MATCH
Communications in Mathematical
and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **58** (2007) 239-280

ISSN 0340 - 6253

History and Progress of the Generation of Structural Formulae in Chemistry and its Applications

(dedicated to the memory of Ivar Ugi)

Ralf Gugisch, Adalbert Kerber,* Reinhard Laue,
Markus Meringer, Christoph Rücker
Department of Mathematics
University of Bayreuth
D-95440 Bayreuth, Germany

(Received July 31, 2006)

Abstract

After a few remarks on the history of molecular modelling we describe certain mathematical aspects of the generation of molecular structural formulae. The focus is on the automatic generation of structural formulae for the purpose of molecular structure elucidation and the examination of molecular libraries. The aim is to give a review and to point to relevant literature. We demonstrate an application in the area of quantitative structure-property/activity relationships. Then, we give a glance on ongoing research in the generation of 3D-structures (stereoisomers and conformers), and finally we mention two problems that should be solved in the near future, the possible use of hypergraphs, and the generation of patent libraries.

*corresponding author, email: kerber@uni-bayreuth.de

1 History

The first level in modeling a molecule is the *arithmetic description* using a **molecular formula,** e.g.

$$C_6H_6.$$

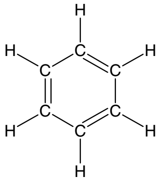
This does not suffice to distinguish molecules, as already Alexander von Humboldt (1769-1859) stated ([1]) in vol. I of his book [2], published in 1797. We quote from page 128:

> – Drei Körper a, b und c können aus *gleichen* Quantitäten Sauerstoff, Wasserstoff, Kohlenstoff, Stickstoff und Metall zusammengesetzt und in ihrer Natur doch unendlich *verschieden* seyn.

Here Humboldt states in a very clear language that chemical compounds (*Körper*) may exist that contain the same quantities of oxygen, hydrogen, carbon, nitrogen or metal while they may be different in infinitely many aspects. On page 127 he even uses the word "Bindung" (bond).

In the 1820s Wöhler and von Liebig found that cyanic acid and fulminic acid have the same atomic constituents, and so they proved Humboldt's statement to be true. In 1830 Berzelius realized this as a general *phenomenon* and called it **isomerism**.

The existence of this phenomenon means that higher precision is needed in distinguishing compounds, that we have to go to a higher level of accuracy. This second level is called the *topological* or *constitutional level.* The topological model of organic molecules is *a graph theoretic interaction model,* expressing the molecule in question in terms of a **structural formula**, e.g.



This is a connected multigraph consisting of 6 nodes of valence 4, they represent the carbon atoms, and 6 nodes of valence 1, the hydrogen atoms. The edges, called *covalent bonds, express interactions between pairs of atoms.* (The situation is a bit more complicated in reality, since there is aromaticity. We neglect this at the moment, but will come back to it later. In fact there is a problem. The graph theoretic model of a molecule apparently needs to be extended!)

A mathematical generator of connected multigraphs with given valences of the nodes produces altogether 217 structures with 6 nodes of valence 4 and 6 nodes of valence 1, and so there are 217 mathematically possible connectivity isomers that have the molecular formula C_6H_6. Among these are exactly six isomers of formula $(CH)_6$, i.e. in these each C atom bears exactly one H atom.

Experience has shown that there are distinct compounds (molecules) even sharing the same connected multigraph. Therefore another (the third) level of detail has to be considered. This is the *geometric level*, where phenomena such as *chirality* and *stereoisomerism* occur. Energy models allow placements of connected atoms in 3D space, they show e.g. that of the 217 C_6H_6 structures fewer than 70 are reasonable in the sense that 3D models containing usual bond lengths, bond angles etc. can be built, and that among these there are exactly 7 for which two distinctly different rather than a single 3D realization are possible: stereoisomers [3]. In 5 of these 7 cases the two stereoisomers are mirror images of each other, the phenomenon of nonidentical mirror images is called chirality. Hence, the problem that arises is the following:

> Construct all these structural formulae, the corresponding connectivity isomers as well as their stereoisomers in an efficient manner free of redundance. Moreover we would like to have them in a canonic form, so that they can be compared!

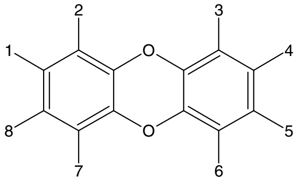
2 Solutions

The most famous paper that describes an early attack to solve these problems is due to G. Pólya ([4],[5], see also [6]). There are, of course, various predecessors, e.g. a paper by Lunn and Senior [7], who were the first to note that group theory plays a role here, and a paper by Redfield [8] that contained even better results. Nevertheless, Pólya's paper is not only a masterpiece, but it gave rise to the development of a whole theory that is nowadays called *Pólya's Theory of Enumeration.*

Pólya's approach to the enumeration of molecules with a given molecular formula is to subdivide the molecule in question into a *skeleton* and a set of *univalent substituents.* It leads to the following problem:

> Evaluate the set of essentially different distributions of the substituents over the sites of the skeleton, with respect to *the given symmetry group of the skeleton.*

The resulting isomers are called *permutational* or *substitutional* isomers. A software package that calculates the number of these isomers using exactly Pólya's approach is due to van Almsick, Dolhaine and Hönig [9]. For example, the 22 permutational isomers of dioxin (tetrachlorodibenzo-p-dioxin) are the essentially different distributions of 4 hydrogen and 4 chlorine atoms over the 8 sites of the skeleton



In order to fix the symmetry group, the skeleton of dioxin *is supposed to be planar and of symmetry group* $D_{2\mathrm{h}}$, which is equivalent to the Kleinian four group V_4. The way how double cosets and ladders of subgroups of the symmetric group can be used in order to construct the 22 isomers is described, for example, in [10].

However, in many isomer generation problems information on the skeleton and its symmetry group is either not available or these concepts are not even applicable, e.g. in generating the C_6H_6 structural formulas above. In fact, skeleton and symmetry group are concepts on the third (geometrical) level, and therefore, as a rule, do not play any role in the solution of problems on the second (topological) level. Nevertheless, even in such problems Pólya's Theory of Enumeration is useful, since it allows to find structural formulas as equivalence classes of multigraphs.

Pólya's approach uses the concept of *group action*. For more details on this notion and its applications to constructive theory of discrete structures see e.g. [10, 11].

Consider two group actions ${}_GX$ and ${}_HY$, i.e. mappings

$$G \times X \to X, (g,x) \mapsto gx,\ \ H \times Y \to Y, (h,y) \mapsto hy,$$

subject to the conditions that $g'(gx) = (g'g)x, h'(hy) = (h'h)y$ and $1x = x, 1y = y$, for any $g, g' \in G, h, h' \in H$, and the identity elements 1 of G and H.

These actions give rise to corresponding actions of $G, H, H \times G, H \wr G$ on the set of mappings

$$Y^X := \{f \mid f \colon X \to Y\}.$$

They are defined as follows:

$$G \times Y^X \to Y^X, (g, f) \mapsto f \circ g^{-1},$$

where $(f \circ g^{-1})(x) = f(g^{-1}x)$,

$$H \times Y^X \to Y^X, (h, f) \mapsto h \circ f,$$

where $(h \circ f)(x) = hf(x)$,

$$(H \times G) \times Y^X \to Y^X, ((h, g), f) \mapsto h \circ f \circ g^{-1},$$

where $(h \circ f \circ g^{-1})(x) = h(f(g^{-1}x))$,

$$(H \wr G) \times Y^X \to Y^X, ((\varphi, g), f) \mapsto \tilde{f},$$

where $\tilde{f}(x) = \varphi(x) f(g^{-1}x)$, since the *wreath product*

$$H \wr G = H^X \times G = \{(\varphi, g) \mid \varphi : X \to H, \ g \in G\}.$$

Many structures in mathematics and sciences can be considered as orbits of such actions, for suitably chosen ${}_GX$ and ${}_HY$. Examples are graphs, molecular graphs, switching functions, and various other notions.

In Pólya's approach to the enumeration of permutational isomers, X is the set of active sites of the molecular skeleton, G means the symmetry group of the skeleton, while Y denotes the set of admissible kinds of ligands that are to be distributed over the active sites of the skeleton. The corresponding action is an action of the form ${}_G(Y^X)$.

A good example for the fourth type of action is the enumeration of stereoisomeric inositols (see Figure 1). Here we also have a skeleton (cyclohexane) with 6 active sites. But now, each ligand (OH) may be connected to the skeleton in two different ways, say up and down. In order to obtain the appropriate group, we have to extend the automorphism group of the skeleton cyclohexane (which we assume to be the dihedral group D_6, or D_{6h} in Schönflies symbolism) by the information of whether or not the direction of each ring atom's OH ligand (up or down) is changed by a particular automorphism. The proper group to consider is a subgroup of the wreath product $S_2 \wr D_6$, action is on the set Y^X with the set Y of admissible configurations (up or down) and X as before the set of 6 active sites of the skeleton, see [12]. (Remark: Wreath products of the form $S_2 \wr G$, G a permutation group,

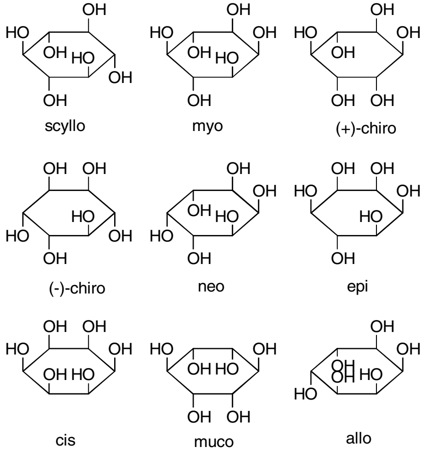


Figure 1: The nine inositols

are also known as signed permutation groups. In reference [12] the latter notation is used.)

The very elegant method of Pólya determines the number of orbits of the above group actions. This is achieved by an application of the Lemma of Cauchy–Frobenius which is obtained by doubly counting pairs (g, f) with $gf = f$, a powerful combinatorial tool. Unfortunately, this approach is non–constructive, but meanwhile a constructive version of that lemma exists ([13]).

The chemist usually wants to *see* the isomers that were counted. So we need a constructive solution of the orbit problem, a transversal of the orbits has to be evaluated!

The first systematic approach towards a construction of the isomers that correspond to a given molecular formula which did *not* assume a knowledge of the skeleton and its symmetry group was the famous **DENDRAL project** [14], run by Lederberg in the sixties/seventies. It was successful since it comprised an efficient generator. It was mathematically sophisticated since its designers used algebraic concepts such as computation of transversals of sets of double cosets, Sims chains etc. The reason for its limited use was that at that time no proper graphics were available, exotic languages were used, and the hardware was very expensive at that time compared with today's standards and efficiency. The project's ambitious motivation was to implement **Automated Molecular Structure Elucidation!** Its idea is the following:

- *Generate in silico* all the structural formulae that fit to given data of an unknown compound, coming from a given chemical spectrum, say NMR, IR, or MS.
- If the resulting set of candidates is too big (which usually is the case), then try to *reduce this search space*, by adding further information coming from the history of the compound etc.
- *Restart the generation* by including these new constraints *interactively* until a suitably small set of structures remains.
- If there is a reasonably small search space or no further information available, simulate for each of the remaining structural formulae its spectrum and *rank the candidates* according to similarity between the simulated and the experimental spectrum.

Based on MS, this is still a difficult problem, and only rather simple cases can routinely be solved using a PC [15, 16]. For a collection of molecular generators that can be used see the special issue on molecular generators

> A. Kerber (ed.): *MATCH Commun. Math. Comput. Chem.* **27** (1992)

A molecular generator that is suitable for mass spectroscopy (since it contains an interpreter of mass spectra based on Varmuza's MSclass [17]) is MOLGEN–MS [18]. For NMR spectroscopy, techniques to extract structural detail from a spectrum are more advanced, and so is the structure elucidation software developed by Elyashberg [19, 20] (generator by Molodtsov [21]).

2.1 A Mathematical Model of Organic Molecules

The molecular model used in MOLGEN can be described as follows: Each atom p of the molecule in question carries the *type of the atom:*

$$AT(p) = (AS(p), val(p), rad(p), chg(p)),$$

where

- $AS(p)$ is the *atom symbol* (e.g. C, O, N, ...),
- $val(p)$ means the *valence* of p,
- $rad(p) \in \{TRUE, FALSE\}$ indicates whether p is a radical center, and
- $chg(p) \in \{-3, -2, ..., +3\}$ indicates the *atomic charge* of p.

Aromatic doublets can be eliminated after construction, but there remains a problem (see Subsection 6.1).

This leads to the following definition that allows to embed this model into Pólya's Theory of Enumeration:

Definition 2.1 *Let $\mathcal{A} := \{\theta_1, ..., \theta_n\}$ denote a set of n atoms, and indicate by*

$$\binom{\mathcal{A}}{2} := \{\{i, j\} \subseteq \mathcal{A}, i \neq j\},$$

the set of pairs *of atoms in $\mathcal{A}$, by $4 := \{0, 1, 2, 3\}$ the set of bond multiplicities.*

- *A molecular graph is a mapping $f : \binom{\mathcal{A}}{2} \to 4$, where $f\{i, j\}$ denotes the bond multiplicity between atoms i and j.*
- *The set $4^{\binom{\mathcal{A}}{2}}$ of all these mappings is the set of all possible molecular graphs,*
- *and* the subset of connected molecular graphs with the prescribed valences *is denoted by $(4^{\binom{\mathcal{A}}{2}})'$.*
- *Since the atoms are numbered, we introduce the equivalence relation $f \sim f'$, if there exists a permutation $\pi \in S_n$ that keeps the type: $AT(\theta_i) = AT(\theta_{\pi(i)})$ and the multiplicities of the bonds:*

$$f(\{\theta_i, \theta_j\}) = f'\left(\{\theta_{\pi(i)}, \theta_{\pi(j)}\}\right).$$

- *The set of orbits of the symmetric group*

$$S_n \backslash\backslash \left(4^{\binom{A}{2}}\right)'$$

  then is the set of structural formulae corresponding to the given molecular formula defined by AS.

□

Hence, in the generator MOLGEN *group theoretic methods* play a decisive role, accompanied by *combinatorial tools,* and a careful concept for the *data structure.* Constraints on the molecular candidates (for example substructures, ring sizes,...) are already used *during the construction* (see [22, 23, 24]).

Pairwise isomorphism tests have to be avoided strictly. For this purpose a *canonical form* of molecular graphs is required [25]. The research on fast algorithms for canonical forms is still ongoing.

There are several specialized versions of MOLGEN available:

- **MOLGEN**, stand-alone, usable online in a reduced form, several versions (see below),
- **MOLGEN-MS**, which interprets mass spectra,
- **MOLGEN-COMB**, for the generation of combinatorial libraries from a set of molecules and reactions,
- **MOLGEN-QSPR**, offers >700 *molecular descriptors* and communicates with statistical software,
- **UNIMOLIS**, is meant for E-learning of the basic notions of isomerism, in particular of stereoisomerism. It is available online

  http://www.unimolis.de/

  and also on CD. It communicates online with MOLGEN.

MOLGEN 4.0, for example (see [26]), allows to put the following constraints in addition to a molecular formula:

- *intervals* for atom numbers,
- *atom types*,

- a *goodlist* of possibly overlapping substructures that are contained in the generated molecules, and
- a *goodlist* of substructures that must not overlap.
- A *badlist* of fragments which are forbidden,
- *surroundings* of fragments, *subunits*,
- H-*distribution*, *hybridization*,
- numbers of *cycles* of various lengths,
- numbers of *bonds* of given multiplicities,
- the number of ^{13}C NMR *signals*, which yields a bound for the symmetry group.

Stereoisomers are obtained by finding stereocenters and systematically inverting their configurations [27, 28, 3, 12]. Dreiding's and Dress' approach [29, 30, 31] using chirotopes (also known as oriented matroids) is under development [32], see section 5.

2.2 The Main Constructive Methods

We list these methods in particular, since *the very same methods apply also to the construction of various other discrete structures,* e.g. to the construction of groups, designs and codes.

2.1 Equivalence classes as orbits: *To construct finite discrete structures defined as equivalence classes, we proceed as follows:*

i) Replace the equivalence relation by a ***group action***

$$G \times X \to X, (g, x) \mapsto gx$$

that has the equivalence classes as orbits

$$G(x) = \{gx \mid g \in G\},$$

so that the set of equivalence classes is the set of orbits

$$G \backslash\!\backslash X = \{G(x) \mid x \in X\}.$$

ii) *The orbit $G(x)$ is essentially the same as the set*

$$G/G_x = \{gG_x \mid g \in G\}$$

of left ***cosets*** $gG_x = \{gh \mid h \in G_x\}$ *of the stabilizer*

$$G_x = \{g \in G \mid gx = x\},$$

since the mapping

$$G(x) \to G/G_x,\ gx \mapsto gG_x,$$

is a bijection.

2.2 The use of double cosets: *Assume an action ${}_GX$ and a subgroup $U \leq G$.*

i) *The set of orbits of U on $G(x)$ is bijective to the set*

$$U \backslash G/G_x = \{UgG_x \mid g \in G\}$$

of ***double cosets***

$$UgG_x = \{ugh \mid u \in U, h \in G_x\},$$

as the mapping

$$U \backslash\!\backslash G(x) \to U \backslash G/G_x,\ U(gx) \mapsto UgG_x$$

is a bijection.

ii) *For example, in Pólya's Theory, the set of equivalence classes of mappings with the same content as $f \in Y^X$ is bijective to*

$$G \backslash S_X/(S_X)_f,$$

where $(S_X)_f$ means the stabilizer of $f \in Y^X$ in the symmetric group S_X on X.

iii) *For the use of double cosets in chemistry, the reader is referred to [33], the review article by Ruch and Klein [34] and [35].*

iv) *Hall used double cosets for the construction of p–groups as early as in 1939.*

Thus, if we want to represent the orbits of U on X, we can break the problem into pieces by choosing a suitable bigger group G which also acts on X and in a way that its action extends the action of U. We restrict the attention to orbits $G(x)$ of G, and in each case we evaluate a transversal of the set of double cosets $U\backslash G/G_x$, from which a transversal of the orbits can be obtained.

2.3 The Homomorphism Principle: *Assume two finite actions of* G, *say* ${}_GM$ *and* ${}_GN$, *together with a surjective mapping* $\Theta\colon M \to N$, *such that* Θ *commutes with the action (such actions are called* homomorphic*):*

$$\begin{tikzcd}{}_GM \arrow[r, "\Theta"] & {}_GN\end{tikzcd}$$

such that $\Theta(gm) = g\Theta(m)$,
for each $m \in M, g \in G$.

Moreover we assume that T *is a transversal of the set of orbits* $G\,\backslash\!\!\backslash N$ *of* G *on* N. *Then the following is true (see e.g. [10, 11]):*

i) Each orbit $\omega \in G\,\backslash\!\!\backslash M$ *intersects the inverse image* $\Theta^{-1}(n)$ *of exactly one element in the transversal:*

$$\forall\ \omega \in G\,\backslash\!\!\backslash M\ \exists_1\ n \in T\colon\ \omega \cap \Theta^{-1}(n) \neq \emptyset.$$

ii) This intersection is an orbit of the stabilizer of the representative:

$$\omega \cap \Theta^{-1}(n) \in G_n\,\backslash\!\!\backslash M.$$

iii) Hence we can obtain a transversal T_G *of* $G\,\backslash\!\!\backslash M$ *as disjoint union of transversals of the inverse images:*

$$T_G := \bigcup_{n \in T} T(n),$$

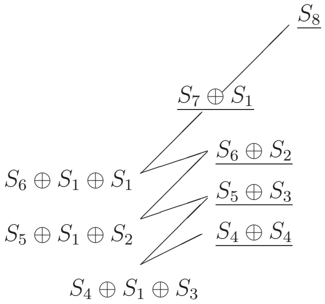
where $T(n)$ *denotes a transversal of* $G_n\,\backslash\!\!\backslash \Theta^{-1}(n)$.

This method can be applied, for example, to actions of the form ${}_G(Y^X)$, and it allows recursively to evaluate transversals, recursive to the order $|Y|$. Thus the Homomorphism Principle allows to reduce the size of the set and of the group.

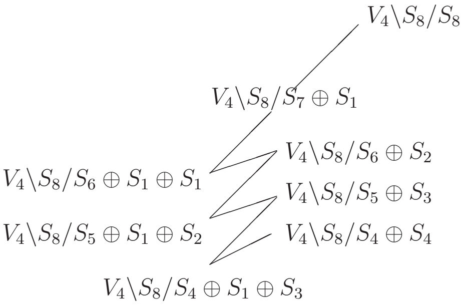
Moreover, the Homomorphism Principle can be applied for a recursive evaluation of orbit representatives of $G \backslash\backslash Y^X$, according to content. The content of $f \in Y^X$ is the sequence of multiplicities $|f^{-1}(y)|, y \in Y$. The recursion uses an up-down-sequence, a *subgroup-ladder* [36], of stabilizers of the form

$$(S_X)_f = \oplus_{y \in Y} S_{f^{-1}(y)} \leq S_X$$

in the symmetric group. Here is the subgroup ladder that can be used in the evaluation of the 22 permutational isomers of tetrachlorodibenzo-p-dioxin:


$\underline{S_8}$
$\underline{S_7 \oplus S_1}$
$\underline{S_6 \oplus S_2}$
$S_6 \oplus S_1 \oplus S_1$
$\underline{S_5 \oplus S_3}$
$S_5 \oplus S_1 \oplus S_2$
$\underline{S_4 \oplus S_4}$
$S_4 \oplus S_1 \oplus S_3$

The underlined subgroups are stabilizers (in the symmetric group) of permutational isomers obtained by distributing chlorine and hydrogen atoms over the free active sites. $S_7 \oplus S_1$, for example, is – up to isomorphism – the stabilizer of an isomer that contains 7 hydrogen atoms and exactly one chlorine atom, while $S_4 \oplus S_4$ is isomorphic to the stabilizer of an isomer that contains 4 hydrogen atoms and 4 chlorine atoms, i.e. a permutational isomer of dioxin. To this subgroup ladder there corresponds the following ladder of sets of double cosets


$V_4 \backslash S_8 / S_8$
$V_4 \backslash S_8 / S_7 \oplus S_1$
$V_4 \backslash S_8 / S_6 \oplus S_2$
$V_4 \backslash S_8 / S_6 \oplus S_1 \oplus S_1$
$V_4 \backslash S_8 / S_5 \oplus S_3$
$V_4 \backslash S_8 / S_5 \oplus S_1 \oplus S_2$
$V_4 \backslash S_8 / S_4 \oplus S_4$
$V_4 \backslash S_8 / S_4 \oplus S_1 \oplus S_3$

To this ladder of sets of double cosets we apply the Homomorphism Principle. For example, in the last but one step, we have a transversal of $V_4 \backslash S_8 / S_5 \oplus S_3$ at hand, i.e. the permutational isomers containing exactly 5 hydrogen and 3 chlorine atoms. From this transversal we obtain, by an application of the Homomorphism Principle, a transversal of the bigger set $V_4 \backslash S_8 / S_4 \oplus S_1 \oplus S_3$. In the final step we evaluate the desired transversal of the smaller set $V_4 \backslash S_8 / S_4 \oplus S_4$ from which we obtain the desired 22 permutational isomers of dioxin that are shown in Figure 2, constructed by MOLGEN. For the sake of clarity, the 4 hydrogen atoms are not shown.

3 Molecular Libraries

Once we have an efficient generator at hand, it is easy to generate molecular libraries. They play a central role in combinatorial chemistry (see e.g. [37]). For instance QSAR/QSPR models can be computed and applied in order to predict physicochemical properties or biological activities. For the generation of virtual combinatorial libraries MOLGEN-COMB was designed ([38, 39]). It is part of MOLGEN-QSPR [40]. The evaluation of molecular descriptors – altogether more than 700 of them – allows to look for correlations between values of descriptors and compound properties, to build and to apply QSAR/QSPR models. A classical example is the search for the boiling points of the compounds in a molecular library, if this property is known for only part of the library.

There is an interface that allows to connect MOLGEN–QSPR with the software package for statistical computing R [41]. See [42, 43] for examples, where boiling points of haloalkanes and in particular of fluoroalkanes, respectively, are predicted.

An interesting byproduct of the application of MOLGEN–QSPR is that we can also look for correlations between molecular descriptors. An application to a library of 13410 diverse chemical compounds exhibited 26 equivalence classes of fully correlated descriptors [44]. Using the same computer program we found that the second Zagreb Index M_2 is half of $mwc^{(3)}$, the number of molecular walks of length 3 (see Section 4).

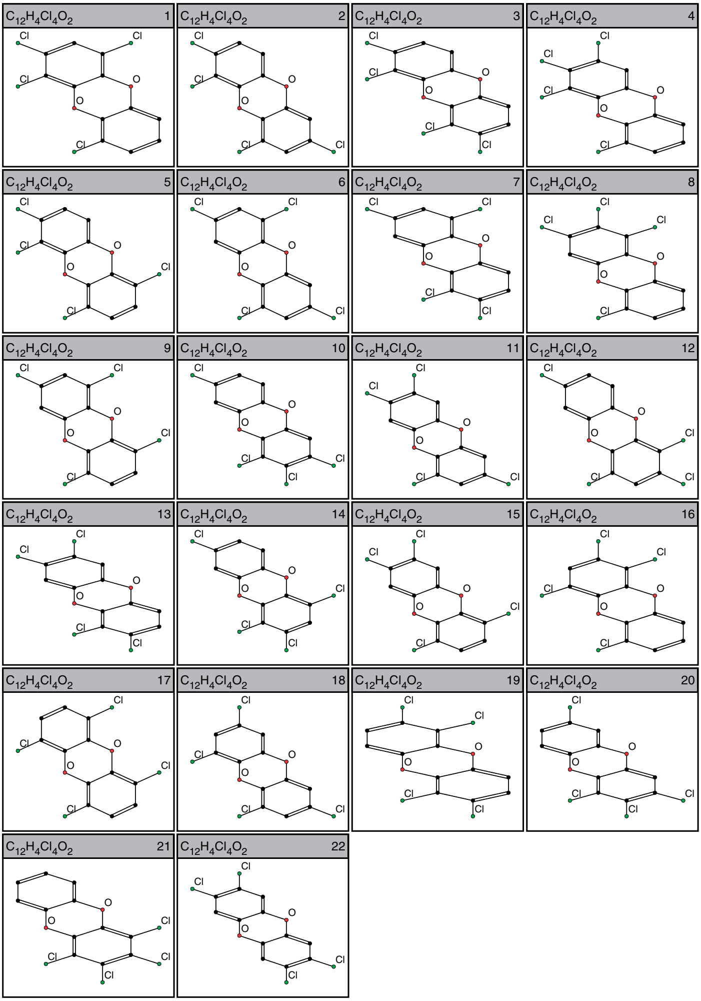


Figure 2: The 22 permutational isomers of dioxin

4 Quantitative Structure–Property/Activity Relationships

A frequently occurring problem in computational chemistry is the prediction of physicochemical properties or biological activities for chemical compounds given by their molecular graphs. A widely applied approach is the establishment of quantitative structure–property/activity relationships (QSPR/QSAR) starting from a set of compounds with known property/activity values. These known values can originate either from databases or from new measurements. We call this initial set of compounds the real library.

The search for a QSPR/QSAR is generally divided into two steps:

i) Using molecular descriptors chemical compounds are mapped onto real numbers. Typically a large number of molecular descriptors is applied, so that after this first step chemical compounds are represented by equal length vectors of real numbers. Together with the known property/activity values these are the input for the second step.

ii) Methods of supervised statistical learning are applied in order to find prediction functions that have the vector representations of the real library compounds as input, and that have output values which well fit the given property/activity values. In terms of statistical learning theory, molecular descriptors serve as independent variables and the property/activity is the dependent variable.

Once a QSPR/QSAR is found, it can be applied in order to make predictions for compounds whose property/activity values are not yet known. These could for instance be part of a virtual library generated by MOLGEN–COMB. Figure 3 shows a simplified flowchart of QSPR/QSAR search and application. Algorithmic parts are highlighted in grey. In the following we will give a short survey on frequently used molecular descriptors, topological indices, followed by a short summary of machine learning techniques offered by MOLGEN–QSPR.

4.1 Molecular Descriptors

For a connected molecular graph f on n atoms let f^s the associated simple graph, E_f the set of edges of f, Ω the set of non–hydrogen atoms, $f|_\Omega$ the subgraph of f induced by Ω, $\mathrm{dist}_f(i,j)$ the distance between atoms i and j

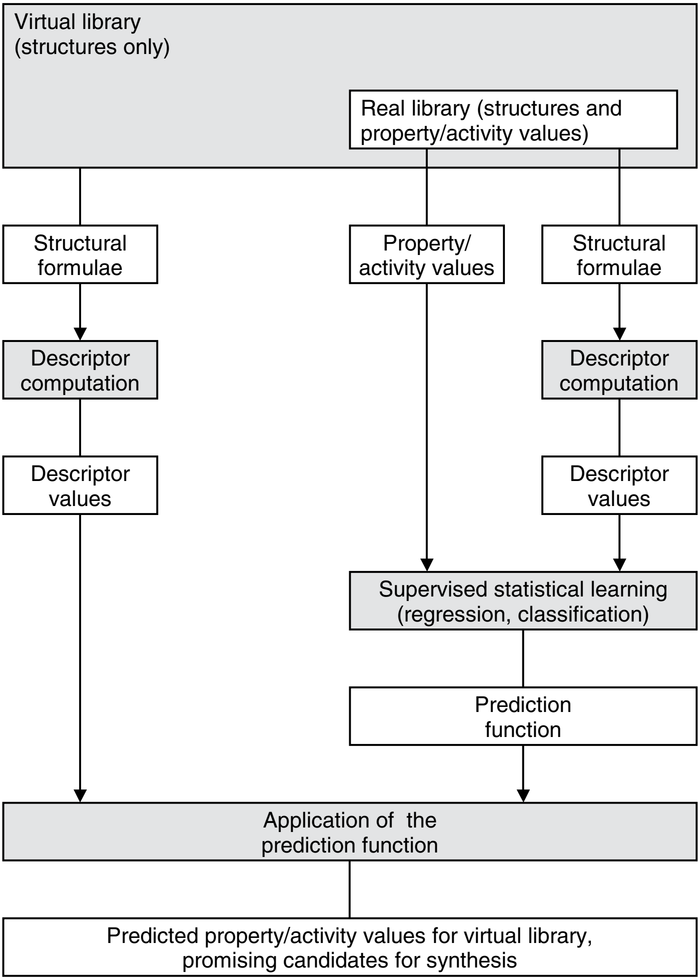


Figure 3: Flowchart of QSPR/QSAR search and application

and $\deg^s_f(i)$ the number of neighbors of i in f, or, in other words, the vertex degree of i in f^s.

One of the first applications of topological indices was developed by H. Wiener [45]. He used the index later named after him

$$W(f) = \frac{1}{2} \sum_{i\in\Omega} \sum_{j\in\Omega} \mathrm{dist}_f(i,j),$$

for modeling boiling points of alkanes (cf. 4.3).

Zagreb indices [46] sum up squares and products of vertex degrees:

$$\begin{aligned} M_1(f) &= \sum_{i\in\Omega} \left(\deg^s_{f|_\Omega}(i)\right)^2, \\ M_2(f) &= \sum_{\{i,j\}\in E_{f|_\Omega}} \deg^s_{f|_\Omega}(i)\cdot\deg^s_{f|_\Omega}(j). \end{aligned}$$

Randic indices [47, 48] of order m are computed by

$$^0\chi(f) = \sum_{i\in\Omega} \left(\deg^s_{f|_\Omega}(i)\right)^{-\frac{1}{2}}$$

if $m = 0$ and by

$$^m\chi(f) = \sum_{\substack{(v_0,\ldots,v_m)\\ \text{path in } f|_\Omega}} \prod_{i=0}^{m} \left(\deg^s_{f|_\Omega}(v_i)\right)^{-\frac{1}{2}}$$

if $m > 0$.

The *vertex distance degree* or *distance sum* of vertex $i \in \Omega$ is defined as

$$\deg^d_{f|_\Omega}(i) := \sum_{j\in\Omega} \mathrm{dist}_f(i,j).$$

It is needed for computing the Balaban index [49, 50]

$$J(f) = \frac{B(f)}{C(f)+1} \sum_{(i,j)\in E_{f|_\Omega}} \left(\deg^d_{f|_\Omega}(i)\cdot\deg^d_{f|_\Omega}(j)\right)^{-\frac{1}{2}}.$$

Herein $B(f)$ denotes the number of bonds of the molecular graph and $C(f)$ represents its cyclomatic number.

The *molecular topological index* by Schultz [51, 52] is defined as

$$MTI(f) = \sum_{i\in\Omega}\sum_{j\in\Omega}\sum_{k\in\Omega} a_{ik}\left(a_{kj} + \mathrm{dist}_f(k,j)\right),$$

where $A_{f^s|_\Omega} = (a_{ij})$ denotes the adjacency matrix of $f^s|_\Omega$.

The *molecular walk count* of length k adds all entries of the k–th power of the adjacency matrix of $f^s|_\Omega$:

$$mwc^{(k)}(f) = \sum_{i\in\Omega}\sum_{j\in\Omega} a_{ij}^{(k)}, \text{ where } (a_{ij}^{(k)}) = (A_{f^s|_\Omega})^k.$$

These indices describe the labyrinthicity and complexity [53, 54, 55] of a (molecular) graph. The *total walk count* sums molecular walk counts over all lengths k:

$$twc(f) = \sum_{k<|\Omega|} mwc^{(k)}(f).$$

The *principal eigenvalue* (largest in absolute value) of the adjacency matrix $A_{f^s|_\Omega}$ can also be used as molecular descriptor. It is denoted by λ_1^A.

There are topological indices that are not purely topological, but which take also the chemical element of the atoms into account.

For a molecular graph f the *valence vertex degree* of atom $i \in \Omega$ is defined as

$$\deg_f^v(i) = \frac{VE(i) - HC(i)}{TE(i) - VE(i) - 1}.$$

$HC(i)$ denotes the number of H atoms attached to atom i, $VE(i)$ is the number of valence electrons of atom i, and $TE(i)$ is the total number of electrons of atom i, i.e. its atomic number.

Valence vertex degrees are used to compute Kier & Hall indices [56, 57, 48]. Similar to Randic indices of order m Kier & Hall indices also sum over all paths of length m, but they use valence vertex degrees instead of vertex degrees:

$$\begin{aligned}
{}^0\chi^v(f) &= \sum_{i\in\Omega} \left(\deg_f^v(i)\right)^{-\frac{1}{2}},\\
{}^m\chi^v(f) &= \sum_{\substack{(v_0,\dots,v_m)\\ \text{path in } f|_\Omega}} \prod_{i=0}^{m} \left(\deg_f^v(v_i)\right)^{-\frac{1}{2}}.
\end{aligned}$$

Another category of topological indices that also take chemical elements of atoms into account, are Basak's information theoretical indices [58, 59]. For computing them at first all atoms have to be classified with respect to their chemical elements and bonds to neighboring atoms up to distance r. With k_r classes and n_{ri} atoms in class i, the following indices can be defined:

$$\begin{aligned} IC_r(f) &= \sum_{i \in k_r} \frac{n_{ri}}{n} \log_2 \frac{n_{ri}}{n}, \\ CIC_r(f) &= \log_2 n - IC_r(f) \text{ and} \\ SIC_r(f) &= (\log_2 n)^{-1} IC_r(f). \end{aligned}$$

These are called Basak's information content, complementary information content and structural information content of order r.

For reviews on these and many other molecular descriptors see the books written by Todeschini and Consonni [60], Karelson [61], and the collection [62] edited by Devillers and Balaban.

4.2 Supervised Statistical Learning

Supervised statistical learning is characterized by the presence of the dependent variable that guides the learning process and acts as a "teacher". Also unsupervised learning techniques, such as cluster analysis, play an important role in cheminformatics. These are typically applied when questions of diversity/similarity within chemical libraries have to be answered.

For the purpose of property/activity prediction two types of supervised learning techniques can be distinguished. If the dependent variable is discrete, classification methods will be applied. In QSPR/QSAR the dependent variable has quantitative character, i.e. is given by real numbers. The appropriate category of learning technique is called regression. A simple type of regression is ordinary least squares regression, based on a QR–decomposition of the design matrix defined by the descriptor values. Then the prediction function is a linear function of the descriptors. Often this type of regression is also called multiple linear regression (MLR).

In order to avoid overfitting in MLR it is necessary to find small subsets of descriptors that allow the calculation of good prediction functions. For this purpose there is an algorithm included in MOLGEN–QSPR that performs an exhaustive search for the best subsets of descriptors for MLR. For problems with large numbers of compounds and descriptors and/or big subsets exhaustive search usually is too time expensive. In order to handle such problems MOLGEN–QSPR offers an algorithm for step-up subset selection.

Besides these methods based on MLR, the current version of MOLGEN–QSPR offers k–nearest neighbor regression, and via an interface to the (freely available) statistical software package R [41] several more sophisticated techniques:

- regression trees [63],
- artificial neural networks [64, 65],
- support vector machines [66] and
- multivariate methods including PLS and PCR [67].

For a comprehensive survey on statistical learning, see [68]. We conclude this section with a small example of a QSPR study extracted from [69].

4.3 Example: Boiling Points of Decanes

Figure 4 shows a real library of 50 decanes together with their boiling points (BP). Structures and BP are extracted from the Beilstein registry, BP are given in °C. We want to find QSPR models for this physical property.

For this purpose we start our examinations with 30 topological descriptors as introduced in subsection 4.1:

$$W,\ M_1,\ M_2,\ {}^0\chi,\ {}^1\chi,\ {}^2\chi,\ {}^0\chi^v,\ {}^1\chi^v,\ {}^2\chi^v,\ {}^3\chi^v,\ J,\ MTI,\ twc,$$
$$mwc^{(2)},\ mwc^{(3)},\ mwc^{(4)},\ mwc^{(5)},\ mwc^{(6)},\ mwc^{(7)},\ mwc^{(8)},$$
$$\lambda_1^A,\ IC_0,\ CIC_0,\ SIC_0,\ IC_1,\ CIC_1,\ SIC_1,\ IC_2,\ CIC_2,\ SIC_2.$$

Since decanes are exclusively built of carbon and hydrogen atoms connected by single bonds, ${}^k\chi$ and ${}^k\chi^v$ have identical values. For this reason we exclude ${}^0\chi$, ${}^1\chi$ and ${}^2\chi$. By definition all decanes have the same molecular formula $\mathsf{C}_{10}\mathsf{H}_{22}$. Thus IC_0, CIC_0, SIC_0 are constant and will not influence the modeling. Tables 1 and 4.3 show values for the remaining 24 indices applied to the 50 decanes of Fig. 4. We see that M_1 and $mwc^{(2)}$ have the same values. This identity holds in general (for a proof see [55]):

$$mwc^{(2)} = M_1.$$

In order to detect pairwise affine dependences between the 24 indices we conducted a correlation analysis. This way we found the dependence between

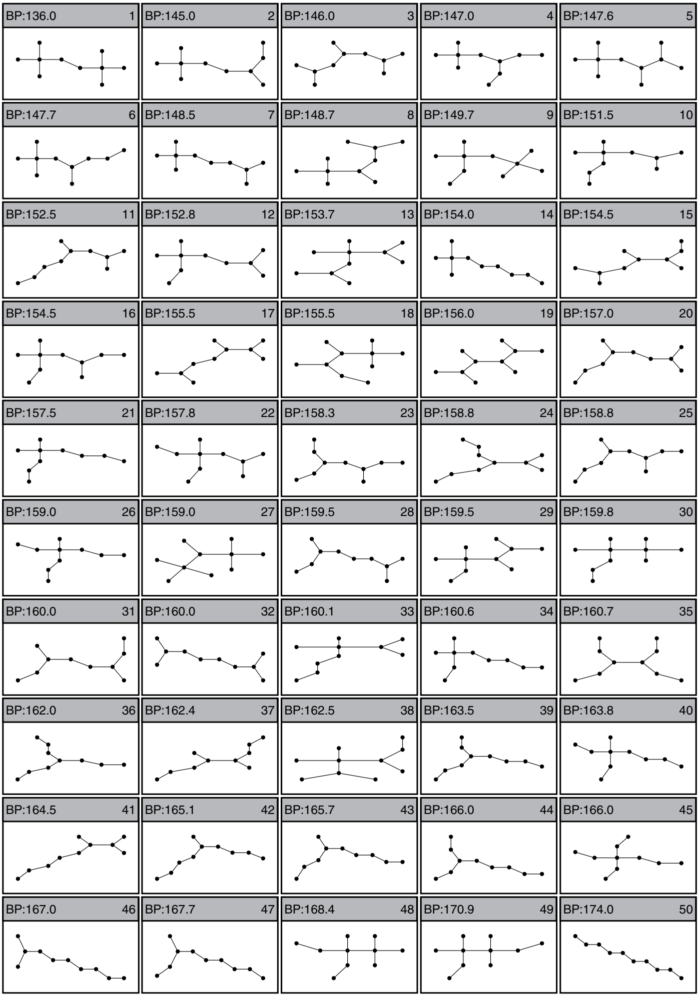


Figure 4: Real library of decanes together with their boiling points

M_2 and $mwc^{(3)}$ by an automated method: The correlation matrix shows an entry 1 for these two descriptors. The exact relation between these two indices can also be computed automatically by simply calculating a linear regression with one of the indices as dependent variable and the other one as independent variable. Thus we obtained

$$mwc^{(3)} = 2M_2.$$

Again, this relation holds in general, see [44].

Being fully correlated defines an equivalence relation on the set of molecular descriptors. Since we will search for MLR models, only one representative of each equivalence class needs to be included in our studies. Further members of the equivalence class will not improve a MLR.

A glance at the correlation matrix shows further dependences between the descriptors. Table 2 represents a part of the correlation matrix. Signs of correlation coefficients were suppressed. The first column shows absolute values of the correlation coefficients between BP and the descriptors. Descriptors were sorted in descending order of these values. The other columns contain absolute values of correlation coefficients of two descriptors each. In this particular example pairs from $\{IC_1, CIC_1, SIC_1\}$ are fully correlated. This is due to the fact that compounds in the decane library have the same number of atoms. More precisely, for decanes we have $CIC_1 = 5 - IC_1$ and $SIC_1 = \frac{1}{5}IC_1$. Also pairs from $\{IC_2, CIC_2, SIC_2\}$ are fully correlated. Thus we exclude CIC_1, SIC_1, CIC_2 and SIC_2 from our considerations.

Using all remaining 18 indices, MLR delivers a model with $R^2 = 0.97439$ and $R^2_{CV} = 0,94191$. In order to avoid overfitting we look for models with fewer descriptors. For $n = 1, ..., 5$ we run through all n–subsets of the 18 topological indices and note the models with highest R^2. Furthermore we give the used descriptors X_j, $j \in n$, followed by the QSPR equation for the prediction function f. Finally also prediction functions for auto–scaled descriptor values (with arithmetic mean 0 and variance 1) are given in order to allow better appreciation of the various descriptors' influence.

$n = 1$ descriptor: ${}^2\chi^v$,

$$\begin{aligned} f &= -8.0356X_0 + 190.74 \\ &= -5.0362X_0^* + 157.85. \end{aligned}$$

$n = 2$ descriptors: $mwc^{(4)}$, $mwc^{(8)}$,

$$\begin{aligned} f &= -1.2961X_0 + 0.026540X_1 + 287.83 \\ &= -42.917X_0^* + 41.312X_1^* + 157.85. \end{aligned}$$

	W	M_1	M_2	${}^0\chi^v$	${}^1\chi^v$	${}^2\chi^v$	${}^3\chi^v$	J	MTI	twc	$mwc^{(2)}$	$mwc^{(3)}$
1	127	46	44	8.4142	4.2071	5.6213	1.6250	3.5630	464	19248	46	88
2	134	42	41	8.1987	4.4545	4.6128	2.0841	3.3555	488	15138	42	82
3	135	40	39	8.1463	4.5197	4.3643	1.7475	3.3374	490	12930	40	78
4	126	42	42	8.1987	4.4925	4.4473	2.0557	3.6308	456	17334	42	84
5	124	44	44	8.3618	4.3272	4.9861	2.0724	3.6842	450	19018	44	88
6	131	42	41	8.1987	4.4545	4.6586	1.7423	3.4695	476	16146	42	82
7	139	42	40	8.1987	4.4165	4.8467	1.7083	3.2055	508	13874	42	80
8	123	44	45	8.3618	4.3372	4.8966	2.3034	3.7348	446	20498	44	90
9	119	46	46	8.4142	4.2678	5.2552	1.9660	3.8876	432	23048	46	92
10	127	42	42	8.1987	4.4772	4.5122	1.8876	3.6256	460	17946	42	84
11	142	38	37	7.9831	4.6639	3.8769	1.9243	3.1600	516	11114	38	74
12	131	42	42	8.1987	4.4772	4.4503	2.3556	3.4647	476	16602	42	84
13	120	44	46	8.3618	4.3599	4.7413	2.4973	3.8656	434	22234	44	92
14	146	40	38	8.0355	4.5607	4.3713	1.7803	3.0438	534	12390	40	76
15	130	40	41	8.1463	4.5746	3.9924	2.4585	3.5027	470	14984	40	82
16	126	42	43	8.1987	4.5152	4.2353	2.5551	3.6419	456	18280	42	86
17	136	40	40	8.1463	4.5366	4.1925	2.3374	3.3014	494	13242	40	80
18	118	44	47	8.3618	4.3921	4.5402	2.8635	3.9418	426	23206	44	94
19	121	42	45	8.3094	4.4641	4.2063	2.9325	3.8140	436	19426	42	90
20	143	38	37	7.9831	4.6639	3.8650	2.0183	3.1244	520	10786	38	74
21	134	40	40	8.0355	4.6213	4.0178	2.1339	3.4175	486	15664	40	80
22	122	42	44	8.1987	4.5378	4.1157	2.6082	3.8026	440	20028	42	88
23	133	38	39	7.9831	4.7399	3.4316	2.5873	3.4123	480	13028	38	78
24	131	38	39	7.9831	4.7187	3.5814	2.2617	3.4999	472	13848	38	78
25	138	38	38	7.9831	4.7019	3.6430	2.2831	3.2686	500	12020	38	76
26	126	40	42	8.0355	4.6820	3.6642	2.5607	3.6903	454	18298	40	84
27	111	48	51	8.5774	4.1547	5.4537	2.5981	4.2311	402	29658	48	102
28	146	38	37	7.9831	4.6639	3.8382	2.1753	3.0333	532	10236	38	74
29	116	44	48	8.3618	4.4147	4.3748	3.1439	4.0341	418	24610	44	96
30	115	46	50	8.4142	4.3107	4.8839	2.9053	4.1018	416	29160	46	100
31	141	38	38	7.9831	4.7019	3.6042	2.5461	3.1682	512	11298	38	76
32	151	38	36	7.9831	4.6259	4.0722	1.8129	2.9095	552	9316	38	72
33	127	42	44	8.1987	4.5040	4.2468	2.7376	3.6334	460	19738	42	88
34	138	40	40	8.0355	4.6213	3.9749	2.4142	3.2770	502	14774	40	80
35	125	38	41	7.9831	4.7948	3.1532	2.7642	3.6982	448	15866	38	82
36	138	36	36	7.8200	4.8461	3.2321	2.0908	3.2951	498	10950	36	72
37	135	38	39	7.9831	4.7187	3.5319	2.4594	3.3759	488	13386	38	78
38	115	44	49	8.3618	4.4248	4.2854	3.3705	4.0893	414	26106	44	98
39	141	36	36	7.8200	4.8461	3.2052	2.2402	3.2055	510	10570	36	72
40	129	40	42	8.0355	4.6820	3.6213	2.8410	3.5755	466	17588	40	84
41	143	38	38	7.9831	4.6807	3.7171	2.4011	3.1296	520	11616	38	76
42	149	36	35	7.8200	4.8081	3.3896	2.1010	2.9984	542	9330	36	70
43	150	36	35	7.8200	4.8081	3.3896	2.0820	2.9680	546	9194	36	70
44	145	36	36	7.8200	4.8461	3.1783	2.3706	3.0869	526	10052	36	72
45	121	40	44	8.0355	4.7426	3.3107	3.0303	3.8748	434	20526	40	88
46	158	36	34	7.8200	4.7701	3.5967	1.8850	2.7732	578	7896	36	68
47	153	36	35	7.8200	4.8081	3.3628	2.2474	2.8862	558	8788	36	70
48	110	46	52	8.4142	4.3713	4.5178	3.3713	4.3283	396	31916	46	104
49	111	46	52	8.4142	4.3713	4.4749	3.5999	4.2818	400	31632	46	104
50	165	34	32	7.6569	4.9142	3.1213	1.9571	2.6476	604	6500	34	64

Table 1: Values of topological indices for the real library of decanes in Fig. 4

	$mwc^{(4)}$	$mwc^{(5)}$	$mwc^{(6)}$	$mwc^{(7)}$	$mwc^{(8)}$	λ_1^A	IC_1	CIC_1	SIC_1	IC_2	CIC_2	SIC_2
1	218	432	1040	2114	4978	2.1987	1.3245	3.6755	0.26489	1.7947	3.2053	0.35895
2	188	376	854	1728	3900	2.1474	1.4227	3.5773	0.28455	2.5354	2.4646	0.50707
3	174	342	764	1506	3366	2.1010	1.3602	3.6398	0.27205	2.2823	2.7177	0.45645
4	198	402	942	1926	4494	2.1889	1.4227	3.5773	0.28455	2.4104	2.5896	0.48207
5	210	430	1012	2098	4894	2.2047	1.3870	3.6130	0.27739	2.2322	2.7678	0.44645
6	194	382	908	1794	4272	2.1753	1.4227	3.5773	0.28455	2.5354	2.4646	0.50707
7	184	356	818	1590	3660	2.1289	1.4227	3.5773	0.28455	2.4729	2.5271	0.49457
8	212	450	1040	2250	5144	2.2361	1.3870	3.6130	0.27739	2.2322	2.7678	0.44645
9	234	472	1198	2422	6140	2.2646	1.3245	3.6755	0.26489	2.0416	2.9584	0.40832
10	200	404	968	1962	4710	2.2089	1.4227	3.5773	0.28455	2.5590	2.4410	0.51179
11	158	312	668	1328	2844	2.0698	1.3716	3.6284	0.27433	2.6945	2.3055	0.53891
12	192	396	896	1874	4214	2.1813	1.4227	3.5773	0.28455	2.4965	2.5035	0.49929
13	218	470	1102	2402	5608	2.2616	1.3870	3.6130	0.27739	2.2169	2.7831	0.44338
14	170	328	738	1436	3242	2.1192	1.3213	3.6787	0.26427	2.3204	2.6796	0.46407
15	180	376	822	1730	3770	2.1455	1.3602	3.6398	0.27205	2.4576	2.5424	0.49151
16	200	416	968	2020	4704	2.2082	1.4227	3.5773	0.28455	2.5590	2.4410	0.51179
17	172	354	754	1566	3326	2.1067	1.3602	3.6398	0.27205	2.3448	2.6552	0.46895
18	222	484	1138	2494	5854	2.2711	1.3870	3.6130	0.27739	2.3183	2.6817	0.46367
19	202	442	986	2170	4826	2.2143	1.1995	3.8005	0.23989	1.7947	3.2053	0.35895
20	156	310	650	1306	2724	2.0529	1.3716	3.6284	0.27433	2.5460	2.4540	0.50919
21	182	372	852	1756	4030	2.1823	1.3213	3.6787	0.26427	2.3675	2.6325	0.47351
22	206	438	1024	2186	5106	2.2361	1.4227	3.5773	0.28455	2.4965	2.5035	0.49929
23	166	346	736	1538	3270	2.1085	1.3716	3.6284	0.27433	2.5460	2.4540	0.50919
24	168	354	760	1614	3456	2.1358	1.3716	3.6284	0.27433	2.5931	2.4069	0.51863
25	162	328	702	1426	3056	2.0886	1.3716	3.6284	0.27433	2.6556	2.3444	0.53113
26	192	410	942	2018	4642	2.2216	1.3213	3.6787	0.26427	2.3439	2.6561	0.46879
27	258	558	1404	3042	7650	2.3344	1.2575	3.7425	0.25151	1.5704	3.4296	0.31407
28	154	304	632	1252	2602	2.0314	1.3716	3.6284	0.27433	2.5695	2.4305	0.51391
29	226	502	1180	2626	6174	2.2882	1.3870	3.6130	0.27739	2.3183	2.6817	0.46367
30	242	552	1310	3038	7156	2.3433	1.3245	3.6755	0.26489	2.0416	2.9584	0.40832
31	158	322	668	1368	2834	2.0615	1.3716	3.6284	0.27433	2.5306	2.4694	0.50613
32	150	288	596	1154	2374	2.0000	1.3716	3.6284	0.27433	2.4056	2.5944	0.48113
33	200	436	986	2174	4916	2.2410	1.4227	3.5773	0.28455	2.5590	2.4410	0.51179
34	178	364	816	1680	3784	2.1679	1.3213	3.6787	0.26427	2.4300	2.5700	0.48601
35	178	386	838	1818	3946	2.1701	1.3716	3.6284	0.27433	2.2806	2.7194	0.45613
36	150	306	642	1314	2760	2.0743	1.3009	3.6991	0.26017	2.3183	2.6817	0.46367
37	166	348	742	1568	3342	2.1268	1.3716	3.6284	0.27433	2.5306	2.4694	0.50613
38	228	522	1208	2778	6424	2.3073	1.3870	3.6130	0.27739	2.3183	2.6817	0.46367
39	148	302	624	1280	2648	2.0642	1.3009	3.6991	0.26017	2.4280	2.5720	0.48560
40	188	404	908	1962	4418	2.2120	1.3213	3.6787	0.26427	2.4064	2.5936	0.48129
41	158	324	674	1394	2904	2.0886	1.3716	3.6284	0.27433	2.6320	2.3680	0.52641
42	142	282	574	1150	2344	2.0285	1.3009	3.6991	0.26017	2.5141	2.4859	0.50282
43	142	280	572	1134	2324	2.0237	1.3009	3.6991	0.26017	2.5141	2.4859	0.50282
44	146	296	604	1230	2516	2.0491	1.3009	3.6991	0.26017	2.3655	2.6345	0.47310
45	200	444	1010	2246	5110	2.2504	1.3213	3.6787	0.26427	2.2189	2.7811	0.44379
46	136	260	520	1000	2000	1.9696	1.3009	3.6991	0.26017	2.4516	2.5484	0.49032
47	140	276	554	1098	2208	2.0066	1.3009	3.6991	0.26017	2.4516	2.5484	0.49032
48	252	586	1402	3286	7826	2.3649	1.3245	3.6755	0.26489	2.0294	2.9706	0.40588
49	250	584	1388	3266	7734	2.3623	1.3245	3.6755	0.26489	1.9669	3.0331	0.39338
50	122	232	444	848	1626	1.9190	1.1216	3.8784	0.22433	1.9056	3.0944	0.38113

Table 1, continued

$n = 3$ descriptors: ${}^3\chi^v$, twc, $mwc^{(5)}$,

$$\begin{aligned} f &= 16.793X_0 + 0.0085894X_1 - 0.69764X_2 + 246.86 \\ &= 7.7409X_0^* + 53.768X_1^* - 59.883X_2^* + 157.85. \end{aligned}$$

$n = 4$ descriptors: ${}^3\chi^v$, $mwc^{(6)}$, $mwc^{(7)}$, $mwc^{(8)}$,

$$\begin{aligned} f &= 10.930X_0 - 0.32884X_1 - 0.042581X_2 + 0.064274X_3 + 229.69 \\ &= 5.0382X_0^* - 79.236X_1^* - 25.319X_2^* + 100.05X_3^* + 157.85. \end{aligned}$$

$n = 5$ descriptors: W, ${}^3\chi^v$, twc, $mwc^{(4)}$, $mwc^{(8)}$,

$$\begin{aligned} f &= 0.44512X_0 + 9.7937X_1 - 0.0038957X_2 - 0.95038X_3 + 0.03649X_4 + 164.25 \\ &= 5.6464X_0^* + 4.5145X_1^* - 24.386X_2^* - 31.468X_3^* + 56.794X_4^* + 157.85. \end{aligned}$$

Table 3 shows statistical characteristics R^2, R^2_{CV}, S, S_{CV} and F, as well as differences between values obtained by resubstitution and leave–one–out crossvalidation (LOO–CV) of the best linear models with $n = 1, ..., 18$ topological indices. For R^2, R^2_{CV} and F the maximum values are underlined, in the other columns the minimum values are marked.

R^2 necessarily grows with increasing number of descriptors n, thus R^2 is not suited for the selection of a particular model. R^2_{CV} achieves its maximum for $n = 12$. However, 12 descriptors certainly are too many for 50 observations. Such a model would surely be overfitted. For the same reason also the model with minimum S including $n = 14$ descriptors should not be chosen for prediction.

A reasonable choice could be the model with $n = 6$ descriptors, supported by the argument that S_{CV} reaches its minimum. In [70] the difference $S_{CV} - S$ is mentioned as a measurement for the stability of a QSPR. This reasoning would suggest the model with $n = 4$ descriptors. This choice would be supported by the minimum difference between R^2 and R^2_{CV}. But also with $n = 3$ descriptors good characteristics are obtained. Especially F is maximal for this model.

Figure 5 shows experimental and calculated BP for this model. Additionally predictions obtained by LOO–CV are included. The good correlation between experimental and calculated values can even be recognized visually, and especially the high consistence of predictions obtained by resubstitution and crosssvalidation.

Altogether there are 75 constitutional isomers with molecular formula $C_{10}H_{22}$. These can be generated using MOLGEN within fractions of a second. Applying our canonical form the 50 structures of the real library can be identified automatically. We call the remaining 25 isomers the purely virtual

	BP	${}^2\chi^v$	${}^1\chi^v$	IC_1	CIC_1	SIC_1	${}^0\chi^v$	${}^3\chi^v$	$mwc^{(2)}$	$mwc^{(4)}$	W	MTI
BP	1.000	0.679	0.587	0.513	0.513	0.513	0.485	0.478	0.447	0.290	0.254	0.237
${}^2\chi^v$	0.679	1.000	0.975	0.297	0.297	0.297	0.892	0.054	0.896	0.768	0.586	0.558
${}^1\chi^v$	0.587	0.975	1.000	0.302	0.302	0.302	0.970	0.163	0.964	0.876	0.732	0.708
IC_1	0.513	0.297	0.302	1.000	1.000	1.000	0.310	0.042	0.272	0.222	0.283	0.281
CIC_1	0.513	0.297	0.302	1.000	1.000	1.000	0.310	0.042	0.272	0.222	0.283	0.281
SIC_1	0.513	0.297	0.302	1.000	1.000	1.000	0.310	0.042	0.272	0.222	0.283	0.281
${}^0\chi^v$	0.485	0.892	0.970	0.310	0.310	0.310	1.000	0.371	0.986	0.951	0.867	0.850
${}^3\chi^v$	0.478	0.054	0.163	0.042	0.042	0.042	0.371	1.000	0.368	0.539	0.641	0.654
$mwc^{(2)}$	0.447	0.896	0.964	0.272	0.272	0.272	0.986	0.368	1.000	0.970	0.862	0.844
$mwc^{(4)}$	0.290	0.768	0.876	0.222	0.222	0.222	0.951	0.539	0.970	1.000	0.943	0.931
W	0.254	0.586	0.732	0.283	0.283	0.283	0.867	0.641	0.862	0.943	1.000	0.999
MTI	0.237	0.558	0.708	0.281	0.281	0.281	0.850	0.654	0.844	0.931	0.999	1.000
$mwc^{(6)}$	0.202	0.710	0.831	0.180	0.180	0.180	0.921	0.602	0.945	0.995	0.948	0.939
λ_1^A	0.196	0.628	0.762	0.245	0.245	0.245	0.875	0.644	0.898	0.969	0.969	0.964
$mwc^{(3)}$	0.175	0.680	0.818	0.195	0.195	0.195	0.922	0.665	0.932	0.986	0.954	0.945
$mwc^{(5)}$	0.142	0.655	0.794	0.172	0.172	0.172	0.902	0.675	0.919	0.983	0.955	0.947
J	0.141	0.553	0.707	0.195	0.195	0.195	0.848	0.696	0.853	0.949	0.990	0.989
$mwc^{(8)}$	0.140	0.675	0.803	0.146	0.146	0.146	0.899	0.635	0.926	0.987	0.940	0.932
$mwc^{(7)}$	0.105	0.637	0.777	0.144	0.144	0.144	0.885	0.684	0.908	0.977	0.945	0.938
twc	0.097	0.642	0.779	0.131	0.131	0.131	0.883	0.674	0.909	0.978	0.937	0.930
IC_2	0.002	0.459	0.500	0.594	0.594	0.594	0.511	0.260	0.540	0.551	0.435	0.423
CIC_2	0.002	0.459	0.500	0.594	0.594	0.594	0.511	0.260	0.540	0.551	0.435	0.423
SIC_2	0.002	0.459	0.500	0.594	0.594	0.594	0.511	0.260	0.540	0.551	0.435	0.423

Table 2: Part of the sign–suppressed correlation matrix for BP and topological indices for the real library of decanes

n	R^2	R^2_{CV}	$R^2 - R^2_{CV}$	S	S_{CV}	$S_{CV} - S$	F
1	0.46101	0.40131	0.059698	5.5019	5.7986	0.29669	41.06
2	0.89336	0.87999	0.013366	2.4732	2.6236	0.15042	196.87
3	0.93721	0.92689	0.010325	1.9183	2.0700	0.15172	$\underline{228.87}$
4	0.95011	0.94126	$\underline{0.008856}$	1.7287	1.8759	$\underline{0.14718}$	214.27
5	0.95814	0.94709	0.011048	1.6015	1.8005	0.19896	201.42
6	0.96339	0.95022	0.013176	1.5149	$\underline{1.7666}$	0.25173	188.62
7	0.96450	0.95043	0.014074	1.5095	1.7838	0.27431	163.02
8	0.96520	0.94761	0.017590	1.5127	1.8561	0.34331	142.13
9	0.96686	0.94794	0.018922	1.4944	1.8731	0.37868	129.67
10	0.97045	0.95468	0.015764	1.4292	1.7699	0.34062	128.07
11	0.97151	0.95542	0.016090	1.4216	1.7783	0.35671	117.81
12	0.97275	$\underline{0.95591}$	0.016840	1.4092	1.7924	0.38323	110.05
13	0.97304	0.95294	0.020097	1.4209	1.8772	0.45629	99.94
14	0.97424	0.95061	0.023625	$\underline{1.4087}$	1.9504	0.54171	94.53
15	0.97426	0.94917	0.025088	1.4287	2.0075	0.57889	85.79
16	0.97438	0.94563	0.028750	1.4468	2.1075	0.66078	78.43
17	0.97439	0.94191	0.032484	1.4688	2.2122	0.74343	71.63
18	$\underline{0.97439}$	0.94191	0.032484	1.4923	2.2476	0.75532	65.53

Table 3: Characteristics of the best linear models with n descriptors for BP of decanes

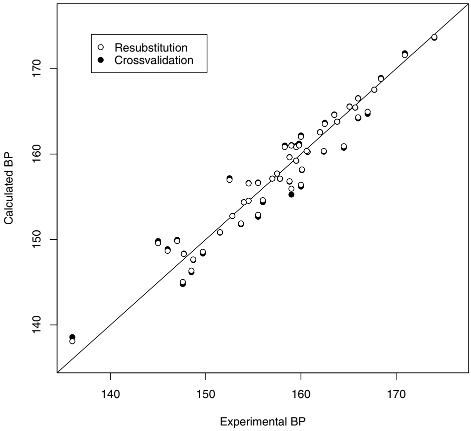


Figure 5: Experimental vs calculated BP (3-descriptor model)

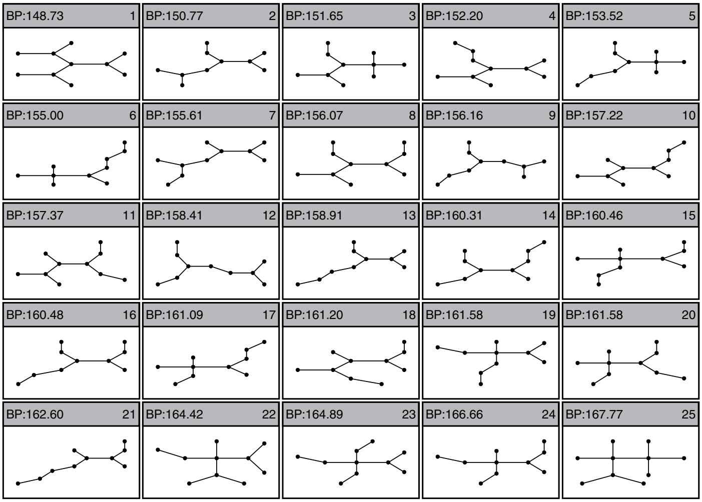


Figure 6: Purely virtual library of decanes with predicted BP

library. For these remaining compounds there existed no data about experimental BP in the *Beilstein* registry. In Figure 6 we give predictions for these decanes, calculated by the 3-descriptor model.

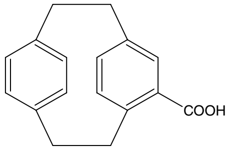
5 Generation of Stereoisomers

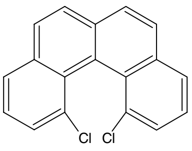
The first approach to computer-based generation of stereoisomers is due to Nourse et al ([27, 28, 3], 1979), based on the notion of *stereocenter*. The orientations of the four neighbors of each stereocenter describe the configuration of a molecule, and Nourse provided algorithms to *identify* all potential stereocenters and to systematically *change their orientation*, in order to *generate all possible configurational stereoisomers up to symmetry*. This method was implemented as CONGEN/STEREO.

In 1992, Zlatina and Elyashberg [71] provided an algorithm to obtain approximate 3D coordinates for the computed configurations, based on a given conformation of one isomer. The expert system RASTR for molecular elucidation contains these algorithms.

Wieland enriched Nourse's approach by group theoretic aspects (stabilizer chains were used for storing the automorphism group; orderly generation) and realized it in MOLGEN in 1994 [72, 73].

These implementations are very efficient and in many cases all stereoisomers are generated. However, depending entirely on the notion of the stereocenter, the approach has its limitations. First, striving not to miss any stereocenter, the algorithm is sometimes too generous in attributing the property of a stereocenter to an atom. This often results in many more stereocenters than a chemist would accept, and thus in excess stereoisomers. These have to be removed by special restrictions [3]. Second, the algorithm is unable to detect stereoisomerism that is not formally due to the presence of stereocenters. For example, chirality and thus the existence of two enantiomers is not detected in the [2.2]paracyclophanecarboxylic acid and the dichlorobenzophenanthrene shown here:



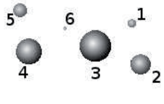


A more general approach, even allowing the generation of conformers, comes from the idea of Dreiding and Dress [29, 30], who used *chirotopes* (also known as oriented matroids) as a tool for describing conformations. Similar ideas are due to Klin, Tratch and Zefirov [74, 75], who used their approach especially in order to examine chirality of molecules [76] and to generate reaction types [77]. In Bayreuth, work is ongoing to develop a stereoisomer and conformer generator based on the chirotope approach [32].

The difference between the use of chirotopes and Nourse's approach is that not only the orientations of the four neighbor atoms of a stereocenter are considered, but orientations of potentially any four atoms can distinguish stereoisomers. Thus, we may consider the chirotope approach as a generalization of Nourse's approach.

As it is impossible to give a comprehensive overview on this topic within the available space, we give a very short introduction with a small example, and refer to a forthcoming article addressing the topic in more detail. For the mathematical background of chirotopes and oriented matroids, the book [78] can be recommended.

Consider a set of points in space. To any sequence of four points an orientation (positive, negative, or zero if the 4 points are coplanar) is assigned. Using the well-known right hand rule, the orientation may be determined even manually. Further, we can assign to any set of n numbered points an *orientation function* $\chi : n^4 \to \{0, \pm 1\}$ which denotes the orientation of each quadruple of points thereof. As χ is alternating, it suffices to specify the function values of all sorted quadruples. Using a suitable order on the set of all sorted quadruples, say the reverse lexical order, we can write an orientation function χ as the sequence of its function values. For example, here is an orientation function for 6 points:



$$\begin{array}{lccccccccccccccc} & 1234 & 1235 & 1245 & 1345 & 2345 & 1236 & 1246 & 1346 & 2346 & 1256 & 1356 & 2356 & 1456 & 2456 & 3456 \\ \chi = & + & + & 0 & - & - & + & + & 0 & - & + & + & 0 & + & + & + \end{array}$$

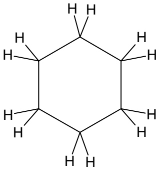
Orientation functions fulfill an oriented version of the base exchange axiom, the so called binary Grassmann-Plücker relations: For any two quadruples $\vec{a} = (a_0, a_1, a_2, a_3), \vec{b} = (b_0, b_1, b_2, b_3) \in n^4$, the following holds:

$$\begin{aligned} &\chi(\vec{a}) \cdot \chi(\vec{b}) = 1 \Longrightarrow \\ &\qquad \exists\ i \in \{0, \ldots, 3\} : \chi(b_i, a_1, a_2, a_3) \cdot \chi(b_0,\ \ldots, \underset{\substack{\uparrow \\ i\text{th position}}}{a_0}, \ldots, b_3) = 1\,. \quad \text{(GP)} \end{aligned}$$

In general, the alternating non-trivial (i.e. not constantly zero) functions $\chi : n^4 \to \{0, \pm 1\}$ fulfilling (GP) are called *chirotopes* (of rank 4). Thus, the orientation function of any sequence of points in 3D space (not all in one plane) is a chirotope.

Note that not every chirotope is an orientation function. We call a chirotope which is the orientation function of a set of points *affinely realizable.* The decision, whether a chirotope is affinely realizable or not, and to find a realization, is a problem shown to be NP-hard. Nevertheless, the more general theory of oriented matroids allows some necessary tests for affine realizability. Finally, we call the chirotope *uniform* if it has no zero function values.

A generator for chirotopes using the general generation techniques described above was developed by one of the authors [32]. It can serve as generator of conformations of molecular structures. We demonstrate this on the example of cyclohexane:



The molecule has 6 non-hydrogen atoms, so we generate chirotopes over 6 elements. In order to avoid doublets, we have to consider the automorphism group of the molecular graph, which is the dihedral group with 12 elements (this is equivalent to the symmetry group D_{6h} of an assumed plane cyclohexane). Using this as acting group on the set of chirotopes, all generated orbit representatives will lead to essentially different conformations of the molecule (provided the chirotope in question is affinely realizable). In order to reduce the complexity of the problem, we assume that no four atoms are coplanar, concentrating in this way to uniform chirotopes only. (Our assumption is not really a restriction, because we could move one atom a little bit out of the plane of three other atoms, if necessary.) This way we get 386 chirotopes. The first few of them are listed below.

1234	1235	1245	1345	2345	1236	1246	1346	2346	1256	1356	2356	1456	2456	3456
+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
+	+	+	+	+	+	+	+	+	+	+	+	+	+	−
+	+	+	+	+	+	+	+	+	+	+	+	+	−	−
+	+	+	+	+	+	+	+	+	+	+	+	−	+	+
+	+	+	+	+	+	+	+	+	+	+	+	−	−	+

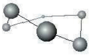
⋮

This amount is quite a lot for such a small example. By giving further restrictions to the generator which will be described in a forthcoming article, this number can be reduced. The main simplification is the following, also giving a lot of freedom in adjusting the level of detail in our investigations: As not each orientation of a quadruple of atoms is of same importance for conformational analysis, we can select a few relevant quadruples and identify all chirotopes that do not differ on the selected quadruples. This way, we get classes of chirotopes, identified by the orientations on the selected set of quadruples, i.e. by a *partially defined chirotope.* If we choose for example to consider only the orientations of quadruples of atoms forming a chain, i.e. if we analyse the conformation of all butane substructures in cyclohexane only, we can reduce the set of generated structures to 13 partially defined chirotopes:

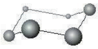
1234	1235	1245	1345	2345	1236	1246	1346	2346	1256	1356	2356	1456	2456	3456
+				+	+				+			+		+
+				+	+				+			+		−
+				+	+				+			−		+
+				+	+				+			−		−
+				+	+				−			+		+
+				+	+				−			+		−
+				+	+				−			−		+

1234	1235	1245	1345	2345	1236	1246	1346	2346	1256	1356	2356	1456	2456	3456
+				+	−				+			−		+
+				−	+				+			+		+
+				−	+				+			+		−
+				−	+				−			+		−
+				−	+				−			−		−
−				−	+				−			+		−

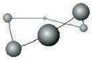
So far, we did not consider coordinates at all, and all our computations used discrete mathematics only. The remaining part is to try to find for each of the generated (partially defined) chirotopes a conformation of cyclohexane fulfilling the prescribed orientations. This was done by restricted optimization of an energy function. We used a very simple energy function similar to MM2, and the prescribed orientations were formulated as restrictions to the optimizer. This way, we found conformations for 7 of the 13 generated chirotopes. For only 3 of these the optimization process found a local minimum. The other conformations could have been optimized further, but not without injuring one of the prescribed orientations, and so we ignored them. The remaining 3 conformations were exactly what we expected: The chair form and two enantiomeric twist forms.



++++−−



+−++++



+−+−−−

Note that restricted optimization is not guaranteed to find an optimum, even if it exists. As already mentioned, the exact decision on affine realizability of chirotopes and of finding a conformation is a very hard problem. Further research has to be done.

There is also the possibility to generate chirotopes up to *negation*, leading to a generation of conformations where enantiomers are considered equal. In the example of cyclohexane, this way we get two conformations, the chair and a twist form.

6 Problems

6.1 Aromaticity

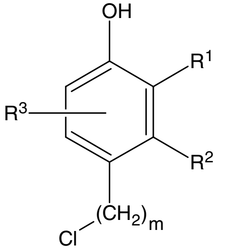
Although powerful generators of molecular formulae have been developed, there remain serious problems. For example the phenomenon of aromaticity shows that in aromatic rings it is not pairs of atoms but all atoms in the whole ring that interact. So we possibly should go from graphs to hypergraphs, which may also be necessary in order to cover compounds such as metal complexes. In hypergraphs a hyperedge consists of a subset of the set of vertices which does not need to be a 2–element subset. An aromatic ring can be considered as such a hyperedge. This new hypergraph approach is described in [79, 80]. It remains to answer the question: Which subsets can interact? Another problem of this approach is that it increases the complexity of calculations. At least in the construction of t–designs some experience has been gathered on constructing systems of subsets, see also [81].

6.2 Patents in Chemistry

What should be done right now is the following (which needs an extension of the present generators of molecular structures to recursively define molecules):

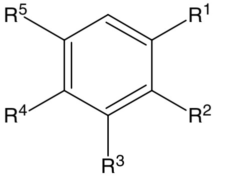
> Generate patent libraries, correponding to Markush formulae, in such a way that compounds in the libraries are generated in a canonical form, so that two libraries can be searched for overlap.

For example, the library of [82]



R^1 : $\mathsf{CH_3}$, $\mathsf{C_2H_5}$
R^2 : alkyl (1–6 C atoms)
R^3 : $\mathsf{NH_2}$
m : 1–3

should be compared with, say,



R^1 : $\mathsf{CH_3}$, $\mathsf{C_2H_5}$, OH
R^2 : alkyl (1–6 C atoms)
R^3 : OH, $\mathsf{OCH_3}$, $\mathsf{OC_2H_5}$, $\mathsf{CH_3}$, $\mathsf{C_2H_5}$
R^4 : OH, $\mathsf{CH_2Cl}$, $\mathsf{NH_2}$
R^5 : H, $\mathsf{CH_3}$, $\mathsf{C_2H_5}$, $\mathsf{NH_2}$

MOLGEN generates 33 alkyl residues for R^2. These 33 structures are stored in a separate library for R^2 that is part of the input for MOLGEN–COMB. MOLGEN–COMB generates libraries of sizes

$$396 \text{ and } 5939,$$

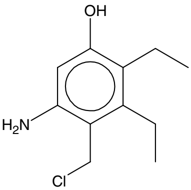
respectively. Note that one size is 5939 and *not* 5940, as would naively be expected. Due to symmetry of the benzene skeleton, the compounds with

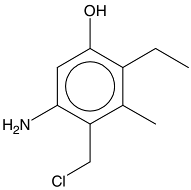
$$\mathsf{R}^1 = \mathsf{OH}, \mathsf{R}^2 = \mathsf{C_2H_5}, \mathsf{R}^3 = \mathsf{CH_3}, \mathsf{R}^4 = \mathsf{OH}, \mathsf{R}^5 = \mathsf{H}$$

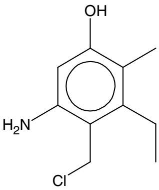
and

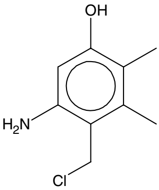
$$\mathsf{R}^1 = \mathsf{OH}, \mathsf{R}^2 = \mathsf{CH_3}, \mathsf{R}^3 = \mathsf{C_2H_5}, \mathsf{R}^4 = \mathsf{OH}, \mathsf{R}^5 = \mathsf{H}$$

are identical, as is easily found by the program. Moreover, since the files of these libraries are in canonical form we get immediately the overlap:









As a rule, Markush formulae appearing in chemistry patents are much more complicated, containing variable groups on several nested levels. Therefore real life problems in this field are much more difficult to solve.

References

[1] E. O. von Lippmann. *Alexander von Humboldt als Vorläufer der Lehre von der Isomerie.* Chemiker–Zeitung, 1:1–2, 1909.

[2] A. von Humboldt. *Versuche über die gereizte Muskel- und Nervenfaser, nebst Vermutungen über den chemischen Prozeß des Lebens in der Tier- und Pflanzenwelt.* 1797.

[3] J. G. Nourse, D. H. Smith, R. E. Carhart, and C. Djerassi. *Computer-Assisted Elucidation of Molecular Structure with Stereochemistry.* J. Am. Chem. Soc., 102:6289–6295, 1980.

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

[5] G. Pólya and R. C. Read. *Combinatorial Enumeration of Groups, Graphs and Chemical Compounds.* Springer, 1998.

[6] A. T. Balaban. *Enumeration of isomers.* In D. Bonchev and D. H. Rouvray, editors, *Chemical Graph Theory. Introduction and Fundamentals*, pages 177–234. Abacus Press – Gordon and Breach, New York, 1991.

[7] A. C. Lunn and J. K. Senior. *Isomerism and Configuration.* J. Phys. Chem., 33:1027–1079, 1929.

[8] J. H. Redfield. *The theory of group–reduced distributions.* Amer. J. Math., 49:433–455, 1927.

[9] M. van Almsick, H. Dolhaine, and H. Hönig. *Efficient Algorithms to Enumerate Isomers and Diamutamers with More Than One Type of Substituent.* J. Chem. Inf. Comput. Sci., 40:956–966, 2000.

[10] A. Kerber. *Applied Finite Group Actions.* Springer, Berlin, Heidelberg, New York, 2. edition, 1999.

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

[12] C. Rücker, R. Gugisch, and A. Kerber. *Manual Construction and Mathematics– and Computer–Aided Counting of Stereoisomers. The Example of Oligoinositols.* J. Chem. Inf. Comput. Sci., 44:1654–1665, 2004.

[13] R. Laue. *Eine konstruktive Version des Lemmas von Burnside.* Bayreuther Mathematische Schriften, 28:111–125, 1989.

[14] 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.

[15] A. Kerber, R. Laue, M. Meringer, and C. Rücker. *Molecules in Silico: The Generation of Structural Formulae and its Applications.* J. Comput. Chem. Jpn., 3:85–96, 2004.

[16] A. Kerber, M. Meringer, and C. Rücker. *CASE via MS: Ranking Structure Candidates by Mass Spectra.* Croat. Chem. Acta. In press.

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

[18] 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.

[19] M. E. Elyashberg, K. A. Blinov, A. J. Williams, S. G. Molodtsov, G. E. Martin, and E. R. Martirosian. *Structure Elucidator: A Versatile Expert System for Molecular Structure Elucidation from 1D and 2D NMR Data and Molecular Fragments.* J. Chem. Inf. Comput. Sci., 44:771–792, 2004.

[20] S. G. Molodtsov, M. E. Elyashberg, K. A. Blinov, A. J. Williams, E. R. Martirosian, G. E. Martin, and B. Lefebvre. *Structure Elucidator from 2D NMR Spectra Using the StruEluc Expert System: Detection and Removal of Contradictions in the Data.* J. Chem. Inf. Comput. Sci., 44:1737–1751, 2004.

[21] S. G. Molodtsov. *The Generation of molecular graphs with obligatory, forbidden and desirable fragments.* MATCH Commun. Math. Comput. Chem., 37:157–162, 1988.

[22] T. Grüner, R. Laue, and M. Meringer. *Algorithms for Group Actions: Homomorphism Principle and Orderly Generation Applied to Graphs*, volume 28 of *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, pages 113–122. American Mathematical Society, 1996.

[23] T. Grüner. *Strategien zur Konstruktion diskreter Strukturen.* PhD thesis, Universität Bayreuth, 1998.

[24] R. Laue, T. Grüner, M. Meringer, and A. Kerber. *Constrained Generation of Molecular Graphs*, volume 69 of *DIMACS Series in Discrete Mathematics And Theoretical Computer Science*, pages 319–332. American Mathematical Society, 2005.

[25] J. Braun, R. Gugisch, A. Kerber, R. Laue, M. Meringer, and C. Rücker. *MOLGEN–CID, A Canonizer for Molecules and Graphs Accessible through the Internet.* J. Chem. Inf. Comput. Sci., 44:542–548, 2004.

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

[27] J. G. Nourse. *The Configuration Symmetry Group and Its Application to Stereoisomer Generation, Specification, and Enumeration.* J. Am. Chem. Soc., 101:1210–1215, 1979.

[28] J. G. Nourse, R. E. Carhart, D. H. Smith, and C. Djerassi. *Exhaustive Generation of Stereoisomers for Structure Elucidation.* J. Am. Chem. Soc., 101:1216–1223, 1979.

[29] A. Dreiding and K. Wirth. *The multiplex. A classification of finite ordered point sets in oriented d-dimensional space.* MATCH Commun. Math. Comput. Chem., 8:341–352, 1980.

[30] A. Dreiding, A. Dress, and H. Haegi. *Classification of mobile molecules by category theory.* Studies in Phys. and Theor. Chem., 8:341–352, 1982.

[31] A. Dress. *Chirotops and Oriented Matroids.* Bayreuther Mathematische Schriften, 21:14–68, 1986.

[32] R. Gugisch. *Konstruktion von Isomorphieklassen Orientierter Matroide.* Bayreuther Mathematische Schriften, 72:1–124, 2005.

[33] E. Ruch, W. Hässelbarth, and B. Richter. *Doppelnebenklassen als Klassenbegriff und Nomenklaturprinzip für Isomere und ihre Abzählung.* Theor. Chim. Acta, 19:288–300, 1970.

[34] E. Ruch and D. J. Klein. *Double Cosets in Chemistry and Physics.* Theor. Chim. Acta, 63:447–472, 1983.

[35] A. Kerber and R. Laue. *Group Actions, Double Cosets, and Homomorphisms: Unifying Concepts for the Constructive Theory of Discrete Structures.* Acta Applicandae Mathematicae, 52:63–90, 1998.

[36] B. Schmalz. *Verwendung von Untergruppenleitern zur Bestimmung von Doppelnebenklassen.* Bayreuther Mathematische Schriften, 31:109–143, 1993.

[37] T. Carell, E. A. Wintner, A. J. Sutherland, J. Rebek Jr., and Y. M. Dunayevskiy. *New Promise in Combinatorial Chemistry: Synthesis, Characterization, and Screening of Small–Molecule Libraries in Solution.* Chem. & Biol., 2:171–183, 1995.

[38] A. Kerber, R. Laue, and T. Wieland. *Discrete Mathematics for Combinatorial Chemistry*, volume 51 of *DIMACS Series in Discrete Mathematics And Theoretical Computer Science*, pages 225–234. American Mathematical Society, 2000.

[39] A. Kerber, R. Laue, M. Meringer, and C. Rücker. *Molecules in Silico: A Graph Description of Chemical Reactions.* Adv. Quantum Chem. In press.

[40] 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.

[41] R. Ihaka and R. Gentleman. *R: A Language for Data Analysis and Graphics.* J. Comput. Graph. Stat., 5:299–314, 1996.

[42] C. Rücker, M. Meringer, and A. Kerber. *QSPR Using MOLGEN–QSPR: The Example of Haloalkane Boiling Points.* J. Chem. Inf. Comput. Sci., 44:2070–2076, 2004.

[43] C. Rücker, M. Meringer, and A. Kerber. *QSPR Using MOLGEN–QSPR: The Challenge of Fluoroalkane Boiling Points.* J. Chem. Inf. Model., 45:74–80, 2005.

[44] J. Braun, A. Kerber, M. Meringer, and C. Rücker. *Similarity of Molecular Descriptors: The Equivalence of Zagreb Indices and Walk Counts.* MATCH Commun. Math. Comput. Chem., 54:163–176, 2005.

[45] H. Wiener. *Structural Determination of Paraffin Boiling Points.* J. Am. Chem. Soc., 69:17–20, 1947.

[46] I. Gutman, B. Ruščić, N. Trinajstić, and C. F. Wilcox Jr. *Graph Theory and Molecular Orbitals. XII. Acyclic Polyenes.* J. Chem. Phys., 62:3399–3405, 1975.

[47] M. Randić. *On Characterization of Molecular Branching.* J. Am. Chem. Soc., 97:6609–6615, 1975.

[48] L. B. Kier, W. J. Murray, M. Randić, and L. H. Hall. *Molecular Connectivity V: Connectivity Series Applied to Density.* J. Pharm. Sci., 65:1226–1230, 1976.

[49] A. T. Balaban. *Highly Discriminating Distance–Based Topological Index.* Chem. Phys. Lett., 89:399–404, 1982.

[50] A. T. Balaban. *Topological Indices Based on Topological Distances in Molecular Graphs.* Pure Appl. Chem., 55:199–206, 1983.

[51] H. P. Schultz. *Topological Organic Chemistry. 1. Graph Theory and Topological Indices of Alkanes.* J. Chem. Inf. Comput. Sci., 29:227–228, 1989.

[52] H. P. Schultz and T. P. Schultz. *Topological Organic Chemistry. 6. Graph Theory and Molecular Topological Indices of Cycloalkanes.* J. Chem. Inf. Comput. Sci., 33:240–244, 1993.

[53] G. Rücker and C. Rücker. *Counts of All Walks as Atomic and Molecular Descriptors.* J. Chem. Inf. Comput. Sci., 33:683–695, 1993.

[54] G. Rücker and C. Rücker. *Walk Counts, Labyrinthicity, and Complexity of Acyclic and Cyclic Graphs and Molecules.* J. Chem. Inf. Comput. Sci., 40:99–106, 2000.

[55] I. Gutman, C. Rücker, and G. Rücker. *On Walks in Molecular Graphs.* J. Chem. Inf. Comput. Sci., 41:739–745, 2001.

[56] L. B. Kier and Hall L. H. *The Nature of Structure–Activity Relationships and their Relation to Molecular Connectivity.* Eur. J. Med. Chem., 12:307–312, 1977.

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

[58] S. C. Basak. *Use of Molecular Complexity Indices in Predictive Pharmacology and Toxicology: A QSAR Approach.* Med. Sci. Res., 15:605–609, 1987.

[59] S. C. Basak. *Information Theoretic Indices of Neighborhood Complexity and their Applications.* In J. Devillers and A. T. Balaban, editors, *Topological Indices and Related Descripors in QSAR and QSPR*, chapter 12. Gordon and Breach, Amsterdam, 1999.

[60] R. Todeschini and V. Consonni. *Handbook of Molecular Descriptors.* Wiley–VCH, Weinheim, 2000.

[61] M. Karelson. *Molecular Descriptors in QSAR/QSPR.* Wiley-Interscience, New York, 2000.

[62] J. Devillers and A. T. Balaban, editors. *Topological Indices and Related Descripors in QSAR and QSPR.* Gordon and Breach, Amsterdam, 1999.

[63] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees.* Wadsworth International Group, Belmont, California, 1984.

[64] B. D. Ripley. *Pattern Recognition and Neural Networks.* Cambridge University Press, Cambridge, 1996.

[65] J. Zupan and J. Gasteiger. *Neural Networks for Chemists.* VCH Verlagsgesellschaft, Weinheim, New York, Basel, Cambridge, Tokyo, 1993.

[66] V. Vapnik. *The Nature of Statistical Learning Theory.* Springer, New York, Berlin, Heidelberg, 1995.

[67] H. Martens and T. Næs. *Multivariate Calibration.* Wiley, Chichester, 1989.

[68] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning.* Springer, New York, Berlin, Heidelberg, 2001.

[69] M. Meringer. *Mathematische Modelle für die kombinatorische Chemie und die molekulare Strukturaufklärung.* Logos–Verlag Berlin, 2004.

[70] S. Nikolić, G. Kovačević, A. Miličević, and N. Trinajstić. *The Zagreb Indices 30 Years After.* Croat. Chem. Acta, 76:113–127, 2003.

[71] L. A. Zlatina and M. E. Elyashberg. *Generation of stereoisomers and their spatial models corresponding to the given molecular structure.* MATCH Commun. Math. Comput. Chem., 27:191–207, 1992.

[72] T. Wieland. *Erzeugung, Abzählung und Konstruktion von Stereoisomeren.* MATCH Commun. Math. Comput. Chem., 31:153–203, 1994.

[73] 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.

[74] S. S. Tratch and N. S. Zefirov. *Combinatorial Models and Algorithms in Chemistry. The Ladder of Combinatorial Objects and its Application to the Formalization of Structural Problems of Organic Chemistry.* In N.F. Stepanov, editor, *Principles of Symmetry and Systemology in Chemistry*, pages 54–86. Moscow, Moscow State University Publ., 1987. (in Russian).

[75] M. H. Klin, S. S. Tratch, and N. S. Zefirov. *2D-configurations and Clique-cyclic Orientations of the Graphs $L(K_p)$.* Rep. Mol. Theory, 1:149–163, 1990.

[76] S. S. Tratch and N. S. Zefirov. *Algebraic Chirality Criteria and Their Application to Chirality Classification in Rigid Molecular Systems.* J. Chem. Inf. Comput. Sci., 36:448–464, 1996.

[77] S. S. Tratch and N. S. Zefirov. *Systematic Search for New Types of Chemical Interconversions: Mathematical Models and Some Applications.* J. Chem. Inf. Comput. Sci., 38:331–348, 1998.

[78] A. Björner, M. Las Vergnas, B. Sturmfels, N. White, and G. M. Ziegler. *Oriented Matroids.* Cambridge University Press, Cambridge, 1993.

[79] E. V. Konstantinova and V. A. Skorobogatov. *Molecular Hypergraphs: The New Representation of Nonclassical Molecular Structures with Polycentric Delocalized Bonds.* J. Chem. Inf. Comput. Sci., 35:472–478, 1995.

[80] E. V. Konstantinova and V. A. Skorobogatov. *Application of Hypergraph Theory in Chemistry.* Discrete Math., 235:365–383, 2001.

[81] M. Wang. *Canonical Forms of Discrete Objects for Databases and Internet Data Exchange.* Bayreuther Mathematische Schriften, 75:1–118, 2006.

[82] 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.