4.14 Chemoinformatics

J. Polanski, University of Silesia, Katowice, Poland

Symbols

a^i	acceptor synthon located at the *i*th atom relatively to functional group heteroatom (i = 0)
d^i	donor synthon located at the *i*th atom relatively to functional group heteroatom (i = 0)
P	physical, chemical, or other properties, where *m and p* in subscript refer to measured and predicted values, respectively
S	structural properties
E	end reaction state matrix
B	initial reaction state matrix
R	reaction matrix

4.14.1 Introduction

Chemoinformatics (cheminformatics) is a term that has been coined recently to describe a discipline organizing and coordinating the application of computers in chemistry. Although computers have been assisting chemists for years, this term did not appear until recently. Thus, it is not surprising that not all chemists are impressed by this fact. Actually, there are a number of controversies over the necessity for the foundation of this relatively novel chemistry branch. Wendy Warr, who surveyed this issue among chemists, concluded this by stating, “some people felt that it was a neologism invented by information professionals who felt that chemical information was not a sexy enough name to safeguard their jobs. Opinion has now shifted towards acceptance of chemoinformatics as a discipline although not everyone agrees about the definition, or even about the syntax: 50% of respondents like chemoinformatics.”[1]

4.14.2 The Origins and Scope of Chemoinformatics

Chemoinformatics, which joins together chemistry and informatics, is evidently related to computer applications in chemistry. However, not all chemical branches that depend on computers should necessarily be included in the field. Clark asks the question: “Does quantum chemistry have a place in cheminformatics?”[2] Even though the author considered a very narrow research area for chemoinformatics, this causes a hesitation on a “possible role of quantum mechanical techniques in chemoinformatics”[2] and suggests the autonomy of quantum chemistry.[2,3] The essence of this discipline is the assumption that we do not need any specific chemical interaction for the explanation of chemical bonding. In principle, it is just the physics of atoms and mathematics that allow the correct modeling of molecular objects, and pure mathematics, hypothetically, can be done without computers. However, even today, such an approach has important limitations; we can investigate rather small molecules. The larger the molecules, the more remote and inaccessible the precise mathematical explanation. In fact, it appears that chemistry is often too elusive for a precise description of the molecular bodies. Because such bodies are the most substantial object of chemical investigations, this even provokes the question “is chemistry a science?”[4]

Philosophers have developed several theories to explain the origins of science. According to conventionalism, logical structures called laws of nature are created or invented, which are then verified by conducting experiments. Inductivism finds the origins in “collecting and classifying sensory input data into a form called observable facts”.[5] Inductive logic is then applied to draw general conclusions or laws of nature. Finally, for deductivists, theories are at the origins of science. Scientists can never prove the theory, but science develops through theory falsification.[5] Independent of the philosophy we would accept, we need data and theories for the development of science. We cite here Brock,[6] who discussed the history of fundamental concepts or theories in chemistry to illustrate the complexity of chemical researches. Consider chemical bonding. The molecular orbitals or valence bonding theories describe atomic scale aggregation into molecules. Both models have been competing with each other and chemists still discuss which is correct and which is better. In physics, it is possible to develop a relatively simple model to explain certain facts of nature. In contrast, in chemistry, a theory often partially interprets some data, also offering partial solutions. Thus, we need several high-quality models for the correct theory. Brock concluded that theoretical chemistry is still an empirical science based on the Schrödinger equation. It however appeared that a general solution of the equation will never be found.

Mathematics is an instrument used for modeling and developing theories, which means that it can be interpreted as a compression tool unifying the facts of nature. Now we do not need individual facts any longer, 'which disappear,' but a single equation that explains the reality. Reductionism is an approach that insists that a system complexity can be explained on another level by such a compressed model. For the illustrative discussion of these problems, the reader is referred to Cohen and Stewart.[7] However, the reality often appears to be too complex or even unavailable for an accurate mathematical description. Alternatively, a model developed can be too complex for a precise solution. Since we still need answers in such situations, we have to rely on simplifications, even if it would be less reliable. Eventually, speculation or educated guess is better than blind guess or no answer and "educated guess is being supported by the computer".[8]

This describes the first application of computers in chemistry, which is to assist a chemist in a calculation or computation that requires the calculation by computers. Why can computations still be possible, flexible, and efficient in data processing when human calculations fail? The efficiency of *in silico* mathematics[9] is achieved, first of all, not by computer intuition or flexibility but by a brute force that preserves mathematical rigor and formalism. This makes mathematical philosophy *in silico* evidently different from the human one. The enormous speed and competence in low-level manipulations coupled with human intelligence allowed computers to solve "formerly intractable problems, and explore areas beyond the reach of human calculation".[9,10] In this context, we can also outline the domain of chemoinformatics preferentially to data processing that cannot do without *in silico* mathematics, that is, those chemistry branches that depend on massive data that cannot be compressed to the standard mathematical models. Oprea suggested that we also do not include into chemoinformatics some traditional chemistry branches that are usually associated with computational chemistry but "generate more numbers than information (. . .), e.g., physical and chemical property calculation."[11] It seems however that in a more general way this refers to such operations that, even though they can be performed efficiently *in silico*, hypothetically, can be done without a computer on the basis of relatively simple mathematical equations.

Data storage systems is the second important field for the application of computers in chemistry. Chemistry starts from data, that is, facts and numbers, which when processed and delivered properly at a proper time and place make up information. Processing information in turn develops chemical knowledge. Chemistry focuses on atoms and molecules and their properties and transformations. A whole lot of matter available in the universe can be arranged to an unbelievably large number of molecules forming chemical data space. To illustrate the numbers, Chemical Abstracts Service (CAS) currently has registered almost 37 million chemical compounds, 60 million sequences, and 15 million single and multistep reaction data entries.[12] The population of chemical space (CS), that is, the number of potential compounds, is estimated between 10^{18} and 10^{200} (the number 10^{60} being cited most often), which can be compared to the factual CS (FCS) of the order of 10^{7} and an estimated number of stars in the universe of 10^{22}.[13,14] The expansion of CS can be even better illustrated if we analyze a single molecule of n-hexane substituted with 150 different substituents. Bringing together all mono- to 14-substituted molecules will give a molecular population of 10^{29}.[15]

The term CS itself is an example of the impact of mathematics on chemistry. In chemistry, this term appeared recently to illustrate the necessity for the control of the structural constitution of such a space or, in other words, the diversity of the molecular population investigated in combinatorial chemistry. In mathematics, a space is a set of a certain structure; in particular, a vector space is a set of multidimensional vectors in a generalized coordinate system. Mathematics demands some further conditions for such a space. Thus, an origin and a base (unit vectors in each dimension) are to be defined to form a space. The term CS used in chemical literature is a synonym of the chemical set including all possible chemical compounds. This often refers to virtual compounds, that is, those that have not already been synthesized. Mapping CS to biological space or to property space is a further borrowing from mathematics. **Figure 1** attempts to further organize chemistry in the form of CS. CS is formed by chemical molecules. A molecule is a vector, elements of which describe the structure (structural properties *S*) and chemical or physical properties, *P*. As shown in **Figure 1**, CS is constructed from two basic moieties, FCS, that is, real molecules forming chemical compounds that

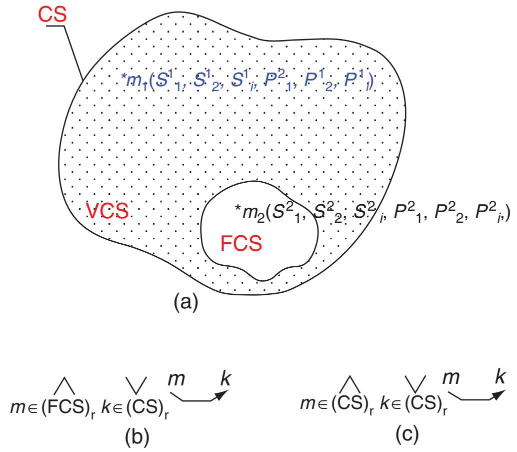

CS
$*m_1(S^1_1, S^1_2, S^1_i, P^2_1, P^1_2, P^1_i)$
VCS
$*m_2(S^2_1, S^2_2, S^2_i, P^2_1, P^2_2, P^2_i)$
FCS
(a)
$\bigwedge_{m \in (FCS)_r} \bigvee_{k \in (CS)_r} m \rightarrow k$
(b)
$\bigwedge_{m \in (CS)_r} \bigvee_{k \in (CS)_r} m \rightarrow k$
(c)

Figure 1 Chemical space (CS) consisting of factual (FCS) and virtual (VCS) spaces, **FCS** ∪ **VCS** = **CS** is formed of the molecules, where each molecule can be given by a vector (a). This provides a base for the definitions of molecular transformation operators capable of mapping molecular objects in CS, for example, organic synthesis *in vitro* operator (b) or reaction prediction operator *in silico* (c) (cf. Section 4.14.4.6). Two symbols *m* and *k* were used for the better illustration that mapping by these operators needs two different molecule types (*m*-reagents, *k*-products); *r* denotes the *r*-reaction domain in CS. The operator *in vitro* starts from FCS, while that *in silico* can work entirely in CS.

have already been obtained and described, and virtual CS (VCS), that is, hypothetic molecular structures. Accordingly, two common chemical problems (chemical synthesis and reaction predictions) are defined using a vector space formalism.

The investigation and construction of new chemical objects cannot be made without an efficient data mining system that allows verification and screening of physical and chemical characteristics among a variety of described compounds. Access to information is a fundamental problem in chemistry. This should enable delivery of proper data to chemists' desks where needed. Therefore, from the beginning, chemists developed documentation systems on chemical compounds. *Chemisches Zentralblatt* appeared as early as 1830; the first edition of Beilstein's *Handbuch der Organischen Chemie* was published in 1881 and contained two volumes, registering 1500 compounds, with more than 2000 pages. This comprehensive encyclopedia of organic structures covers chemical literature from 1771 to date.[16–18] *Chemical Abstracts* have been published since 1907. Unlike in other sciences, data storage systems are discussed in basic chemical handbooks, for example, *March's Advanced Chemistry*.[17] The improvement of data storage could have significantly stimulated the development of chemical sciences. Accordingly, chemical information branches have been keen to profit from computers. It is much more efficient to keep information on the computer desktop than just on the desk. Therefore, besides computations, chemical information formed an important component of chemoinformatics. Gasteiger illustrates this by the fact that, in 1975, the *Journal of Chemical Documentation* changed its name to *Journal of Chemical Information and Computer Sciences*.[19] If we think in a similar vein, we can use the same title to show recent developments in this field, since the journal name has just changed to *Journal of Chemical Information and Modeling*. Computer science is now too far from the chemical core, and chemists believe that they are generating their own tools for computer chemistry investigations. Willet suggests that this journal might today reasonably be entitled *Journal of Chemoinformatics*.[20] Thus, the journal's history briefly illustrates the scope of the discipline.

In fact, modeling is the next important problem in which chemists need computer assistance. The most obvious dictionary meaning of a model is 'a physical representation that shows what an object looks like.' For years, molecules were too small for direct observation and even today we usually watch them indirectly by analyzing measurable data. Therefore, from the very early days, chemists had to assemble physical objects resembling molecular scale shapes. Molecular models can be any physical representation of molecular configuration assigned to molecular objects that are constructed to understand and explain measurable characteristics manifested by molecules.[21] Molecular models (Dreiding, CPK, and so on) are so popular among chemists that our conception of molecules is predominantly shaped by such real-world reproductions. In contrast, macroscopic analogies provide only a model imitation, and simple hard sphere-like molecular representations cannot furnish the exact illustration of the microscopic bodies that can be described only by quantum mechanics. Although modeling is a broad term that describes a variety of methods, its substantial meaning in chemistry involves the construction and visualization of chemical molecules. The development of computer technology provides a virtual reality platform for chemistry that is known under the term molecular modeling.

One way or the other, increasing dependence on computers is a fact in modern chemistry. This has brought a need for better organization of this field. As far we have indicated, computations, data storage and modeling have potential computer applications in chemistry. In fact, these problems are also of fundamental importance for general computer sciences. Computer sciences, a term used in the United States, or informatics, coined as its synonym in Europe (for the discussion of the differences see Roberts[22]), can be defined as "the science of algorithmic processing, representation, storage and transmission of information."[23]

In general, such a definition also describes potential application areas for computers in chemistry. Consequently, a recent definition of chemoinformatics presented by Gasteiger in the *Handbook of Chemoinformatics* points for "the application of informatics methods to solve chemical problems."[24] This includes more specific descriptions of this field. Brown describes this discipline as "the combination of all the information resources that a scientist needs to optimize the properties of a ligand to become a drug."[25,26] According to Paris, chemoinformatics "encompasses the design, creation, organization, storage, management, retrieval, analysis, dissemination, visualisation and use of chemical information, not only in its own right, but as a surrogate or index for other data, information and knowledge."[1] Chemoinformatics should be interpreted as an element of knowledge management. This includes problems such as "compound registration into databases, library enumeration; access to primary and secondary scientific literature (. . .)."[27,28]

Chemical informatics is another term related to the application of computers in chemistry. It is noteworthy to indicate that it is the oldest computer chemistry representation that appears in the literature as early as the 1980s. Formal definition of the branch includes: "computer-assisted storage, retrieval, and analysis of chemical information, from data to chemical knowledge."[29,30] *Chemical Informatics Letters*, an open web access journal published since 2000, brings the latest news in this field. The website, edited by Goodman, is designed in a hypertext format, which makes a great difference to the standard form of a conventional journal.

Cheminformatics and chemiinformatics are synonyms that sometimes replace the term chemoinformatics[29] Finally, computer chemistry also seems to describe a similar chemistry branch. It is noteworthy that the research centers which are being explicitly called computer chemistry laboratories, for example, Labor fur computer Chemie at Technische Universität München, were established in the 1980s and 1990s. The history and operation of the European computer chemistry institutes can be found in Noordik.[31]

Chemistry is not the only science that has developed its own informatics. A variety of multidisciplinary informatics have appeared. Accordingly, bioinformatics relates to genetic information encoding living organisms' structures and processes. Medical informatics focuses on diseases, patients, and drugs. Crystalloinformatics and protein informatics are other examples of interdisciplinary informatics.

4.14.3 Teaching Computers Chemistry: Data Input Problems

Computer-understandable chemistry is required for the machines to process the data. At the same time, it is also required to enable an interaction between chemist and computer. It is not a trivial problem to translate structure data of molecular objects into a machine-readable and -processable system that is clear enough and unambiguous. Chemical molecules are the main object of chemical investigations. Molecules can represent both real chemical compounds that have been obtained previously and described, or virtual structures representing hypothetical compounds under design or speculation. Organic chemistry and inorganic chemistry are disciplines that construct such objects in reality, in the proportions of approximately 1:200 in favor of organic chemistry. To control CS, that is, all possible real or virtual molecules, we need to have efficient machine-searchable databases registering all compounds that have been synthesized by chemists from the very early days to today. This problem, which appeared in the 1960s, can be defined as structure representation and searching.[20] We discuss below several problems referring to structure representation, which is of substantial importance for the organization of chemistry *in silico*. Structure searching as a chemical operator *in silico* will be discussed in Section 4.14.4.3.

What we usually mean in the broadest sense by structure is chemical entity described by constitution and stereochemistry where constitution means "the description of the identity and connectivity (and corresponding bond multiplicities) of the atoms in a molecular entity (omitting any distinction arising from their spatial arrangement, i.e. – molecular stereochemistry."[32]

Atomic composition given by molecular formulae is not sufficient to unambiguously identify a molecule. Chemical entities of the same atomic composition but different constitution and/or stereochemistry are called isomers. The International Union of Pure and Applied Chemistry (IUPAC) defines isomers as "one of several species (or molecular entities) that have the same atomic composition (molecular formulae) but different line formulae or different stereochemical formulae and hence different physical and/or chemical properties." A line formula is constructed by indicating atoms that are "joined by lines representing single or multiple bonds, without any indication or implication concerning the spatial direction of bonds."

The discussed rules allow the chemist to define unambiguously chemical entities that are characterized by certain structure or structure properties, as suggested in **Figure 1**. If we refer to a molecule defined according to **Figure 1**, m(S_1, S_2, S_i, P_1, P_2, P_i), then we can make further discrimination of properties into molecular properties and chemical properties. For example, structure property can refer to both a molecule (molecular surface, molecular volume, 3D structure) and chemical compounds (3D crystal structure). Similarly, chemical or physical property can describe a molecule, for example, polarizability, and a chemical compound, for example, melting point. **Figure 2** illustrates the basic terms that refer to molecular objects in FCS and VCS.

It is worth mentioning that in the majority of chemical applications stereochemical description does not include a precise description of the real 3D molecular structure (which is known relatively rarely), but rather its rough scheme. This is shown, for example, in **Figure 3**, which illustrates two hypothetically possible 3D structures of *trans*-1,2-dibromocyclohexene.

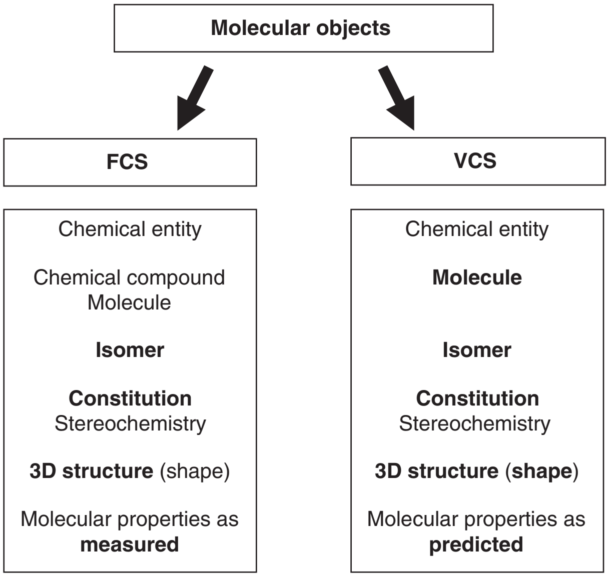


Figure 2 Molecules are substantial objects of chemical investigations both in FCS and VCS. It is not easy to differentiate the terms that are used to describe molecules in these spaces. However, some differences can be definitely indicated, for example, in experimental FCS chemistry we are only very rarely investigating a single molecule. Chemical compound, that is, a population of molecules interacting with each other or agglomerated into a solid or liquid phase, predominantly focuses our attention. In contrast, theoretical methods often focus on a single molecule.

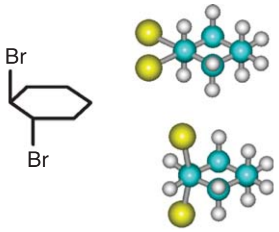


Figure 3 Two hypothetically possible 3D structures for *trans*-1,2-dibromocyclohexene.

4.14.3.1 Computer-Processable Molecular Codes

Kekule was the first who realized the formation of carbon chains and rings. However, with the exception of the so-called sausage formulas, he did not use graphical representation of the molecules. Couper, independent of Kekule, developed the concept of a four-valence carbon atom and presented carbon chains in the form of atoms connected by dotted lines; finally, Crum Brown developed and popularized a molecular notation similar to that used today. The evolution of molecular graphs illustrating molecular objects is briefly outlined in **Figure 4**.

Molecular graphs illustrating 2D atomic arrangement of chemical entities are easy to read for chemists. Although current computer systems are prepared to understand molecular graphs, generally such a form is not computer-friendly and needs a transformation before it can be presented to the computer. Linear notation and connection tables are two systems that enable an efficient coding of molecular graphs. Linear notation is a system that allows a molecule to be represented in the form of a string similar to that of line formulae. The Dyson, Wiswesser (WNL), Sybyl, *R*epresentation of *S*tructure *D*iagram *A*rranged *L*inearly (ROSDAL), which was developed by Beilstein Institute, and Simplified Molecular Input Line Entry *S*pecification (SMILES) notation are several systems used.[34] For a detailed discussion of the chemical structure notation, the reader is referred to a number of monographs available.[20] SMILES is probably the most popular line notation currently applied to a number of environments; for example, **Figure 5** illustrates drawing the structure of benzamide in ACD ChemSketch freeware and explicitly writing its SMILES code into the appropriate window.

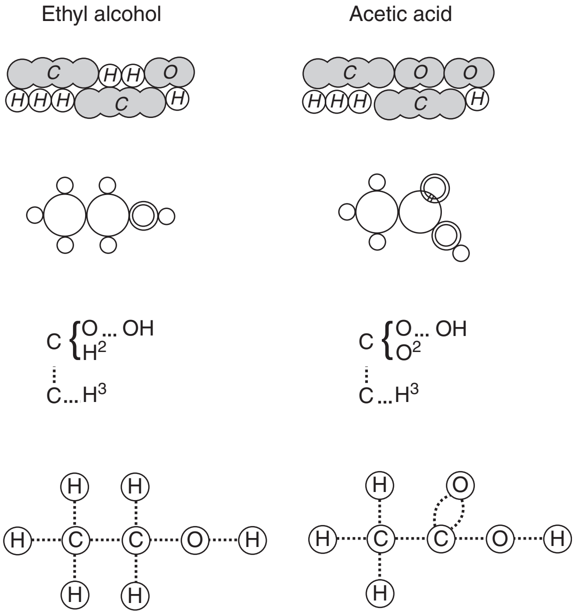


Figure 4 Ethanol and acetic-acid formulae as shown by Kekule, Loschmidt, Couper and Crum Brown, from top to bottom, respectively. Adopted from Ihde, A. J. *The Development of Modern Chemistry*; General Publishing Company, Ltd.: Don Mills; 1984.

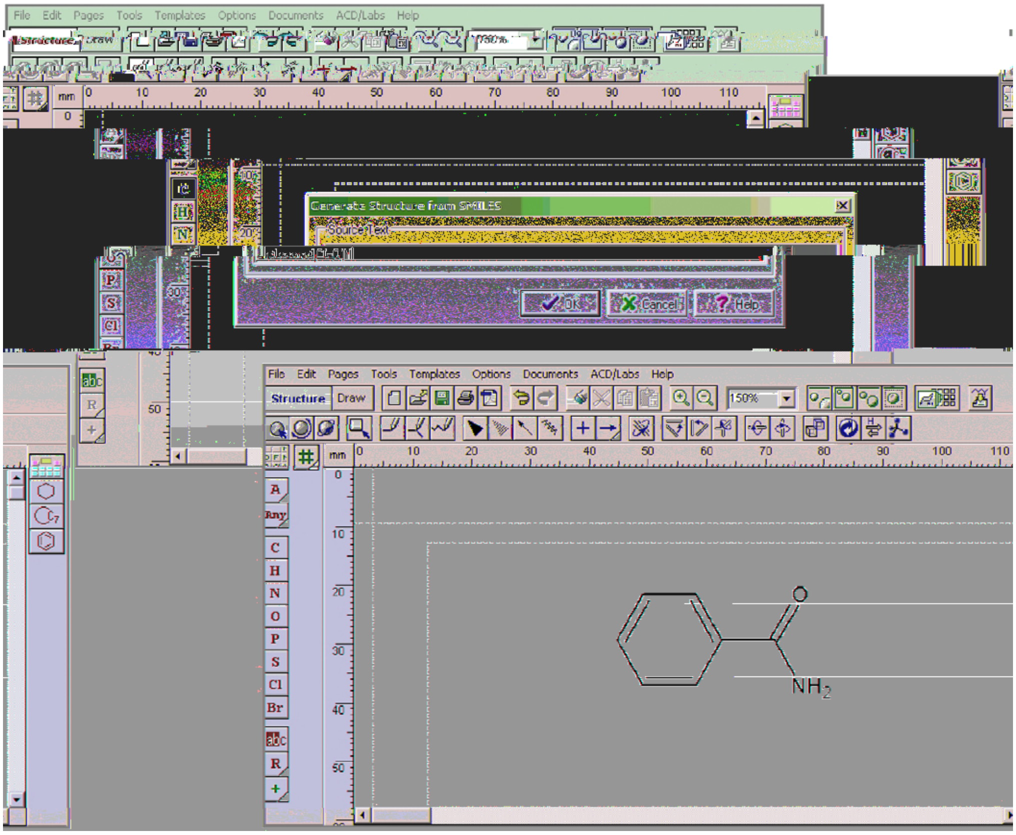
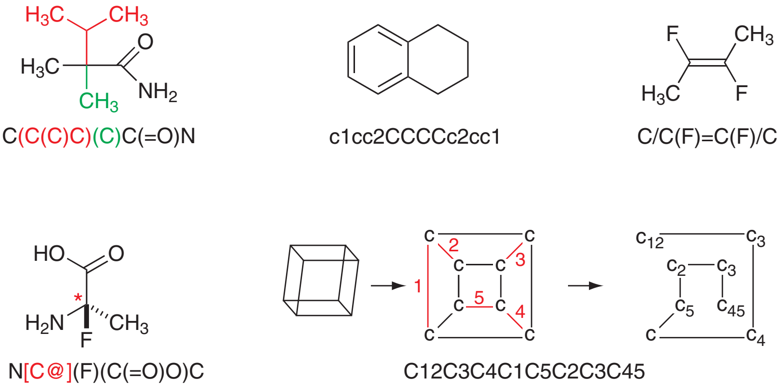

File Edit Pages Tools Templates Options Documents ACD/Labs Help
Structure Draw
150%
Generate Structure from SMILES
Source Text
OK
Cancel
Help
A
C
H
N
O
P
S
Cl
Br
R
O
NH2

Figure 5 Generating a molecular graph from its SMILES notation in ACD ChemSketch.

See SMILES manuals for the detailed code rules.[35] An excellent tutorial is also available online from Daylight Chemical Information Systems.[36] Several illustrative examples for the molecules coded by SMILES are shown in Figure 6.

Chemical graphs can be coded by matrices. Adjacency matrix, atom connectivity matrix, incidence matrix, and bond electron matrix are only few examples of the possible notations.[24]

Connection tables are another possibility for coding molecular structures. Connection tables record, in a tabular form, only the atoms and bonds within a molecule. In contrast to matrix notation, this allows a decrease in the amount of data with increasing molecular size. Figure 7 illustrates an example of a connection table in



C(C(C)C)(C)C(=O)N

c1cc2CCCCc2cc1

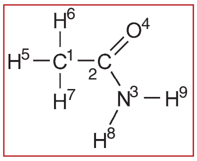
C/C(F)=C(F)/C

N[C@](F)(C(=O)O)C

C12C3C4C1C5C2C3C45

Figure 6 An example of SMILES coding several different molecules.

Atoms	
1	C
2	C
3	N
4	O
5	H
6	H
7	H
8	H
9	H



Bonds		
1st atom	2nd atom	Bond
1	2	1
2	3	1
2	4	2
5	1	1
6	1	1
7	1	1
8	3	1
9	3	1

Figure 7 Connection table coding acetamide molecule.

the form of explicit, redundant, and nonredundant connection table. An in-depth description of the matrix and connection table codes can be found in Gasteiger and Engel.[24]

A connection table or linear notation can be formed arbitrarily. This means that numbers can be assigned to the atoms differently and there is no standard molecular representation. Canonical labeling is a solution for this problem. This provides a unique representation for a certain molecular graph. Unique SMILES is an example of such a canonical labeling system.[36]

Chirality is an important chemical structure property and isomeric SMILES is a system that allows various chiral and isotopic specifications.

4.14.3.2 Molecular Editors

Molecular graphs are an unambiguous, chemist-friendly, and illustrative way for the presentation of constitution and stereochemistry of molecules. Molecular Editor is an interface that not only allows a user to draw professionally presented molecular structures, but also acts as a tool for the translation of such a structure into computer-processable molecular codes.

A number of systems have been developed that are capable of translation of molecular formulas introduced into a computer by its user in the form of direct drawing; examples range from using a mouse to machine-readable code. ISIS,[37] ChemSketch (ACDLAB),[38] JME,[39] and RasMol[40] are molecular editors available free of charge at their respective websites.

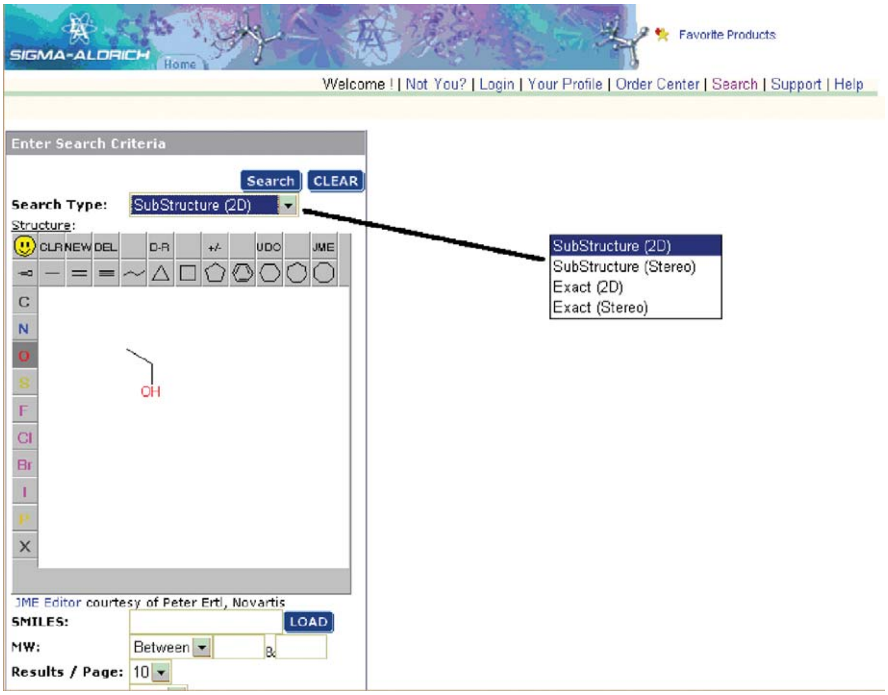
We cannot discuss here all of the above-mentioned software, but we will concentrate on JME editor, which was programmed by Peter Ertl from Novartis. “Since molecular construction and editing are indispensable for chemical information systems, and in 1994 no such tool was available for the WWW, we decided to develop our own WWW-based molecular editor. This editor was based on a clickable map.” Adding atoms, rings, and functional groups, connected by bonds, is achieved by choosing “the desired action from the menu, and then picking the appropriate place on the drawing area.”[39] Currently, JME is a Java applet that allows input of a molecular structure by drawing its graph within hypertext directly on the website operated. A number of organizations using JME can be found at the Molinspiration website.[39] **Figure 8** illustrates the applet mounted at the online catalog of the Sigma-Aldrich fine chemical supplier.

Sometimes, it is helpful for chemical documentation to transform 2D molecular illustration presented on a sheet of paper into the form of a connection table. The Clide program is an optical character recognition (OCR) system that performs such a transformation.[41]

Although we may question drawing a molecule in a molecular editor by writing its code instead of using a mouse, this method is much more convenient in a number of situations, for example, when a number of structures are to be generated via an automatic approach.

4.14.3.3 Computer-Oriented Chemical Compounds Nomenclature

Chemical nomenclature is an example illustratively showing the differences between chemist and computer when acquiring and processing chemical data. Chemical molecules can be complex and it is often impractical to use their explicit structures in the form of molecular graphs, connection tables, or similar notation systems just



Search Type: SubStructure (2D)

SMILES:

MW: Between

Results / Page: 10

Figure 8 The JME molecular editor mounted at the website of Sigma-Aldrich fine chemical supplier. The results of substructure searches using this applet are shown in **Figure 16**.

to designate a proper chemical entity. In particular, this relates also to verbal communication among chemists. Thus, what is needed in chemistry is a human-friendly name system that allows identification of individual chemical molecules. In fact, the identification of compounds by assigning them certain names occurs prior to other designations. For example, water is identified as a substance that is absolutely necessary for human life and found in the environment as a relatively pure chemical compound. Although the name came before we had gained any idea of its chemical constitution and structure, it is adopted into the chemical nomenclature to label a water molecule. The formal designator oxygen dihydride is only very rarely used to name this compound. Faraday was the first to isolate a substance which he named bicarbuet of hydrogen. This was renamed to a simpler and still used name, benzin, or benzene in English, by Eilhard Mitscherlich, who obtained this compound by thermal treatment of lime and acid isolated from gum benzoin, a balsamic resin of tropical Asian trees of the genus *Styrax*. The formal name that explains the compound structure is cyclohexatriene.[42] When chemists obtain a compound, they often use simple names that describe the origin of compound, commemorate an events (for example Olympiadane)[43] or person (for example Buckminster-fullerene)[44] point to other associations, or just use acronyms enabling clear and rapid identification of structures. Some illustrative examples are given in Figure 9.

Although trivial names are unique and clear for people working in the field, their etymology and sense can be completely misty for the general chemical audience, especially after some time has passed from the synthesis and/or isolation of the molecule. Therefore, a systematic nomenclature scheme has been developed by chemists starting from the recommendation of the 1892 Geneva Convention of IUPAC. This system has been continuously improved by IUPAC.[45] Some other nomenclature systems have also been developed but have never come into extensive application.[46] Theoretically, IUPAC rules should relate to all factual and virtual structures, that is, a name can be generated for each molecule, regardless of its structural complexity. However, in reality the complexity of possible molecular structures requires many rule extensions and restrictions. In fact, IUPAC regularly provides recommendations on the nomenclature for the novel classes of compounds that appear. Fullerenes, or phanes, are examples of such relatively novel classes with special IUPAC nomenclature rules published. A perfect nomenclature should not only be unambiguous but also

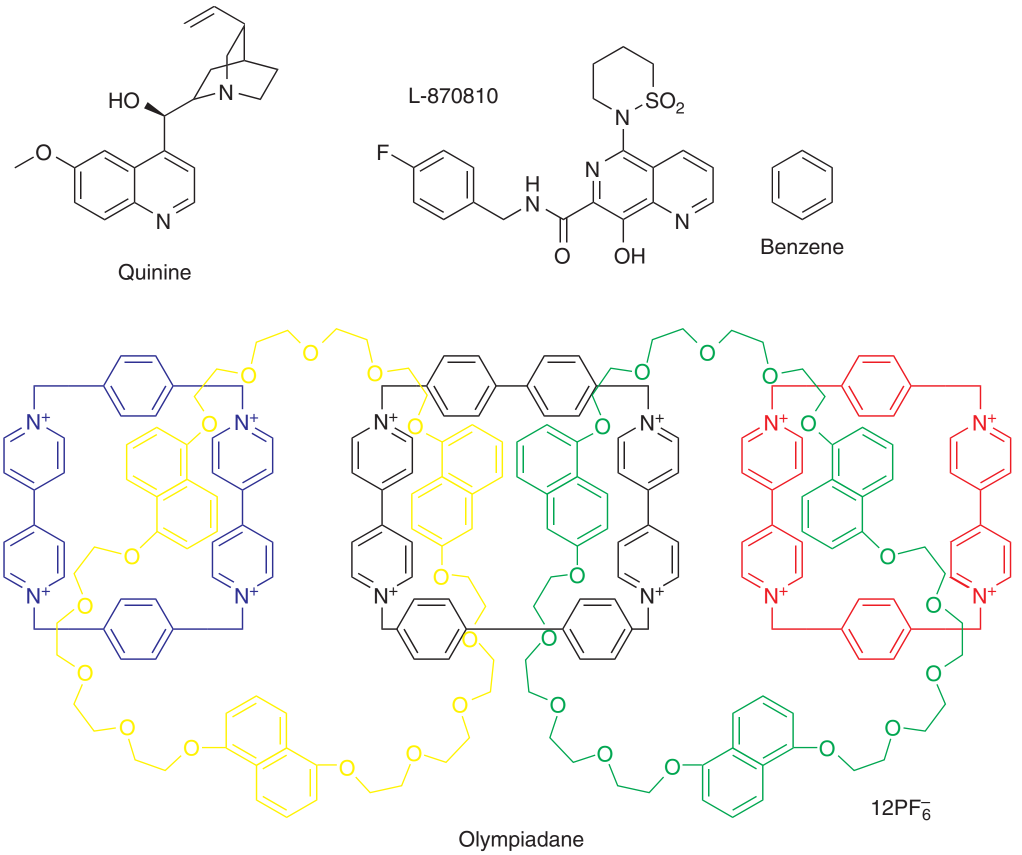


Figure 9 Trivial chemical names: quinine named after the quinine tree bark, benzene (benzin) named after *gum benzoin*, chemical acronym L-870810, and Olympiadane to commemorate the Olympic Games.

unique. This means that we should not only clearly identify the structure given a name label but also a single name label should describe only a single chemical entity. IUPAC rules are human-oriented, which means that easy chemical nameability and name readability by human chemists has been given the highest priority. A human-friendly nomenclature does not necessarily meet the requirements of a computer-oriented system. For example, the IUPAC system does not restrict the names generated for a single structure to a unique value. This problem is still a challenge and is to be solved by the preferred name program (PNP) currently realized by IUPAC.[47] In practice, Beilstein and CAS, two main chemical information suppliers, adopted their own rules to provide different nomenclature systems that obey IUPAC rules but restrict them.[46]

Although it may sound surprising, until recent years a name to structure conversion has not been solved in a general way. The first system that allowed input of chemical structures in the form of their chemical names was developed by Beilstein in 1986. This was operated internally at the Beilstein Institute, and the structures input was restricted to the Beilstein notation subrules.[46]

Several other converters are available now, for example within the Advanced Chemistry Development (ACD) LAB program.[38] However, this system is restricted quantitatively to a name length up to 255 characters including spaces, punctuation marks, and others, and up to 255 heavy atoms in generated structure. Several further limitations concerning the nomenclature require the user to input special name representation for correct recognition. Finally, ACD/Name to Structure software is not present in a freeware package of the ACD ChemSketch but should be purchased as an additional module.

Although the Beilstein Institute allows a user to search its database by the chemical name (CN) field using the IUPAC-based name, that used in the Beilstein Handbook (*Beilsteins Handbuch der organischen Chemie*) (BH) (or the names used in original publication, those generated by AutoNom, the structure to name converting tool (cf. Section 4.14.4.1), available within the database) is preferred. Otherwise, according to the Beilstein database help, 'name searches are not recommended to identify compounds, because names are ambiguous or not systematic in many cases.'

4.14.3.4 Coding Chemical Reactions

Atom bonding systems in molecules can change during a process described as chemical reaction. Chemical reaction involves the breaking and formation of chemical bonds. Chemical compounds or reactants to be converted are transformed during chemical reactions to reaction products.

The problems of chemical reaction nomenclature resemble those of the description of chemical compounds. Many reactions, honoring distinguished chemists, are named after the discoverers. This naming corresponds to trivial chemical compounds nomenclature. In fact, there is no information on the reaction itself within its trivial name. *Merck Index* is a popular compendium book that provides a guided tour through name reaction chemistry.[48] Similarly, the Organic Chemistry Portal offers an excellent web-based name reaction database.[49]

However, the accumulated chemical reaction resources needed more systematic classification and nomenclature that would give more detailed information on the particular molecular transformation. The most substantial classification of organic reactions groups them into four classes: substitutions (exchanges), additions, eliminations, or rearrangements. The IUPAC Commission on Physical Organic Chemistry developed systematic nomenclature for the reaction grouped into several classes.[17] Precisely, this system describes the rules for the nomenclature of eight reaction types, that is, substitutions, additions, eliminations, attachments and deattachments, rearrangements, coupling and uncoupling, insertions and extrusions, and ring openings and closings. This is briefly illustrated in **Figure 10**.

Although the IUPAC system seems to be attractive and universal, officially it has not been used in any single organic chemistry handbook with the exception of the recent issue of March's *Advanced Organic Chemistry*. Such reaction class description is also too rough for the precise coding of the molecular transformations of a certain reactant to individual product.

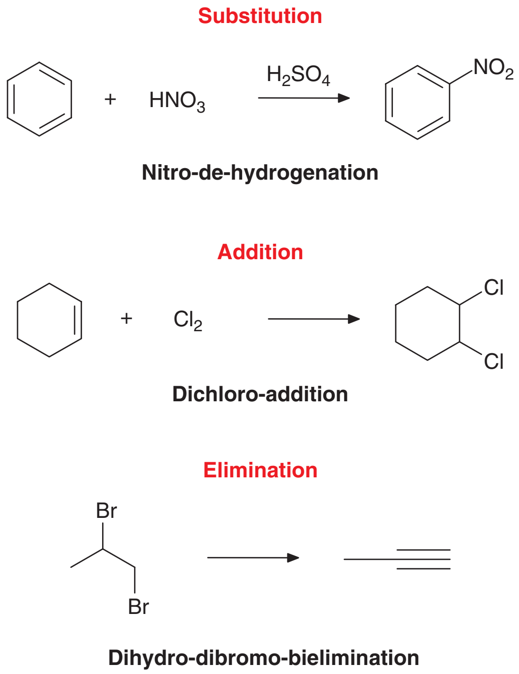
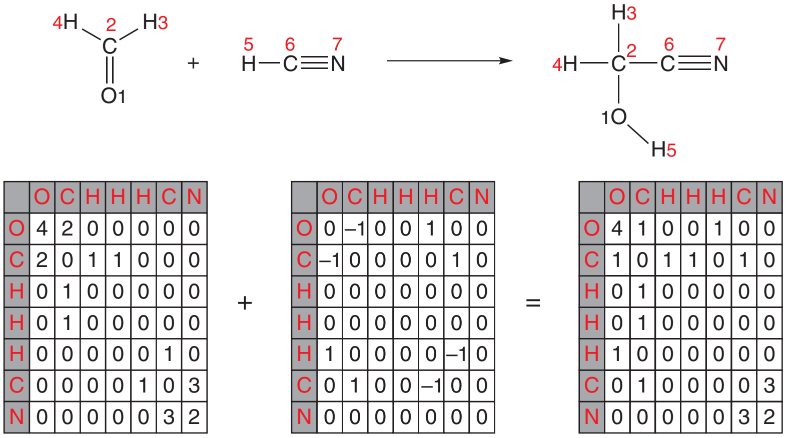


Figure 10 Reactions named according to the rules of IUPAC Commission on Physical Organic Chemistry. Adopted from Smith, M. B.; March, J. *March's Advanced Organic Chemistry Reactions Mechanisms, and Structure*; Wiley: New York, 2001.



$$\begin{array}{c|ccccccc} & O & C & H & H & H & C & N \\ \hline O & 4 & 2 & 0 & 0 & 0 & 0 & 0 \\ C & 2 & 0 & 1 & 1 & 0 & 0 & 0 \\ H & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ H & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ H & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ C & 0 & 0 & 0 & 0 & 1 & 0 & 3 \\ N & 0 & 0 & 0 & 0 & 0 & 3 & 2 \end{array} + \begin{array}{c|ccccccc} & O & C & H & H & H & C & N \\ \hline O & 0 & -1 & 0 & 0 & 1 & 0 & 0 \\ C & -1 & 0 & 0 & 0 & 0 & 1 & 0 \\ H & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ H & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ H & 1 & 0 & 0 & 0 & 0 & -1 & 0 \\ C & 0 & 1 & 0 & 0 & -1 & 0 & 0 \\ N & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array} = \begin{array}{c|ccccccc} & O & C & H & H & H & C & N \\ \hline O & 4 & 1 & 0 & 0 & 1 & 0 & 0 \\ C & 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ H & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ H & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ H & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ C & 0 & 1 & 0 & 0 & 0 & 0 & 3 \\ N & 0 & 0 & 0 & 0 & 0 & 3 & 2 \end{array}$$

Figure 11 Reaction coded by the **B** + **R** + **E** matrix. Adopted from Gasteiger, J.; Engel, T. *Chemoinformatics: a Textbook*; Wiley-VCH: Weinheim, 2003, p 186.

An illustrative algebraic model for the description of molecular transformations has been developed by Ugi and coworkers. This is based on logical connectivity and matrix addition. In this notation, the reaction is represented by the matrix equation $\mathbf{B} + \mathbf{R} = \mathbf{E}$, where **B** (beginning) represents an initial reaction stage, **E** (end) codes the final state, and **R** is a reaction matrix. Figure 11 illustrates an example of the reaction noted in such an approach.[24]

The Dugundji–Ugi (DU) notation not only provides an elegant and clear coding system for molecular transformations, but also reveals some further interesting features. For example, the **R** matrix indicating the distance between **B** and **R** similar to a real reaction designs an important measure describing the extent of valence electron shifts needed for **B** to **R** conversion; therefore, it directly explains the real chemistry of the transformation. Other representations and classifications of chemical reaction have been developed but will not be discussed here further and the reader is referred to Gasteiger and Engel[24] and Chen.[50]

4.14.3.5 Organizing Chemical Facts into Databases

Finally, what we need to enable chemists using computers to perform efficient chemistry is access to chemical information, that is, chemical data represented by chemical facts and numbers. Thus, for example, coding chemical transformation as described in Section 4.14.3.4 does not provide factual information gathered in chemistry on this specific reaction; for example, no information on reaction conditions, solvents, temperatures, catalysts, by-products, can be found in the **B**, **R**, and **E** matrices. These data are, however, fundamental for chemical research. Therefore, a number of chemical databases have been converted into a form compatible with the computer platform. Chemical data organized in searchable databases form a focal point of chemoinformatics. Chemical compounds and reaction databases such as Beilstein and Chemical Abstracts, patent databases such as esp@cenet, chemical substance catalogs, for example, Aldrich, and a variety of chemical journals are the sources that are available online with user-friendly interfaces. The impact of searchable chemical databases on chemical research is discussed in Section 4.14.4.8. Table 1 specifies several databases available online. An extensive list of a number of other databases is available on the web.[51]

4.14.4 *In Silico* Chemistry: Data Processing and Data Output Problems

Computers equipped with chemical information and software capable of understanding chemical data provide a chemoinformatic platform advising and assisting chemists in their research. Some of the problems appearing during an interaction between a chemist and computer in the course of data processing and data output are discussed below.

Table 1 Some chemical databases available online

Provider	*Data available*	*Web address*
Beilstein Institut	Chemical compounds and reaction	www.beilstein-institut.de
CAS	Chemical information (literature bibliography)	www.cas.org
Sigma-Aldrich, Fluka Supelco	Commercially available chemicals	www.sigmaaldrich.com
NIH	HIV therapeutics database	http://chemdb2.niaid.nih.gov
NIH, National Center for Biotechnology Information (NCBI)	PubMed – literature database	www.ncbi.nlm.nih.gov[a]
National Institute of Advanced Industrial Science and Technology (AIST)	Spectral database	www.aist.go.jp[a,b]
NIH	An extensive list of chemistry databases on small molecules	http://cactus. nci.nih.gov/
European Patent Office	esp@cenet patent database	www.espacenet.com
NCBI	Various molecular databases including Pubchem Compound - chemical compounds database	www.ncbi.nlm.nih.gov
eMolecules, Inc.	Chemical molecules, spectra, suppliers, etc.	www.emolecules.com
Elsevier, MDL	Discovery Gate, small molecule database environment that enables the simultaneous searches of several different databases (including Beilstein and patent databases)	www.discoverygate.com

[a] Several other protein, structure, etc, databases are available at this address.
[b] www.aist.go.jp/RIODB/SDBS/cgi-bin/direct_frame_top.cgi?lang=eng.

4.14.4.1 Computer-Generated Chemical Names

Chemical names generators realizing a structure to name transformation are generally supplied with a molecular editor that enables introduction of molecular structure in the form of a 2D graph. The Autonom program developed in the Beilstein Institute was the pioneer in this field. Wisniewski discussed and designed algorithms that included the following components:[46]

- structure initialization,
- functional group identification,
- ring perception and recognition,
- parent structure selection,
- binary name tree processing,
- chemical name assembly.

Functional group identification is a table-driven approach that enables recognition of favored atom groups known as functional groups, which are then ranked according to the rules predefined by IUPAC. Officially, the approach adopted involves "rapid atom by atom connectivity search mechanism" similar to that used in substructure searches.[46]

Cycle systems formed by atoms or their assemblies are important components determining chemical names. Thus, all cycle closures within the smallest sequence of atoms are to be identified. The so-called smallest set of smallest rings (SSSR) algorithm is used for the correct identification of the cycle structures consistent with nomenclature rules. The ambiguity of the cycles' identification within chemical graphs can be illustrated by the topology of a simple tetrahedron having four faces, three rings, but six valid SSSRs.[36,52]

The cycle perception step described above is a preliminary step that allows a program to identify certain ring classes, for example monocyclic alkanes, bicyclic alkanes, monospirocyclic alkanes, or trivial name ring systems whose names are obtained by using a lookup dictionary procedure. A collection of detailed rules and routines describe naming for each individual class. During the parent structure selection step, the candidate structural fragments, mainly rings and chains, are screened. Global regulations that rule the structure of a name generated are a sequence of principles that obey IUPAC nomenclature. Nonparent structure fragments are then introduced as substituents and subsequent substituents on substituents. The so-called binary name tree processing is then performed. During this step, the parent molecular fragment becomes the root of the tree, and other tree nodes represent other identified units that are to be named. Processing of the name tree starting from the root gives the final preliminary name assembly that includes, for example, punctuation and locants. A special control is then applied for the identification of the trivial name blocks that are preferred by IUPAC, for example, AutoNom generates the name benzoic acid and not benzene carboxylic acid. The success ratio of this program amounted to 86.3% when tested for more than 63 000 sample structures. The current version of the AutoNom program allows a user to generate both the name forms consistent with the Beilstein or ACS nomenclature.[46]

ChemSketch, developed by Advanced Chemistry Development Inc., is a freeware part of the extensive software system that can be downloaded directly from the ACD/Labs Internet site.[38] As a freeware version it allows users to generate a name for 'molecules containing no more than 50 atoms, and no more than 3 rings, with atoms from among only H, C, N, P, O, S, F, Cl, Br, I, Li, Na, and K.' The ILAB is an interface for the charged ACD/Labs Online service enabling extension of this in a pay-per-use fashion.[38] Similarly to Beilstein, the ACD generator also provides the names in their Beilstein or ACS version.

4.14.4.2 Molecular Modeling

Molecular modeling is a method that includes a variety of computational schemes that are aimed at simulating molecular structures, their properties and behavior *in silico*. In particular, this should also include molecular manipulations, that is, visualizing molecules on the screen using different modes, merging molecules, superimposing, and rotating molecules in space and bonds within individual molecules, and so on, as well as molecular predictions, that is, predicting molecular shape by 3D structure generation and modeling or forecasting chemical properties or eventual biological activity or effects. In particular, modeling virtual molecular structures themselves is not a trivial problem and can be achieved on the different level of approximation. For a brief introduction into general problems and applications of molecular modeling, the reader is referred to Höltje *et al.*[53]

4.14.4.2.1 Structure generators

4.14.4.2.1(i) 2D structure generators In novel approaches we often sample VCS by systematically changing various molecular moieties in the user-directed mode. This can demand generation of thousands or even millions of structures and this operation can be achieved only by using the automated way. Such an operation can be easily programmed in a variety of environments, for example MATLAB, basing on SMILES codes whose syntax is simple enough. The 2DCOOR program is an example of a 2D structure generator available from Molecular Networks.[54]

4.14.4.2.1(ii) 3D molecular structure In a variety of chemical research, we simplify the real structure of a chemical molecule to its molecular configuration (cf. Section 4.14.2). What we usually mean by molecular configuration is a simplified 3D molecular structure, for example, we are classifying E and Z isomers as two different configuration series, although some other effects such as steric hindrance can further affect individual structures. Actually, in organic chemistry, we often rely on such simplification. However, molecules are 3D objects, which means each atom can be described by its exact space location. We can observe this by applying X-ray diffraction pattern on crystals, which allows us to reveal the 3D structure of the atomic lattice and thus to describe the 3D structure of the molecule. This effect is limited to condensed matter (crystals). Although there are many further approaches that allow chemists to disclose some structural data concerning the 3D atomic pattern, for example, by the application of NMR, current physics and chemistry do not have general technology for the observation of the 3D molecular structure. X-ray crystallography poses problems related to production of crystals, which is not always an easy task, and there is also the question of the relationship between condensed matter atom configuration and configuration in other environments. Even though nowadays we have data for quite a number of structures (**Figure 12**) including peptides or drug–ligand complexes, it is only a small percentage of the compounds described.[55]