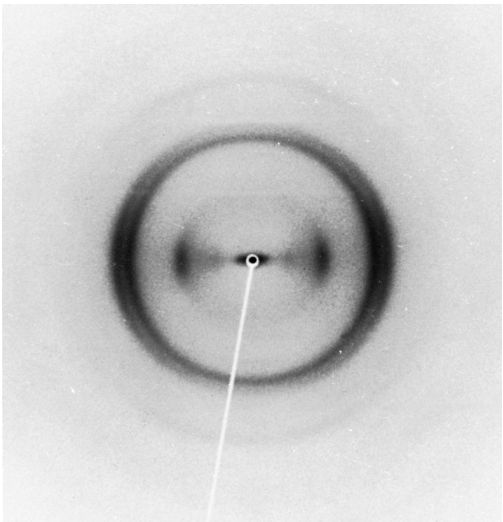


FIGURE 2 X-ray diagram of oriented rayon fibers.

D. The Concept of Scattering Equivalence

A diffuse small-angle scattering pattern, in particular one of the isotropic type, from a macroscopic object containing colloidal inhomogeneities is never subject to an unambiguous structural interpretation. Any particular structural model derived from this method can only be consistent with the observation, but need not be unique in the sense that it is the only one to produce a given scattering pattern. From this fact, models that fit the observations are generally referred to as equivalent in scattering.

The following ambiguities are always present, if not ruled out by ancillary information from other sources:

1. The Babinet principle. It is not possible to distinguish by its scattering a system of particles surrounded by empty space from a complementary one consisting of voids in a continuum of matter, voids and particles being of the same structure (Fig. 3).
2. In particulate systems it is not possible unconditionally to distinguish between the scattering arising from a certain defined particle shape and one arising from polydispersity. Shape determination becomes possible only with particles of uniform size. On the other hand, a certain size distribution may be derived only if all particles have the same shape but differ in their absolute dimensions. The problem of structure analysis becomes unsolvable if both shape variation and size distribution are present.

E. The Limiting Cases of Dilute and Densely Packed Systems

In a discussion of the structures that give rise to small-angle scattering, it is appropriate to distinguish between *dilute* and *dense* systems. By convention, a dilute system is one consisting of separate domains (particles) that do not interfere with each other in their scattering. A system that cannot be brought to this state is a dense system.

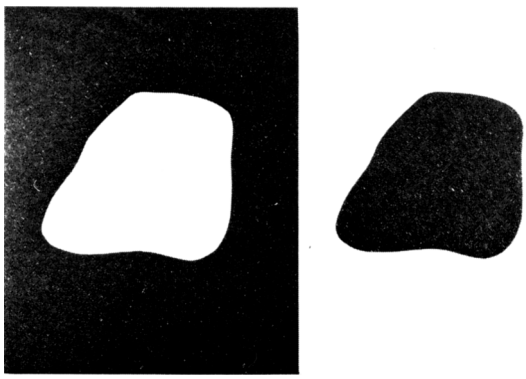


FIGURE 3 The Babinet principle holds that the scattering of the two complementary objects—particle in vacuum and hole within dense matter—give the same scattering pattern.

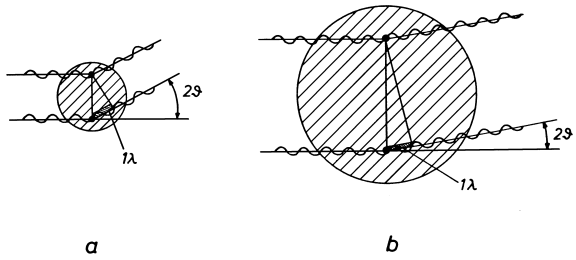


FIGURE 4 Scattering for two points of a wave front on (a) small and (b) large particles.

This differentiation is reflected by the historical development of the quantitative theoretical foundations of the method, which from the very beginning followed along two separate lines. Guinier started from the scattering of the single particle and has compared it to Fraunhofer scattering of visible light on a turbid droplet. One can compare it also to the scattering of X-rays on a single atom, that is, with the atomic scattering factor, with the only difference that the scattering from particles is confined to a much smaller angular range, as can be easily understood from the following consideration.

Take two points within the sphere depicted in Fig. 4a and consider the scattering angle for which the path difference of the secondary waves amounts to 1λ; generalizing this idea to the secondary waves from all points (electrons) of the sphere shows that superposition of waves with random phases will lead to a practically complete anihilation. With decreasing scattering angles, however, the phase differences decrease; the waves start to reinforce each other; and at zero scattering angle, finally, all secondary waves are precisely in phase, where consequently the maximum of the scattering curve is observed: A scattering curve of the general type of curve 1 in Fig. 5 is the result. For the same wavelength and a much larger sphere (Fig. 4b) the path difference of 1λ will be reached at much smaller scattering angles, and hence the entire scattering curve will be confined to a narrower angular range (curve 4 in Fig. 5). In the same proportion as the particles grow larger, the scattering condenses to progressively narrow angular ranges. For example, the scattering curves 1, 2, 3, and 4 in Fig. 5 correspond to particles with linear dimensions in the ratio 1:2:3:4. Finally, if the particles are gigantic as compared to the wavelength, X-ray small-angle scattering will be observed. This reflects the general principle of reciprocity in optics between the dimension of scattering objects and scattering angles.

In a monodisperse dilute solution, in which the mean distances between the particles are large as compared to

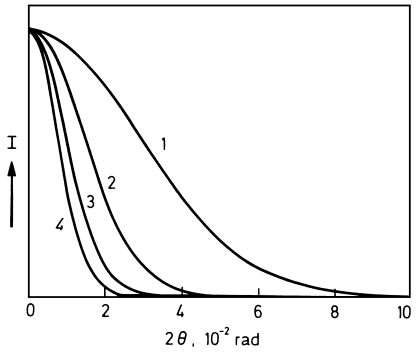


FIGURE 5 Scattering curves from object with the same shape and different sizes.

their size and randomly distributed, the scattering intensities from the single particles are additive. At the same time, the random disorder in a dilute system leads to an average over all spatial arrangements and the result is a scattering curve that is generally referred to as particle scattering. It is characteristic for the behavior of the single particle. Errors arising from interparticular interferences can be eliminated by preparation of a concentration series and extrapolation to zero concentration.

In the course of the development, methods have been found to evaluate about a dozen particle parameters unambiguously from the scattering curve. Moreover, it is possible through trial-and-error methods to perform a more detailed structure analysis. Finally, by introducing electron density markers, distances between the positions of these markers within the particle can be measured.

In the case of polydisperse, dilute solutions, particle parameters can also be determined; however, they represent mean values. Apart from macromolecular solutions, polydisperse dilute systems are frequently found in glasses and alloys. If an approximate shape can be assumed—in the cases mentioned above the particles frequently do not depart very far from isotropic shape—the size distribution can be calculated.

In densely packed systems—for instance, in solid cellulose—the concepts just outlined for dilute solutions of particles cannot be unconditionally applied, since the basic tenet, that the scattering intensities for individual particles simply add to give the scattering of the whole system, no longer holds. Here the entire system has to be in the foreground of the considerations. This is due to the fact that the distances between individual zones of similar electron density are small as compared to their sizes, so that interference effects *between* these zones become dominant over the interferences from *within* the zones. The general treatment of dense systems, therefore, has to consider the whole scattering sample as one huge particle for which only statistical system parameters can be evaluated.

In the early times, workers in this field also performed experiments on densely packed systems and evaluated them according to the theory of dilute systems and thus neglected the possibilities of interference interactions. This has been recognized as incorrect.

However, there are also cases with densely packed systems in which interference effects can be neglected. This occurs if the system contains a sufficiently broad distribution of particle sizes. This approach becomes impractical, however, with systems of very high density, such as in solid high polymers, where the space occupancy is between 99 and 100%. Then it is much simpler to treat the voids as subjects of evaluation, in the sense of Babinet's principle, that are obviously present as dilute systems.

Relatively highly concentrated systems (up to 20% by volume) of very anisotropic particles can also be treated as dilute systems according to the principles of particle scattering, since with elongated or flat particles the interference effects are strongly reduced if the particles are not arranged in parallel. However, it is just this anisotropy that, with very high packing density, enforces a largely parallel arrangement, so that this approach is again bound to fail. Nevertheless, it was possible to treat highly air-swollen cellulose at degrees of swelling of about 6 (corresponding to space filling of about 16%) satisfactorily as a dilute system. If, however, the density becomes excessive, one has to work with idealistic models (such as the "lamellar stack," which has been widely employed in studies on high-polymer synthetics) if an analysis in terms of voids is not successfully applicable. This is treated in more detail in Section IV.

F. The Most General Unconditional Treatment

This applies to systems in which no geometrical regularities can be recognized and not even single particles can be defined. A good model is a spongelike structure where the two phases are mutually interpenetrating in an irregular fashion without the existence of discrete particles of the one or other phase. In this case typical system parameters can be evaluated that are characteristic for the underlying structure. These are treated below.

1. Mean Square Fluctuation of the Electron Density

The relevant general quantify of a system that determines its integral scattering power in a small-angle experiment is its mean square electron density fluctuation:

$$\overline{(\Delta\rho)^2} = \langle(\rho - \rho_{av})^2\rangle, \tag{3}$$

where ρ_{av} is the average electron density over the entire irradiated system and ρ is the local electron density at

a given point of it. Irrespective of the internal structural features of the system, the integral over the scattering intensities $I(h)$ taken over all values of h is directly proportional to the mean square fluctuation $\overline{(\Delta\rho)^2}$. The following integral:

$$Q = \int_0^\infty I(h)h^2\,dh \tag{4}$$

is therefore called the "invariant," since an internal redistribution of the electron densities might well lead to changes in the shape of the scattering pattern but the integral remains constant. The proportionality of the invariant to the mean square electron density fluctuation is given by the scattering intensity of the single electron I_e [see Eq. (1)] and by the thickness t of the sample according to

$$Q = \text{const}\,\overline{(\Delta\rho)^2}, \quad \text{const} = 2\pi^2 I_e\, t. \tag{5}$$

For this reason $\overline{(\Delta\rho)^2}$ is also frequently called the "scattering power" of a system.

2. The Volume Fractions of Multiphase Systems

Equations (4) and (5) become particularly useful for the case of two-phase systems, where $\overline{(\Delta\rho)^2}$ is related to the volume fractions w_1 and w_2 $(= 1 - w_1)$ of the two phases, differing in electron density by $\Delta\rho$, according to the following:

$$\left\langle \left(\rho^2 - \rho_{av}^2\right)^2 \right\rangle = (\Delta\rho)^2 w_1 w_2 \equiv (\Delta\rho)^2 w_1(1 - w_1). \tag{6}$$

Combining Eqs. (1), (5), and (6), we get

$$Q/P_0 = \text{const}(t/a^2)(\Delta\rho)^2 w_1(1 - w_1) \tag{7}$$

From the integral over the scattering intensities one obtains therefore the volume fractions of the two phases. Equation (7) can easily be extended to the case of a system with three or more phase; for three phases it reads:

$$Q/P_0 = \text{const}(t/a^2)\left[(\Delta\rho_{1,2})^2 w_1 w_2 + (\Delta\rho_{2,3})^2 \times w_2 w_3 + (\Delta\rho_{1,3})^2 w_1 w_3\right], \tag{8}$$

where

$$w_1 + w_2 + w_3 = 1.$$

The equation can be solved if the $\Delta\rho$ values and one of the volume fractions are known from other sources.

3. Inner Surface

For two-phase systems, the specific phase boundary area (inner surface per unit volume inverse angstroms) can be obtained from the final slope toward large angles 2θ, which follows an h^{-4} course:

$$I(h) \to K/h^4; \qquad \lim_{h\to\infty} I(h)h^4 = K. \tag{9}$$

This so-called "Porod's law" relates in this form to curves unaffected by collimation influences. For curves measured with a line-shaped primary beam [denoted as $\tilde{I}(h)$ by the superscript tilde], the relationship takes the following form:

$$I(h) \to \tilde{K}/h^3; \qquad \lim_{h\to\infty} \tilde{I}(h)h^3 = \tilde{K} \tag{10}$$

The terms K and $\tilde{K}$, respectively, are called the "tail-end constants."

For two-phase systems, the constant K is related to the inner surface, that is, to the interfacial area between the two phases. Defining S_i as the specific inner surface of a monodisperse solution—as the most important special case—given by the ratio of surface and volume of the particle, we have the following:

$$S_i = \pi K/Q = 4\tilde{K}/\tilde{Q}. \tag{11}$$

4. Intersection Length

The degree of colloidal dispersion is characterized in a particularly transparent way by the so-called "intersection length," also introduced by Porod. Figure 6 schematically shows an irregular two-phase gel system that can be considered as "spongelike" and that does not allow the definition of an average particle size. If, however, arrows

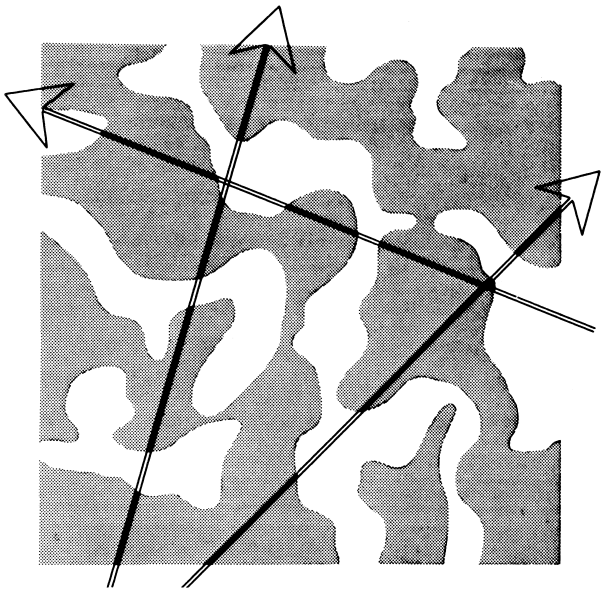


FIGURE 6 Schematic section through a spongelike structure with representation of the intersection length. [From Porod, G. (1951). *Kolloid-Z.* **124,** 83.]

are shot through this system in all possible directions, and an average is taken over all lengths passing through a single phase, one indeed obtains a characteristic parameter for the dispersity of the two phases. In this case, the simple relationship holds for the average values $\bar{l}_1$ and $\bar{l}_2$

$$\bar{l}_1 = 4w_1/S_i; \qquad \bar{l}_2 = 4w_2/S_i. \tag{12}$$

These parameters have particular relevance for solid colloidal systems.

Frequently these system parameters are also useful in the characterization of particles in addition to the more specific particle parameters. The only limit to this approach is the applicability of the definition of a two-phase system with well-defined interfaces. If the system, however, shows a significant deviation in its scattering from the ideal h^{-4} course, which can be easily determined from the slope in a double-logarithmic plot ($\log I$ versus $\log h$), the concepts of specific inner surface or intersection lengths are not applicable. Then, the system can be discussed in terms of fractal geometry (Section VII.A).

II. PARTICLE SCATTERING IN MONODISPERSE SOLUTIONS OF CORPUSCULAR STRUCTURES

In this section the principles of particle scattering and the various ways by which structural information can be obtained are described. Most of the examples used here are from the field of biological macromolecules. Therefore, no separate section is devoted to this area of research, as is the case for other fields to be treated in Sections III, IV, and V.

A. From the Object to the Scattering Curve

1. Electron Density, Contrast

In the discussion above of particle scattering it was tacitly assumed that the particles are surrounded by vacuum. Generally, however, particles exist in solution, and therefore we must answer whether and how the solvent modifies the scattering of a colloidal particle. For this it is necessary to introduce the concept of electron density ρ, which can be expressed by the number of electron moles per cubic centimeter. For homogeneous matter it is obvious that

$$\rho = \frac{\sum \text{ order numbers}}{\sum \text{ atomic weights}}\, d, \tag{13}$$

where d = density. The summation can be extended over any arbitrary quantity, most conveniently 1 mol of the substance.

The effect of the solvent is illustrated by Fig. 7, which shows on the ordinates the electron densities of the solvent (ρ_1) and of the dissolved particle (ρ_2) along a section

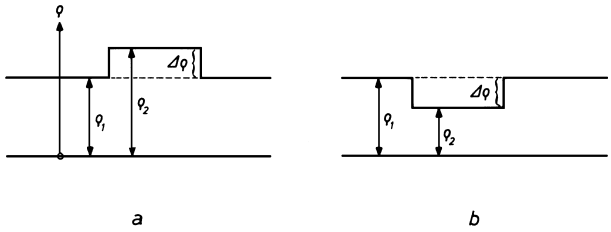


FIGURE 7 Electron density and contrast. For explanation, see text.

through the solvent and the dissolved particle. For the present argument, the electron density within each phase can be considered as constant, since the spatial range of the fluctuations due to the atomic structure is small as compared to the size of the colloidal dimensions. Consequently, the relevant quantity for the scattering of a particle, apart from its volume, is the electron density contrast, $(\Delta\rho) = \rho_2 - \rho_1$, against the solvent background. The remaining electrons of the particle, together with the solvent, form a continuum with no contribution to the scattering. According to Babinet's principle, the scattering effect is unaffected by an interchange of the two electron densities. From this it follows that the arrangement in Fig. 7a is equivalent in scattering to that in Fig. 7b. If the electron densities ρ_2 of the solute and ρ_1 of the solvent are equal, the whole system acts as a homogeneous continuum from which no small-angle scattering arises; this agrees with experience. In this case the X-ray beam does not "see" the particle.

According to this notion, the effect of the solvent on dissolved particles generally is a strong attenuation of their scattering; for example, the contrast $\Delta\rho$ of proteins in aqueous solution is only about 10% of ρ_2.

2. Quantitative Calculation of Particle Scattering

Consider a dumbbell-shaped particle consisting of two point-shaped scattering centers with a mutual distance r: according to Debye, the scattering intensity of such a particle, after averaging over all spatial orientations, is given by the following:

$$i(h) = f_i^2 + f_k^2 + 2f_i f_k\, (\sin hr)/hr, \tag{14}$$

where the f terms are the atomic shape factors and r the distance between the atomic centers. In the calculation of the scattering properties of a large particle, one can assume homogeneous electron density if the ranges of inhomogeneities within the particle are small in comparison to the overall particle dimensions. Consequently, the distance distribution function $\rho(r)$ for the number of distances r between two different volume elements within a particle is a continuous (fully differentiable and integrable) function

of r. Thus, the frequency of distances lying between r and $(r + dr)$ is given by $p(r)\,dr$. According to Eq. (14) for the scattering of the dumbbell, integration over all distances immediately leads to the total scattering curve of a particle:

$$I(h) = 4\pi \int_0^\infty p(r) \frac{\sin hr}{hr}\, dr. \qquad (15)$$

For simple geometric shapes, the distance distribution $p(r)$ can be calculated without great difficulty.

Alternatively, one can start from a combination of the electrons to atoms and obtain, again according to Debye, the relation between the structure of a particle consisting of n atoms and its scattering curve, according to the following equation, which is equivalent to Eq. (15):

$$I(h) = \sum_{i=1}^{n} \sum_{k=1}^{n} f_i f_k \frac{\sin hr_{ik}}{hr_{ik}}. \qquad (16)$$

Hence each atom has to be combined with all the others and with itself, and the summation performed over all pairs.

There exist also other possibilities for the calculation of scattering curves. The scattering curves for many triaxial bodies (ellipsoids, elliptic cylinders, hollow spheres, hollow cylinders, prisms, etc.) with various axial ratios are documented in the literature.

B. From Particle Scattering to the Particle Structure

1. Overview

In the previous section, an outline of the methods for the calculation of scattering curves for a given particle structures has been presented. However, the aim of a small-angle scattering analysis is exactly the opposite: generally one wants to obtain information on the scattering object from the scattering curve. It is possible to formulate relations between certain properties of the particle scattering curve and discrete parameters of the particles, such as mass, volume, and size parameters. Frequently, the determination of only one or a few particle parameters is in itself already an important achievement.

The next step is an attempt to obtain more detailed information on the particle shape. For this, a most useful approach is offered by the possibility of calculating the scattering curves for *assumed particle structures* on the basis of the theory indicated above. It is of advantage if a rough idea on the type of structure to be expected can be obtained from entirely different sources. From a first model scattering curve one further proceeds by trial-and-error variation of the assumed particles until scattering equivalence is reached that is, until the calculated scattering curve agrees with the experimental one. The model found in this way will generally be a good approximation to reality.

So far, structures have been discussed in which the assumption of homogeneous electron density is justified. However, with longer range electron density inhomogeneities within the particle, the above approach can obviously not lead to a correct result irrespective of scattering equivalence between model and experiment. This leads to the additional task of determining the internal electron density distribution, for which there exist three possibilities:

1. The first approach is tied to the *existence of symmetries*. With spherical symmetry, the radial electron density distribution can be calculated. For elongated, rotationally symmetric particles, one can calculate the radial density distribution of the cross section. For lamellar particles with a mirror plane, finally, the electron density distribution vertical to the lamellar plane can be determined. If, however, no such symmetries are present, information on the density distribution can still be obtained by certain modifications of the scattering system. The following two approaches can be taken.

2. Contrast variation by adjustment of the solvent electron density: in principle, this offers the possibility of alternatively making certain domains of the particle invisible and only investigating the rest, which is not matched by the solvent. For the theoretical formalism, see Section II.B.5.

3. Labeling of the dissolved substance. With X-rays one can use heavy atoms or groups of atoms as labels attached to certain domains of the particle and determine their mutual distances (Section II.B.5); with neutron scattering on complex particles, that is, particles consisting of several chemically distinct units, the scattering power of single subunits can be enhanced arbitrarily by selective deuteration. For instance, one can deuterate two subunits and measure their distance; the same can be achieved by deuterating the whole particle and leaving the two subunits of interest unlabeled. This approach can be extended to other pairs of subunits, which leads, in the sense of a "triangulation," to a map of center-to-center distances for all subunits.

2. Prerequisites; Sources of Errors; Corrections

a. Homogeneity, polydispersity. As a rule, the evaluation of scattering data starts from the assumption of monodispersity of the solution. Only in this case can the scattering be considered representative for a certain particle species. This requires a test of monodispersity.

b. Concentration effect. The correct evaluation of particle scattering is possible only under the condition of

infinite dilution. This is due to interferences between different particles, interparticle interferences, which arise if their mutual distances are neither completely irregular nor sufficiently large. Typically, this effect leads to a decrease in scattering intensity at the smallest angles. In general, therefore, it is necessary to measure a series of concentrations and extrapolate to zero concentration.

c. Effect of beam geometry and its elimination. The theoretical scattering curves are derived for the condition of a point-shaped primary beam cross section. In actual experiments, however, one normally uses a flat, ribbon-shaped primary beam to achieve higher intensity, so that in the plane of registration the cross section takes the shape of a narrow rectangle. Obviously, both the length and the width of the primary beam affect the shape of the measured scattering curve. This effect has to be eliminated in order to allow a correct comparison to the theoretical curves. A number of methods have been developed for this correction, often called "desmearing," and they are easily executed by computers.

3. Unambiguously Determinable Parameters and Structure Functions of the Particle in Solution

a. Radii of gyration.

i. Corpuscular particles, radius of gyration R. The radius of gyration R of a particle is the root-mean-square distance of all electrons from their center of gravity. Hence R is defined in complete analogy to the radius of inertia in mechanics, with the only difference being that here the electrons take the place of mass elements. Therefore, R can be easily calculated for simple geometrical bodies (Table I).

The *experimental* determination of the radius of gyration can be done according to Guinier, who found that in its innermost part every particle scattering curve follows an exponential course, according to the following equation:

$$I = I_0 e^{-R^2 h^2/3}, \tag{17a}$$

where I_0 corresponds to the scattering intensity at zero angle. In logarithmic form, this reads as follows:

$$\ln I = \ln I_0 - R^2 h^2/3. \tag{17b}$$

TABLE I Examples for Radii of Gyration of Simple Geometrical Bodies

Sphere with radius r	$R = \sqrt{\frac{3}{5}}r$
Hollow sphere with radii r_1 and r_2	$R^2 = \frac{3}{5}(r_1^5 - r_2^5)/(r_1^3 - r_2^3)$
Three-axial ellipsoid with semiaxes a, b, c	$R^2 = (a^2 + b^2 + c^2)/5$

In a plot of $\ln I$ versus h^2 (Guinier plot) one obtains a curve with the decay of $R^2/3$ in its *innermost part*, from which R can be calculated.

The radius of gyration is an important parameter and is often useful as an indicator for structural changes of a substance. Changes studied through the use of the radius of gyration are, for instance, association and dissociation effects, conformational changes by denaturation, binding of coenzymes, and temperature effects. An instructive example is the melting of transfer ribonucleic acid (tRNA). Both the UV absorption and the radius of gyration were found to increase with increasing temperature (Fig. 8). This effect has been interpreted as "melting" of the base pairs, which leads to a loosening up of the originally closely packed chainlike molecule. The behavior of the radius of gyration offers a good illustration for the concomitant size increase of the coil.

ii. Rodlike particles, radius of gyration of the cross section R_c. A sufficiently dilute solution of extremely long and extremely thin molecules would have the scattering properties of a gas of needles, that is, $1/h$. In reality, however, elongated particles have finite thickness and are not infinitely long. The finite thickness leads to a function I_c, the cross-section factor, which depends only on size and shape of the cross section and by which the needle scattering curve has to be multiplied as follows:

$$I = I_c/h, \qquad I_c = Ih. \tag{18a}$$

In analogy to the determination of the radius of gyration of corpuscular particles from the Guinier plot of the scattering curve, the radius of gyration of the cross section R_c for elongated particles can be determined by the Guinier plot of the cross-section factor $\ln Ih$ versus h^2; tangent $(\tan\alpha)_0$ at small angles is given by the following:

$$(\tan\alpha)_0 = R_c^2/2. \tag{18b}$$

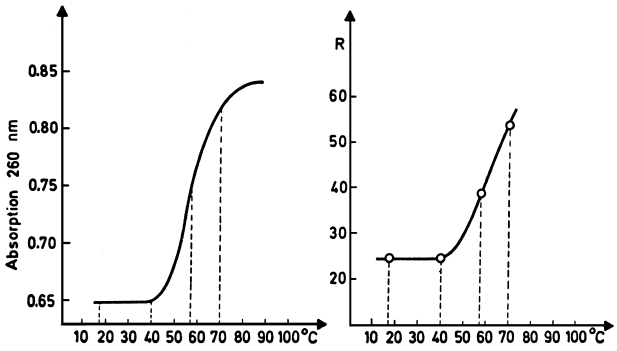


FIGURE 8 "Melting curves" of *t*-RNA^{Phe} (yeast). Temperature dependence of the radius of gyration (right) and of ultraviolet absorption (left). [Reprinted with permission from Pilz, I., Kratky, O., Cramer, F., von Der Haar, F., and Schlimme, E. (1970). *Eur. J. Biochem.* **15,** 401.]

The finite length causes the largest distances to be lacking in comparison to infinitely long needles. According to the law of reciprocity, this lack is reflected by a decrease in intensity toward the very smallest angles.

An instructive example is a study on the enzyme malate synthase, which has a disklike structure in its monomeric form. Upon exposure to a very high X-ray dose, the particles associate laterally, leading to more and more elongated particles. Figure 9 shows that the slope and hence also the cross-section factor are practically independent of irradiation time. Since originally the particles are relatively short, one observes a strong decrease in the cross-section factor toward zero angle after short irradiation periods. With increasing irradiation times and association, however, this effect decreases since the shape approaches that of a long rod. Finally, for very long irradiation times one observes in the inner part an increased cross-section factor to about twice its original value; this can be easily interpreted in terms of the formation of laterally paired rods.

If R and R_c are known, the length of the rod is found according to the following:

$$L = \sqrt{12\left(R^2 - R_c^2\right)}. \tag{19}$$

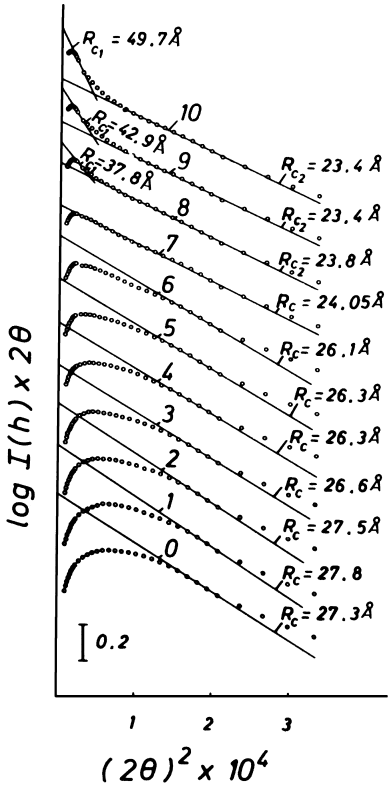


FIGURE 9 Changes in cross-section factor of malate synthase. Time intervals of X-ray irradiation between two subsequent curves, 5.7 hr. [Reprinted with permission from Zipper, P., and Durchschlag, H. (1980). *Mh. Chem.* **111,** 1367.]

iii. Lamellar particles: Radius of gyration of the thickness R_t. A very dilute solution of infinitely large and infinitely thin sheets would have the particle scattering curve $1/h^2$. Since in reality the thickness is always finite, it follows for the scattering curve that

$$I = I_t/h^2; \qquad I_t = Ih^2, \tag{20a}$$

where I_t depends on the thickness only and is therefore called the thickness factor. Its Guinier plot ($\ln Ih^2$ versus h^2) leads to the radius of gyration of the thickness, through the slope $(\tan\alpha)_0$ at smallest angles:

$$\tan\alpha_0 = -R_t^2. \tag{20b}$$

From the general relation between the length of a line and its radius of gyration, it follows that the thickness t of the lamella is connected to R_t by the following:

$$t = R_t\sqrt{12}. \tag{21a}$$

Since for a coinlike particle such as malate synthase the thickness is much smaller than its diameter, it follows from the law of reciprocity that the thickness is expressed at larger angles. If the abscissa range shown in Fig. 9, is extended to larger angles (Fig. 10), the Guinier plot of the thickness factor leads to a radius of gyration corresponding to a thickness of about 1 nm. From the independence of this effect on irradiation time it can be concluded that these particles have constant thickness; that is, their association always occurs laterally.

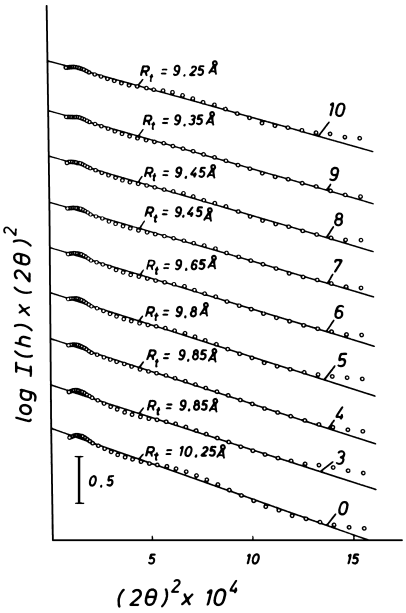


FIGURE 10 Thickness factor of malate synthase. Time intervals of X-ray irradiation between two subsequent curves, 11.4 hr. [Reprinted with permission from Zipper, P., and Durchschlag, H. (1980). *Mh. Chem.* **111,** 1367.]

iv. Parallelepipedic particles. The analysis of the radii of gyration is of particular value in the case of parallelepipedic structures, where the length l is large as compared to the width b, and this in turn is large against the thickness t. In this case, the edge lengths are connected to their radii of gyration according to Eq. (21a), and hence

$$l = R_l\sqrt{12}; \quad b = R_b\sqrt{12}; \quad t = R_t\sqrt{12}. \tag{21b}$$

Furthermore, these individual radii of gyration are related to the radius of gyration of the entire particle R simply by the following:

$$R^2 = R_l^2 + R_b^2 + R_t^2. \tag{22a}$$

The area lb has the radius of gyration R_a, which is the limiting value of R for $t = 0$; therefore

$$R_a^2 = R_l^2 + R_b^2. \tag{22b}$$

The value R_c for the area bt, identical to the already-introduced radius of gyration of the cross section, is the limiting case for $l = 0$, and hence

$$R_c^2 = R_b^2 + R_t^2. \tag{22c}$$

Combining Eqs. (22a) and (22c) yields

$$R_l^2 = R^2 - R_c^2, \qquad l = \sqrt{12\left(R^2 - R_c^2\right)} \tag{22d}$$

and Eq. (22c) can be written as

$$R_b^2 = R_c^2 - R_t^2, \qquad b = \sqrt{12\left(R_c^2 - R_t^2\right)}.$$

Finally the value of t is obtained from $t = \sqrt{12R_t^2}$.

Therefore, from the values of R, R_c, and R_t obtained from Guinier plots of I, Ih, and Ih^2, respectively, the length, width, and thickness of a parallelepiped can be determined.

Particles of this kind are found in aggregates of dyes in solution, for example, of β-naphthol orange (formula shown in Fig. 11).

The degree of association of these molecules increases with salt concentration, as can be immediately inferred from the particle mass. This is shown in Fig. 12, where the degree of association increases up to about 20. The determination of the parameters l, b, and t in the same series as in Fig. 12 leads always to a value of 13.5 Å for t, which is identical to the length of the monomeric molecule (Fig. 11). Therefore, the increase in l and b corresponds to lateral association. The degrees of association and the

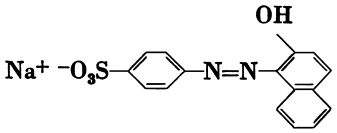


FIGURE 11 β-Naphthol orange.

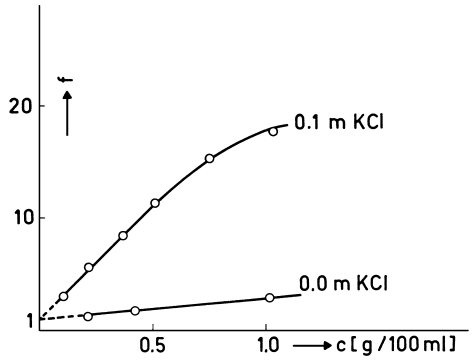


FIGURE 12 Average degree of association f of β-naphthol orange as a function of dye concentration c in aqueous and 0.1 *M* KCl solution. [Reprinted with permission from Kratky, O., Ledwinka, H., and Pilz, I. (1967). *Makromol. Chem.* **105,** 171.]

three dimensions of the particles are listed for some points in Table II. It can be clearly seen that the area a of lb is exactly proportional to the degree of association f; that is, the ratio of a/f is constant. Hence the molecular area within the aggregate has the constant value of 95 Å^2. With the length of the molecule, this leads to a molecular volume of $95 \times 13.5 = 1280$ Å^3, a value that is about four times larger than that calculated on the basis of the molecular weight and the partial specific volume, which is 302 Å^2. An explanation for the cohesion in such obviously very loosely packed aggregates is not straightforward.

A second illustrative example is the azo dye chlorantin light violet 2RLL (formula shown in Fig. 13). Here too, the degree of association increases with concentration and ionic strength. The evaluation in terms of a parallelepipedic structure leads to the values $l = 37$ Å, $b = 9.7$ Å, and a value of t that increases from 8 to 27 Å. The values l and b can be understood in terms of the dimensions of the rather band-shaped monomeric molecule, so that the third dimension results from a lateral stacking on the broad face. Here too, as with β-naphthol orange, the apparent volume occupancy per molecule is larger by a factor of 3 than that calculated from M and $\bar{v}_2$.

For future experiments on dyes, these two results can be summarized as follows. Considering that the molecular weights of these two molecules, 350 for β-naphthol orange

TABLE II Shape Determination of β-Naphthol Orange[a]

c	0.103	0.371	1.037
f	4.5	11.9	23.1
t	13.5	13.5	13.5
a	430	1130	2190
a/f	95	95	95

[a] From Kratky, O., Ledwinka, H., and Pilz, I. (1967). *Makromol. Chem.* **105,** 171–192.

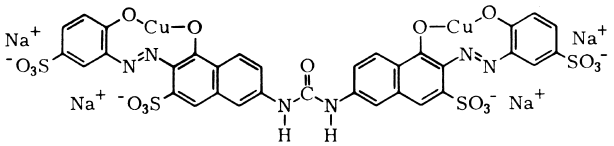


FIGURE 13 Chlorantin light violet 2RLL.

and 1116 for chlorantin light violet, are about two orders of magnitude smaller than those of the macromolecules like the proteins usually studied by this method, these examples demonstrate an enormous extension of the field of application into the domain of medium-sized molecules. It is also noteworthy that no other known method would allow a similarly detailed study of the molecular association in this range of sizes. Optical spectroscopy is limited to low degrees of association, due to the large extinction, and for the usual methods of molecular weight determination the masses are too small. Moreover, with no other method would it be feasible to even approximately yield the entire shape information of the aggregates.

b. Volume of corpuscular particles; cross-sectional area of rodlike particles; thickness of lamellar particles. In a two-phase system of particles and solvent, the invariant

$$Q = \int_0^\infty I(h)h^2\,dh \tag{23}$$

depends only on the square of the net contrast, $(\Delta\rho)^2$ and the volume fractions [Eq. (7)] and is therefore independent of the degree of dispersion with constant weight concentration and hence also of the volume of the single particles. The extrapolated zero-angle intensity I_0, on the other hand, is proportional to the volume of scattering particles, the square of the electron density difference, and the volume fraction, i.e., to $(V\,\Delta\rho^2\,w)$. The ratio I_0/Q is therefore a measure for the particle volume. According to Porod,

$$V = 2\pi^2 I_0/Q. \tag{24}$$

An example of the application of Eq. (24) is the saturation of the apoenzyme of yeast glycerol aldehyde 3-phosphate dehydrogenase with the coenzyme NAD. In the course of this saturation one observes a volume concentration of 7%. From the nonlinear relationship between the degree of saturation and the volume contraction, it was concluded that an allosteric mechanism underlies this saturation process (Fig. 14).

For the cross-sectional area A_c of rodlike particles one obtains a relationship analogous to Eq. (24):

$$A_c = 2\pi(Ih)_0/Q \tag{25}$$

and for the flat leaflets one finds the thickness t according to the follows:

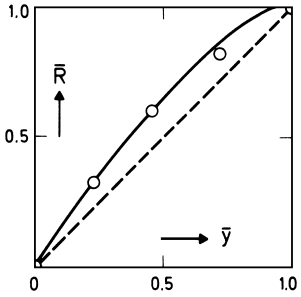


FIGURE 14 Relative volume contraction $\bar{R}$ of glycerolaldehyde-3-phosphate dehydrogenase as a function of the degree of saturation $\bar{y}$ with NAD. [Reprinted with permission from Durchschlag, H., Puchwein, G., Kratky, O., Schuster, I., and Kirchner, K. (1971). *Eur. J. Biochem.* **19,** 9.]

$$t = \pi(Ih^2)_0/Q. \tag{26}$$

Equations (25) and (26) are still applicable with relatively high concentrations.

c. Mass determination.

i. Molecular weight. An important possibility of X-ray small-angle scattering is the determination of the particle mass. This is due to the fact that at zero angle all secondary waves from the electrons within a particle are in phase, and hence the total amplitude is proportional to the number of excess electrons. According to Eq. (1), the single electron gives a scattering intensity of $T_e = 7.9 \times 10^{-26}$ relative to the primary intensity P_0. A simple derivation then leads to the relation for the molecular weight M as follows:

$$M = \frac{21.0 I_0 a^2}{P_0(z_2 - v_2'\rho_1)^2 tc}, \tag{27a}$$

in which 21.0 is the value of $1/N_A I_e$, under the conditions of slit collimation where P_0 is the intensity per 1 cm length of the primary beam, I_0 is the value of the scattered intensity obtained upon "desmearing," a is the distance between sample and plane of registration, z_2 is the number of mole-electrons per gram of solute, ρ_1 is the electron density of the solvent, v_2' is the isopotential specific volume of the solute (normally this can in good approximation be set equal to the partial specific volume $\bar{v}_2$), t (cm) is the thickness of the sample, and c is the concentration (grams per milliliter).

Numerous molecular weights of biologically interesting substances in the range of 10^4 to 10^7 daltons have been determined with the help of this equation. These limits have been considerably extended toward low molecular weights by the above-mentioned studies on dye solutions. The low-molecular-weight record stands presently at 350.

ii. Mass per unit length for rodlike particles. The small-angle scattering method can be considered unique

in its possibility to *weigh* a piece of unit length of an elongated particle. The mass per unit length denominated as M_c is obtained by the following:

$$M_c = \frac{6.69(Ih)_0 a^2}{P_0(z_2 - v_2'\rho_1)^2 tc} \tag{27b}$$

in which 6.69 is the value of $1/\pi N_A T_e$.

iii. Mass per unit area for lamellar particles. In analogy to Eq. (27b), the mass per unit area M_t for lamellar particles is obtained by the following:

$$M_t = \frac{3.34(Ih^2)_0 a^2}{P_0(z_2 - v_2'\rho_1)^2 tc}, \tag{27c}$$

in which 3.34 is the value of $1/2\pi N_A T_e$.

d. Distance distribution function. The intramolecular distance distribution function $p(r)$ is obtained from Eq. (15) simply through Fourier inversion according to the following equation:

$$p(r) = \frac{r^2}{2\pi^2}\int_0^\infty I(h)h^2\frac{\sin(hr)}{hr}\,dh \tag{28a}$$

or

$$p(r) = \frac{1}{2\pi^2}\int_0^\infty I(h)hr\,\sin(hr)\,dh, \tag{28b}$$

which are mutually equivalent.

The maximal distance r_{max} within one particle can be found directly from the abscissa value where the function $p(r)$ vanishes. This frequently allows the recognition of association processes. An instructive example for this application of the $p(r)$ function is the radiation-induced association of malate synthase (Fig. 15). The distance distribution function clearly indicates that in addition to monomers, with their maximum dimensions of about 10 nm, initially particles of twice this size appear, and finally, as the time of irradiation proceeds, even larger particles occur. The process of association is schematically depicted in Fig. 16.

The knowledge of the distance distribution function also offers an independent way for the determination of the radius of gyration by the following:

$$R^2 = \frac{\int_0^\infty p(r)r^2\,dr}{2\int_0^\infty p(r)r\,dr}. \tag{29}$$

However, the major importance of the distance distribution function stems from its more descriptive nature as compared to the scattering curve, which makes it very valuable in shape determinations by trial and error.

e. Electron density distribution of symmetrical particles. For particles with spherical symmetry, it is theoretically possible to calculate the radial electron density

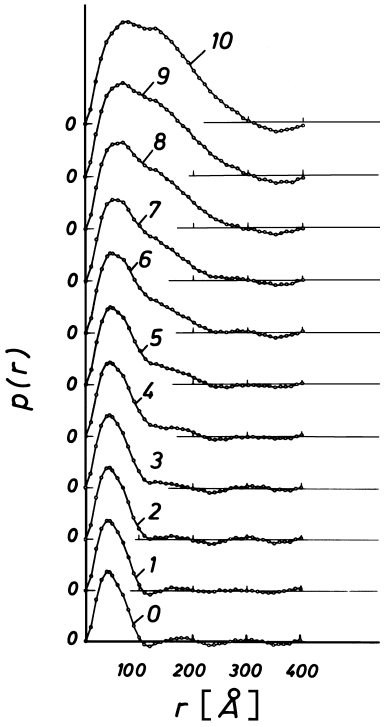


FIGURE 15 Distance distribution function $p(r)$ of aggregating malate synthase. Time intervals as in Figs. 9 and 10. [Reprinted with permission from Zipper, P., and Durchschlag, H. (1981). *Mh. Chem.* **111,** 1.]

distribution $p(r)$ from the scattering intensity $I(h)$, since, as a consequence of symmetry, no information is lost in the process of spatial averaging under these conditions, and one obtains the relationship between the scattering amplitude $A(h) = \pm\sqrt{I(h)}$ and the radial electron density distribution $\rho(r)$ as follows:

$$A(h) = 4\pi\int_0^\infty \rho(r)r^2\frac{\sin hr}{hr}\,dr. \tag{30a}$$

The inverse transform gives $\rho(r)$ according to the following:

$$\rho(r) = \frac{1}{2\pi^2}\int_0^\infty A(h)h\frac{\sin hr}{hr}\,dh. \tag{30b}$$

An exactly centrosymmetrical electron density should cause several subsidiary maxima separated by zeros in the scattering curve. Since in reality the symmetry is never strictly centrosymmetrical, deep minima replace the zeros. In order to apply the above equations, such zeros have to be artificially introduced. In the determination of the amplitudes $A(h)$ from the intensities one has to draw the square root, leading to the problem of the proper choice of signs. Both problems, the artificial production of zeros and the proper choice of signs, can be solved by suitable methods.

An important field of application for this approach is the plasma lipoproteins, which are composite structures

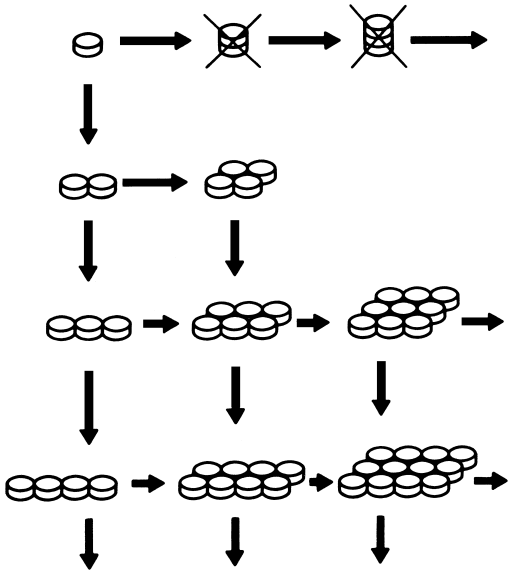


FIGURE 16 Schematic representation of the radiation-induced aggregation of malate synthase. [Reprinted with permission from Zipper, P., and Durchschlag, H. (1980). *Rad. Environ. Biophys.* **18,** 99.]

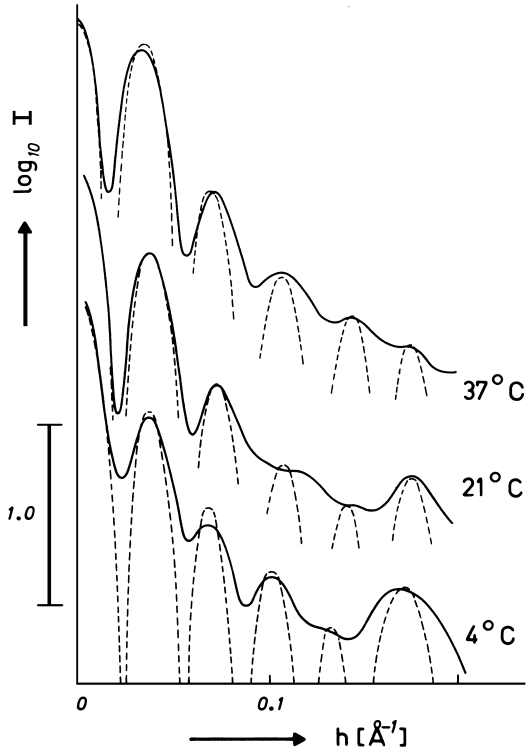


FIGURE 17 Scattering curves of human low-density lipoprotein LpB. [Reprinted with permission from Laggner, P., Degovics, G., Müller, K., Glatter, O., Kratky, O., Kostner, G., and Holasek, A. (1977). *Hoppe-Seyler's Z. Physiol. Chem.* **358,** 771.]

of various lipids and proteins. The scattering curves of a plasma lipoprotein shown in Fig. 17 with their deep minima show the typical appearance of nearly spherical particles. The radial electron density distribution calculated via the scattering amplitudes allows the postulation of a model on the basis of the known chemical composition (Fig. 18). It becomes immediately clear that the core of the particles containing the lipids undergoes a transition from a highly ordered state at low temperatures to a disordered state at high temperatures. This change in structure is reversible between 4° and 40°C.

Another important field where quasispherical particles with an inhomogeneous radial electron density distribution are frequently encountered is detergents in aqueous media. Generally, all detergents share a common feature: they are small, elongated molecules with a clear separation between polar and unpolar parts. Above a certain critical concentration in water (cmc, critical micelle concentration) they form aggregates (micelles) such that the hydrophobic, apolar parts (frequently hydrocarbon chains) form the core and the polar parts, the hydrophilic headgroups [often $-COO^-$, $-SO_3^-$, $-N^+(CH_3)_3$] constitute the surface layer. X-ray small-angle scattering is very well suited for studies on size, shape, and aggregation number—which are often very sensitive to environmental conditions such as temperature and salt concentration—since their micellar sizes are in the range of smaller colloids.

As an example, the X-ray small-angle scattering on sodium dodecyl sulfate (SDS) is briefly described. The scattering curve is very similar to that of a sphere. Fourier transformation of the amplitudes, therefore, leads to a radial electron density distribution with an outer diameter of 70 Å. This $\rho(r)$ function can be understood in terms of a radial arrangement of the hydrocarbon chains surrounded by the electron-dense polar headgroups. It is generally assumed that the micelles grow in size upon raising the concentration and finally transform into cylinders. However, X-ray small-angle scattering shows that the maximal diameter of 70 Å is largely independent of concentration—and also of salt concentration in the aqueous medium—up to a certain limit. The results of absolute intensity measurements indicate that additional detergent molecules enter the 70 Å particles by increasing the packing density until the limit of 140 molecules per micelle is reached. Above this limit—which depends on salt concentration—a further increase in concentration leads to an increase in size and eventually to ellipsoidal shape.

For elongated particles with cylindrical symmetry and for flat particles with a central plane of symmetry, the electron density distribution vertical to the cylinder axis and to the central plane, respectively, can be obtained analogously by Fourier transformation of the thickness and cross-section amplitudes, $[I(h)h]^{1/2}$ and $[I(h)h^2]^{1/2}$.

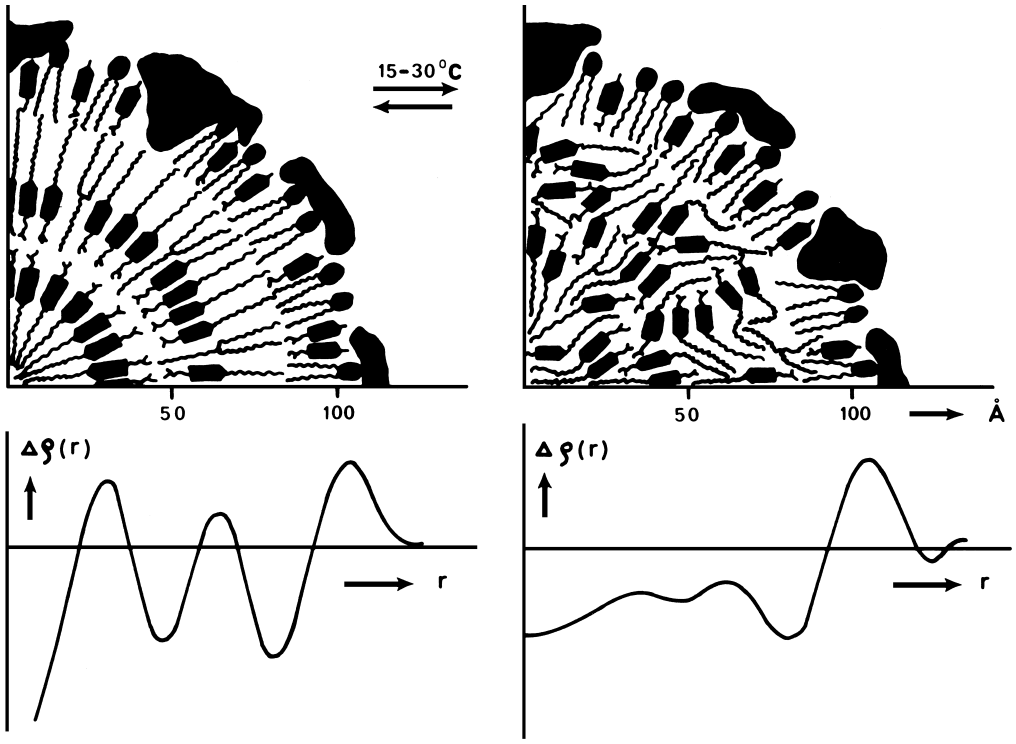


FIGURE 18 Idealized cross-section model of human lipoprotein LpB, derived from the electron density distribution shown in the lower part of the figure. [Reprinted with permission from Laggner, P., Degovics, G., Müller, K., Glatter, O., Kratky, O., Kostner, G., and Holasek, A. (1977). *Hoppe-Seyler's Z. Physiol. Chem.* **358,** 771.]

4. Shape Determination by Trial and Error

a. The two possibilities of comparison. There exist basically two approaches, which are outlined in the following:

1. The theoretical scattering curves calculated for the assumed models are compared to the experimental scattering curve and the models are varied by trial and error until satisfactory agreement is reached, that is, until one of the calculated bodies is "equivalent in scattering" to the real object. This is a comparison in "reciprocal space," since the scattering curve represents a reciprocal picture of reality.
2. The distance distribution function $p(r)$ calculated for a certain model is compared to the "experimental" distance distribution function, calculated according to Eq. (28) from the scattering curve. Again the model is iteratively varied until equivalence in scattering—that is, equality of $p(r)_{\text{theor}}$ and $p(r)_{\text{exp}}$—is obtained. Here we speak of a comparison in "real space," since the distance distribution function represents the real particle.

b. Approximation by simple triaxial bodies. A typical example of trial-and-error curve fitting in terms of simple geometrical bodies is given by the investigation on the pH-dependent structural transition of serum albumin. This protein (MW $\approx$ 70000) shows an increase in radius of gyration from 32 to 38 Å when the pH is lowered from neutral (7.0 and 5.1) to 3.7. The scattering curves for neutral pH (Fig. 19a) show good agreement with prolate ellipsoids of axial ratios of 0.66 : 1 : 2, while at pH 3.7 the best fit is obtained for a flat prism with edge ratio 0.25 : 1 : 1 [Fig. 19(b)]. These axial ratios have been found by analyzing the various radii of gyration in the terms outlined above for parallelepipedic structures. This shape change was interpreted in terms of the model shown in Fig. 20, which implies a partial unfolding of two structural similar domains.

A further instructive example is the shape analysis of hemocyanin of the blue blood pigment of the snail *Helix pomatia*. From electron microscopy, a hollow cylinder had to be expected. Figure 21 shows the theoretical scattering curves for the full and hollow cylinders with different ratios of external to internal diameter r_i/r_a. As with all internally heterogeneous bodies, the intensity of the first side maximum relative to the main maximum increases with the degree of "hollowness." The experimental curve fits well into this series of curves and leads to a ratio of inner to outer diameter of 0.45.

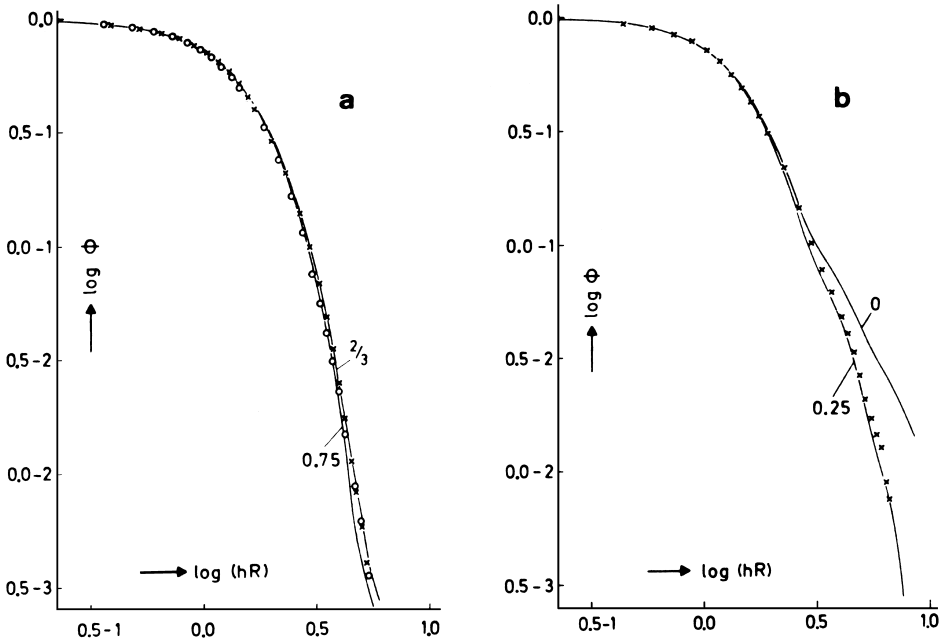


FIGURE 19 Comparison of the experimental scattering curves of porcine serum albumin at different pH values with theoretical curves for geometrical model bodies: (a) at pH 5.1 (◯) and pH 7.0 (×), such that the full lines correspond to theoretical scattering curves of ellipsoids with axial ratios of a:1:2 with $a = 0.75$ and $\frac{2}{3}$; (b) at pH 3.7 (×), such that the full line with 0.25 is the theoretical curve for a rectangular prism of edge-length ratio 0.25:1:1. [Reprinted with permission from Laggner, P., Kratky, O., Palm, W. H., and Holasek, A. (1971). *Mh. Chem.* **102,** 1729.]

c. Approximation by composite structures. Frequently, complex protein structures consist of many approximately spherical subunits of similar size. The calculation of the scattering curve can be performed according to the Debye equation, Eq. (16), in which instead of f_i and f_k the scattering amplitude of the subunits (instead of atomic form factors) is used and for r_{ik} their mutual center-to-center distances are introduced. Although one cannot immediately calculate the entire scattering curve according to Eq. (16) due to the lack of knowledge of the mutual distances, it is clear, nevertheless, that the scattering function contains the product of the squares of the amplitudes of the spherical subunits in all summands. Since the scattering curves of spheres contain zeros, these have to occur also in the scattering curves of the entire particles; that is,

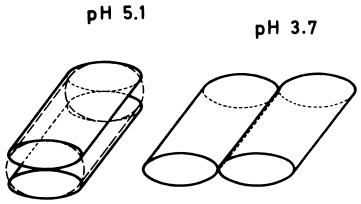


FIGURE 20 Schematic representation of the structural transition of serum albumin at low pH values, as for Fig. 19. [Reprinted with permission from Laggner, P., Kratky, O., Palm, W. H., and Holasek, A. (1971). *Mh. Chem.* **102,** 1729.]

well-developed minima are to be expected if the condition of spherical shape of the subunits is approximately met. From the position of these minima, one can then calculate an average diameter d_s of the subunits:

$$d_s = 3.49/h_{01} = 6.04/h_{02} \tag{31a}$$

where h_{01} corresponds to the *first* and h_{02} to the *second* minimum. This relation leads to the following:

$$h_{01}/h_{02} = 1.73, \tag{31b}$$

thus providing a basis for the decision whether or not two separate minima can have their origin in the spherical structure of the basic units.

As an example, the outer part of the scattering curve of hemocyanin from *H. pomatia* indeed shows two distinct minima (Fig. 22), which fit Eq. (31a), leading to a diameter for the subunits of $d_s = 4$ nm. This information has substantially contributed to solving the problem of the proper filling of the given hollow cylindrical shape with spheres (Fig. 23). Figure 24 also shows the comparison between experimental and theoretical curves calculated on the basis of this model: The agreement is convincing.

d. Models composed of many spheres that do not represent chemical subunits. The construction of models consisting of many small spheres with given center-of-gravity distances, for which the scattering curves

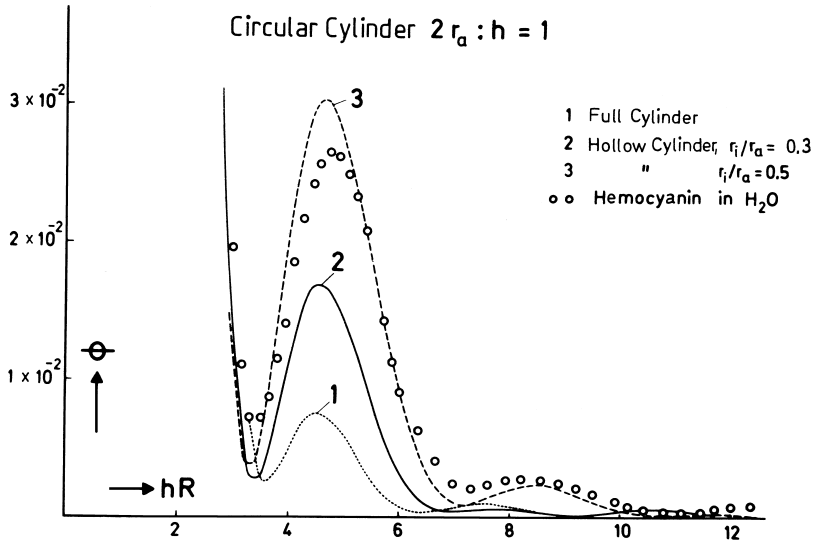


FIGURE 21 Comparison of the scattering curves of circular hollow cylinders with that of hemocyanin in water. [Reprinted with permission from Pilz, I., Kratky, O., and Moring-Claesson, I. (1970). *Z. Naturforsch.* **25b,** 600.]

can be easily calculated according to Eq. (16), is a generally applicable approach in the interpretation. The small spheres, however, may not necessarily relate to true chemical units but only serve to approximate a certain type of structure by simple means. However, this normally requires some knowledge on the type of structure from other sources of information. As an example of this approach, Fig. 25 shows the model for the factor σ of a DNA-dependent RNA-polymerase enzyme subunit of *Escherichia coli*, together with the comparison between the experimental scattering curve and the theoretical curve calculated for this model.

e. Comparison of the experimental and theoretical distance distribution functions. Figures 26 and 27 show the strong dependence of the distance distribution function on the type of structure. With some experience, the approximate shape of a particle can be readily guessed from the appearance of the $p(r)$ function, and a refined model can be obtained by systematic variation of axial ratios and other parameters.

This approach has been taken in an investigation on the structure of the *dimeric complex* between two transfer nucleic acids, $tRNA^{Phe}$ and $tRNA^{Glu}$, which bind to each other by their complementary anticodon tripletts.

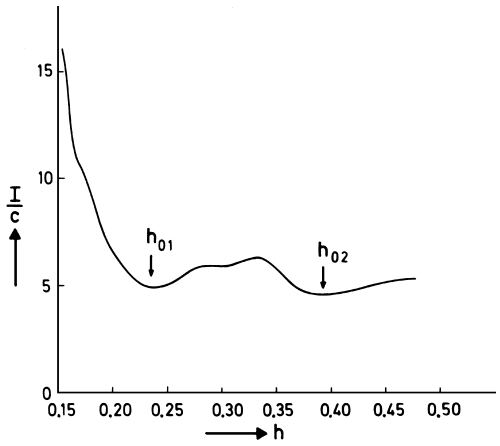


FIGURE 22 Tail end of the scattering curve of hemocyanin from *H. pomatia*. [Reprinted with permission from Pilz, I., Glatter, O., and Kratky, O. (1972). *Z. Naturforsch.* **27b,** 518.]

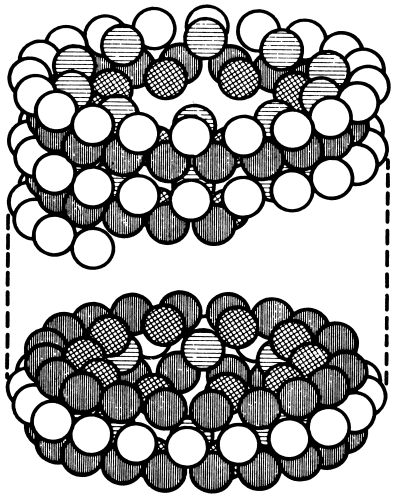


FIGURE 23 Model of hemocyanin consisting of 360 identical subunits, arranged in six similar double layers. [Reprinted with permission from Pilz, I., Glatter, O., and Kratky, O. (1972). *Z. Naturforsch.* **27b,** 518.]

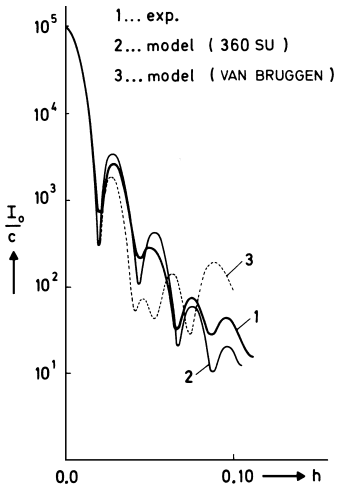


FIGURE 24 Inner part of the experimental scattering curve of hemocyanin (curve 1); theoretical scattering curve of the model depicted in Fig. 23 consisting of 360 subunits (curve 2). (The model of Von Bruggen, to which curve 3 relates, is not discussed further here). [Reprinted with permission from Pilz, I., Glatter, O., and Kratky, O. (1972). *Z. Naturforsch.* **27b,** 518.]

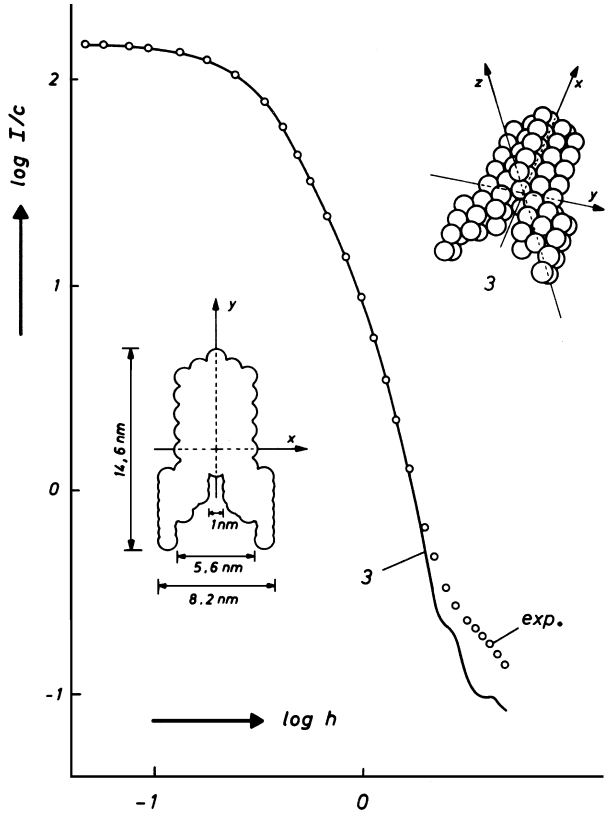


FIGURE 25 Comparison of the experimental scattering curve of the factor σ of DNA-dependent RNA-polymerase of *E. coli* (open circles) to the theoretical curve calculated for the depicted model (solid line). [Reprinted with permission from Meisenberger, O., Pilz, I., and Hermann, H. (1980). *FEBS Lett.* **112,** 39.]

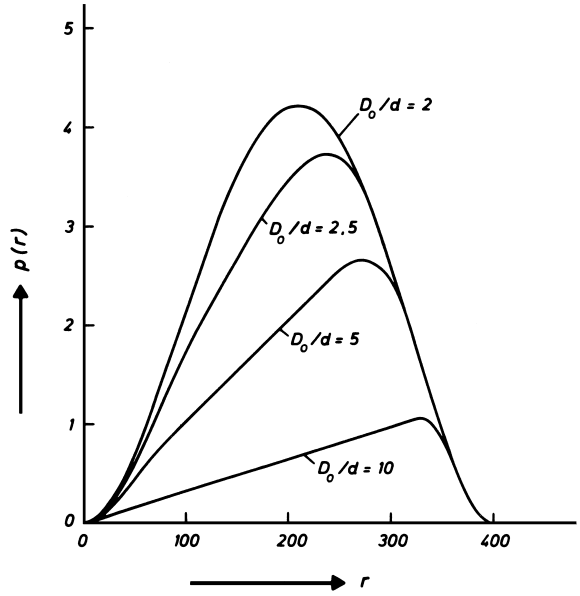


FIGURE 26 Distance distribution function $p(r)$ of hollow spheres, where D_o is the outer diameter and d the thickness of the shell, and $D_o/D = 2$ therefore represents a full sphere. [Reprinted with permission from Glatter, O. (1979). *J. Appl. Crystallogr.* **12,** 166.]

Figure 28 shows the experimental $p(r)$ functions of the complex compared to various forms of model dimers. The $p(r)$ functions of the models have been calculated on the basis of the known L-shaped crystal structure of tRNA. It is evident that the experimental $p(r)$ function lacks the pronounced shoulder at r values of about 100 Å, which are due to the distal "sticking-out" of the open ends, the so-called acceptor arms, of the coil. A better approximation is obtained by a more cigarlike structure, which can be

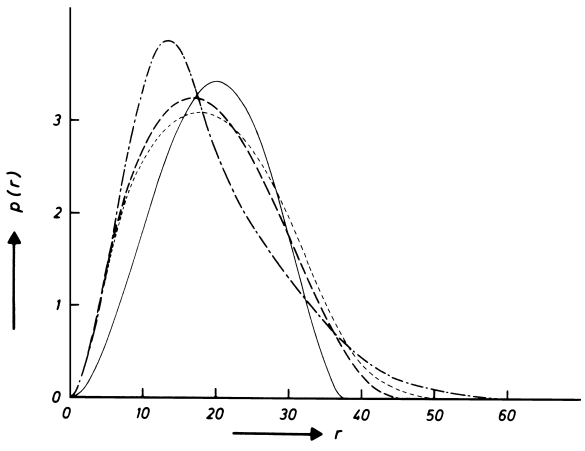


FIGURE 27 Comparison between the distance distribution function of a sphere (——), a prolate ellipsoid of revolution 1 : 1 : 3 (–·–), an oblate ellipsoid 1 : 1 : 0.2 (– – –), and a flat prism 1 : 1 : 0.23 (· · ·), all having the same radius of gyration. [Reprinted with permission from Glatter, O. (1979). *J. Appl. Crystallogr.* **12,** 166.]

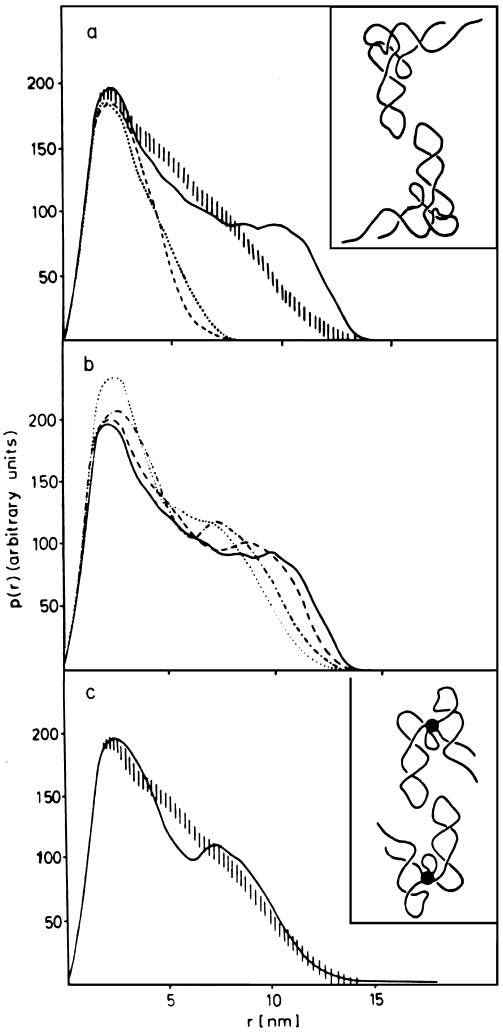


FIGURE 28 Theoretical distance distribution functions for various tRNA models and the experimental $p(r)$ function (±1 standard deviation, shaded curve) of the $tRNA^{Phe}$–$tRNA^{Glu}$ complex. (a) Models based on the atomic coordinates of crystalline tRNA (···), a tRNA having an angle of 30° between the arms of the L (–·–), and a complex of two L-shaped particles, joined as shown in the inset. (b) tRNA dimers having different values of α: —— 90°, – – – 60°, –·– 30°, ··· 0° between the arms of the L. (c) Comparison between the complex with 30° opening angle (insert) and the experimental $p(r)$ function. [Reprinted with permission from Nilsson, L., Rigler, R., and Laggner, P. (1982). *Proc. Natl. Acad. Sci. USA* **79,** 5891.]

achieved by folding the acceptor arms closer to the rest of the structure (Fig. 28b). Clearly the fit thus obtained is better, but some deviations still occur. However, this result shows that the anticodon binding leads to a more compact structure than that obtained by simple longitudinal association of the monomers.

It is noteworthy that a conformational change of this type would not be easily recognized if it only occurred on monomeric particles. The effects would be very small, since the rotation of the two parts of the molecule occurs around the center of gravity. In the dimer, the effect is amplified through the fact that the rotation occurs at distant ends from the center of gravity.

5. Additional Information from Modifications of the Scattering System

a. Contrast variation. With particles consisting of subunits of different electron densities—for instance viruses consisting of nucleic acids and proteins—it would first be desirable to use a solvent that matches exactly the electron density of the protein, making it invisible and showing only the scattering from the nucleic acid; in the second step, the converse should be done, leaving the protein as the only scattering entity.

Stuhrmann and Kirste have developed a general formalism for contrast variation by changes in the solvent electron density. They were able to show that it is possible to separate the scattering function I into three terms by measurements in at least three solvents of different electron densities.

In general, the dependence of the scattering intensities $I(h)$ on the net contrast $\overline{\Delta\rho}$ between solute and solvent can be expressed as a quadratic polynomial as follows:

$$I(h) = I_{\mathrm{v}}(h)\,\overline{\Delta\rho^2} + I_{\mathrm{vs}}(h)\,\overline{\Delta\rho} + I_{\mathrm{s}}(h) \qquad (32)$$

where $I_{\mathrm{v}}(h)$ is the scattering curve of the contour volume, that is, of a homogeneously filled particle with the same shape as the real one. This function is obtained by extrapolation to infinite contrast $\overline{\Delta\rho}$ and can be analyzed in terms of particle shape, as indicated in the previous sections. The term $I_{\mathrm{s}}(h)$ is the scattering curve at zero net contrast and reflects only the internal electron density fluctuations. The cross term is of no simple structural significance and shall not be discussed here in detail.

Due to the limited range of contrasts that can be achieved for X-rays by the addition of salts or other low-molecular-weight solutes (which must not interfere with the integrity of the particles under investigation), the majority of the applications of this approach are in the neutron scattering field, where huge contrast variations are easily achieved by H_2O/D_2O mixtures (to be discussed further).

b. Distance determination with X-rays by heavy-atom labeling. In 1947 a method was developed by Kratky and Worthmann for the determination of the distance between two heavy-atom labels in low-molecular-weight organic compounds. The method shall be discussed for the case of distance determinations between two point-shaped markers. Two different solutions have to be prepared (Fig. 29).

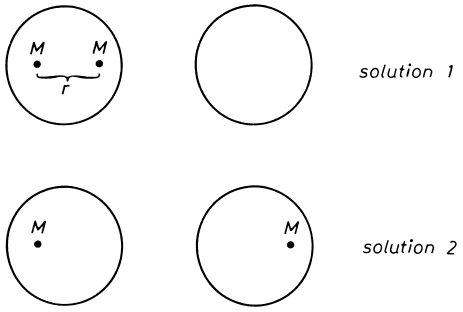


FIGURE 29 Scheme for the determination of the distance r between two equal labels M. For explanation, see text. [Reprinted with permission from Kratky, O. (1983). *Nova Acta Leopoldina, NF 256* **55,** 1–72.]

Solution 1 contains the macromolecule carrying label M (i.e., a heavy atom or a complex of several heavy atoms) on *both* labeling sites. Furthermore, solution 1 has to contain additionally one unlabeled for each double-labeled molecule. In solution 2 each molecule contains only *one* label. The concentrations have to be such that both solutions contain the same number of heavy atoms. For this case it can be shown that the substraction of the scattering curve of solution 2 from that of solution 1 leads to a function determined by the interaction term of the "dumbbell" scattering from the two heavy atoms in solution 1, from which the distance between the two labels can be derived.

An example is given in a study on met-hemoglobin, where Hg labels in cysteine SH groups of the two β-chains were introduced and an average distance of $r = (6.9 \pm 0.2)$ nm between the labels was obtained (Fig. 30).

Obviously, it is of interest to apply this method also to the determination of distances between two equivalent positions in subunits of complex macromolecular structures. This approach has again reached its highest practical importance in neutron diffraction experiments, where the same theory applies and labeling can be achieved by deuteration of entire subunits (to be discussed further).

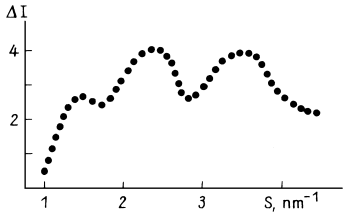


FIGURE 30 Distance determination of Hg labels in methemoglobin. The curve represents the difference between the double- and single-labeled samples. [Reprinted with permission from Vainshtein, B. K., Feigin, L. A., Lvov, Y. M., Grozdev, R. I., Marakushev, S. A., and Likhtenshtein, G. I. (1980). *FEBS Lett.* **116,** 107.]

6. Special Conditions in Neutron Scattering

The basic concepts of elastic neutron scattering are identical to those of electromagnetic scattering, as outlined for X-rays. The wavelength of a neutron is given by the de Broglie relation

$$\lambda = h/mv, \tag{33}$$

where mv is the momentum of the neutron. Typically, neutrons of wavelengths between 1 and 10 Å are used, appropriate to atomic distances. The neutrons are generally obtained from a reactor, equipped with a suitable moderator (cold source) for high yield in thermal neutrons.

Neutrons are scattered by elastic interaction with nuclei. Therefore, the scattering amplitude of an atom depends on its nuclear structure rather than on the electronic shell, as is the case for electromagnetic radiation. The relevant nuclear parameter is the scattering length b, which varies in a nonmonotonic way from one element to another and may differ considerably for different isotopes. This is in contrast to X-rays, where the amplitude increases proportionally to the number of electrons (= order number). The scattering lengths of some of the more important elements and isotopes are listed in Table III.

One of the most fruitful fields of applications for neutron scattering has its origin in the large difference in scattering length between hydrogen and deuterium. In fact, this difference is the most important reason for the application of neutron diffraction on biological macromolecules and synthetic high polymers. Thus, by suitable isotope replacement, the scattering contrast can be varied within wide limits, practically without structural or chemical perturbation. The principal approaches of contrast variation and selective contrast enhancement by deuteration are outlined in the following examples.

a. Differential contrast matching. This approach takes advantage of the fact that the scattering power of a macromolecular domain is proportional to the square of its scattering length density difference to the continuous solvent background, which can be varied within wide

TABLE III Scattering Lengths of Several Elements for Neutrons

Element	Length ($\times 10^{-12}$ cm)
H	0.374
D	0.66
C	0.66
N	0.94
O	0.58
Fe	0.95

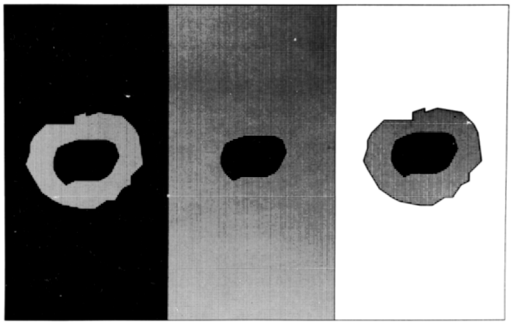


FIGURE 31 Contrast matching in neutron scattering.

limits by changing the H_2O/D_2O composition of the solvent buffer. Thus the scattering of one domain of interest can be enhanced by matching the contrast of the rest of the structure with a suitable H_2O/D_2O mixture. This is schematically depicted in Fig. 31.

Especially in complex biological macromolecules such as viruses or lipoproteins, which are composed of two or more distinct chemical species, this approach has proven highly valuable. The relative scattering length densities of nucleic acids, proteins, and lipids are compared to those of the whole range of H_2O/D_2O mixtures in Fig. 32. The same possibility also exists in principle for X-rays by adding high-electron-density solutes, such as salts or sugar, to the buffer. However, H_2O/D_2O mixtures have several important advantages:

1. There is little danger of changing the structures under investigation by mere H_2O/D_2O variation.
2. A much wider range of contrast variation can be covered.
3. At about 8% D_2O, the scattering length density of the solvent becomes zero, since H_2O has slightly negative and D_2O strongly positive scattering length density, and hence the particle can be observed quasi *in vacuo*.

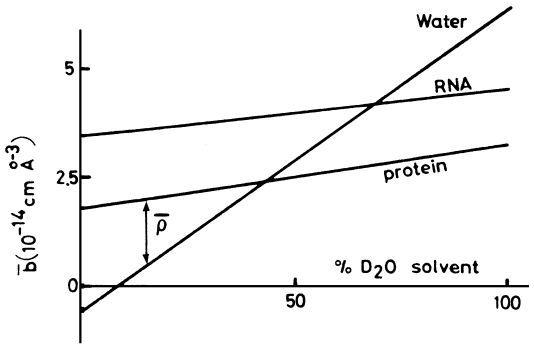


FIGURE 32 Scattering length densities of proteins and nucleic acids in H_2O/D_2O mixtures.

A good example for the contrast matching approach has been provided by a study on tomato bushy stunt virus. With a buffer containing 70% D_2O, the nucleic acids were made invisible, and it was found that the protein is arranged in two concentric shells with a mutual distance of 3 nm. In a buffer of 41% D_2O, the protein was contrast-matched and the scattering showed that the majority of the nucleic acid is arranged between the two protein shells.

b. Evaluation of the contrast dependence of the radius of gyration. The exact formalism for the dependence of R of inhomogeneously filled structures has been developed by Stuhrmann and Kirste. Consider a particle with an internal scattering length density distribution described by the following:

$$\rho(\mathbf{r}) = \bar{\rho}_v + \rho_F(\mathbf{r}), \tag{34}$$

where $\bar{\rho}_v$ is the average density over the particle volume and $\rho_F(\mathbf{r})$ is the deviation from this average at point $\mathbf{r}$, with the center of gravity as the origin for r. Then the radius of gyration R depends on the contrast $\Delta\bar{\rho} = \rho_v - \rho_0$, where ρ_0 is the scattering length density of the solvent in the following way:

$$R^2 = R_c^2 + \alpha/\Delta\bar{\rho} - \beta/(\Delta\bar{\rho})^2. \tag{35}$$

In this equation R_c is the radius of gyration of the contour volume, that is, its value at infinite contrast. The parameters α and β are different moments of the internal density distribution as defined by

$$\alpha = \frac{1}{V}\int_V \rho_F(\mathbf{r})\, r^2\, d^3r \tag{35a}$$

and

$$\beta = \frac{1}{V^2}\int_V \mathbf{r}\rho_F(\mathbf{r})\, d^3r^2. \tag{35b}$$

The term β becomes zero if the centers of gravity of the contour volume and that of $\rho_F(\mathbf{r})$ coincide. Then the contrast variation of the radius of gyration reduces to a linear form as follows:

$$R^2 = R_c^2 + \alpha/\Delta\bar{\rho} \tag{36}$$

and R_c and α can be unambiguously determined by a series of measurements in different H_2O/D_2O mixtures.

In the case of isomorphous deuteration, the value of R_c should be the same for deuterated and undeuterated species. The differences in internal density distribution appear as differences in α. The radius of gyration of the labeled domain R_x can be calculated from $\Delta\alpha$, that is, between deuterated and protonated structure, the value of R_c and the net density difference $\Delta\bar{\rho}$ between the two samples according to the following:

$$R_x^2 = R_c^2 + \Delta\alpha/\Delta\bar{\rho}. \tag{37}$$

Obviously, a comparison of the experimental radii of gyration in one and the same solvent can also lead to this result; however, the above approach is preferable since it is based on more than one pair of experiments, and it also provides the necessary test for the identity of R_c of the two samples.

Further information about one domain can also be obtained if it is accessible for site-specific, isomorphic deuteration. In this case, the comparison between the scattering from deuterated and undeuterated substance yields the desired location of the deuterated domain within the complex structure. To achieve precise results, however, it is always desirable to combine the both methods and perform a contrast variation series on both labeled and unlabeled preparations. As an example for this type of investigation, a study on low-density lipoproteins (LDL) is described. These are highly complex particles consisting of protein, phospholipids, cholesterol, cholesteryl esters, and triglycerides, all having different scattering power (Fig. 18). The surface structure, however, is composed of two species rather similar in their contrast, for both X-rays and neutrons. To solve the question of the surface arrangement it was necessary, therefore, to specifically deuterate one of the two species, which is possible for the phosphorylcholine head groups. Subtraction of the scattering amplitudes, which is permitted if the centers of gravity of the two domains coincide, led to the radius of gyration of the deuterated headgroups, and by difference from the radius of gyration of the entire outer shell, also to that of the protein. It was found that the protein moiety is located on average about 8 Å above the phospholipid head groups.

Another instructive example relates to the question of the orientation of cholesteryl esters, which are arranged radially in the core of LDL particles. Figure 33 shows a schematic view of the possible core structures, which are, except for cases C and E, reconcilable with the radial electron density profiles obtained from X-ray small-angle scattering. A further selection is only possible by selective contrast enhancement of the cholesteryl and fatty acyl residues, respectively, through deuteration. To this end, two different deuterated LDL samples were investigated: one containing cholesterol perdeuteromyristate, and the other containing 25,26,27-d_7-cholesterol oleate (Fig. 34). By comparing either of these samples with unlabeled LDL, the radii of gyration of the fatty acyl and cholesterol domains were obtained individually. It was found that the *radius of gyration* of the cholesterol C-25 isopropyl groups was 70 Å, whereas that of the fatty acyl chains was 60 Å. Thus, all models that contain the fatty acyl moieties at larger average radii than the cholesterol rings could be discarded (models D and F). Models A–C are all qualitatively consistent, and on the basis of molecular packing considerations, model A was proposed to be the most plausible one.

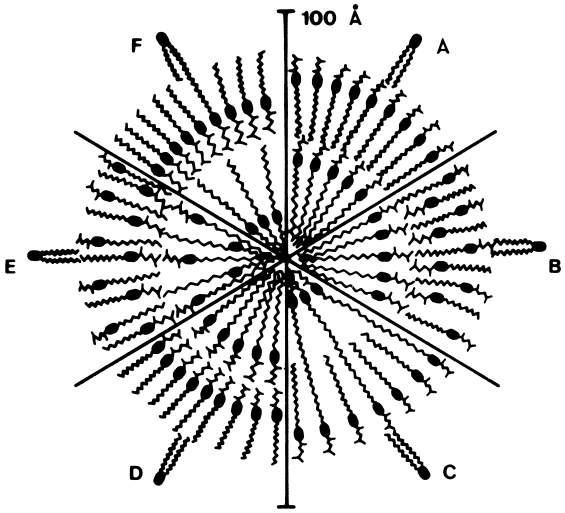


FIGURE 33 Possible models for cholesteryl ester arrangement in LDL. [Reprinted with permission from Laggner, P., Kostner, G. M., Degovics, G., and Worcester, D. L. (1984). *Proc. Natl. Acad. Sci. USA* **81,** 4384.]

c. Label triangulation method for the evaluation of quarternary structures. The quarternary structure of complex biological particles—that is, the relative position of the subunits—can be determined if it is possible to measure an adequate number of center-to-center distances between the subunits. This calculation resembles geodetical triangulation, with the only difference being that in macromolecules the problem is not two-dimensional but three-dimensional. The whole investigation consists of single steps, each involving the attachment of two labels and determination of their mutual distance.

Such a triangulation was first proposed by Hoppe, followed by Engelman and Moore, who performed the first triangulation experiments by neutron diffraction. In this case, however, one is dealing not with heavy-atom labels, but with two entire subunits enhanced in their scattering power. This can be done either by selective deuteration within the complete native molecule or through suitable preparative measures, which leave the two components unaltered within a completely deuterated particle so that the subunits are protonated in contrast to the deuterated surroundings.

The main field of application of such studies are the ribosomal particles of *E. coli*. The 70S particle named for its *S* value of 70 in the ultracentrifuge) has a molecular weight of 2.5×10^6 and can be dissociated reversibly into a 30S and a 50S particle. The triangulation of the 30S subunit has already reached a very advanced state. As shown

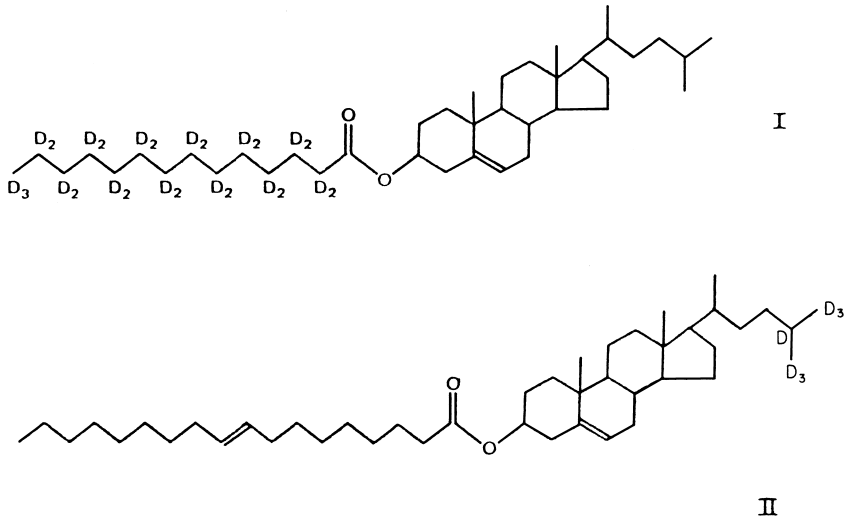


FIGURE 34 Deuterated cholesteryl esters used to study the core of LDL by neutron scattering.

in Fig. 35, the relative positions of six proteins have been determined through their mutual distances.

7. Selective Contrast Enhancement by Wavelength Variation

Near an absorption edge of an element the atomic scattering factor (f) is a complex number that is described as follows:

$$f(\lambda) = f_0 + f'(\lambda) + if''(\lambda), \tag{38}$$

where f_0 is the value far from the edge and f' and f'' are the real and imaginary dispersion corrections. Thus it is possible to selectively change the contribution of one element to the total scattering pattern by scattering experiments at the absorption edge. This approach is formally equivalent to the isotopic labeling technique (selective deuteration), which is so fruitful in neutron scattering.

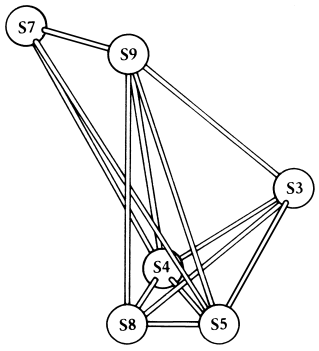


FIGURE 35 Arrangement of six proteins in the 30S subunit of *E. coli* ribosome. [Reprinted with permission from Langer, J. A., Engelman, D. M., and Moore, P. B. (1978). *J. Mol. Biol.* **119,** 463.]

The quantitative formalism for this technique has been described by Stuhrmann and is called "complex contrast variation," in analogy to contrast variation under nonresonant conditions.

Since conventional X-ray tubes emit a strongly discontinuous spectrum, they are not suitable X-ray sources for such investigations. Only with the development of synchrotron radiation sources in the more recent past did such studies become feasible. An early application of this concept to the iron-storing protein ferritin verified the practical applicability of this approach to small-angle scattering in solution. The radius of gyration of ferritin, which is 42 Å in the off-resonance region, increases at the K-absorption edge of iron ($= 1.743$ Å) by about 4% in a sharp peak (Fig. 36). This figure also shows the change in transmission, which can be assumed to be mainly due to changes in absorption. Hence, the relation to the behavior of R reflects the relationship between the real part f' and the imaginary part f'', as expressed by the general Kronig–Kramers relation.

Figure 36 also shows that the monochromator system has to meet very high requirements, since the peak in R is observed over a wavelength interval of only 0.003 Å. Obviously only accurately tunable monochromator crystals with a narrow-wavelength bandpass of $\Delta\lambda/\lambda$ better than 10^{-4} are applicable.

8. Time-Resolved X-Ray Small-Angle Scattering with Synchrotron Radiation

So far, only the possibilities of studying static structures, which are stable over hours or even days of the experiment, have been considered. Structural changes were only discussed insofar as measurements have been made on

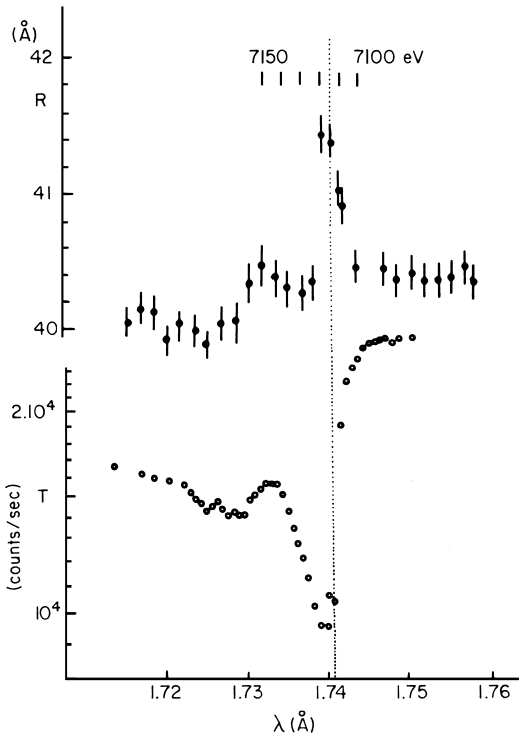


FIGURE 36 Variation of the radius of gyration *R* and the transmittance *T* of ferritin at the K-absorption edge of Fe. [Reprinted with permission from Stuhrmann, H. B. (1980). *Acta Crystallogr.* **A36,** 996.]

different stable states of a system. The direct "visualization" of dynamic processes, in the sense of cinematography, has been neglected because until a few years ago this was impossible due to insufficiently intense X-ray sources. However, the highly intense synchrotron sources now available open such possibilities owing to their brilliance, which surpasses conventional X-ray tubes by at least three orders of magnitude. Some illustrating examples are given next.

The first successful applications of synchrotron radiation for molecular cinematography were in the fields of muscle physiology, where it has become possible to study the transitions in the diffraction pattern of muscle fibers on contraction and relaxation. Today it is possible with suitably fast detectors to obtain a muscle diffraction pattern within exposure times in the millisecond time range, thus visualizing the movement of the so-called cross bridges between filaments of the contractile muscle fibers.

Another fruitful application is in the field of liquid-crystal polymorphism, where the transitions between different structures can be observed in "real time." An example for such an experiment is given in Fig. 37, where the X-ray small-angle powder diffraction pattern of a polymorphic liquid crystalline lipid specimen is shown during a fast temperature scanning experiment. The individual diffraction patterns are obtained from a series of subsequent exposures of 100 msec duration. These results gave the first unambiguous evidence that the transitions between the phases illustrated in Fig. 37 are very fast and highly cooperative. The time-resolution in this type of experiment has recently been increased into the 1-msec range, whereby the existence of structural intermediate states in fast, IR-laser-induced temperature jumps could be demonstrated for the first time.

Similar applications have been reported from the solid high-polymer area. An illustrating example is the thermal

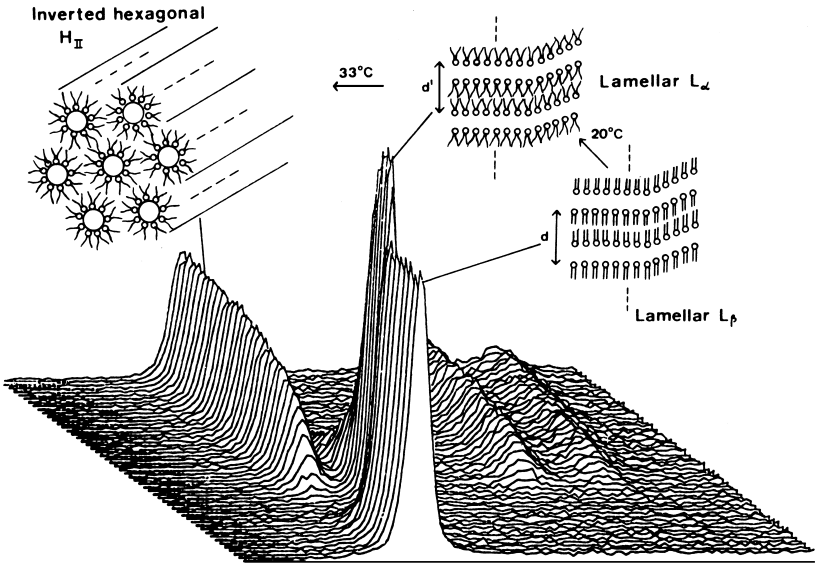


FIGURE 37 Time-resolved X-ray small-angle scattering on a polymorphic liquid-crystalline lipid during a temperature scan.

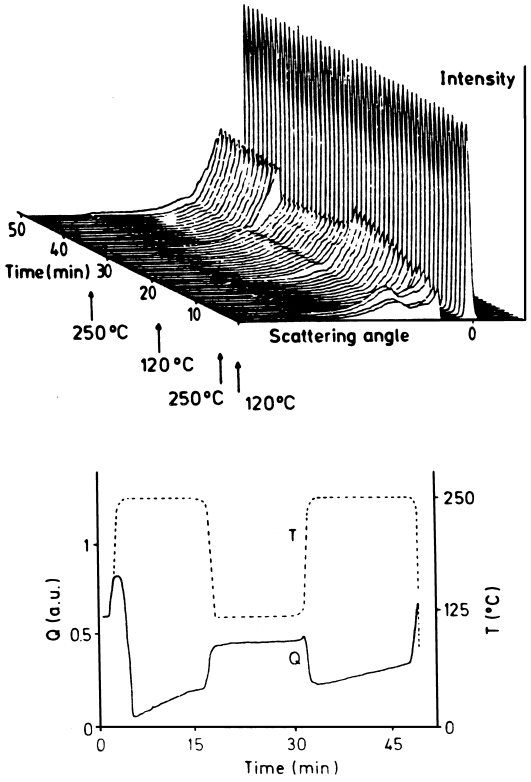


FIGURE 38 Small-angle scattering of unoriented polyethylene terephtalate during stepwise heating and cooling. [Reprinted with permission from Elsner, G., Riekel, C., and Zachmann, H. G. (1985). *Adv. Polymer Sci.* **67,** 1.]

melting and crystallization behavior of polyethylene terephthalate (Fig. 38). Partially crystalline samples were heated up to different temperatures below the melting point and cooled down again with rates of about 100 K/min. Up to 240°C one observes that the invariant $Q = \int I(h)h^2\,dh$ increases simultaneously with heating and decreases with cooling. This can be attributed to changes in density differences $\Delta\rho$ between crystalline and noncrystalline regions due to their different thermal expansion. Above 245°C one observes additionally partial melting and recrystallization. Immediately after heating, Q is seen to increase for the same reason as above, followed by a rapid decrease due to partial melting. This is followed again by an increase caused by recrystallization. In addition to the invariant Q, the small-angle maximum also changes in a characteristic manner. Small-angle scattering with synchrotron radiation thus holds considerable promise for studies on the mechanism and kinetics of polymer crystallization and melting.

However, the method is not confined to ordered structures: some reports on real-time structural studies on macromolecular solutions have also appeared. One example are the studies on the polymerization of the protein tubulin. Tubulin is an important constituent of the cytoskeleton that governs the shape of living cells, and as such plays a crucial role, for instance, in the process of cell division. During this process, fibers of tubulin polymers are built, degraded, and rebuilt. This process can be reproduced in the test tube. A time-resolved X-ray small-angle scattering experiment during thermally induced cycles of polymerization and depolymerization has revealed the existence of rings and small cylinders as intermediate structural units (Fig. 39). Due to the complexity of the system, the structural information directly obtainable from the data is relatively limited. However, the results obtained so far exclude some mechanisms previously proposed. Small-angle scattering with synchrotron radiation fills an important gap between the very crude information from light scattering and the very detailed static results from conventional SAXS experiments and other techniques.

III. LONG-CHAIN MOLECULES IN SOLUTION

A. Special Features of the Scattering Curve

Although the principles outlined for particle scattering are largely also applicable here, the fact that the molecules constitute statistically coiled fibers necessitates additional considerations. The calculation of the scattering curve of such a structure starts from the fact of statistical changes in a direction along the fiber axis. This implies a continuously coiled chain, for which the term "wormlike chain" has been generally adopted. For the innermost part of the scattering curve, below $Rh \sim 1$, the concept of the Guinier approximation remains valid, and the average radius of gyration of the entire coil can be evaluated as described above. For larger angles, however, where due to the general law of reciprocity the smaller structural elements become dominant in scattering, the conformation along the chain determines the interference function. Finally, the curve approaches that of a needle, that is, $I \approx 1/h$, since within very small segments the chain is always more or less straight if the average radius of curvature is large compared to the diameter of the chain.

Figure 40 gives a schematic representation of the scattering curve from a statistically coiled chain. In the case of very long chains, the intermediate portion of the scattering curve follows the course $I \approx 1/h^2$, according to Debye. An important parameter is derived from the abscissa point h^*, where the $1/h^2$ course passes over into the $1/h$ course. This angle h^* depends on the degree of coiling: the more extended the molecule, the larger are the regions along the chain that are still needlelike, and consequently the $1/h$ course will start at relatively small angles. Quantitatively this is characterized by the persistence length a, defined

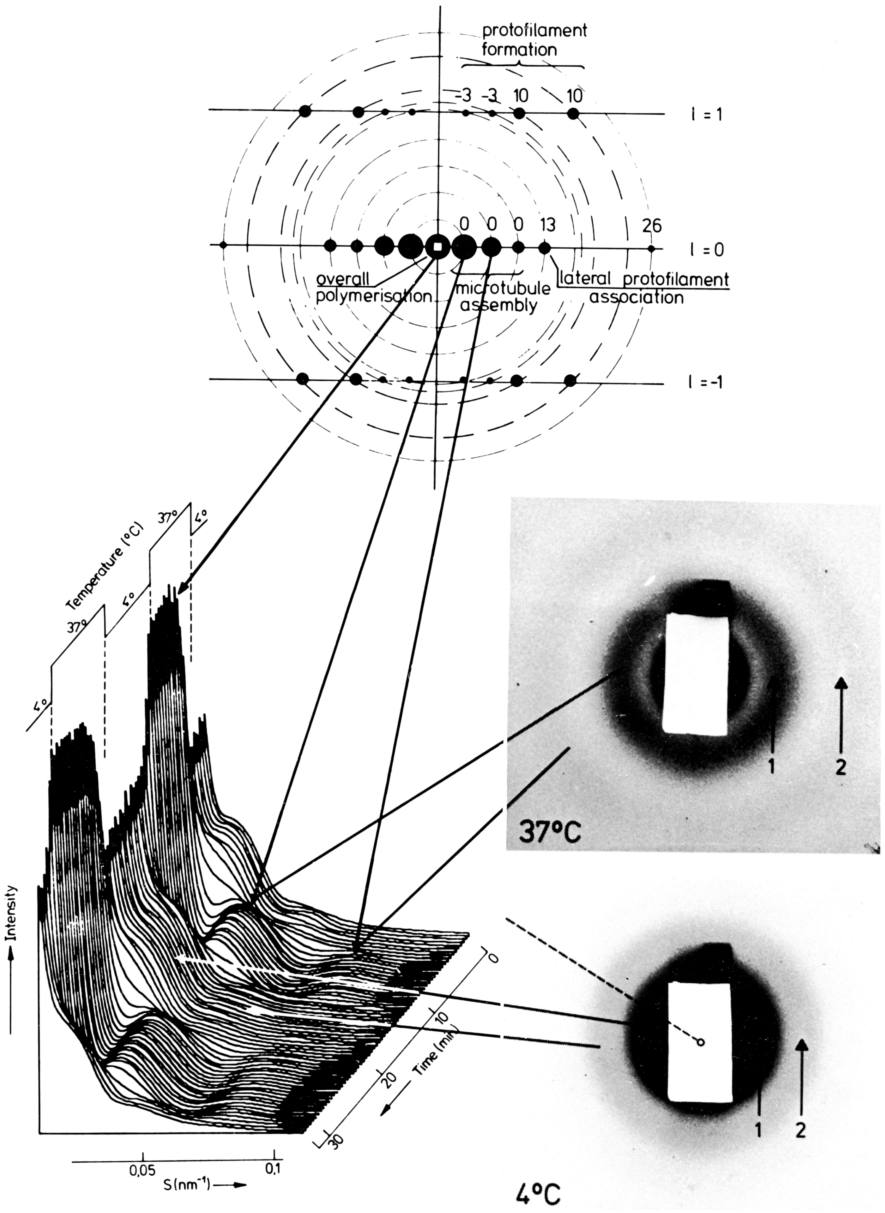


FIGURE 39 (Left) Time-resolved X-ray small-angle scattering of temperature-induced assembly and disassembly of microtubules. The corresponding indexed fiber pattern of oriented microtubules is shown on the top. (Right) Solution scattering patterns of microtubule protein. The side maxima arise from rings and microtubules, respectively. [Reprinted with permission from Mandelkow, E., Mandelkow, E.-M., and Bordas, J. (1983). *Trends Biochem. Sci.* **8,** 374.]

as the average distance, starting from any point along the chain, within which the directional cosine decays to $1/e$ (a is half of W. Kuhn's statistical chain element).

Theoretical calculations have shown that the scattering curves of all very large Gaussian coils become identical when plotted against a quantity μ defined as follows:

$$\mu = ha.$$

In such a plot, the transition point between the $1/h^2$ and $1/h$ courses lies at an abscissa value of $\mu^* = 6/\pi = 1.91$. Therefore, the persistence length can be evaluated from an experimental value h^* according to the following:

$$a = 1.91/h^*.$$

These calculations are based on the model of very long statistical coils, that is, where the length of the whole chain (i.e., its extended, "hydrodynamic" length) is a large multiple of the persistence length and where the persistence length is much larger than the diameter of the chain. These idealizations were convenient in deriving theoretical

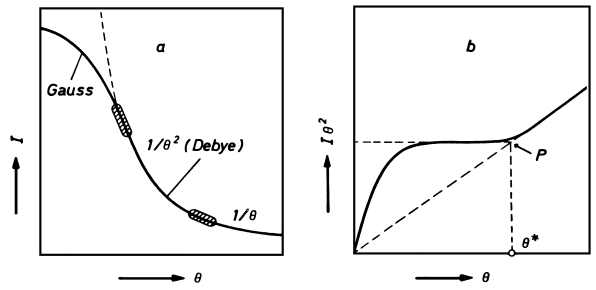


FIGURE 40 Schematic representation of a scattering curve for statistically coiled chain molecules. [Reprinted with permission from Kratky, O. (1966). *Pure Appl. Chem.* **12,** 483.]

scattering functions but are rarely met in nature. The three most important deviations from this ideal picture concern (a) the restrictions to angular and azimuthal variations between chain segments imposed by chemical bonding, (b) the limited extended length, and (c) the finite thickness of the chain. Nevertheless, the analysis of small-angle scattering can provide highly useful information on the structure of dissolved chain molecules.

It shall also be noted that the mass of dissolved long-chain molecules can be determined in the same way as described above [see Eq. (27a)] for corpuscular structures. The absolute intensities in the $1/h$ course of the outer part of the curves can be evaluated in terms of a mass per unit length, similarly to the process for rodlike particles. A comparison of this parameter to the theoretical value for length and mass of the monomeric unit leads to a degree of association.

B. Selected Applications

1. Unlabeled Chain Molecules in Dilute Solution

a. Cellulose nitrate. The first measurements on dissolved chain molecules were carried out in 1942 on cellulose nitrate in acetone. At this early stage it was surprising that small-angle scattering could be observed from the solution, while neither the pure solvent nor the dry film of cellulose nitrate showed any appreciable small-angle scattering. This was the first proof for the fact that colloidal molecules produce readily measurable small-angle scattering even though they have colloidal dimensions in only one direction and low molecular dimensions in the others.

A high-polymer sample with a degree of polymerization of 3500 and a nitrogen content of 14% shows the typical behavior of very large coils (Fig. 41). In these measurements the innermost Gaussian region could not be observed. The absolute intensity in the outer part leads to a mass per unit length—determined in the same way as with rodlike particles—which corresponds precisely to that of the known structure, that is, the degree of association is exactly 1. At the time of these experiments, this too was a heavily disputed finding, although strongly advocated by Staudinger. The intersection between the middle zone and that coinciding with the needle model moves to smaller angles with decreasing concentration, indicating that the molecule becomes increasingly stretched. Extrapolation to zero concentration leads to a value for the persistence length of $a = 110$ Å, which suggests a very stiff conformation. This is in good agreement with other findings from light scattering and viscosimetry.

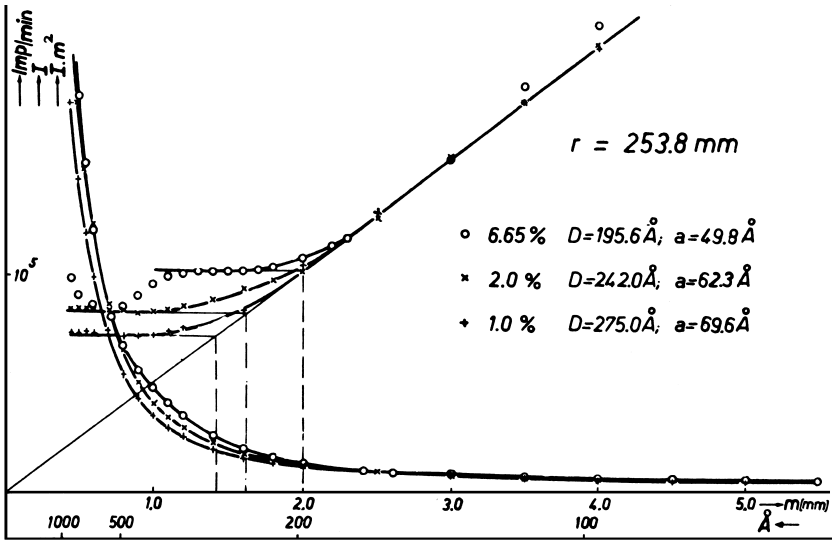


FIGURE 41 Small-angle scattering curves of cellulose nitrate in acetone solution. [Reprinted with permission from Heine, S., Kratky, O., Porod, G., and Schmitz, P. J. (1961). *Makrom. Chem.* **44,** 682.]

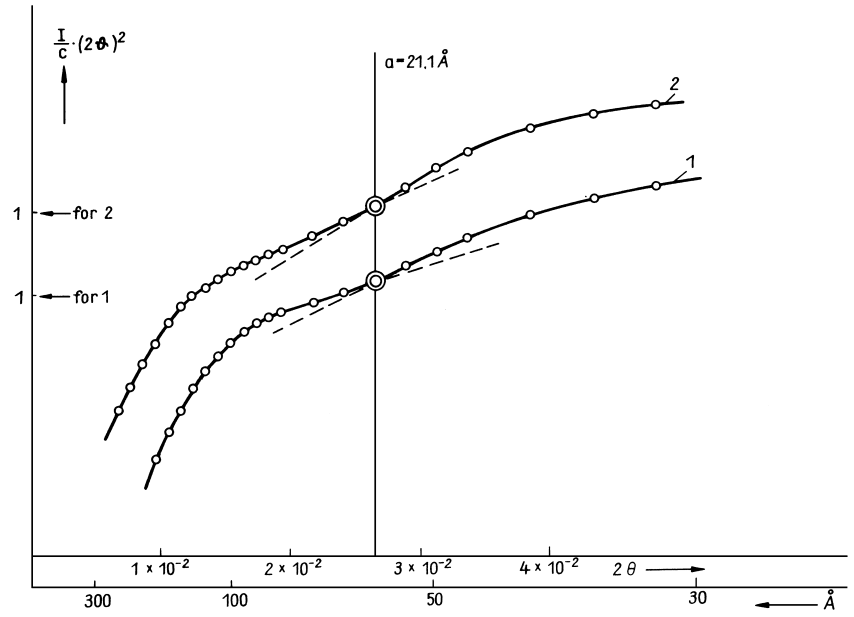


FIGURE 42 Plot of $I \times (2\theta)^2$ versus 2θ for a solution of heparin. [Reprinted with permission from Stivala, S. S., Herbst, M., Kratky, O., and Pilz, I. (1968). *Arch. Biochem. Biophys.* **127,** 795.]

b. Heparin. This substance is pharmacologically used in the prevention of blood coagulation. Molecular weight determinations on the basis of absolute intensity measurements have yielded $M = 12{,}900$; this value agrees well with ultracentrifugal results ($M = 12{,}500$).

The mass per unit length has also been determined on the basis of absolute measurements, $M_c = 54.5$; this is in good agreement with the theoretical value of 52.7 from the known chemical structure. By division of M by M_c, one obtains the length of the molecule: $L = 237$ Å.

The plot of $I \times (2\theta)^2$ versus (2θ), shown in Fig. 42, exhibits a significant kink, leading to a persistence length a of 21.2 Å. The molecular length therefore is the $x = 237/21.2 = 11.2$-fold of the persistence length. Figure 43 shows the theoretically calculated relationship between $x = L/a$ and $y = R/a$, under the condition of a Gaussian coil; hence with the knowledge of x and a, R can be immediately determined. The value of $R_{calc} = 35.9$ obtained in this way is in very good agreement with the Guinier value of $R_{exp} = 33.8$ Å obtained by extrapolation to $c \to 0$.

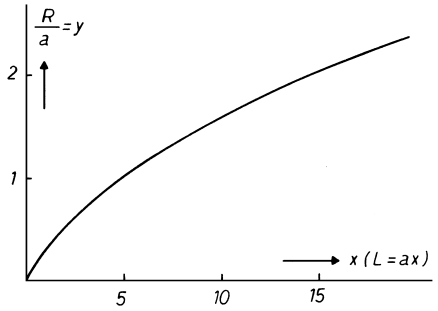


FIGURE 43 The correlation between $R/a = y$ and $L = x$ shown in this graph has been calculated on the assumption of a Gaussian coil. [Reprinted with permission from Heine, S., Kratky, O., and Roppert, J. (1962). *Makromol. Chem.* **56,** 150.]

Heparin can therefore be considered as a prime example of a Gaussian coil, where the mass, the overall dimensions, and the degree of coiling (shape) can be determined from small-angle scattering experiments.

It shall be noted that R and M, and consequently, with known chemical structure, also L can alternatively be determined by light scattering. The determination of the persistence length, however, can only be done through the relation $3R^2 = La$; this determination is based on the assumption that any association is absent and that a Gaussian coil prevails. The determination of a from the position of the kink in the plot of $I \times (2\theta)^2$ versus 2θ by the small-angle method is independent of these two conditions. It can be considered a further advantage of the small-angle method that from the value of M_c it allows us to infer associations, helixlike structures, or the existence of other deviations from statistical coils.

2. Chain Conformation Analysis by the Use of Markers

The method for distance determination outlined in Section II.B.5 can also be applied to chain molecules in solution. If it is possible to prepare samples containing

electron-dense markers on one end as well as on both ends of the chain, the end-to-end distance can be determined. The first investigation of this kind was performed on *n*-dodecane labeled with iodine at the terminal methyl groups. The length determined in this way (12.4 ± 0.3 Å) was practically the same as that of the extended molecule, which lies between 12 and 14.6 Å depending on the configuration of the terminal CH_2J groups.

Despite its correctness in principle, this method is limited to rather short chains since the contrast between markers and background decreases with increasing molecular weight of the polymers. Therefore, the applicability of this method in practice has been rather limited.

IV. POLYMERS IN THE SOLID STATE

A. General

As indicated in the introduction, polymers in the solid state generally have to be treated as densely packed systems for the purpose of small-angle scattering analysis. In this case the concepts outlined above for dilute solutions of particles cannot be unconditionally applied, since the basic tenet, that the scattering intensities for individual particles simply add to give the scattering of the whole system, no longer holds. This is due to the fact that the distances between individual zones of similar electron density are not sufficiently large as compared to their sizes, so that interference effects *between* these zones become dominant over the interferences from *within* the zones. The general treatment of dense systems, therefore, must consider the whole scattering sample as one huge particle for which only statistical system parameters can be evaluated.

With regard to their supramolecular structure, two types can be distinguished: noncrystalline and semicrystalline systems.

In *noncrystalline* polymers, the chain molecules are present in more or less randomly coiled form, similar to their solution structure but in dense packing. Most polymers exist in semicrystalline form, which to a good first approximation can be treated as a two-phase system consisting of regions with crystalline chain packing and other, "amorphous," regions, where the chains are less well ordered. From the arrangement of these regions within the system, two subclasses can be defined: the fringed micellar and the lamellar structures. The first type refers mainly to the natural, "grown" high polymers, while the second type occurs predominantly in synthetic materials. This classification into these three groups, amorphous, semicrystalline natural, and semicrystalline synthetic, is maintained throughout the following presentation, independently of the chemical type of structure. There are basically three possible approaches to a small-angle scattering analysis:

1. The restriction to the general, model-independent system parameters as the scattering power $\overline{(\Delta\rho)^2}$, the inner surface O/V, and the average intersection (chord) length l. While $\overline{(\Delta\rho)^2}$ is universally determinable, the latter two parameters are meaningful only under the assumption of defined interfaces between different phases.
2. Defined alteration of the system into the direction where the concepts of particle scattering become again applicable. Such methods include the contrast enhancement of a small fraction of the polymer by specific markers or by partial swelling or hydrolysis.
3. The determination of model-specific parmeters. If, from other sources, reasonable model assumptions on the polymer solid-state structure can be made, the evaluation of the scattering patterns leads to a quantification of model parameters.

Applications of these methods are given in the following chapters.

B. Amorphous Densely Packed Systems: Chain Conformation Analysis by Markers

The scattering from an amorphous concentrated polymer sample, such as melt or glass, obviously cannot be used to draw information on the chain conformation because the individual chains are closely packed and indistinguishable by their contrast. In order to enhance this contrast, parts of the polymer chains have to be marked by atoms with higher scattering power, so that a *dilute* system of the marked chains arises. Such specimens can, for instance, be prepared by mixing in the melt with subsequent solidification by cooling. A known quantity of polymer is labeled at both ends to an excess of unlabeled polymers; these contrast-enhanced chains thus become individually "visible."

The basic idea is the notion that upon adequate subtraction of the background curve from a system with unmarked molecules, the difference curve is mainly determined by the scattering of the markers. Thus, for a chain containing two markers at a distance x, the scattering function is given by

$$I \approx f^2 (\sin hx)/hx$$

where f is the atomic form factor of the marker.

However, this approach is only an approximation to the scattering of the markers, since these interfere in scattering not only with each other but also with the surrounding atoms. In order to eliminate this latter contribution exactly, one has to use the procedure described in Section II.B.5, in which two different samples have to be studied: one containing double-labeled chains, and another containing single-labeled ones, whereby both samples have to contain the same absolute quantity of labels. This implies the

condition that the second sample has to contain twice the amount of labeled chains. Frequently, however, the much simpler procedure of subtracting the unlabeled polymer is performed, which leads to practically acceptable results.

Another possibility is to label all monomers of the chain, which will lead, for example, to the radius of gyration and the shape of the scattering curve of the entire labeled chain.

An illustrative experiment is the investigation of a linear copolymer of styrene with *p*-iodostyrene with various degrees of iodination within a homologous polystyrene sample. By extrapolating the results both to zero iodination and to zero concentration of the marked chains, it was possible to eliminate both the effects of conformational changes due to iodination and the effects of interparticle interferences. The shape of the scattering curve furthermore supported the conclusion that the chain conformation in bulk polystyrene is a Gaussian coil and that the dimensions are the same as in the unperturbed chain molecules (Flory hypothesis). The concept of chain marking has been developed to its full potential in neutron scattering, where the differences between H and D atoms facilitate a practically nonperturbing and very potent contrast enhancement. The numerous studies conducted so far by this method have all confirmed the Flory hypothesis of mutually independent statistical chains in the bulk state. As a necessary condition for this, it was in particular shown that the radius of gyration R varies linearly with the square root of the molecular weight M, as it does in a θ solvent. Further corroboration of the statistical chain hypothesis came from more detailed studies on the course of the *scattering curves at higher angles* in the range of $R^{-1} < h < l^{-1}$, where l is W. Kuhn's statistical chain element, which is of the order of a few monomer lengths. The expected $1/h^2$ course, following from Debye's theory for statistically coiled chains, has been observed in many examples. Deviations from this behavior found in some cases have been attributed to helical sequences.

C. The Natural Solid High Polymers Discussed on the Example of Cellulose

The model of micellar structure has been developed from studies on cellulose, and it is likely that it represents the dominating structural principle in natural high polymers. The following brief review focuses on the technologically important regenerated cellulose, widely known under the name rayon.

Depending on the conditions of preparation, all grades are possible between very condensed systems (e.g., dry cellulose fibers) and loosened-up states; as pointed out in the Section I.E, the latter can be evaluated formally as dilute systems, despite their relatively high volume concentrations. This broad spectrum of states can be understood on the basis of the model shown in Fig. 44. It shows the existence of two domains: one in which the fibers (symbolized by the lines) are fully extended and another where order is less well developed. The former domains, the "micelles," can be considered in terms of a crystal lattice. They are mainly responsible for the wide-angle diffraction pattern, the "fiber diagram" (Fig. 2). Outside the micelles, all chain molecules continue in the form of "fringes" and most chains enter another micellar domain, so that the micellar domains become interconnected by joint chain molecules. Single chains may extend through several crystalline and amorphous domains, forming micellar strains.

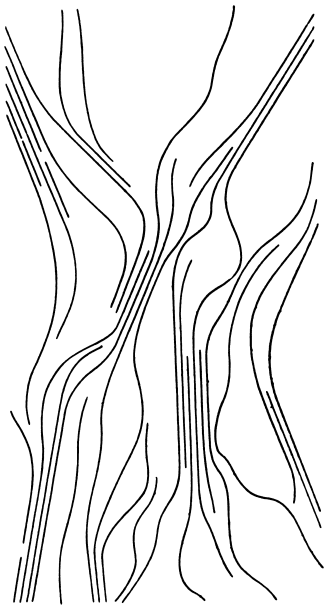


FIGURE 44 Scheme of a micellar network. [Reprinted with permission from Baule, B., Kratky, O., and Treer, R. (1941). *Z. Physik. Chem.* **B50,** 255.]

Generally, the cross section of the micelles is not isotropic since the chain molecules are arranged in a monoclinic lattice, of which the main axis coincides with the direction of the molecular chain. In the two lateral directions, the chain molecules have different tendencies for aggregation. Since the length of the micelles is much larger than the two lateral dimensions, thickness and width, and the latter are rather different, the shapes can be compared to that of a flat ruler.

1. Air-Swollen Cellulose

a. Sample preparation. A highly fruitful idea, introduced by P. H. Hermans and coworkers in 1936, was to loosen up regenerated cellulose in such a way that the micellar strains become isolated from each other so that they obtain the character of independent particles. In practice

this is done by immersion of the freshly precipitated highly hydrated cellulose fibers in an organic solvent for longer periods of time. By subsequent evaporation of the organic solvent, one obtains a porous gel containing the originally water-filled internal phase as air-filled voids. By this method, fibers with a degree of "air-swelling" between 1.06 and 6 can be prepared. In the following, we start from the plausible assumption that during this procedure the parallel order of neighboring micelles, originally existent in the water-swollen state, has largely been lost approaching a disordered state. This feature is of decisive importance to the evaluation of small-angle scattering, since rodlike or planar particles are practically free from interparticle interference even at rather high concentrations, as long as they are present in a disordered array.

b. Determination of the cross section of the micelles from the scattering curve of loosened-up samples. With three samples of the kind just described, there has been an attempt to determine the limiting slope in the Guinier plot of the cross-section factor (Fig. 45). With the sample of highest degree of swelling ($q = 5.98$), an excess scattering occurs in the innermost part apparently originating from smaller associates, which we term *clusters*. The sample with $q = 2.2$ show a remarkable interference effect, which, however, is confined to the innermost part of the scattering curve. For the sample with $q = 1.16$, finally, the interference effects extend to such a large angular range that it becomes impossible to unambiguously define a limiting slope without resorting to the result obtained from the examples with higher swelling. Obviously, an increasing parallel packing order will occur with increasing packing density for sterical reasons, so that increasing interparticular interference effects have to be expected for decreasing degrees of swelling.

From the tangents drawn in the plot in Fig. 45, a value of 4.7 nm is obtained for the radius of gyration of the cross-section R_c according to Eq. (18b). Moreover, the value of $(I \times h)_0$ can be determined from this plot. The cross-section areas A_c can hence be calculated according to Eq. (25). This involves the invariant Q determined from the curves obtained by Guinier extrapolation. Assuming rectangular cross section, the axes w and t can be calculated from R_c and A_c according to

$$A_c = wt, \quad R_c^2 = \tfrac{1}{12}(w^2 + t^2).$$

The approximate values thus obtained are $w = 15$ nm and $t = 6$ nm.

From the comparison of scattering curves in the double-logarithmic plots of the experimental cross section curves to theoretical curves for different axial ratios (Fig. 46) we obtain an axial ratio of approximately 0.4 for the cross section, in good agreement with the above-mentioned values.

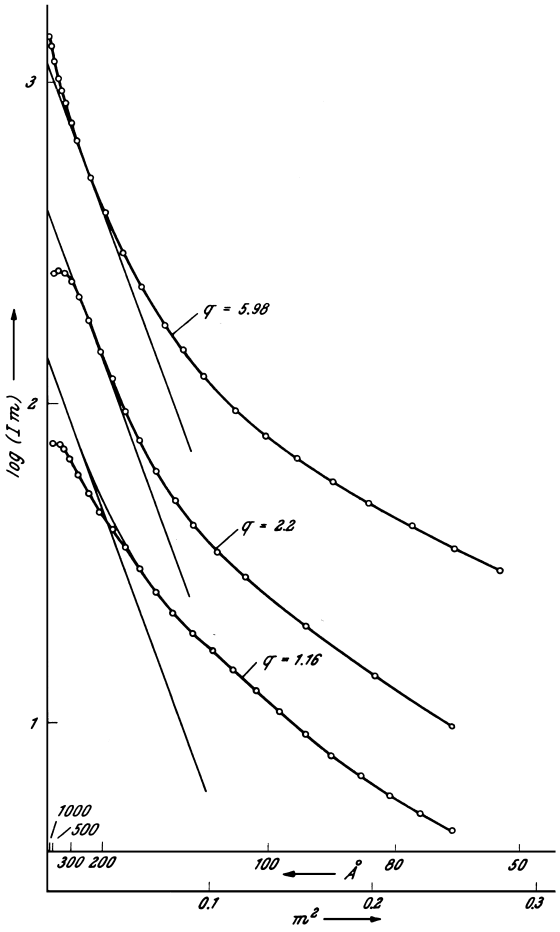


FIGURE 45 Cross-section factor of three "air-swollen" samples of regenerated cellulose in the Guinier plot. [Reprinted with permission from Kratky, O., and Miholic, G. (1963). *J. Polymer Sci.* **C2,** 449.]

So far, too little use has been made of the demonstrated possibility to extract highly informative results from loosened up micellar systems by the method of X-ray small-angle scattering. Apparently, in small-angle scattering research, the tactic familiar to every electron microscopist—that is, to optimally prepare the samples for a given investigation—has not yet found general acceptance.

2. Micellar Shape from Studies on Stretched and Rolled Solid Cellulose

The first clear-cut results obtained from solid cellulose have been concerned with the shape of the micelles as inferred from the anisotropy of small-angle scattering in oriented samples. X-ray scattering patterns obtained from cellulose and other natural fibers with the fiber axis parallel to the plane of the film have long been known (Fig. 2). From the fact that the scattering perpendicular to the long

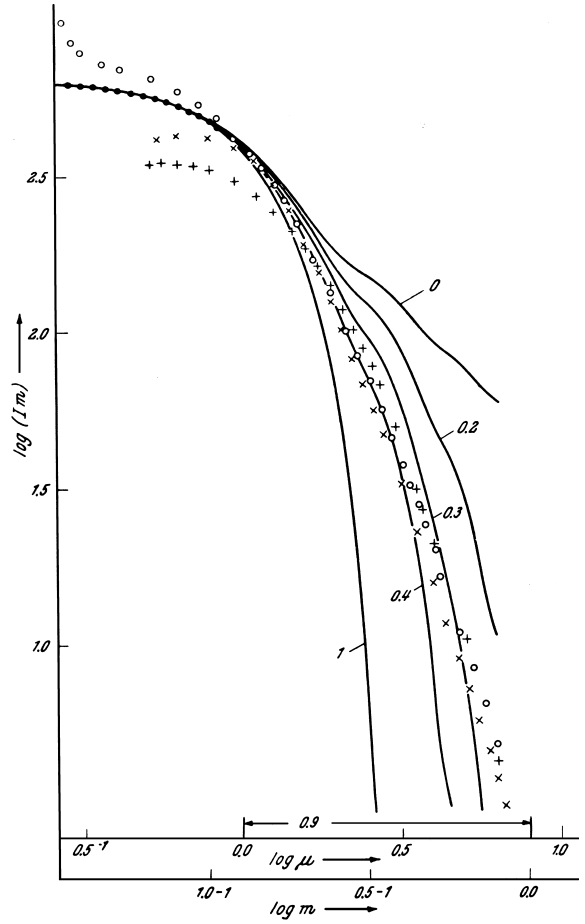


FIGURE 46 Comparison between theoretical cross-section curves for different axial ratios and experimental cross-section curves of air-swollen regenerated cellulose with the following degrees of swelling: (◯) 5.98; (×) 2.2, (+) 1.16. [Reprinted with permission from Kratky, O., and Miholic, G. (1963). *J. Polymer Sci.* **C2,** 449.]

fiber axis extends to much larger angles than does the scattering parallel to it, it can be concluded, according to the principle of reciprocity, that the scattering domains are much larger in the direction of the fiber axis than perpendicular to it.

Figure 47 adds a further highly important aspect. Foils of regenerated cellulose, as present in cellophane, can be swollen in suitable media and stretched and rolled in the same direction. It can be easily imagined that this procedure may lead to a rearrangement of the elongated structures present within the foil, in the sense that their long axis will coincide with the direction of stretching and the lateral extension will be coplanar with the rolling plane. This ordering has been termed "foil structure." A stack of such films can be oriented with respect to a point-shaped primary beam—pinhole collimation—such that the long axis of the particles, that is, the direction of stretching, is

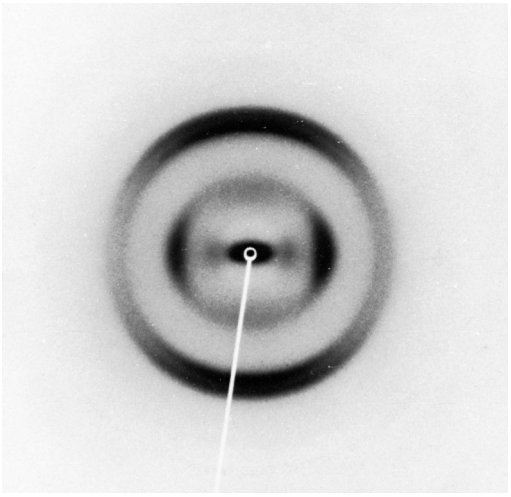


FIGURE 47 X-ray diagram of a stretched and rolled film of regenerated cellulose, irradiated in the stretching direction. The rolling plane is perpendicular to the image plane and intersects it in the vertical line. [Reprinted with permission from Kratky, O., Sekora, A., and Treer, R. (1942). *Z. Elektrochem.* **48,** 587.]

parallel to the beam direction. With the beam falling vertically on the plane of registration, the rolling plane also is vertical to the latter. If the plane of the foils cuts the registration plane in the vertical central line, a scattering pattern as shown in Fig. 47 is obtained. Clearly, both the wide-angle and the small-angle diagrams are anisotropic. The crystallographic lattice planes that coincide with the plane of the foils lead to a wide-angle reflection with its center in the horizontal central line of the image plane and a circular extension depending on the degree of disorder about the ideal position. The wide-angle reflection corresponding to the crystallographic plane parallel to the long dimension and perpendicular to the rolling plane is rotated by 90° with respect to the equator. It is easily understood that the small-angle scattering, which reflects, in the image plane, the width and thickness of the micelles reciprocally, is confined to smaller angles in the direction of the larger width (extending vertically) than at right angles to it in the direction of the smaller thickness. While a normal bundle of fibers leads to an isotropic scattering pattern upon irradiation in the direction of the long axis, since widths and thicknesses of the rulerlike cross section are equally distributed over all directions about the fiber axis, a sample with foil structure shows immediately the difference in dimensions perpendicular to the fiber axis.

3. Longitudinal Periodicity in Cellulose Fibers

Neutron small-angle scattering offers an elegant approach to the semicrystalline nature of cellulose. Partial deuteration can be achieved by the equilibration of a fiber

sample with heavy-water vapor. The extent of hydrogen–deuterium exchange in cellulose depends on the accessibilities of the ordered and less ordered regions within the fibers. Therefore, any periodicities of micellar and amorphous regions become better visible. Such studies can be aimed at (a) detecting the meridional long spacing of native and regenerated cellulose, which is generally not visible in the X-ray small-angle pattern due to insufficient contrast between "amorphous" and "crystalline" regions, and (b) evaluation of the lateral width of the protofibrils and their arrangement in the cellulose fibers.

While native cellulose fibers (Ramie) do not show a long-spacing reflection even upon prolonged equilibration with heavy water vapor, this treatment with samples of regenerated cellulose (Fortisan and rayon fibers) lead to the appearance of a strong meridional reflection around Bragg's spacings of 165 and 193 Å, respectively. The complete two-dimensional intensity distribution of a neutron scattering experiment on Fortisan fiber is shown in Fig. 48.

The diffuse equatorial scattering and the meridional layer line appearing upon deuteration can be quite clearly recognized. This provided unequivocal evidence for a longitudinal periodicity within the supramolecular structure of regenerated cellulose fibers.

The intensity distribution in the horizontal direction of the 165 Å layer line in Fig. 49 can be analyzed in terms of the radius of gyration of the lateral fibril dimensions. This implies the assumption that no interference occurs between neighboring fibrils. An example of such a Guinier plot is shown in Fig. 49. The nonlinear inner part clearly indicates that the system cannot be analyzed as if it were monodisperse. However, the outer linear part yields a radius of 17 Å, in excellent agreement with other, ultrastructural studies. The inner, nonlinear part can be

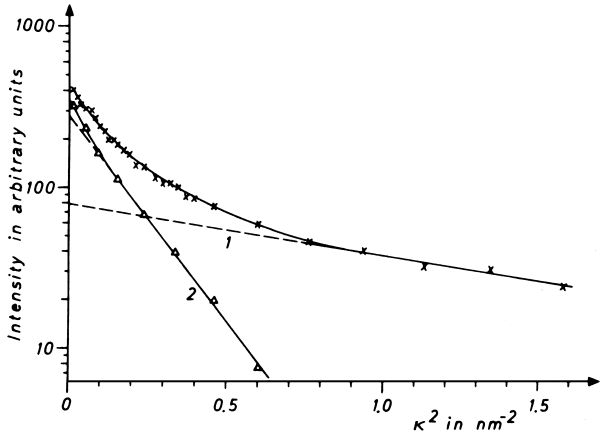


FIGURE 49 Guinier plot of the scattered intensities along the layer line ($\kappa_3 = 0.38$ nm^{-1}) in the horizontal direction. Deuterated Fortisan. [Reprinted with permission from Fischer, E. W., Herchenröder, P., Manley, R. S. J., and Stamm, M. (1978). *Macromolecules* **11,** 213.]

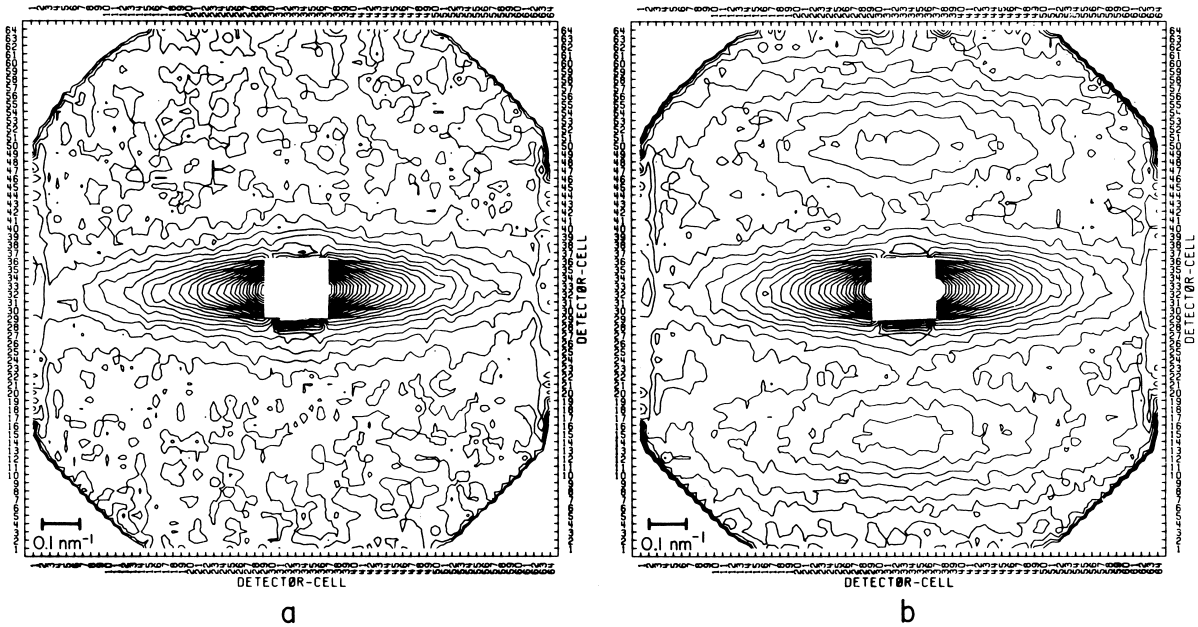


FIGURE 48 Two-dimensional intensity distribution of the neutron scattering pattern of regenerated cellulose (Fortisan; logarithmic scale of equiintensity curves): (a) untreated and (b) deuterated for 5 hr. Fiber direction vertical. [Reprinted with permission from Fischer, E. W., Herchenröder, P., Manley, R. S. J., and Stamm, M. (1978). *Macromolecules* **11,** 213. Copyright 1978 American Chemical Society.]

constructed by adding a second component with a circular cross section of 98 Å diameter. By recognizing the equality $98 \approx 35\sqrt{8}$, it was suggested that this dimension may be due to a superstructure consisting of eight protofibrils. The relative distribution of single and octameric fibrils can be estimated from the intensity ratio at zero angle, taking into account that these intensities should be proportional to the square of the cross section of each fibril. This estimation yielded a relative content of 71% in protofibrils.

A similar analysis of the horizontal intensity decay of the equatorial diffuse scattering has yielded quite different values: a linear outer part corresponding to 22 Å diameters with two additional components of 110 and 246 Å, respectively, to account for the nonlinear inner part. To explain this disagreement, it was suggested that the equatorial diffuse scattering is due to a "dilute" system of microvoids or holes in an otherwise dense system rather than to the microfibril dimensions. By performing a model scattering experiment on a mask consisting of lozenge-shaped holes (see Fig. 50), the complete two-dimensional intensity distribution of cellulose (Fig. 51) has been qualitatively reproduced.

4. Pore Analysis

The small-angle method is very well suited to perform a pore analysis, that is, to determine the fraction of voids within a solid object. This is again demonstrated with the example of cellulose.

In a two-phase system, the absolute value of the invariant Q/P_0, often called the scattering power, is proportional to the volume fractions w_1 and w_2 and the square of the electron density difference, $(\Delta\rho)^2$. It is important to consider that $\Delta\rho$ between crystalline and amorphous domains corresponds to about 10% of the crystalline domain. Obviously the maximum value of Q/P_0 will be found at $w_1 = w_2 = 0.5$. If, however, the experimental value is larger than that, it is tempting to consider voids as a third phase (with $\rho_3 = 0$) to be responsible for this discrepancy. If, for the sake of demonstration, the small difference between ρ_1 and ρ_2 is neglected and the system is considered as two-phase (i.e., cellulose and voids), $(\Delta\rho)^2$ takes about 100 times the value of that between crystalline and amorphous cellulose. It follows that a very small fraction of voids is sufficient to obtain an invariant that is larger than the maximum value for the contrast crystalline—amorphous.

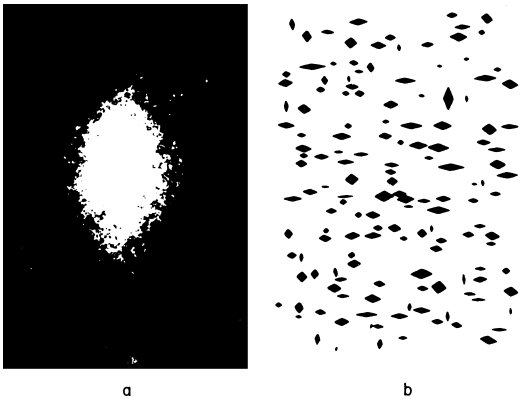


FIGURE 50 Model experiment by laser light scattering: (a) scattering mask and (b) scattering pattern. [Reprinted with permission from Fischer, E. W., Herchenröder, P., Manley, R. S. J., and Stamm, M. (1978). *Macromolecules* **11,** 213.]

The fact that in a two-phase system the scattering power indeed is proportional to the square of the electron density difference was experimentally verified by using air-swollen cellulose in which the voids were filled with different solvents. It is also interesting that at the point of equal electron density of cellulose and solvent, the small-angle scattering completely vanishes: the micelles become "invisible" to the X-ray beam. This is the case if ethyl iodide is used as swelling solvent. The phenomenon is related to the observation that a suspended substance in a liquid can become invisible if the solvent and solute have the same refractive index.

If one extends the treatment to the general case of a three-phase system, Eq. (8) must be used.

In experiments on a series of cellulose samples on an absolute scale, the correlation between macroscopic density and scattering power is shown in Fig. 52. Taking the crystalline volume fraction w_1 as 0.4 and the amorphous w_2 as 0.6 (based upon the analysis of the wide-angle pattern), the point 662 in Fig. 52 corresponds exactly to the scattering power expected for a two-phase system, according to Eq. (7), whereas all other samples show a much higher scattering power, which can only be explained by an additional phase, namely that of the voids. With Eq. (7), a void fraction of 0.78% can be calculated for point 641, while the reduced total density (1.45 versus 1.5) indicates a void fraction of 3.5%. Since the measurements have only been extended up to Bragg distances of 300 Å, it can be assumed that at even smaller angles a strong scattering exists, which would contribute to the absolute value of the invariant. In other words, the largest voids are not expressed by the measured invariant.

Similar experiments have later been performed on 60 types of rayon, extending the angular scale to Bragg values of 2000 Å. Figure 53 shows that the outer part corresponds exactly to $1/h^3$ according to Porod's law. The scattering intensity decays in this region by almost an order of magnitude. The calculated pore fraction varies between 0.04 and 7%.

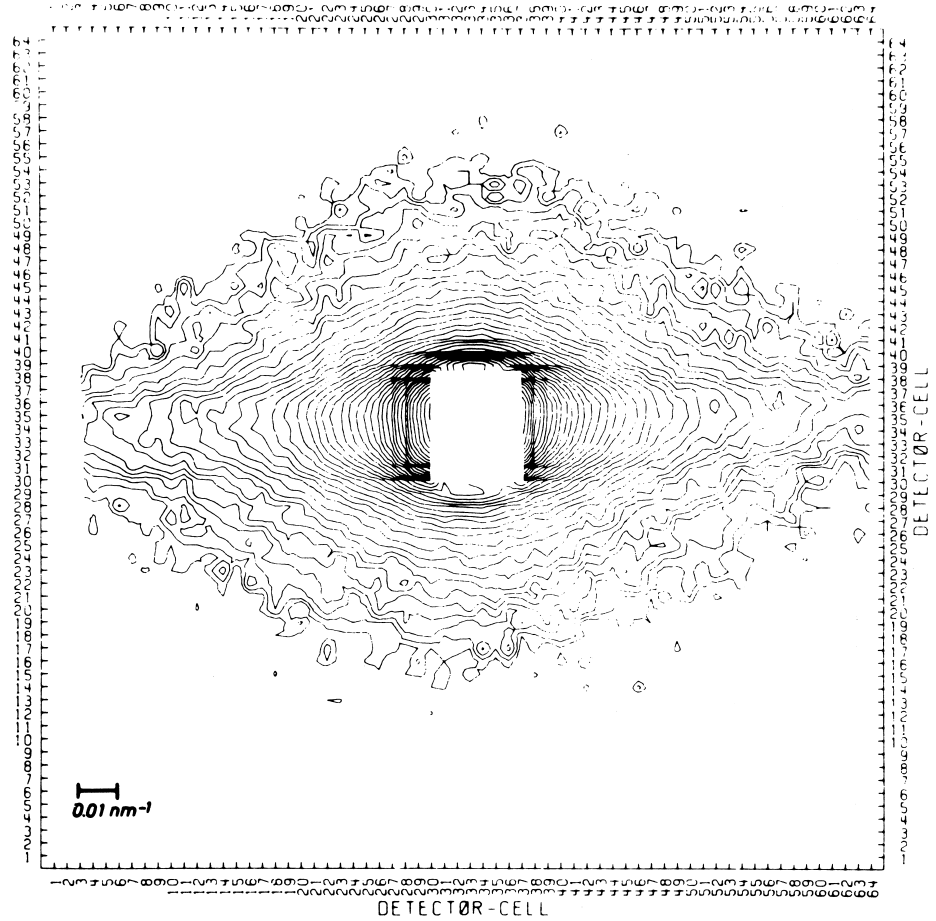


FIGURE 51 Two-dimensional intensity distribution of the diffuse central neutron scattering of native cellulose (Ramie). [Reprinted with permission from Fischer, E. W., Herchenröder, P., Manley, R. S. J., and Stamm, M. (1978). *Macromolecules* **11,** 213.]

Many experiments have been made to demonstrate a correlation between pore content and textile properties, with some success.

D. Synthetic High Polymers in the Solid State

An ideally crystallized low-molecular-weight substance, leads to a multitude of very sharp reflection spots without diffuse background. If any crystal reflections can be seen from a solidified high-polymer sample, it can be concluded that some crystallization has occurred. At the same time, the small number and relatively large half-width of observed reflections indicates that the crystal domains are small and strongly perturbed. From the very presence of a diffuse background, it follows that the crystalline regions are accompanied by amorphous domains. If the melt is allowed to solidify after some parts of the material have crystallized, the mobility of the not-yet-crystallized chains becomes so small that they can no longer lock into ordered structures. Thus the final solid state of a macromolecular substance consists of a two-phase system with crystalline and noncrystalline domains.

Information on this semicrystalline structure can be obtained from X-ray small-angle scattering. A diffuse scattering pattern is observed that decays with increasing angles. Onto this background is superimposed a discrete reflection, which originates from periodic density fluctuations within the material relating to a periodic array of crystalline and noncrystalline regions. From the position of this reflection one can determine the average distance between the centers of the crystallites. This long period generally lies between 50 and 500 Å, hence in the small-angle region.

What models can be used for an interpretation? The oldest model is that of the fringed micelle, which dominates in natural high-polymeric fibers as discussed in Section IV.C.2. With polymers crystallized from the melt or from solution, it appears that the model consisting of lamellae with folded chains achieves higher importance.

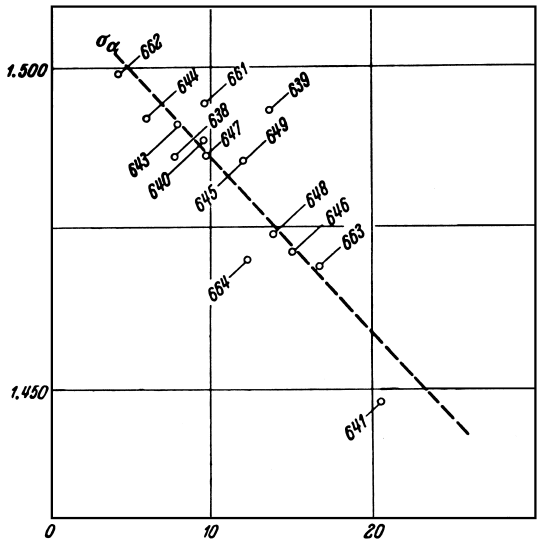


FIGURE 52 Connection between mass density and scattering power for various rayon samples. [Reprinted with permission from Hermans, P. H., Heikens, D., and Weidinger, A. (1959). *J. Polymer Sci.* **35,** 145.]

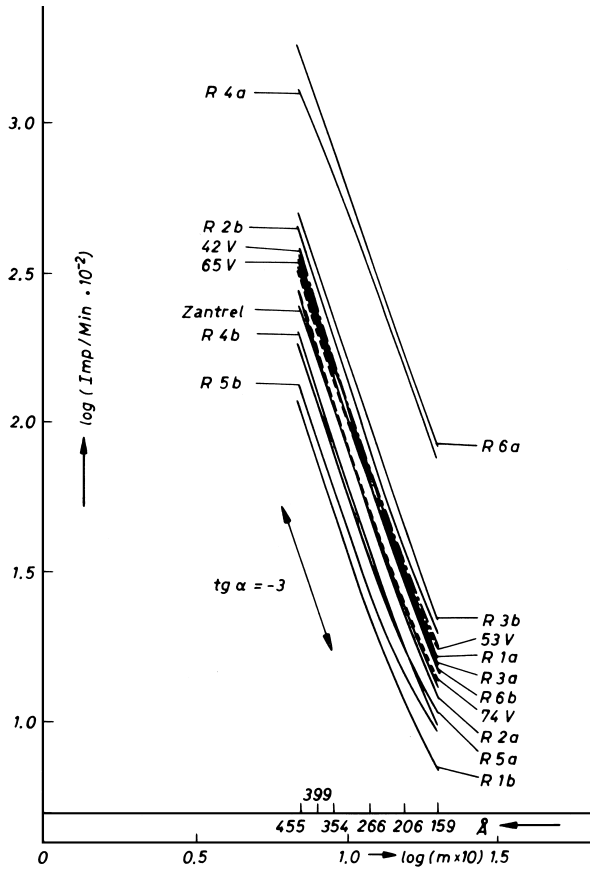


FIGURE 53 Tail end of the scattering curves of rayon fibers, log–log plot. [Reprinted with permission from Kratky, O. (1966). *Pure Appl. Chem.* **12,** 483.]

This model originated from a discovery independently made by E. W. Fischer, A. Keller, and P. H. Till. In electron-microscopic preparations of certain synthetic polymers, they could observe platelets that were very well developed. By X-ray diffraction it was shown that the polymer chain axes are arranged perpendicular to the plane of the platelets. Since the length of the chains, however, is orders of magnitude larger than the lamellar thickness, it must be concluded that the *chains are folded*: they reenter the crystal by U-turns at the surface. Some molecules may extend into the interlamellar space or into the neighboring one, thus forming a link between the lamellae.

1. Introduction to the Quantitative Treatment of the Lamellar Stack Model

The ideas just discussed can also be extended to solid synthetic high polymers and are schematically shown in Fig. 54. As a starting point in this approach, a one-dimensional model consisting of arrays of alternating crystalline and amorphous layers, as shown in Fig. 55, can be used.

Many alternating layers form the "lamellar stack," which is large as compared to the average repeat distance d, and hence the scattering arising from the entire stack falls into a much smaller angular range than that originating from the periodic density variations. In a macroscopic sample these lamellar stacks may be present in all orientations leading to an isotropic, spherically symmetrical scattering pattern, similar to a powder diffraction pattern. The scattering function of the single stack is obtained by multiplying the measured intensities by h^2 (Lorentz correction).

In an ideal, one-dimensional lattice, the scattering is described by Bragg's law, representing the lamellar repeat distance d. However, in reality, the diffraction pattern is significantly broadened, indicating more or less strong deviations from the ideal lattice. The aims of a small-angle scattering analysis are, therefore, the *determination of average structure parameters* and the quantitative description of the deviations.

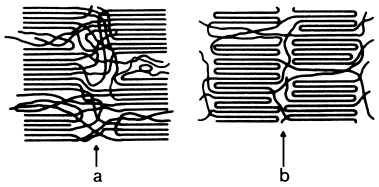


FIGURE 54 Models of supramolecular structures in the case of folded chains. [Reprinted with permission from Zachmann, H. G. (1974). *Angew. Chem.* **86,** 283.]

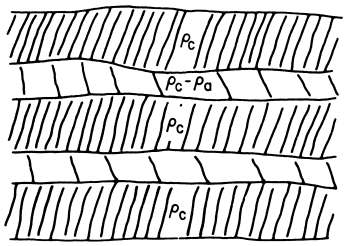


FIGURE 55 Model of a fiber with lamellar crystallites that are separated by amorphous domains of lower density. [Reprinted with permission from Fischer, E. W., Goddar, H., and Schmidt, G. F. (1968). *Makromol. Chem.* **118,** 114.]

Generally, three approaches can be distinguished: (a) the determination of characteristic system parameters or functions from the experimental results, (b) the comparison of model intensity calculations with the experiment, and (c) the analysis of real-space correlation functions.

The average repeat distance d, that is, the sum of the average thicknesses of crystalline and amorphous layers can be obtained from the position $(2\theta)_{100}$ of the first-order maximum in the Lorentz corrected scattering curve according to Bragg's law,

$$D = \lambda/(2\sin\theta)$$

The degree of *crystallinity* can be estimated in a two-phase model from the invariant Q [see Eq. (7)] according to the following:

$$Q/P_0 = w_c(1 - w_c)(\rho_c - \rho_a)^2.$$

If the two phases are connected by a transition layer in which the electron density varies linearly from ρ_c to ρ_a, the above equation can be modified into

$$Q = [w_c(1 - w_c) - (ES/6V)](\rho_c - \rho_a)^2,$$

where E is the thickness of the transition layer and S/V is the specific surface of the phase boundary, which is defined as the plane where $\rho = (\rho_c + \rho_a)/2$.

The specific surface S/V may be determined from *Porod's law* according to Eq. (11).

Finally, the thickness E of the transition layer may be calculated from the one-dimensional correlation function γ obtained by Fourier transformation of the one-dimensional intensity function.

The above parameters can either be useful in monitoring structural changes during thermal or mechanical treatment of the polymer sample, or in parameterizing structural models, which then have to be further fitted to the experimental data by trial and error.

The tedious and error-prone procedure of finding a scattering-equivalent model can be considerably simplified by performing the comparison between experimental result and model in real space through the respective correlation functions $\gamma(x)$ (Fig. 56). The terms $\gamma(x)$ and $I(h)$ are Fourier pairs connected by the following relation:

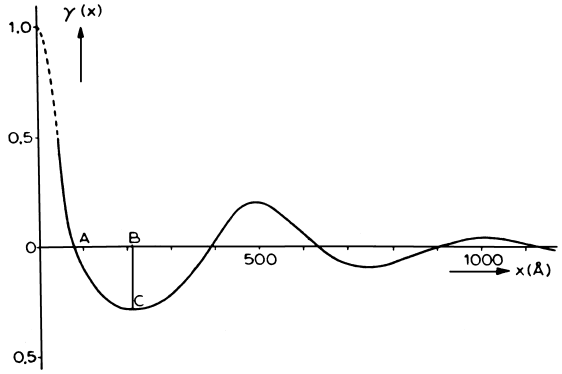


FIGURE 56 One-dimensional correlation function of a sample of bulk polyethylene. [Reprinted with permission from Vonk, C. G., and Kortleve, G. (1967). *Kolloid-Z., Z. Polym.* **220,** 19.]

$$\gamma(x) = \frac{\int_0^\infty Ih^2 \cos hx \; dh}{\int_0^\infty Ih^2 \; dh}$$

and can therefore, be readily interconversed. The correlation function $\gamma(x)$ is the convolution square of the electron density fluctuations, expressed by the following:

$$\gamma(x) = \int_0^\infty \rho(u - x)\rho(u)\, du.$$

Its physical meaning can be visualized easily as follows: consider a measuring rod AB of length u perpendicular to the layers, moving in the x direction. In each position, the product of the electron densities at A and B is determined and the values averaged over all positions; this is repeated for all values of u.

An example of a $\gamma(x)$ function for a sample of bulk polyethylene is shown in Fig. 56. This function must be matched by a theoretical $\gamma(x)$ function of a model. It has been shown that one scale factor and three parameters have to be fitted for this purpose. The scale factor accounts for the average repeat distance, and the parameters are the volume crystallinity w_c and the widths of the distributions of the amorphous and crystalline layer thicknesses, respectively. The particular form of the distribution was found to be of only secondary importance.

It should be noted that the lamellar stack model is not confined exclusively to semicrystalline high polymers but is also important in the structural description of smectic liquid-crystalline phases, such as the hydrated phases of lipids and membranes, where the alternating phases are the lipid bilayer and the water of hydration, respectively. The same basic considerations apply also there, but with the complication that the electron density of the lipid phase is not at all constant but varies strongly and characteristically across the normal of the lamellar plane.

2. Particles in Polymer Glasses and Melts

According to the general model of Flory, amorphous polymers consist of randomly packed Gaussian chains with conformation and dimensions identical to the state in a θ solvent. This implies that there exist no density fluctuations on the macromolecular scale, and therefore no small-angle scattering should be expected. Nevertheless, in some instances of polymer glasses and melts diffuse small-angle scattering has been observed, which poses the question of its structural origin.

As one possible cause for this effect, foreign substances, such as stabilizers or small amounts of catalysts left over from the polymerization process, have been identified. Also, nonidealities in the surface layers, microcracks and scratches, were found to be responsible for the small-angle scattering. The scattering at very small angles can be well described as originating from thermal fluctuations, which may be calculated from the isothermal compressibility at the glass transition point or above, depending on whether glass or melt is considered.

V. INORGANIC SUBSTANCES—MATERIALS SCIENCE

A. General

This field includes many of the common natural or synthetic materials, such as metals, alloys, semiconductors, glasses, colloids, and catalysts. In all cases, the aim of small-angle scattering is the characterization of structural inhomogeneities. These may be of highly practical interest in their relation to specific macroscopic properties, such as mechanical strength, electrical conductivity, magnetic coercive force, and catalytic activity. Control of these inhomogeneities on a more than merely empirical basis can lead to quality improvement. On the other hand, there is also strong theoretical interest, as such data provide reliable information to test theories of cluster or defect formation and phase-separation processes.

In contrast to the substances discussed in previous chapters, most inorganic materials are composed of rather heavy elements. For X-rays this has the disadvantage that their mass absorption, which increases strongly with the atomic number, becomes a limiting factor through the very small optimum sample thickness, that is, $1/\mu$. With Cu K_α radiation the optimum thickness of pure aluminum is 76 μm, and it is less than 10 μm for most heavier elements of interest. For Mo K_α the values are about one order of magnitude larger, but still the optimal values are relatively small and consequently the samples may not be always representative for the bulk material. This disadvantage does not occur in neutron scattering, since neutron absorption does not increase in a systematic way with the atomic number and the absorption cross sections are relatively low as compared to X-rays. Therefore, neutron small-angle scattering has proven particularly fruitful in materials science, despite its limitation that only a few reactors around the world offer a suitably high flux of "cold" neutrons.

Another complication for the use of X-rays in this field is the effect of double Bragg reflection from polycrystalline materials. This occurs if a ray first reflected by one crystallite hits another one with its lattice planes almost but not quite parallel to the first one. The resulting scattering to small angles may in some cases be stronger than the "true" small-angle scattering due to inhomogeneities, and the two effects cannot be distinguished. A way to overcome this is the use of neutrons of suitably long wavelengths, with $\lambda > 2d$ (d is the largest lattice plane distance), so that the geometric conditions for Bragg reflection are altogether avoided.

Owing to the ample diversity of the various materials, a simple systematic treatment is impossible. Only in a minority of cases are the conditions for particle scattering met so that parameters related to size and shape of the inhomogeneities can be reliably evaluated; in most cases, both nonuniform sizes and high concentrations of inhomogeneities render such information unaccessible. In these latter cases, the scattering power or the invariant as outlined above for densely packed systems can be used to quantify the structural or compositional fluctuations. Quite generally, however, the main emphasis of such studies has been on relative changes in scattering behavior with different composition or pretreatment of the samples, rather than on a detailed structural interpretation, which for the reasons already indicated may be a difficult if not impossible task.

B. Physical Metallurgy

The classical example for the utilization of X-ray *small-angle scattering* in this field has been the investigation of phase separation during age hardening of alloys, pioneered by a study of Guinier in 1938 on the systems Al–Cu and Al–Ag. Guinier observed diffuse small-angle scattering at the center of the otherwise normal discrete diffraction pattern of the metallic crystal lattice. Independently and almost at the same time, Preston published similar findings. The consistent interpretation of these results was the segregation of submicroscopic clusters of atoms of the minor component out of the supersaturated solid solution. Owing to their inventors, a certain type of such clusters is generally called *Guinier–Preston zones* or GP zones. Since then, numerous studies by X-rays and neutrons have been published on this theme, with the general goals being to

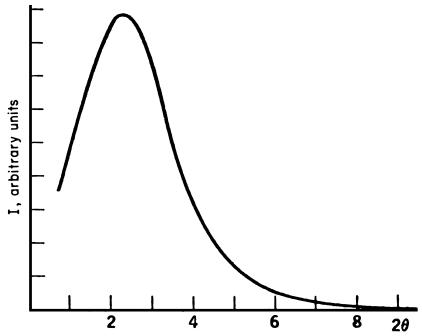


FIGURE 57 Small-angle scattering from Al(20)–Ag(80) after cooling from 520°C. [Reprinted with permission from Guinier, A., and Fournet, G. (1955). "Small Angle Scattering of X-Rays," p. 205, Wiley, New York.]

elucidate mechanisms and kinetics on the submicroscopic structural level of such phase separations and to establish detailed phase diagrams.

As a methodologically important example, we give the analysis of this first stage of age hardening.

After rapid cooling of an Al–Ag alloy from temperatures where the alloy is homogeneous to about 100°C, an isotropic, diffuse, small-angle scattering pattern as shown in Fig. 57 is observed. It shows essentially two interesting features: the scattering intensity decreases toward zero at very small angles and a maximum occurs at Bragg values of about 50 Å.

The interpretation by Walker and Guinier attributed this scattering pattern to independent particles with an internal electron density distribution. Furthermore, it was concluded that the particles have an average spherical symmetry, since the orientation of the sample has no effect on the scattering curve. The analysis can then follow along the lines of Fourier transformation as outlined in Section II.B.3.e to obtain the radial electron density distribution or the correlation function $p(r)$. If the latter is normalized to $p(0) = 1$, then the function gives the probability of finding an atom of silver at a distance r away from another one (Fig. 58).

The physical explanation for this behavior is the following. When the silver atoms cluster around a particular point in the alloy, they migrate by diffusion, but slowly, as the temperature of annealing is low. Therefore, the silver atoms clustered in a nucleus leave a shell that is poorer in silver atoms than the average. Thus the scattering particle is made up of a nucleus of a high electron density surrounded by a shell of density lower than the average throughout the sample. On average, however, the whole "particle" has the same electron density as the total sample volume, since it is only created by internal demixing. Thus the overall net contrast of the particles is zero, and consequently $I(0) = 0$, in agreement with the observation. This particularly simple concept applies, however, only to

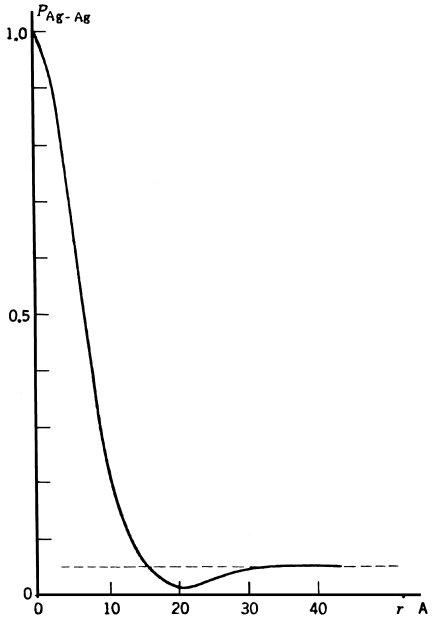


FIGURE 58 The probability of finding a silver atom at a distance r from another one in the Al–Ag alloy, as evaluated by Fourier transformation of the curve in Fig. 57. [Reprinted with permission from Guinier, A., and Fournet, G. (1955). "Small Angle Scattering of X-Rays," p. 207, Wiley, New York.]

the first stage of phase separation at low annealing temperatures, where diffusion is very slow. For longer annealing times and higher temperatures, but still below the miscibility limit in the phase diagram, the small-angle pattern often becomes very complex in that it develops directed streaks of scattering, the directions and number of which depend on the orientation of the specimen. In this case, the small-angle pattern can no longer be discussed separately from the wide-angle diffraction; that is, the scattering is strongly influenced by the crystal lattice and its defects.

It should be noted, however, that X-ray small-angle scattering has been highly useful in obtaining otherwise hard to access information on metallurgically important phenomena, such as vacancy clustering in irradiated materials, growth or dissolution of precipitates, and spinoidal decomposition.

C. Miscellaneous

1. Catalysts

Small-angle scattering provides a noninvasive method to evaluate the specific surface area of catalysts. An analysis of the Porod slopes [see Eq. (9)] of the scattering from the supporting material with and without catalyst may lead to a reliable value for the total catalyst surface.

A typical result of a study in amorphous platinum supported on silica and aluminium, in terms of a size distribution of catalyst particles which are assumed to be spherical,

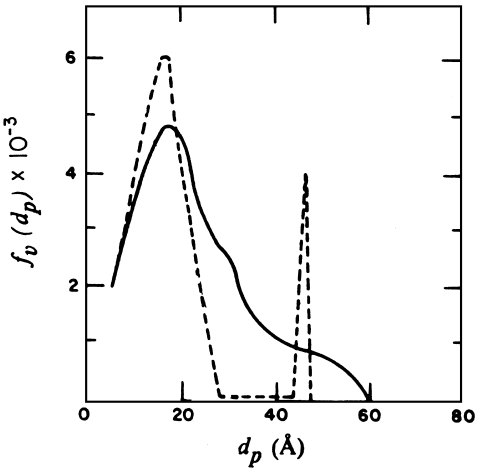


FIGURE 59 The distribution of metal particle diameters in an alumina-supported platinum catalyst containing 0.6 wt.% Pt. The continuous curve is from X-ray small-angle scattering, and the dashed curve is from electron microscopy. [Reprinted with permission from Renouprez, A., Hoang-Van, C., and Compagnon, P. A. (1974). *J. Catalysts* **34,** 411.]

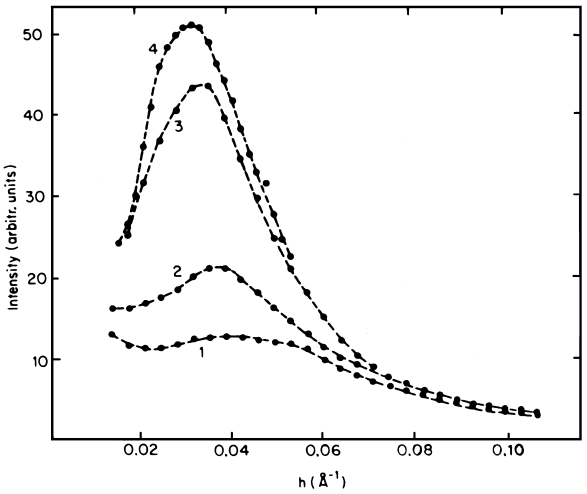


FIGURE 60 Small-angle scattering curves for a set of B_2O_3–PbO–Al_2O_3 samples splat cooled with variable cooling rate ranging from about 2×10^3 (curve 4) to 2×10^4 K/sec (curve 1). [Reprinted with permission from Acuña, R. J., and Craievich, A. F. (1979). *J. Noncryst. Solids* **34,** 13.]

is shown in Fig. 59 together with the size distribution obtained from electron microscopy. It is evident that, at least for the smaller sizes, the distributions are similar with both methods. Even if these distribution functions are not completely correct in every detail, they yield useful information, especially for comparing different metal concentrations, temperatures, and so on. As an economically interesting result, it was found that for platinum and silica, the incorporation of platinum at more than 1% by weight would be uneconomical, as beyond this concentration excess metal is primarily taken up in larger particles with a lower catalytic surface.

2. Glasses and Ceramics

Although visible-light scattering has been most frequently used to study phase-separation processes in multicomponent glasses, X-ray small-angle scattering also proves useful since it resolves compositional fluctuations at a much smaller scale of dimensions and can yield data on the nature of the decomposition process. For instance, the mixture B_2O_3(80)–PbO(15)–Al_2O_3(5) after rapid cooling from the liquid state to room temperature shows small-angle scattering curves (Fig. 60) that are very similar to those observed with the Al–Ag alloy (compare Fig. 57), especially at lower quenching rates. These data form an important basis for a discussion of the decomposition process in terms of the spinoidal decomposition model.

3. Critical Scattering in Liquids

At the critical point of a system, the density fluctuations approach infinity and hence also the small-angle scattering goes through a maximum. An example is shown in Fig. 61, where the scattering from the liquid mixture of Li–NH_3 at 210 K is depicted for different concentrations. Such experiments can be very useful in supplementing

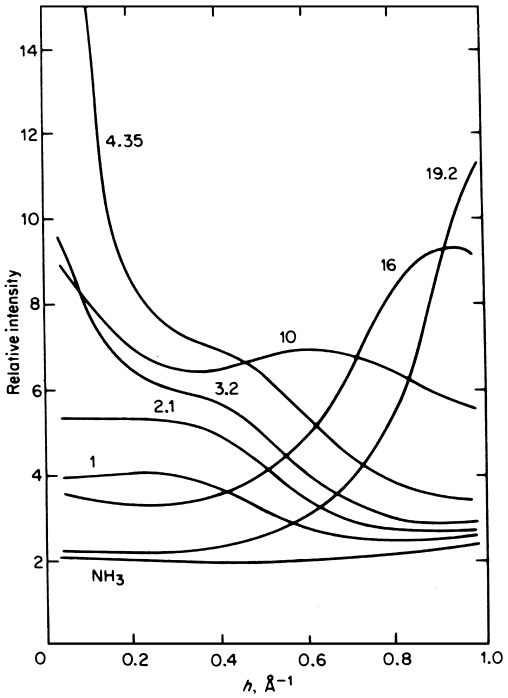


FIGURE 61 Small-angle scattering curves obtained with Mo K_α radiation of Li–NH_3 solutions of 210 K. Lithium concentrations are given in moles per mole NH_3. [Reprinted with permission from Knapp, D. N., and Bale, D. H., *J. Appl. Crystallogr.* **11,** 606.]

thermodynamic data but may be complicated in solids due to competing metastable and stable phase formation.

VI. INSTRUMENTATION

A. X-Ray Source and Camera Setup

1. Choice of Anode Material

For the vast majority of applications, an X-ray tube with a copper anode is used; the wavelength of the K_α line is 1.54 Å.

The use of longer-wavelength radiation would spread the scattering curve over a larger angle, permitting higher resolution. This possibility, however, is only very rarely used, due to the sharp increase in the absorption coefficient μ with increasing wavelength. The intensity I scattered at small angles is proportional to $te^{-\mu t}$ (t being thickness of the sample). This expression has its maximum at $t_{opt} = 1/\mu$. If μ is very high, it is often difficult to obtain sufficiently thin samples.

It must be mentioned, however, that the use of an open X-ray tube combined with a pinhole camera may overcome some of the difficulties of Al radiation ($\lambda = 8.34$ Å). With such instruments, excellent measurements on thin polymer films can be performed.

Usually the resolution obtainable with copper radiation is sufficient. On the other hand, there are needs to use shorter wavelengths, as for samples containing metals with a high absorption coefficient (see Section V.A).

2. Installation of X-Ray Tube and Camera

For this description it is assumed that an X-ray tube with line focus and four windows is used. First, one has to decide whether the tube should be mounted vertically or horizontally. Better mechanical stability and access to all four windows favor the vertical position, but several types of small-angle cameras (e.g., the Rigaku Denki goniometer) require a horizontal tube. In the following discussion the case of a vertically mounted X-ray tube is assumed, as this appears to be preferred by most workers.

Another basic decision concerns the type of window to be used for the camera. Frequently the size of the focus is 10×1 mm and the X-ray path has the usual 6° inclination from the horizontal plane of the focus. In this case the projection of the focus into the plane perpendicular to this X-ray path is a line with dimensions 10×0.1 mm on the "long" side of the focus; in its "narrow" side, the focus appears as a radiating square of dimensions 1×1 mm. This is often termed as working "at the line focus" or "at the square focus," respectively.

Naturally, a "slit camera" (i.e., a camera whose collimation system is designed for a primary beam with the cross section of a long and narrow rectangle) should be placed in front of the line focus, whereas a "pinhole camera" (i.e., one that utilizes a circular primary beam cross section) should be placed at the square focus. If the tube is mounted vertically, the longer dimension of the slit is referred to as its length and the smaller dimension is called width. It should be noted that in the case of the X-ray tube being mounted horizontally and working at the line focus, the larger dimension of the slit runs vertically and is correspondingly called slit height instead of slit length. In both cases (i.e., vertical and horizontal X-ray tube), the smaller dimension of the slit is referred to as the slit width.

3. X-Ray Tubes with Rotating Anode

Several manufacturers offer high-power generators. In order to achieve efficient cooling, the focus is generated at the surface of a water-cooled cylinder, which rotates at high angular velocity to dissipate the heat over the whole of its circumference.

While sealed tubes are normally driven with 2–3 kW, the power of commercially available high-power instruments ranges from 6 to 60 kW. The high primary intensity allows correspondingly shorter exposures, which is particularly useful for the study of time-resolved processes or of unstable materials. Moreover, investigations on extremely diluted solutions demand high power sources.

4. Synchrotron Radiation

The strongest X-ray sources available at present are synchrotrons or storage rings, instruments hitherto mainly used in high-energy physics. They consist of circular or oval rings, in which strong electric and magnetic fields, applied alternatingly, accelerate bunches of charged particles (electrons, positrons, or others) to high energy and keep them in constant orbit. Tangentially to the curved particle flight path high-energy photons are emitted. This so-called synchrotron radiation has several unique properties that open new possibilities in many fields of research. For X-ray small-angle scattering the most important ones are the following: (a) extremely high source brightness, (b) continuous spectrum from the hard X-ray to the far-infrared range, and (c) sharp collimation of the high energy part of the spectrum.

The intensity depends in a well-defined and computable way on the energy of the circulating particles and on the radius of curvature at the tangent point. In the region of $\lambda = 1$ Å, as an example, the storage ring DORIS II in Hamburg provides a source that is more than three orders of magnitude brighter than a rotating copper anode tube. Using special devices like "wigglers" or "undulators," which are arrays of dipole magnets forcing the electrons

or positrons to move on periodically wiggling paths, the intensities can be even further increased. Owing to the special conditions at synchrotron radiation sources (i.e., white beam of extreme intensity), very specific and stringent requirements exist for the design of cameras, especially with respect to monochromatization, focusing, and detectors.

The second important feature of synchrotron radiation, its continuous-wavelength spectrum, makes it an interesting tool for measurements at different wavelengths. Instruments for small-angle scattering with continuous-wavelength tuning have been developed and employ sophisticated arrangements of mirrors and monochromators. Basically, there are two types of arrangements: (a) double monochromator systems in which the beam reflected on the first crystal at certain angles is reflected back to the original direction by a second, parallel crystal, so that the beam direction does not change in the course of wavelength scanning; or (b) systems consisting of focusing mirrors (which also serve as premonochromators, eliminating wavelengths shorter than about 1 Å) and a single monochromator. With rotation of the monochromator by an angle of θ, the camera has to be rotated through 2θ to follow the primary beam path. Presently such systems exist for wavelengths up to approximately 7θ, which allows experiments around the K-absorption edges of elements as light as sulfur, phosphorus, and silicon.

B. Small-Angle Cameras with Slit Collimation

This subsection is confined to a discussion of cameras with slit collimation, which utilize a primary beam with line-shaped cross section. These are the most widely used type of camera in the field of diffuse small-angle scattering. Their advantage over point-focusing or pinhole cameras generally lies in the much higher primary beam intensity, which is particularly important with conventional X-ray tubes and which outweighs the disadvantage that the measured scattering curves have to be corrected ("desmeared") for the geometry of the beam. The mathematical procedures for desmearing are at a state of high development and pose little problems with modern computing facilities. In the following, a selection of cameras of this type are described.

1. Camera with Three Slits

The simplest collimation system consists of two parallel slits. The narrower these slits are and the larger the distance between them, the higher is the attainable resolution. This simple design has the disadvantage that the slits emit secondary scattering (parasitic scattering) into the small-angle region, which makes it impossible to use this camera for high-resolution work. A considerable improvement of the design, however, can be achieved by a third slit (Fig. 62, slit S_3), which is adjusted in such a way that it is just not hit by the direct beams 1 and 2, but shields off as much as possible from the parasitic scattering originating from slit S_2. Consequently, the parasitic scattering is intense between lines 3 and 4, while the "diagonal scattering," limited by lines 5 and 6, is much less intense and can be neglected.

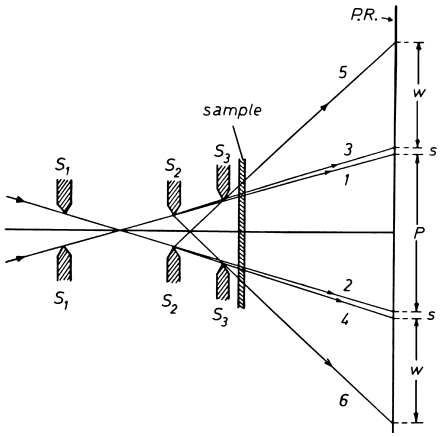


FIGURE 62 Slit camera with three slits, S_1, S_2, and S_3, each consisting of a pair of edges running perpendicular to the plane of paper. The dimensions in the vertical direction are greatly enlarged.

2. Beeman Camera

A slit camera that was used very successfully is the one introduced by Beeman and coworkers (Fig. 63). Two stationary tantalum slits C collimate the beam incident upon the scatterer. Two additional slits C analyze the angular distribution of the radiation leaving the sample. The second pair of slits, together with specimen holder and Geiger counter, are attached to an arm that may be rotated, by means of a calibrated spindle, about an axis through the center of the specimen.

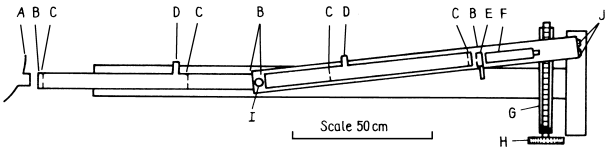


FIGURE 63 Diagram of the Beeman camera: A, X-ray tube exit window; B, mica window; C, tantalum slits; D, vacuum connections; E, Ross filters on slide; F, Geiger counter; G, high-precision screw; H, calibrated wheel; I, pivot and sample holder; J, rollers. [Reprinted with permission from Anderegg, J. W., Beeman, W. W., Shulman, S., and Kaesberg, P. J. (1955). *J. Am Chem. Soc.* **77,** 2927. Copyright 1955, American Chemical Society.]

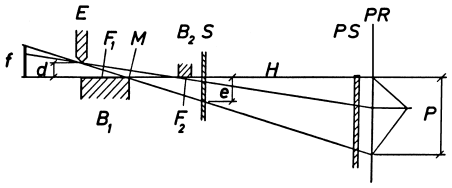


FIGURE 64 Section through block collimation system of the "Kratky camera." The dimensions in the vertical direction are greatly enlarged: S, sample; P, primary beam profile; PS, primary beam stop; PR, plane of registration. [Reprinted with permission from Kratky, O. (1954). *Z. Elektrochem.* **58,** 49; Kratky, O. (1958). *Z. Elektrochem.* **62,** 66; and Kratky, O., and Skala, Z. (1958). *Z. Elektrochem.* **62,** 73.]

An auxiliary device capable of considerably reducing the slit-length smearing is occasionally used in combination with slit cameras: the *Soller slit*. It consists of a set of parallel, thin lamellae, whose planes are arranged parallel to the primary beam axis and perpendicular to the plane of the beam. It is inserted into the camera behind the primary beam stop.

3. The Block Camera

a. The collimation system. The problem of parasitic scattering can be largely removed with the arrangement depicted in Fig. 64: it shows a section parallel to the propagation direction of the beam and perpendicular to the length direction of the focus. The radiation source is represented by the projection f of the focal spot into a plane perpendicular to the beam axis. Collimation of the beam is achieved by three construction elements, the two blocks B_1 and B_2 and the edge E, which all run perpendicular to the plane of the paper. It is essential that the plane defined by the polished upward-directed surface F_1 of B_1 coincide exactly with the downward directed surface F_2 of B_2. This plane is called the main section H. The width of the entering beam is defined by the distance d between the edge E and the main section H.

Figure 65 shows schematically how the instrument is designed to ensure that planes F_1 and F_2 coincide: block B_1 is the center piece of a U-shaped body, while block B_2, called the bridge, is pressed on to its side pieces from above. It is easy to see that there should be no parasitic scattering at all above the main section. In fact, there is a very small amount of parasitic scattering above H due to mechanical imperfections, and this can practically be neglected.

It lies in the nature of this design that by extreme care in the finish of the planar surfaces (by polishing and lapping) the parasitic scattering can be strongly reduced, and resolutions (i.e., smallest measurable scattering angles) are achieved that cannot be matched by any of the other camera types already described. However, this advantage has to be weighed against the fact that due to the asymmetrical arrangement, the scattering can only be measured at one side, that is, above the primary beam. In the downward direction, the parasitic scattering from edge M of block B_1 and from the edge of block B_2 facing the X-ray focus is very strong. The zero-angle position of the scattering curve, therefore, has to be determined by measuring the vertical intensity profile P and finding its central line of gravity.

b. Overall construction of the camera. Figure 66 shows a recent version of the camera mounted in front of the X-ray tube R, and Fig. 67 gives a vertical section through the camera in the direction of the X-ray beam. The collimation system C, the sample holder SH, and the primary beam stop PS are all mounted inside an evacuated housing V, with front window W_1 and end window W_2.

The radiation is recorded outside the vacuum tube. If a film or a position-sensitive detector is used, a stable suspension at the camera housing is sufficient. For the pointwise recording of the scattering curve, the simultaneous movement of detector slit DS and of the detector D is guided by the slits S_1 and S_2.

Figure 68 shows an application of the camera in the range of medium resolution; it shows the scattering curve of a solution of lipoprotein. The blank scattering is

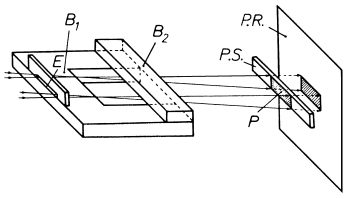


FIGURE 65 Schematic drawing of the collimation system: P, primary beam. Explanations of the other symbols are given in the legend to Fig. 64.

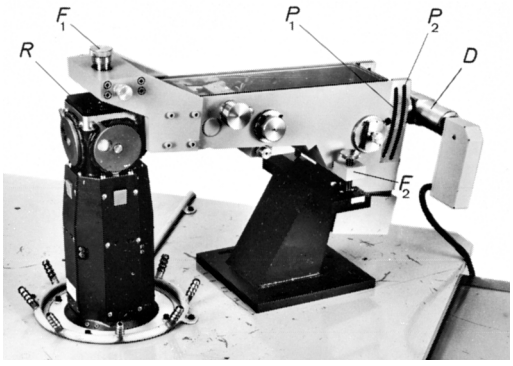


FIGURE 66 Design according to H. Stabinger and O. Kratky. [Reprinted with permission from Stabinger, H., and Kratky, O. (1984). *Colloid Polymer Sci.* **262,** 345.]

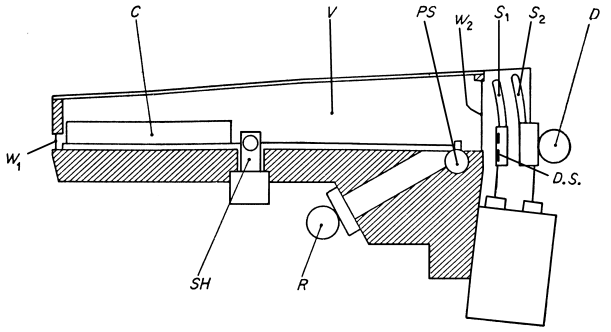


FIGURE 67 Section through the small-angle camera. Design according to H. Stabinger and O. Kratky. [Reprinted with permission from Stabinger, H., and Kratky, O. (1985). *Colloid Polymer Sci.* **262,** 345.]

completely negligible down to a Bragg Value of 500 Å, even compared to the blank scattering of the solvent-filled capillary.

4. The Bonse–Hart Camera

The instrument is based on multiple reflections of the primary beam from opposite sides of a groove in an ideal germanium crystal (2 in Fig. 69). The divergence of the beam decreases every time it is reflected from one of the walls, leading to a strictly monochromatic beam with a divergence of only several arc seconds. After penetrating the sample 3, the beam is again reflected several times from the inner walls of a second crystal 4 and finally enters the detector 5. Crystal 4 can be turned about an axis perpendicular to the plane of the paper with a precision spindle. This rotation of crystal 4 about its center allows the recording of the small-angle scattering originating

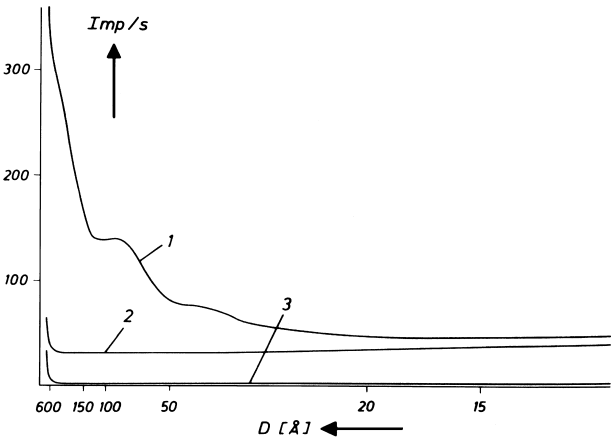


FIGURE 68 Measurements with the block camera: 1, scattering of a lipoprotein solution, 5.1%; 2, blank scattering of the Mark capillary with solvent; 3, blank scattering of the empty camera, corrected for absorption (multiplied by 10).

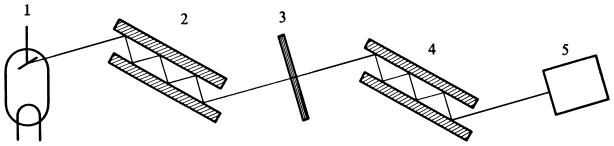


FIGURE 69 Schematic drawing of the Bonse–Hart camera. [Reprinted with permission from Bonse, U., and Hart, M. (1965). *Appl. Phys. Lett.* **2,** 155.]

from the sample: it is easy to see that if crystal 4 is turned by an angle α from its position parallel to crystal 2, only the radiation scattered by the sample to the same angle can reach the detector. The appealing feature of this ingenious design is the fact that one can measure down to the smallest angles without using a narrow entrance slit: all other techniques have to deal with the fact that the primary intensity drops rapidly due to the narrow entrance slits required to measure down to very small angles, while the Bonse–Hart system uses the same primary intensity for all angles. On the other hand, with decreasing resolution, the other methods offer the option of increasing the size of the entrance slit with concomitant increase in primary intensity, while the Bonse–Hart system does not offer such an option.

A quantitative experimental comparison with the block collimation system has yielded the following result. While the Bonse–Hart system yields constant intensity, the primary intensity of the block collimation system increases with the third power of the width of the entrance slit e (which, in turn, is inversely proportional to the highest attainable Bragg's value). The point of equivalence lies at an angle corresponding to a Bragg spacing of about 7000 Å. If it is sufficient to obtain a resolution of 1000 Å, the block camera yields a 7^3 ($=343$) times higher intensity than the Bonse–Hart camera. Alternatively, if one aims at a resolution of, say, 3×7000 Å ($=210{,}000$ Å), the Bonse–Hart system is more sensitive by a factor of 3^3 ($=27$). Since the majority of applications of the small-angle technique require resolutions far below 7000 Å, the Bonse–Hart technique has until now only been used in a few special problems, despite its truly ingenious design.

Certain modifications of the original Bonse–Hart design, such as the reduction of the number of reflections and the use of asymmetrically cut crystals, may lead to a considerable increase in intensity so that the point of equivalence comes down to a Bragg value of about 1000 Å.

C. Monochromatization

Quantitatively correct interpretation of diffuse small-angle X-ray experiments requires the knowledge of the scattering curve corresponding to monochromatic radiation. Polychromatic effects have to be eliminated, either

experimentally or numerically. Several of the existing methods are described below.

1. Pulse Height Discriminator, Alone or Combined with a K_β Filter

So far, this seems to be the most widely used method. The pulse height discriminator, which is tuned to the K_α line, is connected to a proportional or scintillation counter. The efficiency in suppressing white radiation is inversely related to the channel width. The wavelength of the K_β line is too close to the K_α wavelength to be sufficiently attenuated by the pulse height discriminator alone [for example: $\lambda\,(\mathrm{Cu\,K}_\alpha) = 1.54$ Å; $\lambda\,(\mathrm{Cu\,K}_\beta) = 1.39$ Å]. This necessitates the use of a K_β filter. In the following, the discussion will be limited to copper radiation, for which the K_β filter consists of a nickel foil. The absorption edge of nickel lies between the Cu K_α and the Cu K_β wavelengths; thus nickel absorbs K_β much more strongly than K_α.

2. Balanced Filters

This old monochromatization technique by Ross requires two exposures with different filters in front of the collimation system under otherwise identical conditions. For copper radiation, the two filters consist of nickel and cobalt, respectively. The wavelength of the absorption edge of cobalt ($\lambda = 1.604$ Å) is slightly above the Cu K_α wavelength, and that of nickel ($\lambda = 1.483$ Å) is just below. Let t_{Ni} be the thickness of the nickel filter and t_{Co} that of the cobalt filter; we call the two filters "balanced" if t_{Ni} and t_{Co} have a certain optimal ratio, namely $t_{\mathrm{Ni}}/t_{\mathrm{Co}} = 1/1.0711$. Under this condition, substraction of the two scattering curves leaves only the contribution from radiation whose wavelength lies between the two absorption edges. Contributions from radiation with $\lambda < 1.483$ Å and with $\lambda > 1.604$ Å cancel. Since the total intensity of continuous radiation in this narrow range is negligible compared to the Cu K_α intensity, the difference curve can be regarded as monochromatic. The maximum intensity is obtained with $t_{\mathrm{Ni}} = 6.99\ \mu$m and $t_{\mathrm{Co}} = 7.48\ \mu$m.

3. Crystal Monochromator and Total-Reflection Mirrors

In the majority of applications the monochromator crystal is positioned in front of the collimation system so that the sample is hit only by the purely monochromatic radiation. We refer to the most important types of monochromators, namely the flat quartz monochromator, the bent quartz plate (Johann), and the monochromator consisting of a bent quartz plate whose surface is ground to half the bending curvature (Johannson).

In connection with the continuous spectrum of synchrotron radiation, the techniques of monochromatization have reached a very high state of development. A focusing monochromator now widely applied consists of a bent triangular germanium crystal. The triangular shape leads to a perfect bending geometry.

The biggest disadvantage accompanying the use of most crystal monochromators is the considerable loss in intensity. Pyrolytic graphite overcomes this disadvantage to an appreciable degree and is now frequently used in small-angle scattering.

While Bragg diffraction selects a discrete wavelength, total reflection on polished glass surfaces cuts off the lower-wavelength part of the spectrum. This partial monochromatization is sometimes found sufficient for purposes of small-angle scattering experiments.

D. Detection

There exist three alternative methods that are suitable for the detection and registration for small-angle scattering curves: (a) photographic film, (b) proportional counters with detector slit and precision goniometer (step-scanner), and (c) position-sensitive proportional counters. Photographic film detection, despite its excellent resolution, is not widely used at present, mainly due to the inconvenience of densitometric evaluation; with improved film material and modern densitometers, however, plus the advantages of films (i.e., compact documentation and high resolution), this detection method might gain in importance. In the step-scanning method, a proportional counter with a precisely positioned detector slit is moved step by step through the angular range of interest. This allows for high flexibility in the adaptation to the specific resolution requirements of a given scattering pattern; its disadvantages lie in the sequential measurement of individual points at the scattering curve, which necessitates utmost stability of the primary beam and intensity averaging over several scans. Consequently this method is precise but time-consuming. The recent development of position-sensitive proportional counters combines the advantages of both above-mentioned methods: it allows "simultaneous" electronic registration of the whole scattering pattern. The resolution is generally still not as good as with photographic film, but this is in most cases not a limiting factor. Its immediate compatibility with electronic data storage media and computers is a feature that makes this method increasingly attractive.

E. Absolute Intensity

Mass determinations require knowledge of the absolute intensity, that is, the ratio of scattered intensity to primary

intensity (see Section II.B.3.c). While the scattered intensity can be directly determined with commercially available instruments, direct determination of the primary intensity is not possible due to the rapid succession of quanta in the primary beam that cannot be resolved even with the most advanced detectors.

Several methods have been described to overcome this difficulty. One of them involves the determination of the absorption of several thin nickel foils, which are then stacked to yield sufficiently strong attenuation of the primary beam. The method requires monochromatic radiation, since even a small contamination with short-wavelength radiation will lead to errors due to inverse dependence of absorption and wavelength.

Another method is based on a defined mechanical attenuation of the primary beam. The instrument uses the principle of a sector diaphragm. Although the method was only used in few laboratories, it has served as a convenient instrument to calibrate secondary standards, which are in wide use today.

Such a secondary standard consists of a platelet of polyethylene ("Lupolen"), whose absolute intensity (i.e., the quotient of scattered and primary intensity) is determined for an angle corresponding to a Bragg spacing of 150 Å. Rapid and simple determinations of the primary intensity are then possible in every laboratory with access to such a calibrated Lupolen.

The recently developed "moving slit method" offers another alternative method for the direct determination of the primary energy. Two slits are installed perpendicular to the plane of the primary beam; one with width L_r is located in the plane of registration, and the other one (width L_f) is located near the focus. The second slit can be moved with velocity v across the length of the beam. The primary intensity P_0 (energy per centimeter per second) can be calculated from the following:

$$P_0 = Nv/L_f L_r,$$

where N is the number of quanta reaching the counter during one passage of slit L_f.

Primary standards, such as gaseous Freons or silica gel, have also been developed for which the ratio of primary intensity and scattered intensity can be evaluated theoretically.

So far only the filter method and the secondary standard Lupolen are in routine use. The moving slit method is still too new to have found widespread application.

When films are used for the recording of radiation, the determination of the primary intensity is comparatively simple. A convenient method is to move the film perpendicular to the primary beam's length direction during the exposure of the (unattenuated) beam, thus producing a broad band.

F. Experimental Elimination of the Effect of Intensity Fluctuations; Monitor

No other type of X-ray scattering experiments requires stable X-ray sources as small-angle investigations: The measurements frequently take many hours, during the course of which the scattered intensity is recorded at many different angles; intensity fluctuations of the primary beam during that time, therefore, lead to a deformation of the observed curve. The stringent requirement for maximum stability in the primary intensity can only be relaxed if a position-sensitive detector is used, since it records the whole scattering curve simultaneously.

The most important reason for intensity fluctuations seems to be the relative movement of the camera with respect to the anode; such a movement may be caused either by mechanical or by temperature effects, which cannot be avoided even by careful stabilization of the room temperature: The X-ray tube, which is heated and cooled at the same time, never allows a constant temperature of the entire system. The suspension of the camera at the top of the X-ray tube near the anode is capable of reducing these effects.

A monitor, however, would be a useful tool. The basic idea is to divide the scattered intensity by the (simultaneously recorded) primary intensity or a quantity proportional to it.

This obviously eliminates errors in the scattering curve caused by fluctuations in the primary intensity. Several ways have been suggested and tried for the experimental implementation of this idea. These include the use of an ionization chamber within the collimation system. Other ways are given by the Bragg reflection or the fluorescence emission of the primary beam on the beam stop.

VII. RECENT DEVELOPMENTS

A. Scattering from Fractal Objects

The theoretical relationships between scattering patterns and real-space structure as discussed in the previous sections have so far largely been used to derive structural models in terms of Euclidean geometry. In this approach, the structure of complex bodies is described by analytical geometric elements (lines, planes, spheres, etc.). This may be sufficient for a first approximation to an overall shape and internal structure, but it has become obvious that many real physical objects are irregular to an extent that renders a Euclidean description unsatisfactory.

The possibilities for the discussion of irregular, disordered structures have been substantially enriched through the concept of fractal geometry and its recent penetration into physical sciences. In this formalism, structures

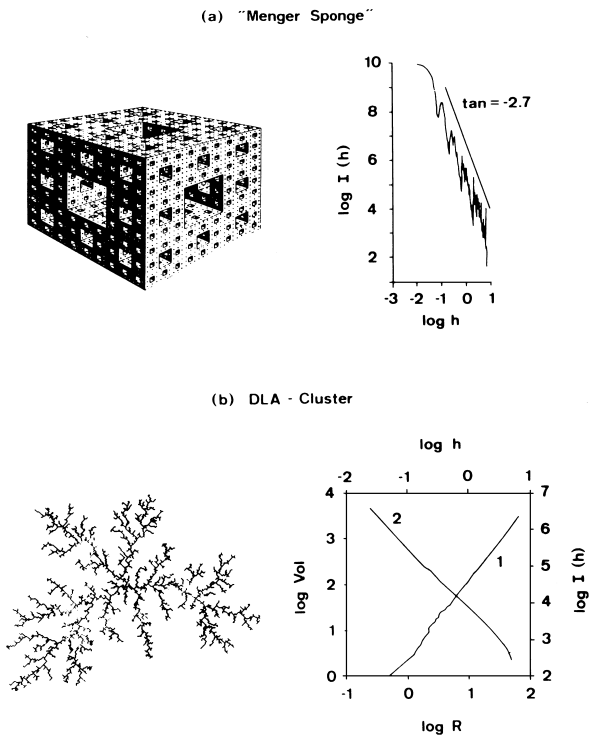


FIGURE 70 Computer-simulated mass fractals and their corresponding small-angle scattering curves. (a) The "Menger sponge" has a fractal dimension $D = 2.7$, which also corresponds to its pore-size distribution $N(r) = r^{-2.7}$ and can be read from the slope in the log–log scattering curve. (b) The diffusion-limited aggregation (DLA) cluster, a model for kinetic growth from monomers, has a fractal dimension $D = 1.7$. This value corresponds to the slope in the log–log functions of radius of gyration versus volume (1) and intensity versus scattering angle (2).

as shown in Fig. 70 can be characterized with the single parameter D, the fractal dimension, which is defined as the exponent in the scaling law that relates the mass M of an object to its size R as follows:

$$M \sim R^D.$$

Such objects, called mass fractals, possess as the common geometrical property their invariance to dilation or contraction, which means that the structures appear similar under different length scales. The normal objects of Euclidean geometry, rods, planes, or spheres, are in this sense only special cases for which the exponent D is equal to 1, 2, and 3, respectively, in accord with our common notion of dimensionality. For mass fractals, the fractal dimension can assume any noninteger value between 1 and 3.

The same concept can be also applied to rough surface structures of objects with internally uniform density, i.e., where $D = 3$. For such "surface fractals," the self-similarity means that the geometric features do not change as the surface is magnified or contracted. In this case, a fractal dimension of the surface D_s can be defined, which

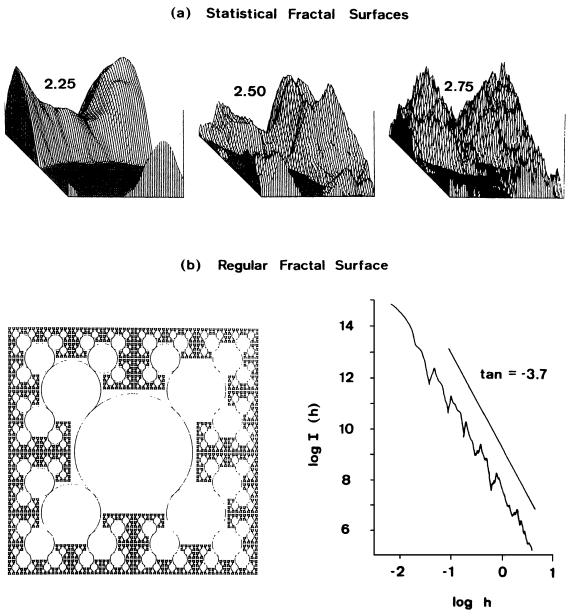


FIGURE 71 Surface fractals. (a) Examples of surfaces with different fractal dimensions, D_s, generated by a spectral synthesis algorithm. (b) Two-dimensional cut in the x–y plane of a three-dimensional surface fractal of 7th order, consisting of 117.187 spheres; the radius of spheres in the subsequent orders decreases by $r/2$. The resulting surface–fractal dimension is $\log(5)/\log(2) = 2.3$. The log–log slope of the analytically calculated scattering curve is -3.7, and hence $D_s = 2.3$. (Computer simulation by M. Kriechbaum.)

is the scaling exponent in the power law relating the surface area S and the length scale unit R by

$$S \sim R^{Ds}$$

Examples for surface fractals are shown in Fig. 71. For a smooth object, $D_s = 2$, while for increasingly rough surfaces D_s approaches the value of 3. This concept of self-similarity can also be extended beyond strictly topological meaning to statistical size distribution laws, e.g., of pore or particle sizes, if they are the same at different resolutions (statistical fractals).

Scattering methods in general, and especially X-ray small-angle scattering, are most useful tools to determine experimentally the fractal dimensions D and D_s, since they probe the dimensionality in reciprocal space by the dependence of the Fourier coefficients on the reciprocal distance, and the decay exponent of scattering intensity is therefore directly related to the fractal dimension in real space. This conceptual relation is expressed by the generalization of Porod's exponential law [Eq. (9)] in the following form:

$$I(h) \sim h^{-p}. \tag{39}$$

For mass fractals, where $1 < D < 3$, it holds

$$p = D$$

and for surface fractals, with $2 < D_s < 3$,

$$p = 6 - D_s.$$

The classical case, where the Porod exponent approaches the value of -4, is thus representative of a three-dimensional system ($D = 3$) with smooth, discontinuous interfaces ($D_s = 2$), as, for example, schematically depicted in Fig. 6 and as found in many smooth colloids. On the other hand, fractality of a structure is hence indicated by deviations of the decay exponent from its ideal value -4. This can be conveniently determined from the slope of an extended linear region in the log I versus log h plot of the scattering curve. Equally, the lower and upper length limits L_{min} and L_{max} for fractality or self-similarity, can be read from the boundaries of h within which the log–log representation is linear, by $L = 2\pi/h$. In certain cases, where mass and surface fractality apply at different length scales, both fractal dimensions can be obtained from one experiment. With normal instrumentation of X-ray small-angle cameras, it is possible to probe length scales between 10 and 1000 Å in orders of magnitude. This may be extended to values on the order of 1 μm by using the Bonse–Hart system (Section IV.B.4) into a length range overlapping with that accessible to light scattering.

Despite the undoubted value of the fractal dimension in differentiating and classifying irregular structures, it does not allow to make exclusive and unambiguous statements in terms of specific structural models. It has been shown that, for example, a fractally rough surface (D_s) can be considered equivalent to a polydisperse system of pores, where the pore size distribution follows a power law with the exponent equal to $[-(1 + D_s)]$. In general, power-law scattering can result either from polydispersity or from a structure composed of mass or surface fractals.

As a practically very important consequence to the analysis of fractally rough or porous materials by X-ray small-angle scattering, the repetitive self-similarity of the structure at different scales of magnification makes it meaningless to define scale-invariant surface areas or intersection lengths (Sections I.F.3 and I.F.4). This is most pertinent in the analysis of porous, surface-rich materials (e.g., catalysts, adsorbents, gels, ceramics) where in the prefractal era it has been customary to determine such values for specific inner surfaces, for example, for catalysts, powders, ceramics, or gels, and to relate them to technological properties or functions. As an example, Fig. 72 shows the fractal scattering behavior of hydrated Portland cement during the process of age hardening, where the scattering curves indicate both mass and surface fractalities. Instead, it will be necessary now to establish new empirical relationships between fractal dimensions of materials and the relevant technologic properties, to attribute predictive value to this new kind of geometrical parameterization. Finally, it may be emphasized that this approach is of course not limited to inorganic materials but increasingly finds important applications also in the fields of biological or synthetic polymer systems.

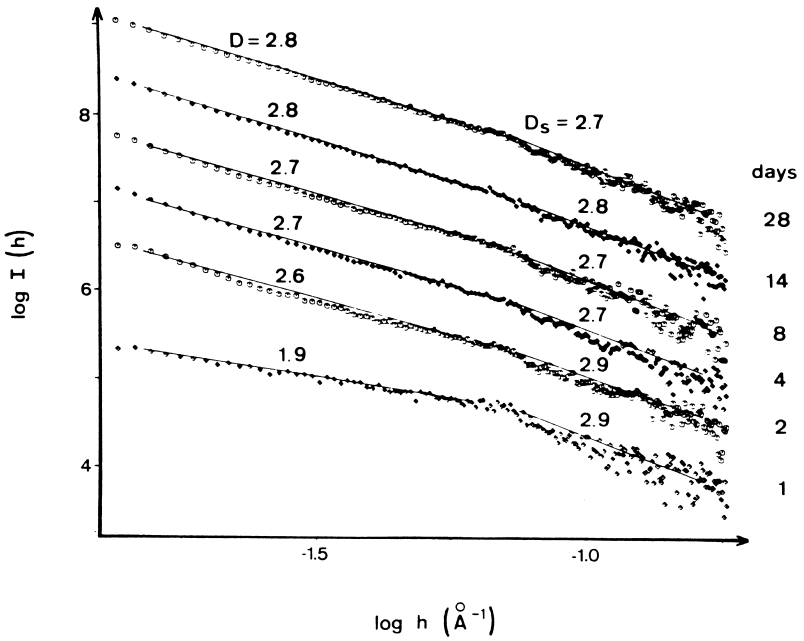


FIGURE 72 Scattering curves of Portland cement stone at different stages of age-hardening. Mass and surface fractalities, respectively, can be read from the slopes in the log–log plots of the scattering curves. [Reprinted with permission from Kriechbaum, M., Degovics, G., Tritthart, J., and Laggner, P. (1989). *Progress Coll. Polym. Sci.* **79,** 101.]

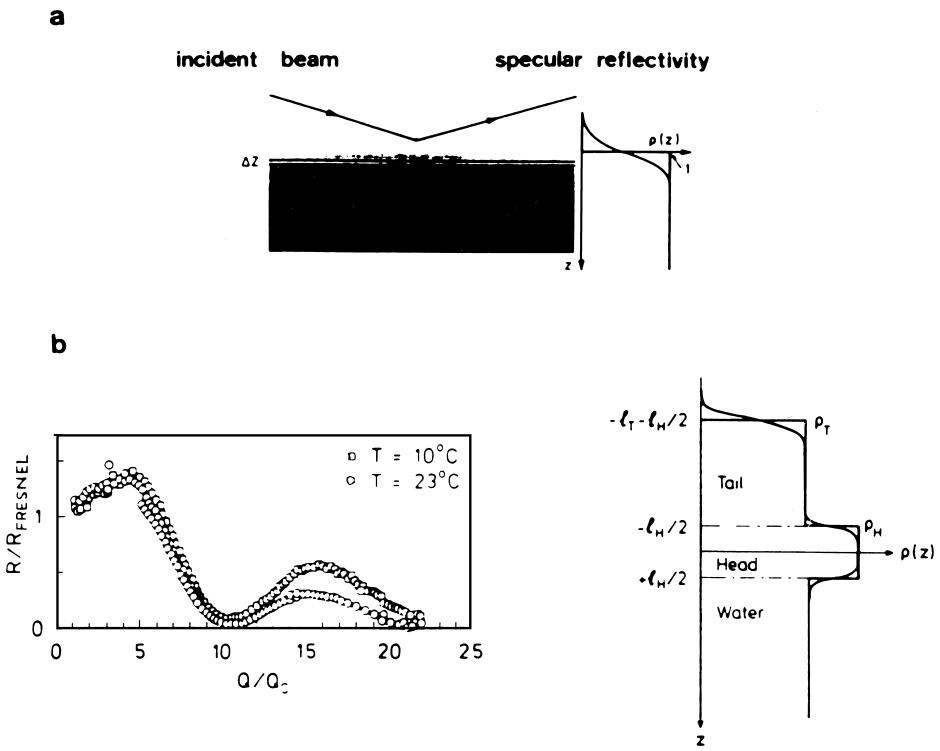


FIGURE 73 Determination of surface monolayer electron density profile from X-ray reflectivity experiment. (a) The angular dependence of specular reflectivity is related to the electron density profile $\rho(z)$ through Fourier transformation. (b) Specular reflectivity of a phospholipid monolayer at air–water interface. (Right) The resulting electron density profile. [Reprinted with permission from Als-Nielsen, J., and Möhwald, H. (1990). "Handbook of Synchrotron Radiation," Vol. IV (S. Ebashi, E. Rubenstein, and M. Koch, eds.), North–Holland, Amsterdam.]

B. Thin Molecular Films

The recently grown interest in thin, monomolecular films as elements in the design of artificial sensing and switching devices has created the demand for specialized methods suitable for their structural characterization. One promising strategy in the preparation of such devices involves the Langmuir–Blodgett technique, where monomolecular films from an air–water interface are deposited onto solid surfaces. Diffraction from such surface layers is naturally very weak and necessitates novel instrumental approaches frequently involving synchrotron radiation. In the following, some examples for the progress made in this field are presented.

1. Monolayers at Liquid Surfaces

Measurement of X-ray reflectivity of a liquid surface at angles beyond the critical angle for total reflection (Fig. 73) yields detailed information about the electron density profile across the surface. The treatment of the data to extract the electron density profile is formally very similar to the one used with particle scattering from radially symmetric systems (Section II.B.3.e). Here, the ratio $R(h)$ between the observed reflectivity and the Fresnel reflectivity of an infinitely sharp interface measures the Fourier transform of the electron density gradient in the molecular surface film. Figure 73(b) shows an example of such "specular reflectivity" data from a phospholipid monolayer at different temperatures. The main attraction of this type of experiment lies in the fact that structural data can be obtained from a film under different lateral pressures applied in a Langmuir trough, and thus macroscopic pressure/area isotherms become directly interpretable in terms of molecular structure. In combination with wide-angle diffraction data obtained under the same conditions in the plane of the film, the two-dimensional molecular packing can be analyzed at high resolution. This method has so far been used to investigate such important aspects as thermal roughness of simple liquids, smectic layering of liquid crystals, and monolayer structures of surfactants, phospholipids, and fatty acids.

2. Oligo- and Multilayers

By careful penetration of liquid–air monolayers with flat, solid supports, it becomes possible to transfer and pick up the monomolecular films. Repeating this process, for example, with monolayers of different molecules,

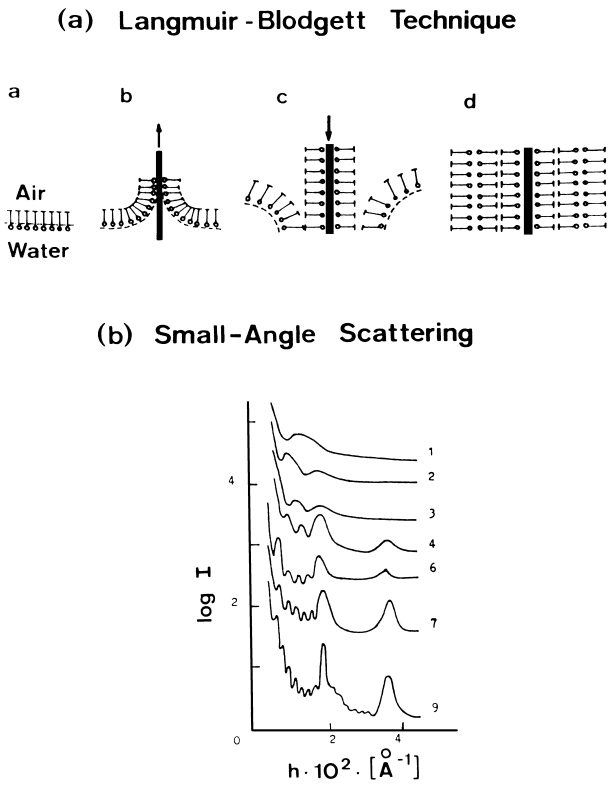


FIGURE 74 (a) Principle of molecular oligolayer formation by the Langmuir–Blodgett Technique. (b) Small-angle X-ray scattering from Langmuir–Blodgett films consisting of different numbers of bilayers. From the dominant reflections, the period distance, and from the side peaks, the number of repeation layers can be determined. [Reprinted with permission from Feigin, L., and Lvov, Y. (1988). *Makromol. Chem., Macromol. Symp.* **15,** 259–274.]

supramolecular layered structures can be purpose designed. Diffraction analysis of such objects is useful both for the aim of structure investigation of compounds or complexes which do not form well-developed single crystals and for monitoring the quality of stacking of layers as desired for specific functions. Figure 74 shows an example of small-angle scattering experiments on such systems. It is evident that good diffraction data can be obtained already from a small number of layers, so that the determination of the layer repeat distance and the electron density profile across the layers can be determined.

SEE ALSO THE FOLLOWING ARTICLES

LIQUIDS, STRUCTURE AND DYNAMICS • MACROMOLECULES, STRUCTURE • X-RAY ANALYSIS • X-RAY, SYNCHROTRON RADIATION, AND NEUTRON DIFFRACTION

BIBLIOGRAPHY

Alexander, L. E. (1969). "X-Ray Diffraction Methods in Polymer Science," pp. 280–356, Wiley, New York.

Damaschun, G., Müller, J. J., and Bielka, H. (1979). "Scattering studies of ribosomes and ribosomal components." *Methods Enzymol.* **59,** 706.

Engelman, D. M., and Moore, P. B. (1975). "Determination of quaternary structure by small-angle neutron scattering." *Annu. Rev. Biophys. Bioeng.* **4,** 219.

Feigin, L., and Svergun, D. I. (1987). "Structure Analysis by Small-Angle X-Ray and Neutron Scattering," Plenum, New York/London.

Gerold, V., and Kostorz, G. (1978). *J. Appl. Crystallogr.* **11,** 367.

Glatter, O., and Kratky, O., eds. (1982). "Small-Angle X-Ray Scattering," Academic, London. (Overview on the entire field of X-ray small-angle scattering, presented in 15 chapters, on 515 pages, by 13 authors.)

Guinier, A., and Fournet, G. (1955). "Small-Angle Scattering of X-Rays," Wiley, New York.

Hendrix, J. (1982). "Position sensitive X-ray detectors." *In* "Synchrotron Radiation in Biology" (H. B. Stuhrmann, ed.), pp. 285–319, Academic, New York.

Jacrot, B. (1976). "The study of biological structures by neutron scattering from solution." *Rep. Prog. Phys.* **39,** 911.

Kratky, O. (1963). "X-Ray small-angle scattering with substances of biological interest in diluted solution." *In* "Progress in Biophysics," Vol. 13, pp. 105–173, Pergamon, New York.

Kratky, O. (1966). "Possibilities of X-ray small angle analysis in the investigation of dissolved and solid high polymer substances." *Pure Appl. Chem.* **12,** 483.

Kratky, O., and Pilz, I. (1972). "Recent advances and applications of diffuse X-ray small-angle scattering on biopolymers in dilute solutions." *Q. Rev. Biophys.* **5,** 481.

Kratky, O. (1983). Die Welt der vernachlässigten Dimensionen und die Kleinwinkelstreuung der Röntgen-Strahlen und Neutronen an biologischen Makromolekülen. "Nova Acta Leopoldina, NF 256," Vol. 55, pp. 1–72, Johann Ambrosius Barth, Leipzig. [Authorized translation: The world of neglected dimensions, small-angle scattering of X-rays and neutrons of biological macromolecules, pp. 1–103.]

Laggner, P. (1988). "X-ray studies on biological membranes using synchrotron radiation." *Top. Curr. Chem.* **145,** 174.

Laggner, P., and Müller, K. W. (1978). "The structure of serum lipoproteins as analyzed by X-ray small-angle scattering." *Q. Rev. Biophys.* **11,** 371.

Pessen, H. T., Kumosinsky, F., and Timasheff, S. N. (1973). "Small-angle X-ray scattering." *Methods Enzymol.* **27,** 151.

Pilz, I. (1973). "Small-angle X-ray scattering." *In* "Physical Principles and Techniques of Protein Chemistry," Part C, pp. 141–243, Academic, New York.

Pilz, I., Glatter, O., and Kratky, O. (1979). "Small-angle X-ray scattering." *Methods Enzymol.* **61,** 148.

Porod, G. (1951). *Kolloid-Z.* **124,** 83; **125,** 61.

Schmidt, P. W. (1989). "Use of scattering to determine the fractal dimension," *In* "The Fractal Approach to Heterogeneous Chemistry, Surfaces, Colloids, Polymers" (D. Avnir, ed.), pp. 67–79, Wiley, Chichester.

Wunderlich, B. (1973). "Macromolecular Physics," Vol. I, Academic Press, New York.