3

Search Challenges and Methods

A successful product must have the following elements: a market and customers with needs for a product that is available in quantity and at suitable prices, with a set of properties that are suitable for the application, containing appropriate material that can be produced by a suitable technology (figure 3.1).

The design of a successful product for the marketplace can be described as the creative synthesis of many elements together, with optimization and harmony. The product innovators usually start with some of the required elements, but other key elements are missing and have to be found: for example, when we needed a good refrigerant but did not know what material would have the right properties, or when we found an interesting nonstick polymer but did not know what products would benefit from it and what markets would appreciate it. At later stages, we need to find ways to optimize the elements to make a better product, such as by learning to make nylon from the raw materials adipic acid and hexamethylenediamine, and by increasing the solubility of taxol by emulsifying with castor oil. Thus, a successful design involves many searches for missing elements, as well as for ways to improve existing elements.

The search from a material to its properties is called the forward search, since handbooks and tables of properties are organized and listed by the materials, so one

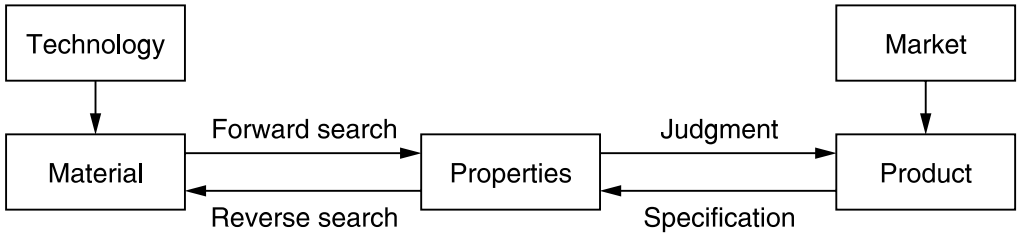


Figure 3.1 The design process for bringing a successful product to the marketplace

looks up the boiling points and the flammability of the compounds by their names. The invention of nylon by Carothers and associates at DuPont in 1928 can be represented by the set of arrows from the left to the right in figure 3.1. Carothers discovered the science and technology of condensation polymerization, which has the capability to make polyester material in possession of properties that are promising but not entirely suitable, as well as the capability to make many other materials. They decided that the product should be a silken fiber, suitable for the market of ladies stockings. From this "lead compound," they searched to find ways to modify the structure to obtain material with superior and desirable properties that were fine-tuned for the stocking market.

The search from properties to a material that has them is called the reverse search, since handbooks are not organized this way. The search for a safe refrigerant by Midgley and associates in 1928, which resulted in the discovery of CFCs, can be represented by the set of arrows from the right to the left in figure 3.1. Kettering told Midgley that there was a market demand for a safe refrigerant and left it to Midgley to specify quantitatively all the parameters of the properties required for this product. Midgley decided that it should have a boiling point of -30 to 0 °C, and should be nonflammable and nontoxic. The next challenge was to search for substances that had these properties, to make modifications for greater convenience in their use, and to find methods to make them on a large scale at a competitive cost.

The most important product design parameters can be grouped into three sets:

1. *Structure parameters*. For a single compound, the structure parameters include: the proportion of atoms and their connectivity, the geometric and energetic parameters of bonds, angles, and conformation, and the electronic parameters of electron distribution and polarization. For multicomponent systems of solutions, microstructural material, and composite material, the additional structure parameters include: the proportion of the various components, and the relations of their phases as solutions, colloids, or composite solids.
2. *Property parameters*. The physical property parameters include: state of matter, phase equilibrium, thermal, mechanical, optical, and electromagnetic properties. The chemical property parameters include: preparation, reactivity, reactants and products, kinetics, flash point, and explosion limit. The biological property parameters include: toxicity, physiological and pharmaceutical effects, nutrition value, odor, and taste.
3. *Market and environment parameters*. The market parameters include: application areas, sales volume, price, and growth potential. The environment parameters include: environmental dispersion, transportation and transformation, areas of concentration, safety, health, and environmental impact.

3.1 Forward Search from Material to Property

The forward search starts from a newly discovered or currently underutilized technology that can be used to make a variety of materials, or can serve as a starting point for improved technologies. The investigators need to discover what material can be made by this technology, the properties of these materials, and what useful products can be designed and made from them. Let us represent the material structure by a vector $\boldsymbol{x} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$, which stands for parameter values such as bond length, bond angle, conformation, dipole moments, etcetera. Let us represent the properties by the vector $\boldsymbol{y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_m\}$, which represents a set of properties such as boiling points and densities. The task of forward search is to find the function $\boldsymbol{y} = f(\boldsymbol{x})$. The "domain" of $\boldsymbol{x}$ is the set of all the material structures of interest, and the "range" of $\boldsymbol{y}$ is the set of all the property values that these materials can assume. Sometimes the forward search function takes the form of $\boldsymbol{y} = f(\boldsymbol{y}'')$ where $\boldsymbol{y}''$ is a set of properties that are more readily available than $\boldsymbol{y}$, such as using the boiling points at 1 atm to predict the critical temperature. Some of the forward search questions that we seek answers to include:

1. Find the structure parameters $\boldsymbol{x}$ and the properties $\boldsymbol{y}$ of a given compound at room temperature and pressure. What are the properties of the compound at elevated or lowered temperatures and pressures, as well as of other environmental variables, such as electromagnetic field and solar radiation, so that this search can be written as $\boldsymbol{y} = f(\boldsymbol{x}, T, P)$?
2. When the lead compound is subjected to various chemical and physical modifications, what are the various families of derivative materials that can be made with this technology, where the derived structure can be written as $\boldsymbol{x} + \Delta\boldsymbol{x}$?
3. What are the properties of these derivative materials, $\boldsymbol{y} + \Delta\boldsymbol{y}$? The derivative $\Delta\boldsymbol{y}/\Delta\boldsymbol{x}$ is of particular interest, as it represents the property modifications resulting from structure modifications. This knowledge would be useful in exploring ways to improve product properties.

The study of the function $\boldsymbol{y} = f(\boldsymbol{x})$ starts from measurements of the properties of many compounds, and subsequent compilations into tables and databases. An analysis of these empirical observations can lead to useful associations and trends, and generalizations that may have predictive power. When a sufficiently large and systematic database has been accumulated, researchers will try to find correlations between a property $\boldsymbol{y}$ and "predictors," which are parameters relating to molecular structure $\boldsymbol{x}$, or other more easily available properties, $\boldsymbol{y}''$.

Product engineers need to be familiar with the search methodologies outlined in the following sections.

Literature Search

The first step in the forward search is to see whether we can find the desired property of the material in a database, which may be a textbook, a handbook, a research journal, or, increasingly, an electronic resource. This method is, in principle, the fastest and least expensive. Properties of materials are measured and first published in research

journals, after having passed a review process. Subsequently, these primary publications are further reviewed and placed in secondary publications of handbooks and textbooks. This system of vetting is designed to weed out unsubstantiated claims, and to reconcile the inconsistencies of different authors. Some databases, such as the Beilstein, elect to list all the measured values in the research literature, instead of a single recommended value for the boiling point of a compound. There are also many printed and electronic sources that have not been vetted by independent reviewers, and so should be used with caution.

The forward search starts from the name of a chemical compound, proceeds to finding its molecular structure, and then its physical and chemical properties, such as the boiling point, melting point, density, etcetera, in a handbook. Many databases for single compounds are also organized by classes and families of similar structures. Fluid solutions represent the next level of complexity. For the most important fluids, such as water, air, and some refrigerants, we can find extensive tables for the thermal properties of mixtures. For complex fluids, such as paint and emulsion, which are difficult to characterize and to reproduce, specialized books and journals should be consulted. The properties of some crystalline solids can be found, but usually not for multicrystal composite and amorphous solids.

Biological and environmental properties are generally not systematically measured and tabulated, until the recent compilation activities sponsored by the U.S. EPA, and by the Occupational Safety and Health Administration (OSHA). Properties such as prices and concentrations regarded as safe are changeable and can only be found in newspapers and journals, such as the *New York Times* and *Chemical Marketing Reporter*.

Prediction from Theory

When the desired properties are not found in available databases, the next move might be to see whether they can be estimated from theory. This is becoming easier due to the advances of the last 80 years in computational chemistry and the availability of fast computers and of improved software. It is possible to calculate the structure of a small molecule from theory, which yields a wealth of structural information on bond angles and energies, on charges and dipole moments, etcetera. There are also a few physical properties that are so thoroughly understood that they can be computed from theory, including the translational and rotational heat capacities of diatomic molecules. There is a much larger class of molecular properties that can be computed by molecular modeling of the behavior of assemblies of molecules.

The computational methods for the structure of a molecule are divided into: the ab initio, the semi-empirical, and the molecular mechanics methods.

The Ab Initio Method

The ab initio method begins by solving the Schrödinger equation for the orbitals of electrons around a molecule, using as little simplification and approximation as is practical. This exact method is available only for small molecules with few atoms. The Schrödinger equation for a system with only one nucleus and a single electron

is given by

$$H\psi = E\psi$$

$$H = -\frac{h^2}{8\pi^2 m}\left(\frac{\partial^2\psi}{\partial x^2} + \frac{\partial^2\psi}{\partial y^2} + \frac{\partial^2\psi}{\partial z^2}\right) - \frac{Ze^2}{r}\psi$$

The solutions are the wave functions ψ, called the atomic orbitals. The molecules are built from the atomic orbitals, and the most stable configuration is the one with the lowest energy.

Semi-empirical Methods

The semi-empirical methods ignore all the core electrons and consider only the outer shell of valence electrons. Many other types of simplification are also made to make the computation feasible in a reasonable length of time. These simplifications make the results less accurate, but make it possible to study larger molecules.

Molecular Mechanics

Molecular mechanics is the modeling of the molecule not as electrons obeying quantum mechanics, but as a number of atoms connected together by flexible bonds, which move according to Newtonian mechanics and empirical "force fields." For instance, in a simple diatomic molecule like H_2, the length of the bond between the two atoms is assumed to have an energy equal to

$$E_{\text{bonds}} = \frac{k}{2}(l - l_0)^2$$

This describes a parabolic energy well, and the lowest energy is attained for the normal length of the bond at l_0, which is 0.74 Å. The energy of the molecule is the sum of all these bond length, bond angle, bond torsion, dipole, and van der Waals energies. The most stable configuration of the molecule is that one with the lowest energy. The parameter values of these force fields are taken from quantum mechanical calculations of similar molecules, and from observations such as spectroscopy.

Prediction by Correlation

Quantitative correlations can be constructed when there is a sufficiently large and systematic database of the properties $\boldsymbol{y}$ and of the predictors $\boldsymbol{x}$. The domain of the correlation is the set of material with a similar structure, where a correlation can be expected to be reasonable correct, such as the boiling points for the normal paraffin. The predictor can be a discrete variable, such as the number of carbons on an alkane and the group position of an element; it can also be a continuous variable, such as temperature, pressure, and mole percent in a solution. There may be only one or several predictors in a correlation, and the function can be linear or nonlinear.

The most reliable choice for the $\boldsymbol{x}$ parameters and for the functions is based on theoretical understanding of the cause and effect between the parameter and the property, and it is more risky to use parameters and functions based on experience without theoretical justifications. When there is no appropriate theory to rely on, the simplest method is a linear function with only one predictor, such as $y = c_0 + c_1 x$, or a nonlinear function such as $y = c_0 \exp(-c_1 x)$. When there are several predictors, the simplest function is a multilinear combination $y = c_0 + c_1 x_1 + c_2 x_2 + \ldots$. It is more reliable to interpolate within the domain with experimental data than it is to extrapolate beyond the domain.

The group contribution method (GCM) begins with the observation that a functional group such as Br– increases the boiling points of many organic compounds much more than the group Cl–, which in turn has a greater effect than the group F–. The GCM decomposes a molecule into a number of "groups," such as H–, K–, CH_3–, NH_2–, OH–, Cl–, C_6H_5–, etcetera. The property of a molecule, such as the boiling point, is assumed to be the sum of the contributions of each type of group, so that $T_b = 198 + \sum \Delta_b$. The method of quantitative structure–activity relations (QSAR) can be considered a branch of this effort, except that the predictors can be structural parameters such as the connectivity among the atoms.

Estimation by Associations and Trends

Situations are often encountered where there are some fragmentary experimental observations to provide for suggestions and speculations, but not enough for a systematic and quantitative study. For instance, we have the toxicity information for many polychlorinated compounds that are used as pesticides and for electrical transformers, but many more species have never been synthesized or measured; their properties are not easily organized into a series with toxicity increasing with the number of chlorine atoms. There are also many situations where the diverse structural variations do not lend themselves to quantitative treatment. For instance, there are seven isomers of the heptane molecule: one has a straight seven-carbon backbone, three have a single side branch, four have two side branches, and one has three side branches. It is a challenge to create a quantitative formulation that can be generalized to any number of carbon atoms on the paraffin chain.

One tries to study and discover patterns of behavior in a domain where data are available, to form hypotheses of associations and trends, and to export them to a larger domain where there are no data. For instance, the water solubility at 25 °C for straight-chain hydrocarbons with six carbons is in the order

$$\text{aromatics} > \text{acetylene} > \text{cycloparaffin} > \text{1-olefin} > \text{paraffin}$$

We can make the hypothesis that this ranking is also valid for a paraffin with any number of carbon atoms, and the even more bold hypothesis that this ranking is valid for any organic compound. This hypothesis might suggest fruitful areas that should be investigated further.

Synthesis, Collection, Measurements

After all the theoretical and empirical estimation work is done, the only way to make sure that a substance actually has the property estimated is to synthesize the material and to make experimental measurements. This method is much more costly in time and money, and should be undertaken after a great deal of library and theoretical searches have already been done. This process of synthesis and measurements also has the effect of enlarging the database and improving understanding in the domain relevant to this development purpose.

3.2 Reverse Search from Properties to Material

The motivation of the reverse search is often to serve a market need that is not satisfied by currently available products. The investigators begin with a study of the marketplace and the business opportunities in introducing a new product by talking to the customers, equipment builders, government regulatory agencies, and experts. Then a hypothetical material is proposed with a given set of required properties, and the problem is where to find such a material.

The starting point of a reverse search is often from a lead product that is currently on the market but in need of improvements, with a given set of properties that should be modified. The search questions include:

1. What physical and chemical modifications of the lead compound have the best potential to improve its properties in the desired directions $\Delta \boldsymbol{y}$? The modified properties can be written as $\boldsymbol{y} + \Delta \boldsymbol{y}$, and the modified structure can be written as $\boldsymbol{x} + \Delta \boldsymbol{x}$. The derivative $\Delta \boldsymbol{x}/\Delta \boldsymbol{y}$ represents the structure change required to achieve a particular property change.
2. What other types of material would have a set of properties in the range of the desired properties? For instance, which class of compounds smells even better than musk, and where would we find a cure of lung cancer? This is symbolically written as $\boldsymbol{x} = g(\boldsymbol{y})$.

The reverse search is the most often encountered search question in product engineering, but there are few organized and convenient search engines.

Literature and Database Search

The printed databases are usually not designed and organized for a reverse search, as tables of densities and boiling points are normally listed according to the types and names of compounds. It is a laborious task to go through a handbook to find all the metal elements that melt between 400 and 500 °C. Many electronic databases, such as CD-ROM and Web sources, have a "search" capability, and this permits a search from a set of properties to a list of the substances that have them. For a single property such as boiling point, the database can sort according to this variable and display all the substances from the lowest to the highest boiling points, which makes it easy to find all compounds that boil between −40 and −39 °C. For several variables, this search would take the form of a Boolean search with "AND" and "OR" between criteria,

such as "boiling point between −40 and −39" AND "density between 1.1 and 1.2." For the currently available databases, the reverse search capability is available only for some of the most common physical parameters, such as boiling points and densities, and not available for the less common parameters such as toxicity and global warming potential.

Forward Searches Plus Interpolation

For most properties that can be predicted from theory, the usual formulation is the forward search from structure to property in the form $\boldsymbol{y} = f(\boldsymbol{x})$. However, a planned series of forward searches can be used to simulate a reverse search. For instance, in response to the need for a reverse search for the sweetest substances we restructure the search by compiling the sweetness of existing substances by their structures, which would facilitate a number of forward searches from structure to sweetness, and then use the structures of the sweetest substances to suggest areas for future explorations. In figure 3.2, three structures $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, and $\boldsymbol{x}_3$ are mapped to three property values $\boldsymbol{y}_1$, $\boldsymbol{y}_2$, and $\boldsymbol{y}_3$. When the desired property $\boldsymbol{y}^0$ is contained in a triangle of the y space, it may be surmised that the relevant structure $\boldsymbol{x}^0$ is also contained in the corresponding triangle in the x space. In fact, if $\boldsymbol{y}^0$ can be given as a linear combination of $\boldsymbol{y}_1$, $\boldsymbol{y}_2$, and $\boldsymbol{y}_3$, in the form of $\boldsymbol{y}^0 = c_1\boldsymbol{y}_1 + c_2\boldsymbol{y}_2 + c_3\boldsymbol{y}_3$, then one can estimate that the desired structure $\boldsymbol{x}^0$ can be approximated by the linear combination of $\boldsymbol{x}^0 = c_1\boldsymbol{x}_1 + c_2\boldsymbol{x}_2 + c_3\boldsymbol{x}_3$.

Thus, we execute three forward searches plus interpolation, which has the effect of a reverse search.

An association or a trend can also be used in the forward or the reverse directions. The Midgley search for a safe refrigerant relied on the qualitative trend that, among the set of simple compounds, the compound tends to be less flammable when one moves from the elements of group 14 to group 17, and they tend to be less toxic when one moves from the elements of period 5 to period 2 in the periodic table. Analogy is another method often used. If one makes the empirical observation that a reduction of bond saturation by removing hydrogen increases the water solubility of a hydrocarbon with six carbon atoms, then one may also try to decrease bond saturation of another

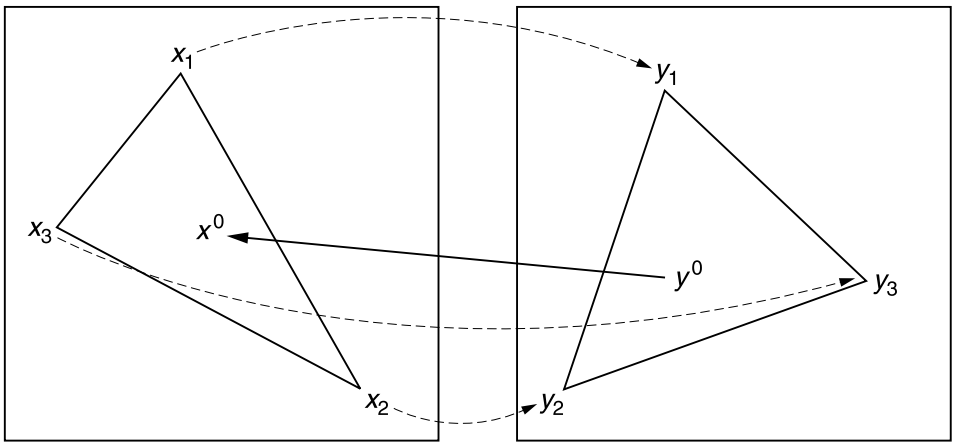


Figure 3.2 Mapping structures to property values

organic compound in order to increase water solubility. Associations and trends are sources of suggestions on fruitful areas that deserve further investigations.

Random Searches

The random search is usually the last resort when there are no good leads on where to start and what the fruitful directions to explore are. This approach is costly in time and money, and success is not certain, so that a strong motivation and a good budget are required. When a compound with unusual properties is discovered, such as when the morphine structure was first identified, synthetic chemists descended with gusto to make many derivatives, which is to say the creation of a set of random $\Delta \boldsymbol{x}$, so that nearly all possible derivates were made and tested for the property changes $\Delta \boldsymbol{y}$. Paul Ehrlich had a set of arsenic compounds that had therapeutic properties for syphilis and unacceptable toxic properties, so he searched for better arsenic compounds by changing their structures more or less by random. Thomas Edison had a carbon filament that worked reasonably well as a lamp filament, and he tried many other sources of carbon for his incandescent lamp. These are incremental random searches, by making small perturbations from a promising starting point, with a high probability of finding something interesting.

A search is much more speculative when bold leaps and "wildcat" searches in uncharted territories are undertaken. The initiative may be based on rumors of uncertain validity or an inspired hunch. There are no solid reasons to believe that the target can be found in this area; thus, there are no explorers and one can purchase property rights at very low costs. Success is unlikely, but when the improbable takes place we would be crowned with glory and profit.

3.3 Searches from Technology to Markets

When we have a powerful technology that can make a material with interesting properties, and we can keep out competition for a while with unique technology and patents, we ask which market will appreciate a material with these properties. An ideal market is also one where the potential buyers are numerous and affluent, where this material will help them to fulfill one or more significant needs, where there are no very satisfactory competitive products, where there are no threats to safety and the environment, and where it can be manufactured and make a good profit. What is the search engine that can systematically discover many or all of the major potential markets? Let us look at the historic cases mentioned in Chapters 1 and 2.

There were a number of adaptation cases that can be classified as fairly obvious, as the inventors and their management had experience with the market to be served. The concept of adapting nylon to women's hosiery was relatively obvious for DuPont, as they had been involved for many years in the manufacturing and marketing of rayon used as textile fibers. The high price and small quantity needed for a pair of hose also fits in well for a novel technology that has never been mass produced in the past. The concept of using penicillin for antibiotics would also be relatively obvious to Arthur Fleming, as he worked in a hospital and was in the army medical services.

The adaptation of CFCs for air conditioning to reduce moisture and temperature would not require a large leap in creativity, although the adaptation of CFCs for blowing air and polymer foam, for cleaning computers, and for fire extinguishers is less obvious. William Perkin had no experience with the dye industry, so it is to his credit that he recognized the potential market for mauve. The gluing of particles to surfaces to make mending tapes and memory recording surfaces are in the borderline area of being less obvious.

There are many more examples of new technologies that required many years of wandering in the wilderness before they found appropriate markets. People would exclaim in amazement at the discovery that the deadly poison botulin could be used to erase skin wrinkles and be reborn as botox. Roy Plunkett could not have anticipated the many uses of Teflon used for diffusion barriers in making the atomic bomb; the DuPont company could not anticipate the nonstick frying pans, nor the Gor-Tex sports garments. Spencer Silver could not find a use for his weak glue until his friend Arthur Fry used it to keep his place in conducting music. We have many powerful and wonderful technologies waiting for the discovery of a suitable market, such as buckminsterfullerene and electrically conducting polymers. At this moment, we do not have systematic and effective search engines from technology and material to marketplaces.

3.4 Literature and Databases

When we need the properties of a chemical compound, such as the boiling point of benzene, the fastest and least expensive method is a forward search from the structure to the properties by consulting a database. The sources of such experimental information are first published in primary research journals, and then pass through professional editors and panels to make their way to secondary textbooks and handbooks.

Printed Database

These are the traditional passive databases of books, handbooks, reference books, journals, and catalogs. They have passed the inspection of editors of committees, and are generally reliable and accurate.

Books, Handbooks, and References

A general reference often consulted today for the physical and chemical properties of common chemicals is *Lange's Handbook of Chemistry* (Dean 1999), which lists many chemical compounds and their most important properties. It is organized into separate chapters of "Physical constants of organic molecules" with 4300 compounds and "Physical constants of inorganic molecules," and lists each compound alphabetically by name. Some of these properties are very sensitive to temperature, but less sensitive to pressure, and they are listed as tables, or more compactly as equations of the form $f(T)$; for example, liquid heats of evaporation, heat capacities of multi-atom gases, vapor pressures over liquids, liquid and solid solubilities in liquids, and liquid viscosities. Some of these properties are sensitive both to temperature and pressure,

and are listed as tables or as equations of the form $f(T, P)$; for example, gas densities and gas solubilities in water. There is much less data available for mixtures, except for the most important ones, as the variations and combinations are very large.

Other important handbooks include *Chemical Properties Handbook* (Yaws 1999) and the *CRC Handbook of Chemistry and Physics* (Weast and Lide 1989); *Tables of Physical and Chemical Constants* (Kaye and Laby 1986) is a more compact handbook of physical and chemical data. One should be on guard that sometimes, when experimental results are not available, the editors may list estimated values in a handbook, which are of less certain accuracy. A printed handbook normally has only limited reverse search capability, of going from a set of properties to the structures that have these properties.

Specialized properties that are not covered in these standard databases can be found in specialized books. The properties of food can be found in *Physical Properties of Foods* (Peleg and Bagley 1983). The properties of many petroleum products can be found in *Petroleum Products Handbook* (Guthrie 1960). The *Merck Index* (1996) lists chemicals, drugs, and biologicals.

The toxic properties of chemicals can be found in a reference such as *Hazardous Chemicals Desk Reference* (Sax and Lewis 1987), which lists a number of chemicals alphabetically by name. For instance, "*n*-butane" is classified (by the Department of Transportation) as a flammable gas, is moderately toxic via inhalation, causes drowsiness, is an asphyxiant, poses very dangerous fire hazard when exposed to heat, flame, or oxidizers, and is highly explosive when exposed to flame. Information about safety and environment is also provided and updated by websites maintained by government agencies, such as the U.S. EPA, the National Institute of Occupational Safety and Health, and the National Safety Data Sheet.

A very handy sourcebook for data on social, economical, and political information is the *Statistical Abstract of the United States*, which is published by the U.S. Department of Commerce through the Census Bureau. It contains information on: population, vital statistics, health and nutrition; education, law enforcement, geography and environment, parks and travel; elections, state and local government finances, federal government finances; national defense, social insurance; labor force and employment, income and wealth, prices; banking and finance, business enterprises; communications and information technology, energy, science and technology, transportation by land, air and water; agriculture, natural resources, construction and housing, manufacturing; domestic trade and services, foreign commerce; comparative international statistics.

Journals and Catalogs

The research journals are the primary place where research results are first published. The prices of compounds change rapidly, and are to be found in journals instead of books. The *New York Times* finance page lists a number of commonly traded metals (aluminum, antimony, copper, gold, iron, lead, mercury, platinum, silver, zinc), food (corn, soya, wheat, rice, sugar), fuels (fuel oil, gasoline, natural gas), and textile (cotton, wool). The *Chemical Market Report* is a weekly journal that lists the prices of many chemical substances in a supplement, which depend strongly on purity and intended use.

The current prices of fine chemicals can be found in the Sigma-Aldrich *Handbook of Fine Chemicals and Laboratory Equipment* catalog, and in Fluka's *Laboratory Chemicals and Analytical Reagents*.

Electronic Databases

These are the modern active databases that have interactive capabilities, and are often searchable for a set of properties. The floppy disks and CD-ROMs have definite dates on them, and do become obsolete with time; the Internet databases can, in principle, be updated at very frequent intervals. A note of caution about Internet databases is that some of them have not been vetted for accuracy and can be unreliable.

Spreadsheets and Databases

A simple spreadsheet, such as Microsoft Excel, can serve as the foundation of a database that has forward and reverse search capabilities. For instance, a table of normal alkanes, together with their densities, boiling points, and melting points, can serve as the starting point. If we want to know all the "normal paraffins that boil between 0 and 40 °C," all we have to do is to do a "sort" operation on the boiling-point column and obtain the result that the only paraffin that is in the range is normal heptane with a boiling point of 36.1 °C. For the more advanced Boolean search of normal alkanes that "boil between 0 and 40 °C" AND "melt between −40 and 0 °C," it would be a far more laborious task in a spreadsheet.

A database such as Microsoft Access would be able to do this Boolean search with ease. The Filter by Form, the Filter by Selection, and the Advanced filter/sort function allow the user to specify: "density >0.6 AND <0.7" or "bp >−40 AND <0." This search yields three results:

Compound	*B.p. (°C)*	*Density (g/mL)*
Cyclopropane	−33	0.619
Methylamine	−6.5	0.655
Propylene	−48	0.691

CD-ROMs and Floppy Disks

Many publishers make available their databases on floppy disks or CD-ROMs, which are often found inside the back cover of the book. An example is the *Properties of Organic Compounds* (Lide and Milne 1999), which is a CD-ROM list of 27,500 known organic compounds and their formulas, molecular weights, melting points, boiling points, densities, refractive indexes, colors, solubility scales, vapor pressures, and threshold limit values (TLVs) that workers can be exposed to in a normal 8 h workday and 40 h workweek. This reference gives only a limited number of properties, and does not go into flammability or toxicity. Solubility in water is given only as a

qualitative scale: 1 = insoluble; 2 = slightly soluble; 3 = soluble; 4 = very soluble; 5 = miscible; 6 = decomposes. It can be searched by the following variables:

Name	*Variable*
MF	Molecular formula
CASRN	Chemical Abstract Service Registered Number
BRN	Beilstein Registered Number
MERCK	Merck Number
BP	Boiling point, °C
MP	Melting point, °C
DENS	Density in g/cm^3
MW	Molecular weight
IR	Infrared peaks, wavenumbers
UV	Ultraviolet peaks, nm
Raman	Raman peaks, cm^{-1}
HNMR	Hydrogen NMR shift, δ ppm
CNMR	Carbon chemical shift, δ ppm
MS	Mass spectrum, the most abundant and the parent peaks

One can also do a reverse search, such as for all compounds that boil between −40 and 0 °C, to speed up the search a great deal. Each of these variables can be searched according to the following criteria:

Equal to (*value*)
Less than (*value*)
Greater than (*value*)
Between (*lower limit*) and (*upper limit*)

If Thomas Midgley had today's tools in his search for a safe refrigerant, then he might start his search for all the known compounds that satisfy the conditions

$$[0 > \text{bp} > -40\,^{\circ}\text{C}] \quad \text{at 1 atm}$$

This would have results in 77 hits: from 1,2-[propadiene, 1,1,2,2-tetrafluoro-] with a boiling point of −38 °C, to [propane, 1,1,1,2,3,3,3-heptafluoro-2-(trifluoromethyl)-] with a boiling point of 0 °C. Out of 77 hits, 49 of them contain elements that include B, Si, N, P, As, O, S, Cl, Br, and I, and perhaps too toxic to be considered as refrigerants seriously. There are 11 hits that are hydrocarbons, such as butane, which would be too flammable to be considered. Perhaps we would eliminate the six hits that have double or triple bonds, as they tend to be less stable and could polymerize. The remaining ones are all hydrofluorocarbons (HFCs) without chlorine, and the prime candidates are $C_2H_2F_4$, $C_3H_3F_5$, and C_4F_{10}.

Internet Databases

A most convenient Internet source of chemical data is the *Chemistry WebBook* of the National Institute of Science and Technology (NIST; http://webbook.nist.gov/chemistry). Finding data in the *WebBook* can be done by one of the 14 types of direct search for chemical species:

- formula search
- chemical name search
- CAS registry number search
- ionization energy search
- electron affinity search
- acidity search
- appearance energy product
- vibration energy search
- electronic energy level search
- applet-based structure search
- file-based structure search
- structure class search
- molecular weight search
- author search.

The types of data available include:

- gas-phase thermochemistry data (enthalpy of formation, molar entropy, ideal gas heat capacity)
- condensed-phase thermochemistry (liquid and solid enthalpy of formation, molar entropy, heat capacity)
- phase change (normal boiling point, normal fusion point, critical temperature, enthalpies of vaporization, fusion, and sublimation, entropies of vaporization, fusion, and sublimation, Antoine equation vapor pressure parameters)
- reaction thermochemistry (enthalpy of combustion, properties of specific reactions)
- gas-phase ion energetics (enthalpy, entropy and Gibbs free energy of reaction)
- ion clustering
- gas-phase IR spectra
- mass spectra
- UV–visible spectra
- vibrational and electronic spectra
- constants of diatomic molecules (vibration and rotation parameters)
- Henry's law data (solubility of gases in water).

The data can also be displayed as a spreadsheet or data table, or as a graphical display, if the browser is Java capable. In addition, much more extensive information is available for selected fluids:

> water; He, Ne, Ar, Kr, Xe; H_2, D_2, N_2, O_2, F_2, CO, CO_2; NH_3, NF_3; CH_4, C_2H_6, C_2H_4, C_3H_8, C_3H_6, butane, isobutane, pentane, hexane, heptane; CH_2F_2, $CHClF_2$, CH_2F_2; $CHCl_2$–CF_3, C_2HF_5, CF_3–CH_2F, R143a, R152a.

The data available includes the temperature and pressure dependence of these fluids for:

> density, specific volume, heat capacity at constant pressure, heat capacity at constant volume, enthalpy, internal energy, entropy, and the speed of sound.

Beilstein is a much more comprehensive database, marketed under the name of *MDL Crossfire*, and requires a subscription fee for access. It contains far more organic substances and lists many more properties. The search commands include:

- text search for key words
- structure editor to draw structure
- tabular search by fields
- bibliographic search by citations, journals
- identification data, by name, formula
- physical data, by melting point, boiling point, pK
- reaction data, by starting material, product, yield, solvent
- solubility data, by solubility, Pow, Henry's law, solution behavior
- spectral data, NMR, ESR, IR, UV, MS.

It has extensive capability for forward search and reverse search from properties to substances. The properties available include:

Chemical data
reactions, purifications

Physical data, single component
structure and energy, bond distance, angles
physical state

crystals—melting point, density
liquid—boiling points
gas—critical temperature, vapor pressure
others—surface tension

transport—viscosity, diffusion, thermal conductivity
thermochemical—enthalpy, heat capacity
safety—flash point, explosive limit

Physical data, multicomponent
solution, solubility, Henry's law constant, partition water–octanol
liquid–liquid systems, liquid–vapor, liquid–solid

Pharmacological and ecological data.

A tabular search using the Fact Editor for the conditions: [bp > -40] AND [bp < 0] yields 462 hits in 15 s. Since most of the hits contain heteroatoms other than fluorine, it would be better to reissue the search command for compounds that do not contain any nitrogen: [bp > -40] AND [bp < 0] NOT ["*N*"], which yields 317 hits in 156 s. The asterisk indicates the wild card that stands for any number of characters of any kind, and the " " is for a text string. When we refine it further to exclude chlorine

as well, we issue the search command: [bp > −40] AND [bp < 0] NOT ["*N*"] NOT ["*Cl*"], which yields 247 hits in 204 s.

At this moment, the NIST *Chemistry WebBook* and the Beilstein are the two most useful search tools for molecular structure and property relations. There are many more types of properties available in the research literature of interest only to a smaller set of readers; these will not be listed in the general-purpose handbooks, but are to be found in specialized books and journals. Our ability to search for such properties in the research literature on the Internet is increasing rapidly, and students should practice doing such searches. Many of the required properties have not been tabulated in a searchable electronic database; for example, flammability, toxicity, ozone hole potential, and greenhouse gas potential. Information on health, safety, and the environment that is publicly available on websites is listed in the references.

References

Dean, J. A., ed. 1999. *Lange's Handbook of Chemistry*. 15th edition. New York: McGraw-Hill.

Guthrie, V. B., ed. 1960. *Petroleum Products Handbook*. New York: McGraw-Hill.

Kaye, G. W. G. and T. H. Laby. 1986. *Tables of Physical and Chemical Constants*, 16th edition. London: Longman.

Merck Index: An Encyclopedia of Chemicals, Drugs & Biologicals. 1996. 12th edition. Merck & Co.

Peleg, M. and E. B. Bagley, eds. 1983. *Physical Properties of Food*. Westport, CT.: AVI Publishing Company.

Sax, N. I. and R. J. Lewis. 1987. *Hazardous Chemicals Desk Reference*. New York: Van Nostrand Reinhold.

Statistical Abstract of the United States. 2003. Washington, DC: U.S. Government Printing Office.

Weast, R. C. and D. R. Lide, eds. 1989. *CRC Handbook of Chemistry and Physics: A Ready Reference Book of Chemical and Physical Data*. 70th edition. CRC Press.

Yaws, C. L., ed. 1999. *Chemical Properties Handbook*. New York: McGraw-Hill.

Further Reading

Books and Journals

General Database

Damm, H. C., P. K. Besch, and A. J. Goldwyn. 1966. *The Handbook of Biochemistry and Biophysics*. Cleveland: The World Publishing Company.

Journals and Catalogs

Aldrich. *Handbook of Fine Chemicals and Laboratory Equipment, 2000–2001*.

Chemical Market Reporters. Current prices of commodities, metals, fuels, and chemicals.

Fluka. *Laboratory Chemicals and Analytical Agents, 1999–2000*.

New York Times. Financial page on the prices of commodities, metals, etcetera.

Electronic Databases

Floppy Disks and CD-ROMs

Beilstein/CrossFire. 1999. Properties of organic chemicals. Frankfurt, Germany.
"Cranium", Molecular Knowledge, Bedford, New Hampshire, 1998. Group contribution methods (GCM) for estimation of range of molecular properties.
Lide, D. R. and G. W. A. Milne, eds. 1996. *Properties of Organic Compounds*. CD-ROM, CRC Press. 27,000 compounds, b.p., m.p., density, spectroscopic data.

Websites

ChemFinder Online: http://ChemFinder.camsoft.com
Columbia University chemistry department: http://www.columbia.edu/itc/chemistry/orgolab/cheminfo.html
EPA, Integrated Risk Information System: http://www.epa.gov/iriswebp/iris/index.html
Material Safety Data Sheet (MSDS): www.camd.lsu.edu/msds/jssearch.htm
National Institute of Occupational Safety & Health (NIOSH), Pocket Guide to Chemical Hazards: www.cdc.gov/niosh/npg/pgdstart.html
National Institute of Science and Technology (formerly National Bureau of Standards) properties of chemicals: http://webbook.nist.gov/chemistry.
Occupational Safety & Health Administration (OSHA): www.osha.gov
Statistical Abstract of the U.S.: http://www.census.gov/statab/.

Exercises

This set exercises concerns research on finding relevant information, as well as on preliminary organization. Start with simple search engines, then switch to more specialized search engines, and then go to the original references when necessary. Document all the sources that you have looked into and note any information you cannot find.

1. Do a forward search from materials to properties. Find the boiling points and melting points of 1-alcohols from C_1 to C_{20}.

   (a) Make a table for these properties.
   (b) Make a plot of these properties versus the number of carbons N_C.
   (c) Compare this plot with the normal paraffins.

2. Make a table of the densities, boiling points, and melting points of the first 54 elements.

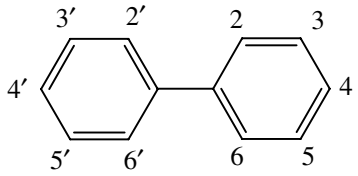
(a) How many of these elements melt below room temperature? In fact let us divide all the elements into sets with the melting point bin of [−273 0 500 1000 1500 2000 2500 3000 3500]. Report the population distribution of elements within each bin.

(b) For these 54 elements, is there a good correlation between melting point and boiling point?
(c) For these 54 elements, is there a good correlation between density and melting point?

3. There are two forms for butane C_4H_{10}: normal butane and isobutane. There are three forms for pentane, C_5H_{12}: pentane, 2-methylbutane, and 2,2-dimethylpropane.

(a) How many forms are there for hexane, for heptane, and for octane?
(b) Find all the melting and boiling points of paraffins containing from one to eight carbons, and organize them into a table according to whether they are: (i) straight chain, (ii) with one side branch (the longest chain is assumed to be the main chain), (iii) with two side branches, (iv) with three side branches, (v) with four side branches.
(c) We observe the general pattern that the boiling and melting points do increase with the number of carbon atoms. Is there a general pattern associated with the number of side chains on the boiling point and melting point?

4. Polychlorinated biphenyls (PCBs) were once used extensively to fill electrical equipment, such as transformers and relays, but they are now banned because they are toxic when discharged to the environment. The basic biphenyl $C_{12}H_{10}$ is two benzene rings connected by a single bridge.



The base biphenyl has 10 hydrogen atoms: four of them (2, 6, 2′, 6′) are near the bridge and are named the α positions, four of them (3, 5, 3′, 5′) are away from the bridge and are named the β positions, and two of them (4, 4′) are opposite the bridge and are named the γ positions. There are three forms for mono-chloro biphenyl, depending on the position of the chlorine atom, and they each have different properties. The PCBs are a set of 210 compounds with 0 to 10 chlorine atoms in a molecule. Monsanto made them under the trade name Aroclor, and they came in a number of mixtures: the light Araclor 1221 is basically mono- and di-chloro, the medium Aroclor 1248 is principally tri-, tetra- and penta-chloro, and the heavy Aroclor 1260 is principally hexa- and hepta-chloro. We are concerned with the transportation and concentration of these chemicals, particularly in fish and humans, which are controlled by the three parameters of vapor pressure (breathe in air), solubility in water (drink water), and octanol–water partitions (accumulation in fat).

(a) Find these relevant transportation and concentration parameters for biphenyl.

(b) Find the parameters of the three mono-chloro biphenyls, and point out the differences among the three forms of α, β, and γ.
(c) How many forms of di-chlorinated biphenyls are there? Name them, and find as many of their parameter values as you can.
(d) How many forms of tri-chlorinated biphenyls are there? Name them, and find as many of their parameter values as you can.
(e) Which type of PCB seems the most dangerous, and what is the nature of the threat?

5. Do a reverse search from properties to materials. What are all the hydrocarbons with boiling points between 0 and 5 °C, and melting points below −20 °C?
6. Let us repeat the journey taken by Thomas Midgley in his discovery of CFCs, in the light of superior modern databases and search engines.

   (a) What are all the compounds that boil between 0 and 2 °C when the pressure is 760 Torr?
   (b) Which of the compounds in (a) are nonflammable and nontoxic?
   (c) Which ones have no chlorine to react with ozone in the stratosphere?
   (d) Which ones are not greenhouse gases?

7. Let us imagine a dialog within the corporate headquarters of Lockheed. CEO: Our attack fighters require fuselage material of the highest melting point and the lowest density. What should we use? By the way, the Defense Secretary does not worry about price.

   Chief Engineer: Carbon is the best deal and lead is the worst deal.

   CEO: Do you recommend a carbon fuselage then?

   Chief Engineer: There are actually other required properties.

   Question for students: What is a complete set of property specifications, besides melting point and density? What procedure would you use to search for the best material, which databases would you apply the procedure to, and what is the best idea that you can come up with?