MOLecular Structure GENeration with MOLGEN, new features and future developments

C. Benecke[*], T. Grüner[†], A. Kerber, R. Laue, T. Wieland[*]
Lehrstuhl II für Mathematik, Universität Bayreuth,
D-95440 Bayreuth, Germany

Abstract

MOLGEN is a computer program system which is designed for generating molecular graphs fast, redundancy free and exhaustively. In the present paper we describe its basic features, new features of the current release MOLGEN 3.5, and future developments which provide considerable improvements and extensions.

1 Introduction

MOLGEN [1–7] is a generator for molecular graphs (=connectivity isomers or constitutional formulae) allowing to generate all isomers that correspond to a given molecular formula and (optional) further conditions like prescribed and forbidden substructures, ring sizes etc. The input consists of

- the *empirical formula,* together with
- an optional list of *macroatoms,* which means prescribed substructures that *must not overlap,*
- an optional *goodlist,* that consists of prescribed substructures which *may overlap,*
- an optional *badlist,* containing *forbidden substructures,*
- an optional *interval for* the minimal and maximal *size of rings,*
- an optional number for the maximal *multiplicity of bonds,*
- an optional prescription if only *tree like structures* should be constructed, or rings are allowed, or if only *cyclic structures* should be generated.

[*]supported by BMBF under grant 03-KE7BAY-9
[†]supported by DFG under grant KE201/16-1

The generation is fast (several thousand molecular graphs are generated per second, depending of course both on the computer and the problem), redundancy-free and exhaustive. The manual describes how the user should carefully use the input possibilities since the efficiency and the speed do clearly depend on the way the input is prepared. For example, the clever use of macroatoms is very important for the speed of the generation, since the goodlist as well as the badlist do work as filters only, i.e. after generation. Each isomer is checked if it contains each element of the goodlist and if it does not contain any of the elements of the badlist.
Moreover additional parts of MOLGEN allow to show the result of the generation, to compute a 3D placement (using a simplified MM2 energy model [5,8] together with numerical optimization and in a non-deterministic way so that repeated calculations may evolve different local minima).
MOLGEN is also capable of generating all possible stereoisomers (configurational isomers) to a given constitutional formula, again exhaustive and redundancy-free (which, of course, also implies the consideration of symmetries) [6,9]. Spatial realizations of the stereoisomers constructed geometrically are displayed.
MOLGEN can import and export files in MDL MolFile-format and detect aromatic mesomers.

2 New features

2.1 The hydrogen distribution

The analysis of NMR spectra often provides information on the degree of substitution of a compound, i.e. on the number of H-atoms connected to carbon or hetero atoms. This information can now be used in MOLGEN 3.1 in order drastically to reduce the number of isomers constructed from the given molecular formula. So the user may optionally

- input the *hydrogen distribution* of the carbon, nitrogen and oxygen atoms.

For example, In the previous version it was possible that a prescribed CH_2 subgroup with two free valences could get another hydrogen leading to a CH_3 subgroup during the construction process.
The input dialog is shown in Fig. 1.

2.2 Hybridization

MOLGEN 3.5 furthermore allows to make use of another important result of NMR spectroscopy, i.e. to

- optionally input *the hybridization states of carbon and hetero atoms,*

which also considerably reduces the number of isomers that must be generated and checked [2]. It is again be possible to give the exact distribution of the hybridization states or to enter just intervals. A draft of the input window is depicted in Fig. 2.

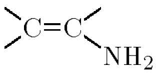
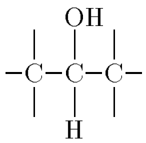
The following table shows the effect of using these new features of hydrogen distribution and hybridization:

Restrictions	**No. of Isomers**
Chemical formula $C_6H_8O_6$ only	2 558 517
No triple bonds	2 434 123
No O-O	360 594
Hydrogen distribution 1 CH_2, 2 CH, 3 C, 4 OH	35 058
Hybridization: 3 Csp^3, 2 Csp^2, 5 Osp^3, 1 Osp^2	990

2.3 Conversion of goodlist into macroatoms

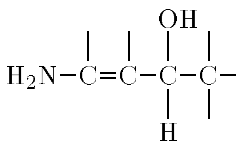
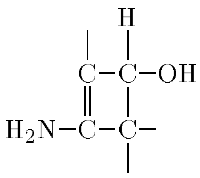
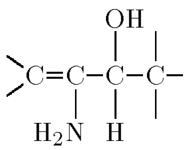
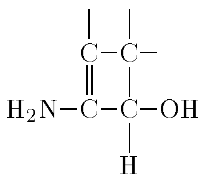
Since macroatoms are most useful, it is reasonable to construct all possible overlappings of the goodlist entries and use these as macroatoms. This can be carried out by another module currently under development [10].

For example, the empirical formula $C_6H_{11}NO$ has 13,982 isomers. If we prescribe the substructures

as goodlist entries, all these 13,982 isomers have to be generated in order to detect that just 71 of them contain both structures.

The construction of all possible overlappings yields four reasonable substructures

Starting the generator with each of these four substructures as well as both initial structures as macroatoms, 78 structures are constructed only, from which the redundancy checker of MOLGEN extracts the correct 71 compounds. The difference between the initial 13,982 and the final 78 structures makes the advantages of this approach obvious.

There are further methods under development that should improve the use of goodlist and macroatoms.

3 Combinatorial chemistry

Other efforts are directed towards mathematical modelling and computer simulation of combinatorial chemistry. This new technique for synthesizing organic compounds has attracted great interest (see e.g. [11–14]). It does not aim at the classical objective of synthesizing *one* substance as pure as possible, but it deliberately utilizes the structural variety to produce a large number of compounds simultaneously.
Typically a set of *building-blocks* is taken that is systematically combined with a *core* structure in all the combinatorially possible ways where the actual reactions make use of chemical, biological or biosynthetical procedures. The resulting set of molecules is called a *combinatorial library.*

A crucial issue in combinatorial chemistry is *diversity* [15]-[17]. A large combinatorial library may fulfill the demand for making many compounds available; for an efficient analysis, however, it should be certified that the elements of a library are not too similar in order to avoid one pharmacological class being tested over and over again. Thus the elements of a library should be as diverse as possible to cover a broad variety with the screening – without requiring too many single substances.
So in connection with combinatorial chemistry the notion of *similarity* has come into the focus [15], [16], [18]- [20]. There is a vast number of ways to define and determine similarity of chemical entities. It turned out that similarity – unlike isomorphy, e.g. – cannot be defined generally. It depends, in fact, on the structure as well as on the studied activity what must be considered similar and what must not. Quite often "similarity" is, however, used in the sense of "structural similarity"; we will also make use of this view in the following.

So the main procedure in combinatorial chemistry can be summarized in the following steps:

1. *Selection of building-blocks*
2. *Generation of the library*
3. *Screening of the library for the required activity*

It is a challenge for mathematics and computer science to model and to implement programs that show beforehand what can be expected from a possible experiment using the powerful methods of combinatorial chemistry.
There are several possibilies for selecting the building-blocks [16,17,21]. Their application mainly depends on the objective that is sought by the combinatorial library. Here we present two methods based on graph theory in conjunction with statistical analysis.

3.1 Molecular graphs

In this paper we consider graphs as mappings

$$\gamma : \underline{\mathbf{p}}^{[2]} \to \{0, \ldots, m-1\}, \quad \text{in short } \gamma \in m^{\underline{\mathbf{p}}^{[2]}}$$

where $\underline{\mathbf{p}}^{[2]}$ is the set of pairs of points of the graph, i.e. the set of all 2-subsets of the set $\underline{\mathbf{p}} := \{1, \ldots, p\}$ of points (or, to be exact, the set of the numbers of the p points), $\gamma(\{i,j\}) = k$ means that there is an edge of degree k – a k-fold bond – between the vertices i and j, and $\gamma(\{i,j\}) = 0$ if the two vertices are not connected.
For molecules we take the usual model, identifying atoms with vertices and bonds with edges. The atomic types are defined by an additional mapping $\beta : \underline{\mathbf{p}} \to \{E_1, E_2, \ldots\}$ with the E_i representing chemical elements such that a molecular graph is a pair (γ, β) consisting of a graph γ and a coloring β of the vertices with atomic types.
Furthermore we call

$$\eta : T^{[2]} \to \{0, \ldots, m-1\} \text{ with } T \subseteq \underline{\mathbf{p}},\ \forall i,j \in T :\ \eta(\{i,j\}) = \gamma(\{i,j\})$$

a *subgraph* of γ, which we indicate by $\eta \subseteq \gamma$.

3.2 Topological indices

A large number of studies have been carried out on the search for quantitative structure-activity relationships (QSAR), i.e. the search for empirical or theoretical parameters that are directly correlated to some biological response [18], [22]- [27]. Since empirical data are not always available [22,26] and experiments or quantum chemical calculations are expensive for larger sets of compounds, a lot of interest lies currently in the use of *topological indices* (or graph invariants) [18], [23], [28]-[31] as discrimination criteria and prediction tools.
We used *connectivity indices* ${}^k\chi$ and ${}^k\chi^b$ for $k = 0, 1, 2$ after [18] which are sums over all paths of length k in the graphs, varying by the use of the adjacency or the connectivity matrix. These indices give a good characterization of the structure, especially with respect to shape, volume, and surface, and have thus been used for a large number of correlations [16,18,27].

Another important class of indices is based on the distance matrix D, where each entry $d_{i,j}$ denotes the length of the shortest path from vertex i to vertex j. The distance matrix can be calculated by the Floyd algorithm [32].
We also used the *Wiener Index* [33], the *Balaban-Index* [34], the *mean square distance index* [35], which is due to Balaban and Motoc, *information-theoretic* indices [20,28], and the *mean information content of distances* that was developed by Bonchev and Trinajstić [36].
We calculated the index values for the 20 natural amino acids (for the table of results and more details see [37]). Principal Component Analysis yields three factors explaining 93.2 % of the original variance. The results obtained show that structures which differ only slightly also have similar factors, e.g. asparagin (asn) and asparagin acid (asp) – a confirmation of our initial assumption that topological indices are suitable for similarity analysis.
After determining the Euclidian distances of the independent variables, we obtained a distance matrix which reveals the quantitative differences between the single molecules more obviously.

For building-block selection, we can group the molecules together according to these results by putting all structures with an euclidian distance below a certain threshold ε in one group. The value of this threshold and thus the coarseness of the decomposition must be determined in agreement with the requirements of the experiment to be simulated. For $\varepsilon = 1.0$, say, we get the following groups:

{gly}, {ala, ser}, {arg}, {asn, asp, gln, glu, met, thr},
{cys}, {his}, {ile, leu, lys, val}, {phe, tyr}, {pro}, {trp}

A first attempt in experiment therefore may consist of taking one representative from each group; afterwards only those groups need to be examined more precisely the representative of which has shown the desired activity.

3.3 Binary property vectors

For the characterization of diversity lists with binary properties can be considered, too. We took a subset of 120 descriptors from the structure codes of the mass spectra information system *MassLib* [38], the same as used in K. Varmuza's program *ToSIM* [39–41]. The following classes of properties are taken into account:

- Aromatic compounds (e.g. substructure phenyl)
- branches in chains and rings (e.g. carbons with three C-neighbors that are not part of a ring)
- Cyclic compounds (e.g. 6-rings)
- double and triple bonds in chains and rings (e.g. a double bonds in a 6-ring)
- elements (e.g. hetero atoms)
- functional groups (e.g. aldehyds)
- special classes of compounds (e.g. methyl ester)

Many of these properties can be described in terms of substructures, and so a substructure search is one of the main algorithms for their determination. We used a procedure from [42]. Other properties were calculated by methods described in [6,5,7].

In the evaluation the *Tanimoto* coefficient [43] is employed which measures the similarity of two bit strings as

$$T_{i,j} = \frac{2C_{i,j}}{E_i + E_j}$$

where $C_{i,j}$ is the number of properties that are common in the i-th and in the j-th structure, and E_i is the total number of properties of the i-th molecule. Thus T lies in the range $0 \leq T \leq 1$, showing 1 for complete identity and 0 for maximal dissimilarity.

Again we combined all results to a distance matrix D with $d_{i,j} = 1 - T_{i,j}$. From this matrix we computed spatial coordinates by *multidimensional scaling*. In this method, the geometric distances are chosen to reflect the pairwise distances as close as possible. By putting together the structures in the same spatial regions, we have an alternative method for grouping the building-blocks at hand.

3.4 Generation of combinatorial libraries

We would like to start the presentation of our method with a syntax describing the underlying chemical reactions formally, especially the two-component synthesis like

$$\mathrm{A} + \mathrm{B} \rightarrow \mathrm{C}$$

In most cases, subgraphs determine the course of the reaction.
Let (η_1, β_1) and (η_2, β_2) be molecular graphs containing r, s atoms, respectively. A *reaction scheme* is defined as the triple $((\eta_1, \beta_1),\ (\eta_2, \beta_2),\ \rho)$ (or (η_1, η_2, ρ) if the atomic type coloring is clear), where ρ is a mapping that associates with a pair (i, j) of vertex numbers i (in η_1) and j (in η_2) the following "bond multiplicity":

$$\rho(i,j) = \begin{cases} k & i \text{ and } j \text{ are connected by a bond of degree } k \\ 0 & i \text{ and } j \text{ remain unconnected} \\ -\infty & \text{one of the atoms } i \text{ or } j \text{ is dropped} \end{cases}$$

By means of this definition[1] many two component reactions can be described sufficiently. Our main interest in such a reaction is in fact the change of the graphs, and not the experimental aspects (like reaction conditions, catalysts or equilibria).
A corresponding algorithm could be formulated for linking two graphs by a reaction scheme over all reacting subgraphs. As we do not explicitly need such a procedure for library generation, we omit a deeper discussion and just present an example:

3.1 Example Peptids are protein molecules built from at least two amino acids which play a central role in biochemistry. The joining of the single amino acids is performed by condensation of the acid group (COOH) and the amid group (NH_2). Thus the decisive reaction structure is the acid amid group, which must be contained in both reaction partners.

$$\begin{array}{ccccc} & {}^{3}\mathrm{N} & & \mathrm{O}^{4} & \\ & | & & \| & \\ - & {}^{1}\mathrm{C} & - & \mathrm{C}^{2} & -\mathrm{O}^{5} \end{array}$$

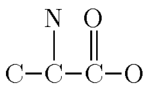
I: $\eta_1 = \eta_2$

[1]This definition is a simplification of the situation and is only used for a formalization of the construction problem discussed below. For more sophisticated purposes more comprehensive approaches like the algebra of *be*- &*r*-matrices of [44] are necessary.

The condensation is represented by the mapping

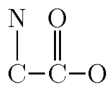
$$\rho = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ -\infty & -\infty & -\infty & -\infty & -\infty \end{pmatrix}$$

We consider the amino acids



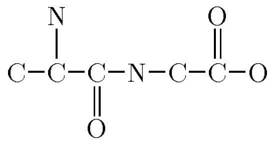
II: Alanin

and

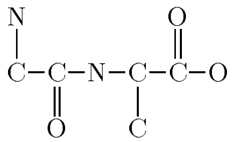


III: Glycin

Obviously both contain the subgraph η_1. Despite the equality of the subgraphs in the reaction scheme, the order of the initial graphs is essential. Taking alanin as the first one, we obtain:



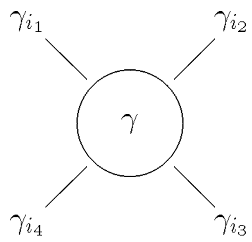
Glycin as γ_1 yields:



3.5 Multiple attachments to a core structure

A special type of reaction scheme is given by a core structure with several reaction sites and a number of ligand compounds.

Let $((\eta_1, \chi_1),\ (\eta_2, \chi_2),\ \rho)$ denote a reaction scheme, (γ, β) a molecular graph containing k substructures isomorphic to (η_1, χ_1) with $k > 1$, and $(\gamma_1, \beta_1), \ldots, (\gamma_n, \beta_n)$ a number of molecular graphs; for sake of simplicity we assume that each of them contains exactly one substructure isomorphic to (η_2, χ_2).
The first task is to determine all attachments of the ligands to the sites of the core, where the sites are given by the substructures of the reaction scheme. For $k = 4$ e.g., the situation is:



Topological equivalence among the sites is described by the permutation group $P_\gamma \leq S_k$ which is induced by the automorphism group $Aut(\gamma, \beta)$ of the molecular graph. So P_γ acts on the k sites which shall be assigned with n different ligands. Using the well known theory of group actions (see e.g. [45]) one obtains:

3.2 Lemma *The essentially different possibilities to attach n ligand structures, which contain the corresponding subgraph of the given reaction scheme exactly once, to the k different reaction sites of a core structure (γ, β) correspond "one-to-one" and "onto" to a transversal of the orbits induced by the action of P_γ on the set of mappings from the set of k sites into the set of n ligands.*

This brings us in a position to formulate a strategy:

3.3 Attachment of ligands to a core structure

1. *Determine the group P_γ as well as all subgraphs $\zeta_1, \ldots, \zeta_k \subseteq \gamma$ which are isomorphic to η_1.*

2. *Compute for $i \in \underline{\mathbf{n}}$ the subgraphs $\zeta^{(i)} \subseteq \gamma_i$ which are isomorphic to η_2.*

3. *Evaluate a system of representatives f of the orbits of P_γ on the set of mappings from the set of sites into the set of ligands.*

4. *Determine the total graph which yields from the attachment of the ligands*

$$\gamma_{f(1)}, \ldots, \gamma_{f(k)}$$

   to γ according to ρ, i.e. by combining the graphs, eliminating vertices which have to be dropped and adding the necessary edges.

5. *If there are further orbit representatives, go to step 3.*

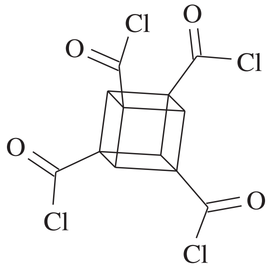
Due to orderly generation in step 3. and the uniqueness of the subgraphs of the ligands we only obtain non-isomorphic solutions.

3.6 Single step generation of libraries

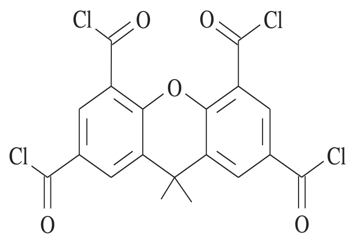
For the generation of a combinatorial library from given building-blocks algorithm 3.3 is perfectly suited, since the basic situation of combinatorial chemistry is just that of this method.[2]

For practical use it is moreover relevant that the multiplicity of a certain building-block can be restricted, i.e. that a (γ_i, β_i) occurs in all compounds of the library at least r and at most s times. This can be reached by an additional test in 3.3 between step iii and step iv. In laboratory, this restriction can be satisfied by an appropriate modification of the reaction conditions.

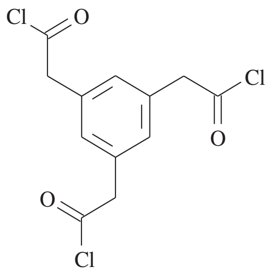
As an example we consider the combinatorial libraries from [14]. The authors used as building-blocks the twenty natural amino acids and as core structures some acid chlorides:



IV



V



VI

a cubane-derivative (structure **IV**), xanthene (**V**) and a benzene triacid chloride (**VI**). The reaction scheme consists of the substructures

and the matrix

$$\rho = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ -\infty & -\infty & -\infty & -\infty & -\infty \end{pmatrix}.$$

The cubane-derivative **IV** is the structure with the highest symmetry, i.e. the largest automorphism group (derived from the symmetry group of the cube), which has 24 elements. Since each automorphism includes a movement of the reacting substructures, we also have $|P_\gamma| = 24$. As there are just four sites, it turns out that $P_\gamma = S_4$.

[2] We assume that each building-block is admissible for each site. In the other case additional rules must be formulated.

The xanthene **V** has an automorphism group with four elements. Two of them include the exchange of the methylene groups on the carbon bridge atom, and thus $|P_\gamma| = 2$. Besides the identity, this is the reflection of the rings on the vertical symmetry axis.
The benzene triacid chloride **VI** has cyclic symmetry. Thus P_γ equals the cyclic group C_3, having three elements.

Even though the symmetry situation is a little complicated for one of the three cores only, the advantages of the mathematical concept behind algorithm 3.3 are obvious.
The general *Ansatz* with an arbitrary permutation group and the efficient orderly generation (cf. [46,47]) allows a very rapid generation of the combinatorial libraries in all three cases. The computing speed is about 40 structures per second on a Pentium 90 MHz PC. Fig. 3 shows six molecules from each of the three libraries.[3]
It is also possibile to carry out several reactions one after the other using the products of one reation as core structures of the next one. The mathematical situtation is very similar to that of the previous section; due to the differences of the cores the irredundancy remains guaranteed. An example is left out here for brevity; see [37] for details.
The following table provides an overview over the sizes of libraries depending on the number of building-blocks used:

n	**IV**	**V**	**VI**
1	1	1	1
2	5	10	4
3	15	45	11
4	35	136	24
5	70	325	45
6	126	666	76
7	210	1225	119
8	330	2080	176
9	495	3321	249
10	715	5050	340
11	1001	7381	451
12	1365	10440	584
13	1820	14365	741
14	2380	19306	924
15	3060	25425	1135
16	3876	32896	1376
17	4845	41905	1649
18	5985	52650	1956
19	7315	65341	2299
20	8855	80200	2680

Although there are equally many sites in **IV** and **V**, the libraries with the first one are considerably smaller due to the higher symmetry of the core.

[3] The 2D placements were automatically calculated by the drawing module of **MOLGEN** [1,5,6]. These pictures reveal the current inacurracies of the employed placement algorithm [5,48] for combinatorial libraries; further work in this respect is in progress.

3.7 Screening

One of our aims is to investigate the use of simulations in combinatorial chemistry. So a few words about the third step, the *screening of the libraries* are necessary. The libraries can be put into a database which we have developed [7] and which allows to apply user defined *screening functions,* e.g. similarity approaches [15,16,18–20,37] etc. This aspect is indeed the most difficult one. In the view of the present level of research a purely *virtual screening* – only in the computer and without experiment – is impractical due to a large computational expenditure.
QSAR should be helpful (see also sec. 3.2). Here a set of sample substances is used to empirically determine one (or several) activity parameters. From the results a correlation with computable structural properties is established by statistical methods. This correlation is afterwards employed for extrapolation on a larger set of compounds (like a combinatorial library, cf. [18,23,27,31]). More elaborate techniques are *3D QSAR* [49,50] and *comparative molecular field analysis (CoMFA)* [50,51]. Here the library elements are first converted to three-dimensional structures by an appropriate method (like distance geometry programs [52], conformation analysis methods [53], expert systems [54,55] or force field calculations, e.g. [8]. The crucial feature such a program is required to have is that the computed conformation must be reasonable for the active site, as the usual software packages produce conformations *in vacuo* or in solution.) Then a superposition of the ligand and the binding site is calculated according to the physical fields (steric and electrostatic) of the molecules in order to be able to estimate the activity of the ligand.

So virtual screening and thus *virtual combinatorial chemistry synthesis* is today not in competition with high-throughput screening robots but – as there are many simulation methods in development – may be the reality of tomorrow.

4 Classification of 3D-placements

An additional module which is currently under development ([42]) for the use in the MOLGEN environment allows a *classification* of 3D-placements of molecules. There is an urgent need for this since there are mostly various 3D-placements, obtained by an optimization procedure that starts from a random distribution of the atoms in space. These 3D-placements therefor represent local minima of the energy function, and the aim is to classify them, and possibly to find the global minimum (in order to have at least a chance to find it, we willingly start from a random distribution of the atoms, otherwise, i.e. using a deterministic algorithm, there is no guarantee and maybe even no chance to find the global minimum). This means, that from a given set of 3D-placements for a molecule, significantly different structures can *automatically* be extracted by applying this module to, say, 500 3D-placements obtained from the optimization.
The module under development uses a *discrete method* and therefore it works very efficiently. It has already been tested for several examples. So let us take a look at tetramethylcyclobuthane (see figure 4)

Using the optimization built into MOLVIEW, 500 3D-placements were computed yielding a set of structures that no longer can be compared manually (see figure 5)
Using the described classification module, four different structures can be extracted which all other structures are similar to (see figure 6)
These are - of course - the four possible stereo isomers of tetramethylcyclobuthane.
As you can see, this classification is also a powerfull tool for full structure search of molecules. Even substructure search can be pursued on base of these algorithms.
The method used for the classification is completely discrete. The algorithms used are based on a backtrack algorithm in combination with iterated classification. The object-oriented implementation allows an easy application to a various set of different problems.

References

[1] Grund, R., A. Kerber, and R. Laue. MOLGEN, ein Computeralgebra-System für die Konstruktion molekularer Graphen. *MATCH*, **1992**, 27, 87–131.

[2] Grund, R. Konstruktion molekularer Graphen mit gegebenen Hybridisierungen und überlappungsfreien Fragmenten. *Bayreuther Mathem. Schr.*, **1995**, 49, 1–113.

[3] Benecke, C., R. Grund, R. Hohberger, A. Kerber, R. Laue, and T. Wieland. MOLGEN, a Computer Algebra System for the Generation of Molecular Graphs. In Fleischer, J., J. Grabmeier, F.W. Hehl, and W. Küchlin, editors, *Computer Algebra in Science and Engineering*, 260–272. World Scientific, Singapore, 1995.

[4] Benecke, C., R. Grund, A. Kerber, R. Laue, and T. Wieland. Chemical Education via MOLGEN. *J. Chem. Edu.*, **1995**, 72, 403–406.

[5] Benecke, C., R. Grund, R. Hohberger, A. Kerber, R. Laue, and T. Wieland. MOLGEN+, a generator of connectivity isomers and stereoisomers for molecular structure elucidation. *Anal. Chim. Act.*, **1995**, 314, 141–147.

[6] Wieland, T., A. Kerber, and R. Laue. Principles of the generation of constitutional and configurational isomers. *J. Chem. Inf. Comput. Sci.*, **1996**, 36, 413–419.

[7] Kerber, A., R. Laue, and T. Wieland. Erkennung, Beschreibung und Visualisierung molekularer Strukturen. In *Proceedings des Statusseminars der anwendungsorientierten Verbundprojekte auf dem Gebiet der Mathematik mit Förderung durch das BMBF.* Springer Verlag, Heidelberg, Berlin, 1996. (In Druck).

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

[9] Wieland, T. Erzeugung, Abzählung und Konstruktion von Stereoisomeren. *MATCH*, **1994**, 31, 153–203.

[10] Wieland, T. Konstruktionsalgorithmen bei molekularen Graphen und deren Anwendung. *Bayreuther Mathem. Schr.*, **1997**, 51, 1–243. (In print).

[11] Gallop, M.A., R.W. Barrett, W.J. Dower, S.P.A. Fodor, and E.M. Gordon. Applications of combinatorial technologies to drug discovery. 1. Background and peptide combinatorial libraries. *J. Med. Chem.*, **1994**, 37, 1233–1251.

[12] Gordon, E.M., R.W. Barrett, W.J. Dower, S.P.A. Fodor, and M.A. Gallop. Applications of combinatorial technologies to drug discovery. 2. Combinatorial organic synthesis, library screening strategies, and future directions. *J. Med. Chem.*, **1994**, 37, 1385–1401.

[13] Ugi, I. Fast and permanent changes in preparative and pharmaceutical chemistry through multicomponent reactions and their 'libraries'. *Proc. Eston. Acad. Sci. Chem.*, **1995**, 44, 237–273.

[14] Carell, T., E.A. Wintner, A.J. Sutherland, J. Rebek, Jr., Y.M. Dunayevskiy, and P. Vouros. New promise in combinatorial chemistry: synthesis, characterization, and screening of small-molecule libraries in solution. *Chem. & Biol.*, **1995**, 2, 171–183.

[15] Johnson, M.A. and G.M. Maggiora, editors. *Concepts and Applications of Molecular Similarity*, New York, NY, 1990. J. Wiley & Sons.

[16] Martin, E.J., J.M. Blaney, M.A. Siani, D.C. Spellmeyer, A.K. Wong, and W.H. Moos. Measuring diversity: experimental design of combinatorial libraries for drug discovery. *J. Med. Chem.*, **1995**, 38, 1431–1436.

[17] Pavia, M.R. The chemical generation of molecular diversity. *NetworkScience*, **Aug. 1995**.

[18] Kier, L.B. and L.H. Hall. *Molecular Connectivity in Structure-Activity Analysis.* Research Studies Press, Chichester, 1986.

[19] Johnson, M.A., Naim. M., V. Nicholson, and C.C. Tsai. Unique mathematical features of the substructure metric approach to quantitative molecular similarity analysis. In King, R.B. and D.H. Rouvray, editors, *Graph Theory and Topology in Chemistry*, 219–225. Elsevier Science Pub., Amsterdam, 1987.

[20] Basak, S.C., V.R. Magnuson, G.J. Niemi, and R.R. Regal. Determining structural similarity of chemicals using graph-theoretic indices. *Discr. Appl. Math.*, **1988**, 19, 17–44.

[21] Sheridan, R.P. and S.K. Kearsley. Using a genetic algorithm to suggest combinatorial libraries. *J. Chem. Inf. Comput. Sci.*, **1995**, 35, 310–320.

[22] Auer, C.M., J.V. Nabholz, and K.P. Baetcke. Mode of action and the assessment of chemical hazards in the presence of limited data: use of structure-activity relationships (SAR) under TSCA, Section 5. *Environ. Health Persp.*, **1990**, 87, 183–197.

[23] Basak, S.C. Use of molecular complexity indices in predictive pharmacology and toxicology: a QSAR approach. *Med. Sci. Res.*, **1987**, 15, 605–609.

[24] Debnath, A.K., G. Debnath, A.J. Shusterman, and C. Hansch. A QSAR investigation of the role of hydrophobicity in regulating mutagenicity in the Ames test: mutagenicity of aromatic and heterocyclic amines in Salmonella typhimurium TA98 and TA100. *Environ. Mol. Mutagen.*, **1992**, 19, 37–52.

[25] Richards, W.G. *Quantum Pharmacology.* Butterworth, London, 2nd edition, 1983.

[26] Stuper, A.J., W.E. Brugger, and P.C. Jurs. *Computer-Assisted Studies of Chemical Structure and Biological Function.* John Wiley & Sons, New York, NY, 1979.

[27] Basak, S.C., S. Bertelsen, and G.D. Grunwald. Application of graph-theoretical parameters in quantifying molecular similarity and structure-activity studies. *J. Chem. Inf. Comput. Sci.*, **1994**, 34, 270–276.

[28] Bonchev, D. *Information Theoretic Indices for Characterization of Chemical Structures.* Research Studies Press, Chichester, 1983.

[29] Randić, M. *J. Chem. Inf. Comput. Sci.*, **1984**, 24, 164–175.

[30] Rouvray, D.H. *Sci. Am.*, **1986**, 255, 40–47.

[31] Wieland, T. The use of structure generators in predictive pharmacology and toxicology. *Arzneim.-Forsch./Drug Res.*, **1996**, 46 (I), 223–227.

[32] Floyd, R.W. Algorithm 97, Shortest path. *Comm. Assoc. Comp. Mach.*, **1962**, 5, page 345.

[33] Wiener, H. Structural determination of paraffin boiling point. *J. Amer. Chem. Soc.*, **1947**, 69, 17–20.

[34] Balaban, A.T. Highly discriminating distance-based topological index. *Chem. Phys. Lett.*, **1982**, 89, 399–404.

[35] Rouvray, D.H. Should we have designs on topological indices? In King, R.B., editor, *Chemical Applications of Topology and Graph Theory*, 159–177. Elsevier, Amsterdam, 1983.

[36] Bonchev, D. and N. Trinajstić. Information theory, distance matrix and molecular branching. *J. Chem. Phys.*, **1977**, 67, 4517–4533.

[37] Wieland, T. Mathematical simulations in combinatorial chemistry. *MATCH*, **1996**, 34, 179–206.

[38] Henneberg, D. and B. Weimann. *MassLib, Evaluation of low resolution mass spectra series*. Max-Planck-Institut für Kohlenforschung, Mülheim/Ruhr, 1992. Version 7.2.

[39] Scsibrany, H. and K. Varmuza. *Handbuch zu ToSIM (Software zur Untersuchung von topologischen Ähnlichkeiten in Molekülen)*. Technische Universität, Wien, 1994.

[40] Varmuza, K. and H. Scsibrany. Clusteranalyse isomerer chemischer Strukturen basierend auf binären Deskriptoren und der Hauptkomponentenanalyse. In *9. CiC-Workshop*, Bitterfeld, 1994. (Posterpräsentation).

[41] Varmuza, K., W. Werther, F. Stancl, A. Kerber, and R. Laue. Computer-assisted structure elucidation of organic compounds, based on mass spectra classification and exhaustive isomer generation. In Gasteiger, J., editor, *Software-Entwicklung in der Chemie 10*, 303–314. GDCh, Frankfurt am Main, 1996.

[42] Benecke, C. *Algorithmen zur Klassifizierung diskreter Strukturen*. PhD thesis, Universität Bayreuth, 1996. (in Vorbereitung).

[43] Willett, P. *Similarity and clustering in chemical information systems*. J. Wiley & Sons, New York, NY, 1987.

[44] Dugundji, J. and I.K. Ugi. An algebraic model of contitutional chemistry as a basis for chemical computer programs. *Top. Curr. Chem.*, **1973**, 39, 19–64.

[45] Kerber, A. *Algebraic Combinatorics Via Finite Group Actions*. BI-Wissenschaftsverlag, Mannheim, Wien, Zürich, 1991.

[46] Laue, R. Construction of combinatorial objects – a tutorial. *Bayreuther Mathem. Schr.*, **1993**, 43, 53–96.

[47] Grüner, T., R. Laue, and M. Meringer. Applications for group actions applied to graph generation. In Finkelstein, L. and C. Kantor, editors, *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, Providence, RI, 1995.

[48] Shelley, C.A. Heuristic Approach for Displaying Chemical Structures. *J. Chem. Inf. Comput. Sci.*, **1983**, 23, 61–65.

[49] Kubinyi, H. Der Schlüssel zum Schloß. II. Hansch-Analyse, 3D-QSAR und De novo-Design. *Pharmazie in unserer Zeit*, **1994**, 23(5), 281–290.

[50] Kubinyi, H., editor. *3D QSAR in Drug Design. Theory, Methods and Applications*, Leiden, 1993. ESCOM Science Publishers.

[51] Cramer III., R.D., D.E. Patterson, and A. Vittoria. *J. Amer. Chem. Soc.*, **1988**, 110, 5959–5967.

[52] Crippen, G.M. *Distance Geometry and Conformational Calculations*. J. Wiley, New York, NY, 1981.

[53] Hoflack, J. and P.J. De Clercq. The SCA program: An easy way for the conformational evaluation of polycyclic molecules. *Tetrahedron*, **1988**, 44, 6667–6676.

[54] Pearlman, R. Rapid generation of high quality approximate 3D molecular structures. *Chem. Design Autom. News*, **1987**, (2), 5–6.

[55] Sadowski, J. and J. Gasteiger. From atoms and bonds to three-dimensional atomic coordinates: Automatic model builders. *Chem. Rev.*, **1993**, 93, 2567–2581.

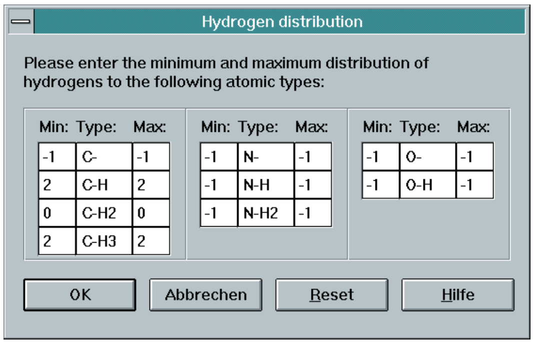


Figure 1: The input dialog for hydrogen distributions in MOLGEN.

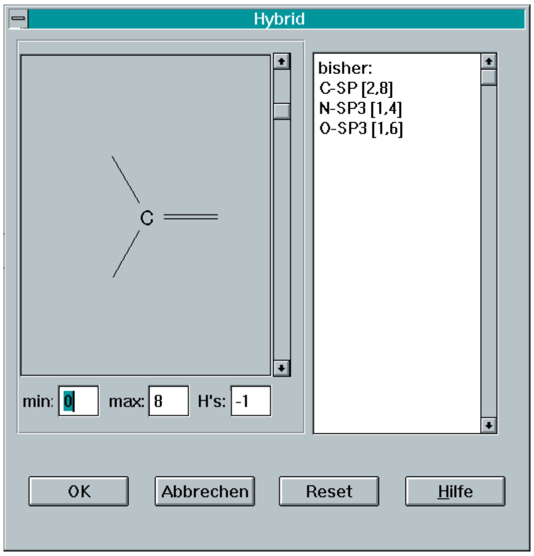


Figure 2: The input dialog for hybridization states in MOLGEN.

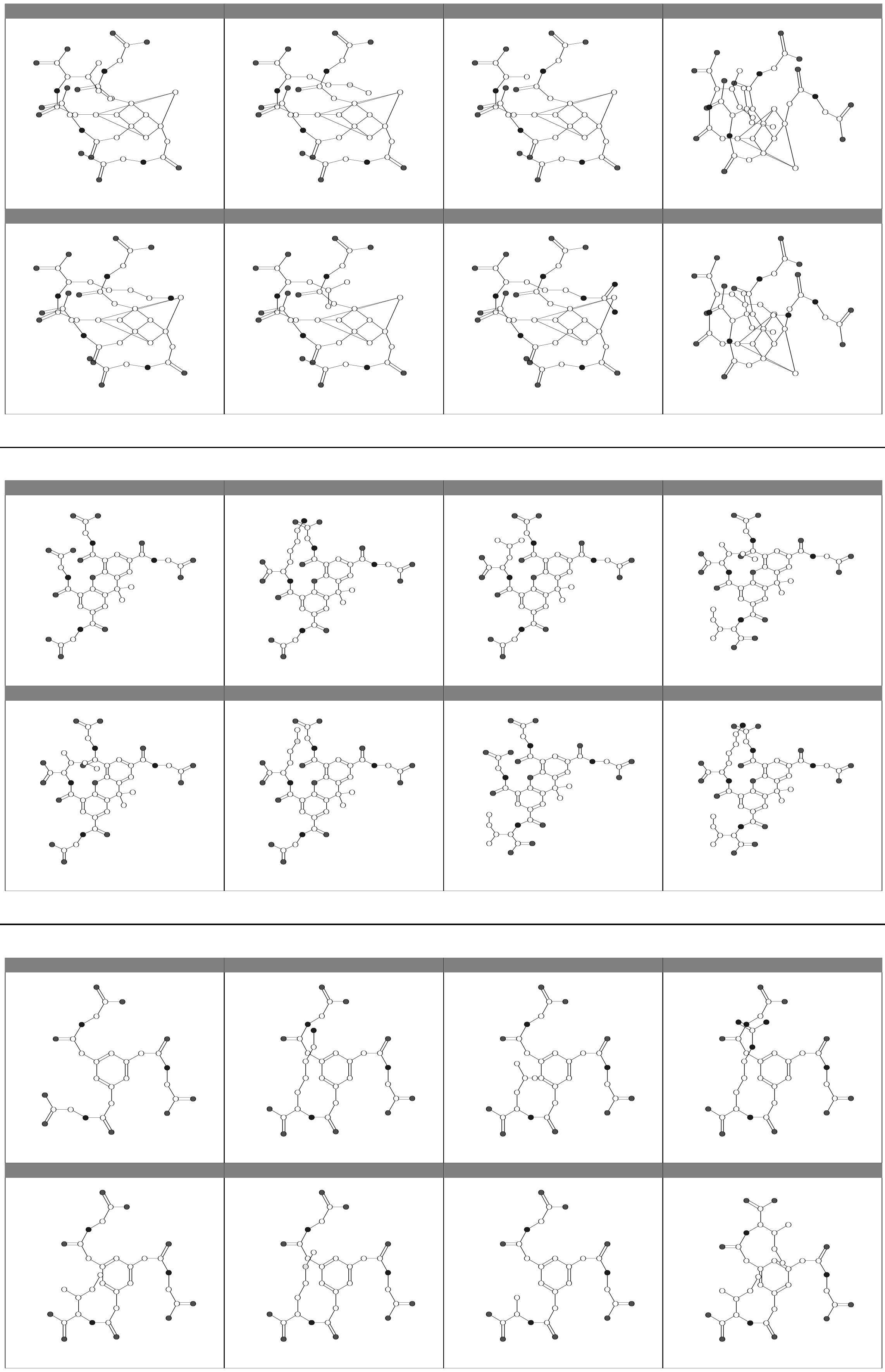


Figure 3: Extracts from the combinatorial libraries produced from the structures **IV**, **V** and **VI** and the natural amino acids

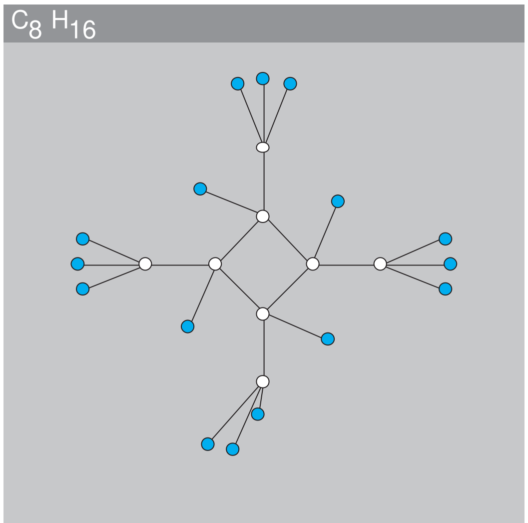


Figure 4: Tetramethylcyclobuthane

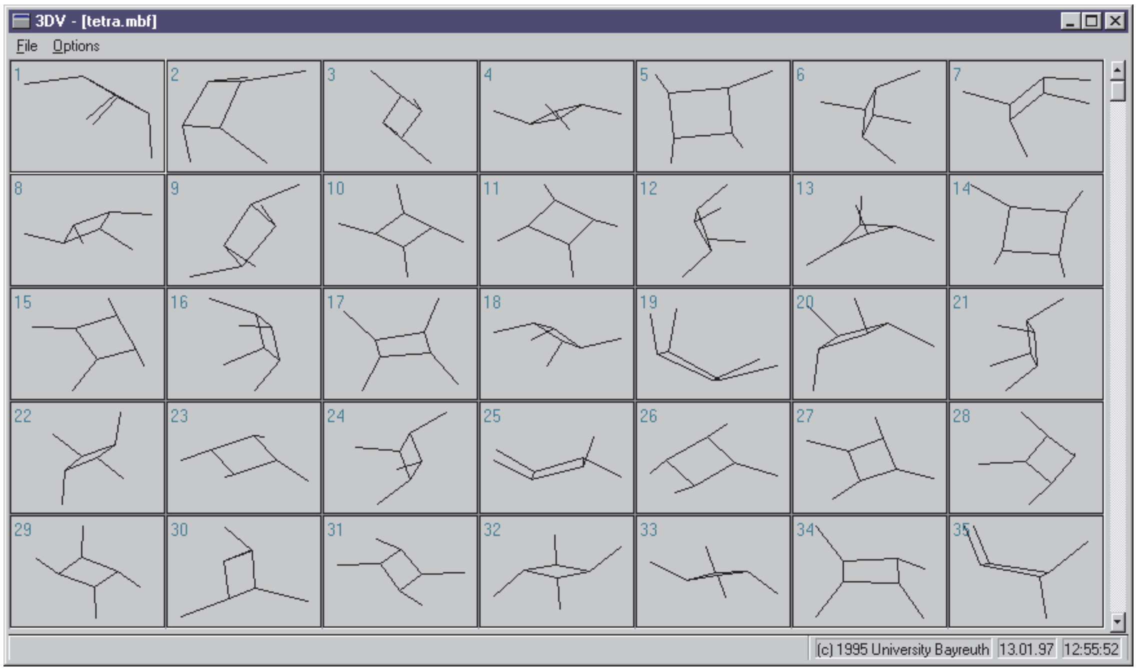


Figure 5: Different 3D placements of Tetramethylcyclobuthane

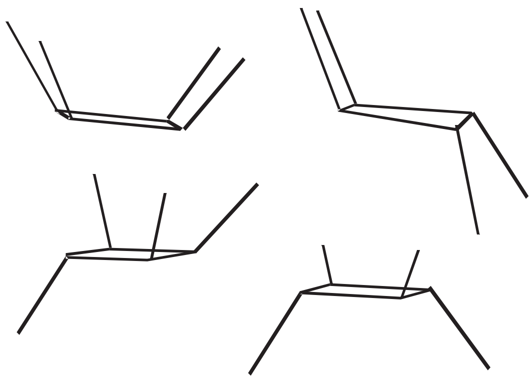


Figure 6: Base of 3D placements of Tetramethylcyclobuthane