Molecules in Silico: The Generation of Structural Formulae and Its Applications

Adalbert Kerber, Reinhard Laue, Markus Meringer, Christoph Rücker

June 25, 2004

Abstract

Computer Chemistry is of quickly increasing importance, in particular since the flood of data is rapidly growing with the introduction of Combinatorial Chemistry using methods of synthesis in large quantities (libraries) and high throughput screening. The accompanying software that allows to optimize such experiments in advance and to economize the cost of measurement afterwards uses in particular mathematical models of molecules and their description. Such models will be described here and the present state of the generation of molecular models in the computer will be discussed, from the mathematical point of view. A brief description of several applications is given.

1 Molecules

Molecules are entities consisting of atoms that interact, their descriptions are approximations on different levels of exactness. We can easily distinguish the following levels of increasing accuracy:

- The lowest level is the *arithmetic* one, where a molecule is described by its *molecular formula*, i.e. by a list of atoms contained in it. For example, the molecular formula of benzene is $\mathsf{C_6H_6}$, it says that a benzene molecule is made of six carbon and six hydrogen atoms.
- The next level is the *topological* or *constitutional* one. A molecule's structural formula (sometimes simply called structure) describes pairwise interactions between atoms, called covalent bonds. For example, to the molecular formula $\mathsf{C_6H_6}$ there correspond altogether 217 (mathematically possible) interaction models of this kind (constitutional isomers).
- A higher level of accuracy is the *geometric* level, where the molecule is placed in 3D space and is described by the coordinates of its atoms. This is the level of *stereoisomerism*. According to a useful traditional distinction made in organic chemistry we distinguish two sublevels.
  - The first sublevel is that of *configuration*. Configuration is what remains constant under small to moderate changes of the atom coordinates. More exactly, a configuration describes the sense of orientation of four positions in 3D space (enantiomerism) or in a plane (E/Z isomerism). A sentence such as "There exist exactly three stereoisomers of tartaric acid" makes sense because usually for a particular constitution a limited number of such configurations exist despite an infinite number of possible arrangements of the molecule's atoms, and despite a finite but usually very high number of conformers. In other words: It is possible to enumerate all stereoisomers of any constitution without knowing the atoms' exact coordinates. For this purpose, all that is required is the knowledge of the rough

geometric environment of a saturated C atom (approximately tetrahedral), of a C atom engaged in a double bond (approximately planar trigonal), etc. For example, for the 18 C_8H_{18} constitutional isomers we obtain 24 stereoisomers.

– The second sublevel is that of *conformation*. Here the exact numerical coordinates of the atoms are considered. For every particular constitution and configuration an unlimited number of conformations exists, many of which are (local) minima on the energy hypersurface (conformers).

At present we are able to handle the arithmetic and topological levels reasonably well. On the level of configuration there is some progress, though there are still several problems to be solved. Generation and classification of conformers is still largely unsolved.

2 The Arithmetic Level

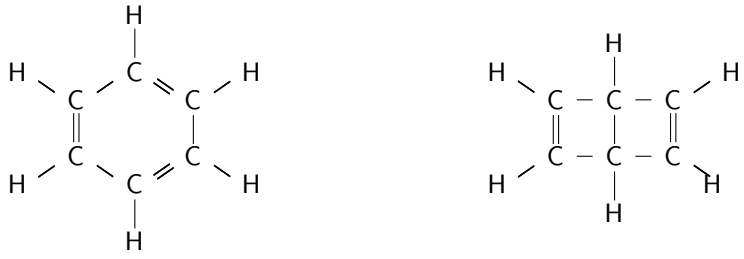
A molecule is described on this level by its *molecular formula*, a list of atoms of which it is made. Thus, C_6H_6 is the molecular formula for benzene and its 216 constitutional isomers. C_6H_6 is thus a valid molecular formula, in contrast to an invalid formula, one that does not correspond to a molecule capable of existence, for example, C_6H_5. Obviously there are restrictions on molecular formulae, and we will consider these in the next section. We mention that along with the usual well–defined molecular formulae there are fuzzy formulae, those that consist of intervals for the occurrence numbers of elements.

3 The Topological Level

On this level of approximation a molecule is described by an *interaction model*, which means that we emphasize interactions between pairs of atoms in the molecule. Mathematical structures that can be interpreted as interaction models are in particular the *unlabelled multigraphs*. The vertices of a graph indicate the objects involved, while the edges connect the interacting ones, and different strengths of interactions are expressed by different multiplicities of the edges.

3.1 Definition *A* structural formula *is a (usually connected) multigraph, the vertices of which are colored by element symbols. Moreover, the degree of each vertex, i.e. the number of edges to which this vertex belongs, agrees with the prescribed valence of the corresponding atom.*

For example, the structural formulae for C_6H_6 are the connected multigraphs consisting of six vertices of degree 4 and six vertices of degree 1. The former are colored by the letter C, the latter by H. Here are two such formulae (out of 217), those of benzene and of Dewar benzene:



All 217 are the interaction models of constitutional isomers of benzene, and they can easily be obtained (in a fraction of a second) using the molecule generator MOLGEN. Clearly most of them will not be structural formulae of existing molecules, but there are no strict rules known to distinguish between molecules merely not yet synthesized and molecules not capable of existence. Attempts were made to define the latter in terms of "forbidden" (too high in energy) substructures, but this approach met with failure in that on inspection of the Beilstein database almost any "forbidden" substructure was eventually found to occur in a known compound. *So whenever completeness matters, e.g. in structure elucidation, all mathematically possible graphs should be constructed.*

Accepting the notion of connected multigraph as an interaction model for molecules we can impose restrictions on molecular formulae: If γ is a multigraph, then we can use the sequence of its vertex degrees

$$\lambda_\gamma := (\lambda_\gamma(0), \lambda_\gamma(1), \ldots),$$

where $\lambda_\gamma(i)$ means the number of vertices of degree i in γ. It is a partition of the number n of vertices in γ. We abbreviate this fact by

$$\lambda_\gamma \models n.$$

To begin with, the following expression gives the number $e(\gamma)$ of edges of γ,

3.2 $$\sum_i i \cdot \lambda(i) = 2 \cdot e(\gamma).$$

This is a formulation of the simple fact that each bond connects two atoms. Assume that the degree sequence of an interaction model of a molecule with molecular formula $\mathsf{C_6H_5}$ is $\lambda_\gamma = (0, 5, 0, 0, 6, 0, \ldots)$. Then we obtain $\sum_i i \cdot \lambda(i) = 29$, an odd number, which violates 3.2. Hence $\mathsf{C_6H_5}$ is not a valid molecular formula.
Moreover, if we assume that the interaction models are always *connected* graphs, then we can use the following necessary and sufficient condition for the existence of at least one connected multigraph in terms of the partition λ_γ :

3.3 $$\sum_i i \cdot \lambda(i) = 2 \cdot e(\gamma), \text{ and } e(\gamma) \geq n - 1,$$

where the inequality on the right says that there are at least $n - 1$ edges required in order not to leave a vertex isolated. So we obtain the following test for the validity of a given molecular formula:

3.4 Corollary *A given molecular formula and a corresponding sequence λ of degrees together describe a valid molecular formula only if they satisfy condition 3.3.*

4 Molecular Graphs

The above definition of structural formula needs to be refined to the notion of molecular graph that we are going to introduce now.

Chemical compounds are described by multigraphs consisting of particular vertices representing *atoms* and edges representing *covalent bonds*. These bonds may be single, double or triple bonds. The vertices are colored by the name of a *chemical element* and an *atomic state*.
A chemical element is identified by its *atomic number* which is the number of positive elementary particles contained in the atom, the *protons*. In its elementary state, the atom contains the same number of *electrons*, in this situation it does not bear a charge.
A certain number of electrons of the atom are able to interact with electrons of other atoms of the molecule in question. Electrons with this property are called *valence electrons*. Their number depends on the element, and the interactions are called chemical (covalent) bonds. An interaction between two electrons (from two atoms) is called a single bond and it is denoted as a single line, an interaction between four or six electrons (from two atoms) is called a double or triple bond and is drawn as a double or triple line, respectively. There are also forms of interactions not amenable to this simple scheme (e.g. mesomerism). A single valence electron that does not participate in a bond is an unpaired electron, two valence electrons on one atom that are not involved in a bond form a free electron pair (a lone pair).
The sum of the number of electrons engaged in covalent bonds, of those in lone pairs and of an unpaired electron (if any) for an atom in a molecule may differ from the number of valence electrons in the isolated atom. The difference is the charge of the atom. For this reason, we define the state of the atom as follows:

4.1 Definition *An* atomic state *is a quadruple*

$$S = (v_s, p_s, q_s, r_s),$$

where

- *the natural number v_s means the valence of the atom,*
- *the natural number p_s indicates the number of free electron pairs,*
- *the natural number q_s denotes the charge of the atom,*
- *while $r_s \in \{true, false\}$ shows that there is an unpaired electron or not so.*

Such a state is called a ground state *if $q_s = 0$ and $r_s = false$.*

The valence of an atom in a molecule is the number of covalent bonds in which it is involved, each bond counted whith its multiplicity, and so it is the degree of the corresponding vertex in the multigraph. For example, the valence of H atoms is 1, for O atoms it is 2, for N atoms it is 3 and for C atoms we have 4. But we should carefully note that this is true only if these atoms are in their ground state, i.e. if there is neither a charge nor an unpaired electron. There are elements such as phosphorus and sulfur that even in the ground state can exhibit more than on valence. For example, there are molecules with 3- or 5-valent phosphorus atoms. Sulfur may have valences 2, 4 or 6, differing the number of free electron pairs. If we skip the assumption that the atoms are in their ground state, further valences can show up.
For this reason we introduce for each chemical element X a set $\mathcal{S}_X$ of *admissible atomic states.* Its definition clearly depends on the particular situation of the molecule in question. For example, the most important elements in organic chemistry are gathered in the following set of elements:

$$\mathcal{E}_4 := \{\mathsf{H}, \mathsf{C}, \mathsf{N}, \mathsf{O}\}.$$

We shall refer in the following to this set, and also to its extension

$$\mathcal{E}_{11} := \{\mathsf{H}, \mathsf{C}, \mathsf{N}, \mathsf{O}, \mathsf{F}, \mathsf{Si}, \mathsf{P}, \mathsf{S}, \mathsf{Cl}, \mathsf{Br}, \mathsf{I}\}.$$

Table 1 contains, for the elements $X \in \mathcal{E}_{11}$ their atomic number TE_X, the number of valence electrons VE_X and a list of atomic states [1]. The states listed are those relevant for structure elucidation using mass spectroscopy.

The set $\mathcal{S}_X$ of admissible states of the element X depends on the chemistry that we are willing to use in a particular situation. A hierarchical classification of the corresponding topological models, introduced in [2], can be described in terms of these states.

4.2 Definition

- Restricted Chemistry *(RC) considers atoms without charge or unpaired electrons that obey the* octet rule

  $$2v_Z + 2p_Z = 8.$$

  As an exception, hydrogen is included, for which trivially $v_Z = 1$ must be fulfilled. On this level, valences of atoms are uniquely determined and called standard valences.
- *On the next level, the level of* Closed Shell Chemistry *(CSC) the octet rule is skipped and the assumption made that $q_Z = 0$ and $r_Z = false$. Therefore we also call this level* ground state chemistry.
- *If we skip the assumption that $q_Z = 0$ and $r_Z = false$ we reach the* Integral Chemistry *(IC). On this level we still suppose that multiplicities of covalent bonds are natural numbers.*
- *A refined description is needed when we want to deal with mesomerism or multicenter bonds. This kind of chemistry is called* Multicenter Chemistry *(MC).*

Summarizing, we obtain the following chain of inclusions:

$$RC \subset CSC \subset IC \subset MC.$$

In terms of these notions, the structure generator *MOLGEN* (up to version 3.5, see [3]) is able to generate chemical compounds from RC. From version 4.0, cf. [4], it is possible to generate molecules for IC. Using an algorithm that identifies aromatic systems MOLGEN covers the most important part of MC, aromatic compounds, as well.

X (TE_X, VE_X)	v_Z	p_Z	q_Z	r_Z	RC	CSC
H (1, 1)	1	0	0	0	x	x
	0	0	1	0		
	0	0	0	1		
C (6, 4)	4	0	0	0	x	x
	3	0	1	0		
	3	0	0	1		
	2	0	1	1		
N (7, 5)	4	0	1	0		
	3	1	0	0	x	x
	3	0	1	1		
	2	1	0	1		
O (8, 6)	3	1	1	0		
	2	2	0	0	x	x
	2	1	1	1		
	1	2	0	1		
F (9, 7)	2	2	1	0		
	1	3	0	0	x	x
	1	2	1	1		
Si (14, 4)	4	0	0	0	x	x
	3	0	1	0		
	3	0	0	1		
	2	0	1	1		
P (15, 5)	5	0	0	0		x
	4	0	1	0		
	4	0	0	1		
	3	1	0	0	x	x
	3	0	1	1		
	2	1	0	1		
S (16, 6)	6	0	0	0		x
	5	0	1	0		
	5	0	0	1		
	4	1	0	0		x
	4	0	1	1		
	3	1	1	0		
	3	1	0	1		
	2	2	0	0	x	x
	2	1	1	1		
	1	2	0	1		
Cl (17, 7)	2	2	1	0		
	1	3	0	0	x	x
	1	2	1	1		
Br (35, 7)	2	2	1	0		
	1	3	0	0	x	x
	1	2	1	1		
I (53, 7)	2	2	1	0		
	1	3	0	0	x	x
	1	2	1	1		

Table 1: Admissible states for the elements in $\mathcal{E}_{11}$ occurring in mass spectroscopy

4.3 Definition (molecular graph) *Let $\mathcal{E}$ denote a set of chemical elements and assume that $\mathcal{S}_\mathcal{E}$ indicates the set of all the admissible atomic states of the elements in $\mathcal{E}$. In formal mathematical terms,*

$$\mathcal{S}_\mathcal{E} := \bigcup_{X\in\mathcal{E}} \mathcal{S}_X.$$

A molecular graph *for a molecule consisting of n atoms from $\mathcal{E}$ is a triple*

$$(\varepsilon, \zeta, \gamma),$$

where ε is a sequence of length n, consisting of element symbols, i.e.

$$\varepsilon(i) \in \mathcal{E} \ (i = 1, \ldots, n).$$

The second component ζ is a sequence of n atomic states, where the $i-th$ component is an admissible state of the $i-th$ atom,

$$\zeta(i) \in \mathcal{S}_{\varepsilon(i)}. \tag{1}$$

The third component γ is a connected multigraph consisting of n vertices and edges that are at most 3-fold, for short,

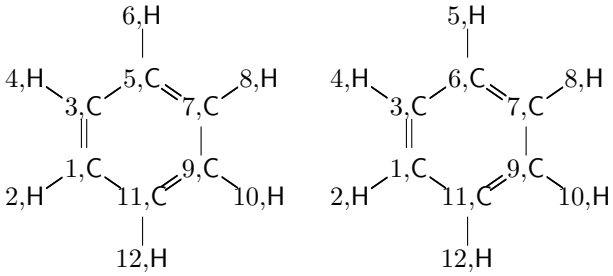
$$\gamma \in \mathcal{G}^c_{n,4}.$$

Its vertices are numbered from 1 to n and are colored by the atom names $\varepsilon(i)$, the components of ε. The degree of vertex i of the graph is equal to the valence of the $i-th$ atom $\varepsilon(i)$,

$$\deg(i) = v_{\zeta(i)}. \tag{2}$$

Let $\mathcal{M}_n$ denote the set of molecular graphs on n atoms. $\diamond$

A problem with molecular graphs is that there are usually many molecular graphs that represent the same molecule, differing in vertex numbering. Here are two examples of differently numbered molecular graphs:



Vertex numbering is unavoidable since also the entries $\varepsilon(i)$ of the element distribution ε, as well as the entries $\zeta(i)$ of the atomic states are numbered. Hence, two such molecular graphs$(\varepsilon, \zeta, \gamma)$ and $(\varepsilon', \zeta', \gamma')$ describe the same molecule if and only if they are the same up to renumbering, which means that there is a permutation π such that

$$(\varepsilon, \zeta, \gamma)^\pi = (\varepsilon', \zeta', \gamma'),$$

where

$$(\varepsilon, \zeta, \gamma)^\pi = (\varepsilon^\pi, \zeta^\pi, \gamma^\pi),$$

defined by

$$\varepsilon^\pi(i) = \varepsilon(\pi(i)),\ \zeta^\pi(i) = \zeta(\pi(i)),\ \gamma^\pi(\{i,j\}) = \gamma(\{\pi(i), \pi(j)\}).$$

$\gamma(\{i,j\})$ denotes the multiplicity of the covalent bond that connects atoms i and j, i.e. $\gamma(\{i,j\}) \in \{0,1,2,3\}$. In mathematical terms, we are faced with the following *action of the symmetric group*:

$$\begin{gathered}\left(\mathcal{E}^n \times \mathcal{S}^n_\mathcal{E} \times \mathcal{G}^c_{n,4}\right) \times S_n \longrightarrow \mathcal{E}^n \times \mathcal{S}^n_\mathcal{E} \times \mathcal{G}^c_{n,4}, \\ ((\varepsilon, \zeta, \gamma), \pi)) \longmapsto (\varepsilon, \zeta, \gamma)^\pi.\end{gathered}$$

This action, like every action of a group on a set, induces an equivalence relation, the classes of which are called *orbits*, for example

$$S_n((\varepsilon, \zeta, \gamma)) = \{(\varepsilon, \zeta, \gamma)^\pi \mid \pi \in S_n\}$$

is the orbit of $(\varepsilon, \zeta, \gamma)$. Properties (1) and (2) of definition 4.3 are preserved by this operation. Therefore S_n also operates on $\mathcal{M}_n$.

4.4 Corollary *A structural formula of a molecule with n atoms contained in $\mathcal{E}$ corresponds to an orbit of S_n on the set $\mathcal{M}_n$ i.e. the set of structural formulae of molecules built from n atoms in $\mathcal{E}$ is the set of orbits*

$$S_n // \mathcal{M}_n$$

Hence the problem of construction of all structural formulae, i.e. all constitutional isomers of the molecular formula in question, amounts to the evaluation of a complete system of representatives of these orbits of the symmetric group. This is obviously an algebraic problem, the efficient application of group theoretic methods is the method of choice. Double cosets can be used, as pointed out by G.Pólya in his seminal paper already. Another useful tool is orderly generation, as indicated by R. C. Read, see [5, 6, 7, 8, 9].

5 Applications

Having described the mathematical tools which form the basis of an efficient, systematic and complete generation free of redundance of all structural formulae that correspond to a given molecular formula and (optional) further restrictions, we are now in a position to list a few software packages that use these methods.

To begin with, we mention three kinds of structure generation problems:

- Generation of structures *based on a molecular formula*,
- generation *based on a given set of reactions and reactants*,
- generation *based on a generic structural formula*, e.g. a Markush formula.

They will be discussed in the following subsections.

5.1 Generation based on a bruttomolecular formula

MOLGEN is a generator of structural formulae corresponding to a given (well-defined or fuzzy) molecular formula and (optional) further restrictions imposed by the user. For example, depending of the version of MOLGEN used, the following numbers can be restricted by upper and lower bounds, i.e. we can force MOLGEN to generate just those constitutional isomers for which the following numbers belong to user-defined intervals:

- the total number of atoms, of heteroatoms, of covalent bonds, numbers of single, double and triple bonds, double bond equivalents, numbers of rings of specified sizes, molecular mass, total charge, number of unpaired electrons, numbers of atoms of specified atomic states, hydrogen distribution (number of C, CH, CH_2, CH_3 groups, etc.), and occurrence numbers of particular substructures.

Substructures can be prescribed, e.g. hydroxyl groups etc., or forbidden by the user in various ways:

- A *goodlist* of substructures that may overlap can be prescribed, as well as a *goodlist* of substructures that must not overlap,
- together with a *badlist* of forbidden and not overlapping substructures.

Moreover, the user can force the generator to

- produce *all* the corresponding constitutional isomers, or just those that contain at least one ring, or, alternatively the isomers that do not contain any ring.

Further features of MOLGEN allow to check restrictions after the generation, e.g.

- *aromatic bonds* can be identified, and correspondingly *aromatic duplicates* can be eliminated.
- *Symmetry aspects* with respect to the symmetry group of the structural formula (which may, of course be bigger than the geometric symmetry group) can be used, for example, to give a lower bound for the *number of carbon signals* in a ^{13}C NMR spectrum.

MOLGEN applies a lot of *algebra* (groups and double cosets of groups, for example), as well as of *combinatorics* (orderly generation etc.). One of the crucial points is the following one. It is extremely important for applications (e.g. for the generation of large combinatorial libraries, for the generation of patent libraries, or for the use of inhouse databases):

- The molecular graphs are constructed in a canonical form, so that it is easily checked whether a generated structural formula is already contained in the current file of molecules — or in any other database generated by or imported into MOLGEN.

We shall return to this in the relevant subsections. Further details can be found in the MOLGEN home page (`www.mathe2.uni-bayreuth.de/molgen4/`). The structure generator can be used (in a restricted version) also online. MOLGEN is quite fast, of course, depending on the conditions the user imposes. For example, if just the molecular formula is given, MOLGEN produces within a second several thousand isomers of moderate size on a standard PC. Table 2 lists for all molecular formulae based on $\mathcal{E}_4$ with mass 146 and at least one C Atom the number of structural formulae (RC) together with the CPU time (in seconds) on a 2.53 GHz Pentium 4 PC.

5.2 Education

For the purpose of chemical education, we developed an interactive online course on molecular symmetry and isomerism including stereoisomerism, called UNIMOLIS. The course is freely accessible in the internet (`www.unimolis.uni-bayreuth.de`) in English or German. A somewhat limited version of MOLGEN is available within UNIMOLIS for the generation of constitutional isomers for a molecular formula entered by the student. The course is also available on CD, in this case to use the generator an internet connection is required.

5.3 Generation of combinatorial libraries

Suppose a combinatorial library is described in terms of a set of building blocks and a set of chemical reactions that link building blocks by means of their functional groups. This corresponds to linking molecular graphs by means of well-defined procedures acting on well-defined subgraphs (see [10, 11]). So we can generate the complete library quickly, completely and free of redundance. A prominent example are the libraries described by Carell et al in [12], where there is a central molecule containing carboxylic acid chloride functions, to which various amine starting materials are attached via amidation.
Already here, during the generation of the library, the importance of the canonical form becomes obvious. If we admit 20 different amines to be attached to 4 carboxylic acid chloride functions that cover, in a tetrahedral arrangement, the cubane skeleton, then a purely combinatorial approach would result in $20^4 = 160000$ seemingly distinct products. However, due to the high symmetry of the central molecule, no more than 13700 of these are in fact distinct. Such a generation is an algebraic problem rather than simply a combinatorial one, group theory is intensively used (see e.g. [8, 7, 9]).

5.4 Examination of molecular libraries, QSPR

The basic problem of QSPR (quantitative structure property relationship) work is to describe the numerical values of some experimental property of compounds in terms of their molecular structures. The aim is to predict, by means of such a relationship, the property values for some other

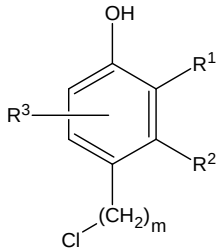
Molecular formula	Structural formulae	CPU time	Beilstein database	NIST MS database
$CH_2N_6O_3$	76720	0.2	0	0
CH_6N_8O	97234	0.3	0	0
$C_2H_2N_4O_4$	216893	0.6	0	0
$C_2H_6N_6O_2$	971399	2.4	1	0
$C_2H_{10}N_8$	57508	0.2	0	0
$C_3H_2N_2O_5$	137656	0.4	0	0
$C_3H_6N_4O_3$	2429018	6.2	10	1
$C_3H_{10}N_6O$	749873	2.1	0	0
$C_4H_2O_6$	9986	0.1	1	0
$C_4H_6N_2O_4$	1432731	3.9	22	0
$C_4H_{10}N_4O_2$	2125930	5.9	33	1
$C_4H_{14}N_6$	68990	0.2	0	0
$C_5H_2N_6$	7055345	14.8	1	0
$C_5H_6O_5$	95870	0.3	28	2
$C_5H_{10}N_2O_3$	1360645	3.8	153	9
$C_5H_{14}N_4O$	311390	1.0	6	0
$C_6H_2N_4O$	26123593	49.9	3	0
$C_6H_{10}O_4$	97394	0.3	345	25
$C_6H_{14}N_2O_2$	257122	0.8	249	3
$C_6H_{18}N_4$	6742	0.0	7	2
$C_7H_2N_2O_2$	17388955	34.1	0	0
$C_7H_6N_4$	96024197	196.1	94	10
$C_7H_{14}O_3$	22151	0.1	672	36
$C_7H_{18}N_2O$	9780	0.0	52	2
$C_8H_2O_3$	1187784	2.7	2	0
$C_8H_6N_2O$	109240025	217.7	177	14
$C_8H_{18}O_2$	1225	0.0	334	28
$C_9H_6O_2$	9660231	20.4	45	4
$C_9H_{10}N_2$	46024195	98.6	411	22
$C_{10}H_{10}O$	7288733	17.2	421	34
$C_{11}H_{14}$	950064	2.7	450	52
$C_{12}H_2$	3571212	65.0	1	0

Table 2: Molecular formulae with mass 146 and at least one C Atom, numbers of structural formulae together wit the CPU time for structure generation (in seconds), numbers of structural formulae included in the Beilstein and the NIST MS database

compounds in the same compound class, or even for all compounds in a certain structure space. The software package MOLGEN–QSPR was developed to assist the scientist in all steps of this endeavor. MOLGEN–QSPR allows to import, to generate or to manually edit the structures of the learning set of compounds (the *real* library), to import or to manually input property values, to calculate numerical values of quite a lot of molecular descriptors, to derive, using various methods of statistical learning, mathematical models for the property of interest (QSPR equations), and to apply such a model to a list of structures or to all structures in a somehow defined class of compounds (the *virtual* library), that again are produced completely and free of redundance by the generator. For applications see [13] and [14].

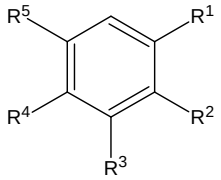
5.5 Patent libraries

Patents in chemistry often claim a whole library of compounds, a *patent library*, defined by a generic structural formula, a *Markush formula*. We present a particularly simple example of two patent libraries to be compared, in order to illustrate the problems to be overcome [15].
The first formula is taken from [16]:



R^1 : CH_3 oder C_2H_5		(variation of substituents)
R^2 : Alkyl		(variation of homology)
R^3 : NH_2		(variation of position)
m : 1–3		(variation of chain length)

In order to obtain a finite library, we restricted substituent R^2 to inclu 1–6 C atoms.
The second Markush formula was constructed by us in order to demonstrate in an easy way the crucial points:



R^1 : CH_3, C_2H_5, OH
R^2 : Alkyl (1–6 C atoms)
R^3 : OH, OCH_3, OC_2H_5, CH_3, C_2H_5
R^4 : OH, CH_2Cl, NH_2
R^5 : H, CH_3, C_2H_5, NH_2

MOLGEN-COMB constructs the corresponding libraries $\mathcal{L}_1$ and $\mathcal{L}_2$ in a few seconds, using a reaction-based generation. These libraries are of the order

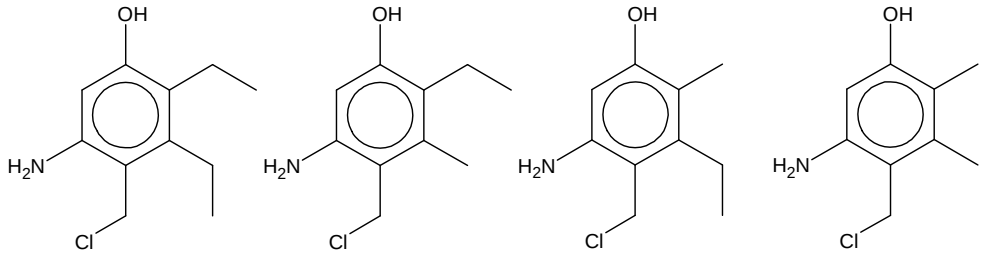
$$|\mathcal{L}_1| = 396,\ |\mathcal{L}_2| = 5939,$$

which are the numbers of compounds contained therein. The first point we should like to emphasize is that in the $\mathcal{L}_2$ case a purely combinatorial approach yields the number $3 \cdot 33 \cdot 5 \cdot 3 \cdot 4 = 5940$ of possible combinations of the admissible substituents. However, due to the symmetry of the benzene skeleton, one structure appears twice ($R^5 = H$, $R^1 = R^4 = OH$ and $R^2 = C_2H_5$, $R^3 = CH_3$ or $R^2 = CH_3$ $R^3 = C_2H_5$), and MOLGEN automatically eliminates the duplicate.

The second important and absolutely crucial point is the canonical form in which the structural formulae of the members of these libraries are produced. It shows in a few seconds that there is an overlap. If either Markush formula represented a claim in one of two patents, the two patent assignees would face a problem, since the intersection of these two libraries is not empty, it consists of four elements:

$$|\mathcal{L}_1 \cap \mathcal{L}_2| = 4.$$

Here is the overlap found:



5.6 Structure elucidation

Another important application of structure generation is in structure elucidation. Here a chemical structure best fitting a given set of spectroscopic data for an unknown compound is to be found.
A database search for the spectra of an unknown is more or less likely to find hits if the unknown was obtained and examined previously. Columns 4 and 5 of Table 2 give the number of compound entries for a particular molecular formula in the Beilstein database[1], and the number of corresponding mass spectra in the NIST MS database[2]. The Beilstein database is the largest collection of known organic compounds worldwide, and the NIST MS database is one of the most comprehensive of its kind. Comparison of entries in the two columns shows that for most known compounds a mass spectrum is not available in the NIST database, even for a molecular mass as low as 146.
Comparison of the "Beilstein" and the "Structural formulae" columns of Table 2 shows how minute a fraction of mathematically possible structural formulae exist as known organic compounds. In fact, since column 2 refers to RC, it gives a lower bound of possible structures (nitro compounds or nitrates, for example, are not included). Beilstein, on the other hand, does of course register nitro compounds and nitrates, and furthermore registers stereoisomers and isotopomers separately. So column 4 gives an upper bound of known structural formulae (constitutions). Thus the ratio of known existing constitutions to possible constitutions is even lower than would be expected from the numbers in columns 4 and 2.
Most structure elucidations, in particular the non-trivial ones thus deal with new compounds. Classically, structure elucidation is done in three steps [17]:

i) Structural features are extracted from spectral data.

ii) All structural formulae compatible with these structural properties are generated.

iii) For the generated structures virtual spectra are calculated and compared to the experimental spectrum, the spectra/structures are then ranked according to goodness of fit.

While step ii) is essentially solved by structure generators such as MOLGEN, steps i) and iii) still pose challenging problems.
Bearing in mind the numbers from column 2 in Table 2, it is obvious that in step i) we have to find restrictions that are highly selective, so as to efficiently downsize the library of potential hits. At the same time restrictions must not be overselective, so as not to exclude the correct structure.
For example, a restriction highly efficient in the case of polycyclic compounds is the limitation to graph-theoretically (gt) planar compounds. A survey of the Beilstein database found that very few known compounds are gt-nonplanar [18], whereas many or even most of the structures generated for a polycyclic compound are gt-nonplanar. However, if the unknown happens to be gt-nonplanar, its correct structure will be missed under this restriction.
In the case of a synthetic product, the chemist often is able to provide some guidance for step i) (starting materials, experimental conditions, etc.), but in the case of a new natural product equivalent information is obviously not available. Spectroscopic methods providing information are

[1]Beilstein database BS0302PR with MDL CrossFire Commander Server–Software, Version 6.0, MDL Information Systems

[2]NIST/EPA/NIH Mass Spectral Library, NIST '98 Version, U.S. Department of Commerce, National Institute of Standards and Technology

numerous, it is, however, everything but easy to automatically and reliably translate this information into useful restrictions for the generation process.
As was demonstrated above, the most important input for a generator is a molecular formula. The method of choice to obtain this information nowadays is MS or the combination GC/MS, thanks to its high sensitivity and resolution. The method is applicable automatically even for large combinatorial libraries. In case of a low-resolution MS only being available, the software package MOLGEN-MS [19, 20] can give suggestions to identify the molecular ion, and then it provides possible molecular formulae for that molecular mass.
Further, using the tools developped by Varmuza [21, 22], MOLGEN-MS is able to identify substructures present or absent from the mass spectroscopic peak patterns.
As to step iii), MS simulation is presented in [23], and first results on the quality of structure ranking according to MS fit are reported in [24]. These procedures are also incorporated in MOLGEN-MS.

6 The Geometric Level

Molecules live in 3D space, and so the final aim is to construct all distinct stereoisomers (configurations) corresponding to a given constitutional formula. Unfortunately, a stereo generator able to automatically construct stereoisomers efficiently, completely, and free of redundance is not yet available. At hand are energy models such as Allinger's molecular mechanics programs [25], that allow to find some local minima of the particular energy function, corresponding to some conformers. Other software packages such as Gasteiger's CORINNA [26] arrive at similar results by a different procedure. However, these packages do not find systematically all conformers, there is even no guarantee that the very lowest (in energy) conformer is found in every case. More importantly, often the chemist is not interested in the conformers but in the stereoisomers, as said above. Work on this problem is going on in this laboratory.

References

[1] W. Werther. *Versuch einer Systematik der Reaktionsmöglichkeiten in der Elektronenstoß–Massenspektrometrie (EI–MS).* Unpublished, 1996.

[2] J. Dugundji and I. Ugi. *An Algebraic Model of Constitutional Chemistry as a Basis for Chemical Computer Programs.* Topics In Current Chemistry, 39:19–64, 1973.

[3] C. Benecke, R. Grund, R. Hohberger, R. Laue, A. Kerber, and T. Wieland. *MOLGEN+, a Generator of Connectivity Isomers and Stereoisomers for Molecular Structure Elucidation.* Anal. Chim. Acta, 314:141–147, 1995.

[4] T. Grüner, A. Kerber, R. Laue, and M. Meringer. *MOLGEN 4.0.* MATCH — Commun. Math. Comput. Chem., 37:205–208, 1998.

[5] G. Pólya. *Kombinatorische Anzahlbestimmungen für Gruppen, Graphen und chemische Verbindungen.* Acta Mathematica, 68:145–253, 1937.

[6] R. C. Read. *Everyone a Winner.* Annals of Discrete Mathematics, 2:107–120, 1978.

[7] A. Kerber. *Applied Finite Group Actions.* Springer, 1991.

[8] S. Fujita. *Symmetry and Combinatorial Enumeration in Chemistry.* Springer, 1991.

[9] R. Laue. *Construction of Combinatorial Objects — A Tutorial.* Bayreuther Mathematische Schriften, 43:53–96, 1993.

[10] T. Wieland. *Mathematical Simulations in Combinatorial Chemistry.* MATCH — Commun. Math. Comput. Chem., 34:179–206, 1996.

[11] R. Gugisch, A. Kerber, R. Laue, M. Meringer, and J. Weidinger. *MOLGEN–COMB, a Software Package for Combinatorial Chemistry.* MATCH — Commun. Math. Comput. Chem., 41:189–203, 2000.

[12] T. Carell, E. A. Wintner, A. Bashir-Hashemi, and J. Rebek Jr. *Neuartiges Verfahren zur Herstellung von Bibliotheken kleiner organischer Moleküle.* Angew. Chemie, 106:2159–2161, 1994.

[13] A. Kerber, R. Laue, M. Meringer, and C. Rücker. *MOLGEN–QSPR, a Software Package for the Search of Quantitative Structure Property Relationships.* MATCH — Commun. Math. Comput. Chem., 51:187–204, 2004.

[14] C. Rücker, M. Meringer, and A. Kerber. *QSPR Using MOLGEN–QSPR: The Example of Haloalkane Boiling Points.* J. Chem. Inf. Comput. Sci., 2004. Submitted.

[15] A. Kerber, R. Laue, and M. Meringer. *An Application of the Structure Generator MOLGEN to Patents in Chemistry.* MATCH — Commun. Math. Comput. Chem., 47:169–172, 2003.

[16] J. M. Barnard and G. M. Downs. *Use of Markush Structure Techniques to Avoid Enumeration in Diversity Analysis of Large Combinatorial Libraries.* http://www.daylight.com/meetings/-mug97/Barnard/ 970227JB.html, 1997.

[17] R. K. Lindsay, B. G. Buchanan, E. A. Feigenbaum, and J. Lederberg. *Applications of Artificial Intelligence for Organic Chemistry: The DENDRAL Project.* McGraw–Hill Book Company, New York, St. Louis, San Francisco, 1980.

[18] C. Rücker and M. Meringer. *How Many Organic Compounds are Graph–Theoretically Nonplanar?* MATCH — Commun. Math. Comput. Chem., 45:153–172, 2002.

[19] T. Grüner, A. Kerber, R. Laue, M. Meringer, K. Varmuza, and W. Werther. *MASSMOL.* MATCH — Commun. Math. Comput. Chem., 38:173–180, 1998.

[20] A. Kerber, R. Laue, M. Meringer, and K. Varmuza. *MOLGEN–MS: Evaluation of Low Resolution Electron Impact Mass Spectra with MS Classification and Exhaustive Structure Generation*, volume 15 of *Advances in Mass Spectrometry*, pages 939–940. Wiley, 2001.

[21] K. Varmuza, P. He, and K.-T. Fang. *Boosting Applied to Classification of Mass Spectra.* J. Data Sci., 1:391–404, 2003.

[22] K. Varmuza and W. Werther. *Mass Spectral Classifiers for Supporting Systematic Structure Elucidation.* J. Chem. Inf. Comput. Sci., 36:323–333, 1996.

[23] J. Gasteiger, W. Hanebeck, and K.-P. Schulz. *Prediction of Mass Spectra from Structural Information.* J. Chem. Inf. Comput. Sci., 32:264–271, 1992.

[24] M. Meringer. *Mathematical Models for Combinatorial Chemistry and Molecular Structure Elucidation.* PhD thesis, Universität Bayreuth, 2004. In German.

[25] N. L. Allinger. *MM2. A Hydrocarbon Force Field Utilizing V_1 and V_2 Torsional Terms.* J. Am. Chem. Soc., 99:8127–8134, 1977.

[26] J. Sadowski and J. Gasteiger. *From Atoms and Bonds to Three–dimensional Atomic Coordinates: Automatic Model Builders.* Chem. Reviews, 93:2567–2581, 1993.