Molecules in Silico: A Graph Description of Chemical Reactions

Adalbert Kerber[*,†], Reinhard Laue[†],
Markus Meringer[‡], Christoph Rücker[§]

[†]Department of Mathematics, University of Bayreuth, 95440 Bayreuth, Germany

[‡]Kiadis B.V., Zernikepark 6–8, 9747 AN Groningen, The Netherlands

[§]Biocenter, University of Basel, Klingelbergstrasse 70, 4056 Basel, Switzerland

June 27, 2005

Abstract

A general mathematical description — mostly in terms of graph theory — is given for reactions of organic chemistry. The corresponding computer program MOLGEN–QSPR[1], given a set of starting materials and a set of reaction schemes, generates all products that can result from starting materials (and intermediates) interacting according to the specified reactions. Example reactions from combinatorial chemistry, synthetic organic chemistry, and mass spectroscopic structure elucidation are considered in detail.

1 Introduction

Mostly building on the Dugundji–Ugi theory [1] several authors devised systems for representing organic reactions in silico, focussing on various purposes. Databases for reaction retrieval such as Beilstein Crossfire contain reactions that are pairs of specific starting materials and specific products. Synthesis planning programs working in the retrosynthetic direction are well-known, for a review see [2]. An early program generated intermediates to link given starting materials and products [3]. Of more recent origin are programs that use kinetic information, obtained directly from experiments [4], or via a QSPR approach [5] or from quantum-chemical calculations [6], for semiquantitative prediction of reaction mixtures. There are attempts to classify known general reactions and even to predict novel reaction types [7].

The aim of our approach is to generate, from a given set of starting materials and a given set of general reactions, all possible products, including those derived from intermediates interacting according to the prescribed reactions. The purpose of the present paper is to detail the mathematical concepts that underlie our reaction module incorporated in MOLGEN-COMB [8] and MOLGEN-QSPR [9, 10, 11].

[*]Corresponding author e–mail: kerber@uni-bayreuth.de

[1]see http://www.mathe2.uni-bayreuth.de

2 Molecular Graphs

The graph model for a description of chemical reactions, following the approach of Fujita [12] and Temkin et al. [13], is based on our formulation [14, 15] of molecular graphs. These are multigraphs consisting of vertices representing atoms and edges representing *covalent bonds*. The bonds may be single, double or triple bonds. The vertices are coloured by the symbols of *chemical elements* and of *atomic states* defined as follows:

2.1 Definition (atomic state) An atomic state is a quadruple

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

where

- the positive integer v_s identifies the valence of the atom i.e. the number of covalent bonds in which the atom is involved, with a double and triple bond contributing 2 and 3, respectively,
- the non-negative integer p_s indicates the number of free electron pairs (lone pairs),
- the integer q_s denotes the charge associated to the atom,
- while $r_s \in \mathbb{B} := \{true = 1, false = 0\}$ shows whether or not the atom bears an unpaired electron.

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

For each chemical element X we introduce the set $\mathcal{Z}_X$ of *admissible atomic states.* Its definition clearly depends on the particular chemistry under investigation. For example, the default atomic state for carbon is $(4, 0, 0, 0)$, which allows to construct most non-ionic organic compounds that are not free radicals. If one is interested in carbenium ions, carbanions, or free radicals, the corresponding atomic states should be admitted, $(3, 0, 1, 0)$, $(3, 1, -1, 0)$, and $(3, 0, 0, 1)$, respectively. If one is interested in carbenes or isonitriles, e.g. in the context of Ugi multicomponent reactions, the atomic state $(2, 1, 0, 0)$ should be allowed. Note that a particular atomic state may include various bond patterns. Thus, $(4, 0, 0, 0)$ is the state of a neutral carbon atom involved in four single bonds, or in one double and two single bonds, or in two double bonds, or in a single and a triple bond, i.e. any saturated, olefinic, aromatic, central allenic, or acetylenic carbon atom, or a carbon in a functional group including carbonyl or nitrile.
The elements are gathered in sets $\mathcal{E}$ such as

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

or its extension

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

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

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

A *molecular graph* describing a molecule of n atoms, numbered from 0 to $n-1$ and taken from $\mathcal{E}$, is a triple

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

where ε is a sequence of length n, consisting of element symbols, i.e.[2]

$$\varepsilon = (\varepsilon(0), \ldots, \varepsilon(n-1)) \in \mathcal{E}^n.$$

The second component ζ is a sequence $\zeta = (\zeta(0), \ldots, \zeta(n-1))$ of n atomic states, where the i–th component is an admissible state of atom i,

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

The third component γ is a connected multigraph consisting of n vertices and edges that are at most 3–fold, i.e. elements of the set $4 = \{0, 1, 2, 3\}$, for short[3],

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

Its vertices are numbered from 0 to $n-1$ and colored by the element symbols $\varepsilon(i)$, the components of ε. The degree of the i–th vertex of the graph is equal to the valence of atom i

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

By $\mathcal{M}^c_n$ we denote the set of connected molecular graphs on n atoms, by $\mathcal{M}^c$ the set of all connected molecular graphs.
Summarising, we can abbreviate our main definition as follows: A molecular graph, modelling a molecule consisting of n atoms taken from the set of elements $\mathcal{E}$, and with their states in $\mathcal{Z}_\mathcal{E}$, is a triple

$$(\varepsilon, \zeta, \gamma) \in (\mathcal{E}^n \times \mathcal{Z}^n_\mathcal{E} \times \mathcal{G}^c_{n,4})$$

that fulfills equations 1 and 2. ◇

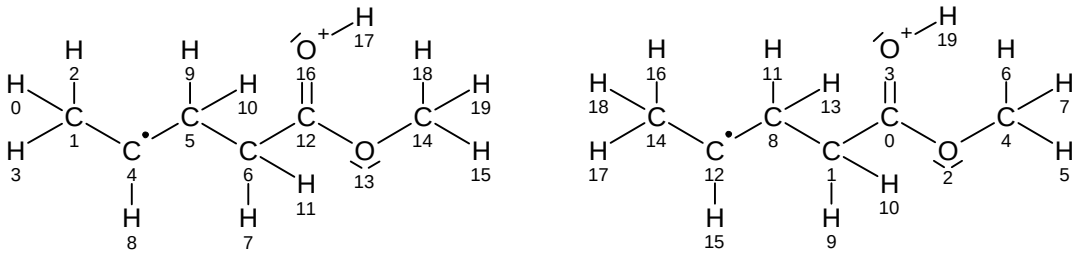
[2]We use in this notation that the natural number n is recursively defined by $n = \{0, \ldots, n-1\}$ and that Y^X is a standard notation for the set of mappings from the set X to the set Y, $Y^X = \{f \colon X \to Y\}$. Here, in order to define molecules consisting of n atoms in $\mathcal{E}$, we take $X = n = \{0, \ldots, n-1\}$, $Y = \mathcal{E}$, and $Y^X = \mathcal{E}^n$.

[3]The index n of $\mathcal{G}^c_{n,4}$ is the set of (numbers of) atoms in the molecule, while the second index $4 = \{0, 1, 2, 3\}$ is the set containing all possible multiplicities of covalent bonds. Moreover, the set $\mathcal{G}_{n,m}$ of multigraphs with vertex set n and set of multiplicities m can also be considered as a set of mappings Y^X. We take for X the set $\binom{n}{2}$ of subsets $\{i, j\} \subseteq n$, the set of pairs of vertices, and for Y the set m of admissible multiplicities, since, for $\gamma \in \mathcal{G}_{n,m}$ we can interprete $\gamma(\{i, j\}) = k$ as the existence of a k-fold bond between vertices i and j. Thus

$$\mathcal{G}_{n,m} = m^{\binom{n}{2}}.$$

$\mathcal{G}^c_{n,m}$ means the subset of $\mathcal{G}_{n,m}$ consisting of connected multigraphs.

Here is an example of two molecular graphs modeling an ion obtained from methyl pentanoate:



In this drawing pecularities of atomic states are represented by several symbols: a plus '+' for a single positive charge, a bar '−' for a free electron pair, and a dot '·' for an unpaired electron.

Two such molecular graphs $(\varepsilon, \zeta, \gamma)$ and $(\varepsilon', \zeta', \gamma')$ describe the same molecule if and only if they are the same up to renumbering, which means that there is a permutation π such that

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

where

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

defined by[4]

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

In mathematical terms, we are faced with the following *action of the symmetric group* S_n:

$$\begin{array}{rcl} \left(\mathcal{E}^n \times \mathcal{Z}_\mathcal{E}^n \times \mathcal{G}_{n,4}^c\right) \times S_n & \longrightarrow & \mathcal{E}^n \times \mathcal{Z}_\mathcal{E}^n \times \mathcal{G}_{n,4}^c, \\ ((\varepsilon, \zeta, \gamma), \pi) & \longmapsto & (\varepsilon, \zeta, \gamma)^\pi. \end{array}$$

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

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

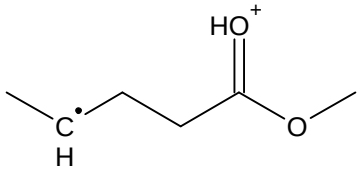
is the orbit of the molecular graph $(\varepsilon, \zeta, \gamma)$. The conditions 1 and 2 in the definition of molecular graphs are preserved by this operation. Hence, a structural formula of a molecule with n atoms from $\mathcal{E}$ corresponds to an orbit of S_n on the set $\left(\mathcal{E}^n \times \mathcal{Z}_\mathcal{E}^n \times \mathcal{G}_{n,4}^c\right)$, i.e. the set of structural formulae of molecules built from n atoms in $\mathcal{E}$ is the set of orbits of the symmetric group

$$S_n // \mathcal{M}_n^c = \{S_n((\varepsilon, \zeta, \gamma)) \mid (\varepsilon, \zeta, \gamma) \in \mathcal{M}_n^c\}.$$

Hence the problem of constructing all structural formulae, i.e. the structural isomers according to a certain molecular formula, amounts to finding a complete system of

[4]Recall that $\gamma(\{i,j\})$ denotes the multiplicity of the covalent bond that connects atoms i and j, i.e. $\gamma(\{i,j\}) \in \{0,1,2,3\} = 4$.

representatives of these orbits of the symmetric group. MOLGEN [16, 17] is a software package that solves this problem efficiently.
It is easily seen that the two molecular graphs shown above belong to the same orbit. Hence a representative of this orbit can be drawn as follows



where for the sake of simplicity we erased the atom numbering, the symbols for free electron pairs, hydrogen atoms adjacent to carbon atoms in the atomic ground state, as well as the symbols of such carbon atoms.

3 Molecular Substructures

It is an old experience in organic chemistry that a reaction takes place at a specific position rather than anywhere in a molecule. Such reactive sites are, e.g., the classical functional groups, or other particular structural elements called *substructures*. For this reason in the following a precise definition of the term substructure is given[5].

3.1 Definition (subgraph) Let $\gamma \in \mathcal{G}_{n,m}$ be a multigraph on the set n of vertices, the edge multiplicities less than m, and $V \subseteq n$ a non–empty subset of n. $\gamma' \in \mathcal{G}_{V,m}$ is a *subgraph* of γ, if

$$\forall\ e \in E_{\gamma'}: \quad \gamma'(e) \leq \gamma(e),$$

where $E_{\gamma'}$ denotes the set of edges of γ'. We write $\gamma' \subseteq \gamma$ to indicate this.
If the stronger condition

$$\forall\ e \in E_{\gamma'}: \quad \gamma'(e) = \gamma(e)$$

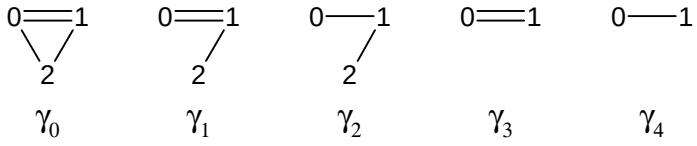
holds, γ' is called a *closed subgraph* of γ ($\gamma' \subseteq^c \gamma$).
If finally

$$\forall\ e \in \binom{V}{2}: \quad \gamma'(e) = \gamma(e),$$

we call γ' an *induced subgraph* of γ ($\gamma' \subseteq^i \gamma$). ◇

For $\gamma \in \mathcal{G}_{n,m}$ and $\emptyset \neq V \subseteq n$ the induced subgraph $\gamma' \in \mathcal{G}_{V,m}$ of γ is uniquely determined. Therefore we call γ' the *induced subgraph* of γ on V and we write $\gamma' = \gamma|_V$. The set of induced subgraphs $\gamma|_{V_0}, ..., \gamma|_{V_{k-1}}$ on the connectivity components $V_0, ..., V_{k-1}$ of γ is denoted by $\mathrm{Con}(\gamma)$.

3.2 Example We illustratively examine the following multigraphs for subgraph relations to γ_0:



[5]Substructure counts play an important role in QSPR [11].

We have $\gamma_1 \subseteq^c \gamma_0$, but γ_1 is not an induced subgraph of γ_0; $\gamma_2 \subseteq \gamma_0$, but γ_2 is not a closed subgraph of γ_0; $\gamma_3 \subseteq^i \gamma_0$; $\gamma_4 \subseteq \gamma_0$, but γ_4 is not a closed subgraph of γ_0. $\diamond$

3.3 Definition (embedding) Let $\gamma \in \mathcal{G}_{n,m}$, $V \subseteq n$ be a non–empty subset of n and $\gamma' \in \mathcal{G}_{V,m}$. Let furthermore n^V_{inj} denote the set of injective mappings from V to n. Such a mapping ϕ is called an *embedding* of γ' in γ

- as a subgraph, if

$$\forall\ \{i,j\} \in E_{\gamma'}: \quad \gamma'(\{i,j\}) \leq \gamma(\{\phi(i),\phi(j)\}),$$

- as a closed subgraph, if

$$\forall\ \{i,j\} \in E_{\gamma'}: \quad \gamma'(\{i,j\}) = \gamma(\{\phi(i),\phi(j)\}),$$

- as an induced subgraph, if

$$\forall\ \{i,j\} \in \binom{V}{2}: \quad \gamma'(\{i,j\}) = \gamma(\{\phi(i),\phi(j)\}).$$

We write $\gamma' \subseteq_\phi \gamma$, $\gamma' \subseteq^c_\phi \gamma$, or $\gamma' \subseteq^i_\phi \gamma$ respectively. Furthermore sets of embeddings are denoted as follows:

$$\begin{aligned}
\text{Emb}_{\subseteq}(\gamma',\gamma) &:= \{\phi \in n^V_{\text{inj}} \mid \gamma' \subseteq_\phi \gamma\},\\
\text{Emb}_{\subseteq^c}(\gamma',\gamma) &:= \{\phi \in n^V_{\text{inj}} \mid \gamma' \subseteq^c_\phi \gamma\},\\
\text{Emb}_{\subseteq^i}(\gamma',\gamma) &:= \{\phi \in n^V_{\text{inj}} \mid \gamma' \subseteq^i_\phi \gamma\}.
\end{aligned}$$

$\diamond$

In order to apply the concept of subgraphs to molecular graphs, we define the molecular substructure:

3.4 Definition (substructure) Let $M = (\varepsilon,\zeta,\gamma) \in \mathcal{M}_n$ be a molecular graph and $k \leq n$. A triple

$$S = (\varepsilon',\zeta',\gamma') \in \mathcal{E}^k \times \mathcal{Z}^k_{\mathcal{E}} \times \mathcal{G}_{k,4} =: \mathcal{S}_k$$

is

- a *substructure* of M, if

$$\exists\ \phi \in \text{Emb}_{\subseteq}(\gamma',\gamma)\ :\ \forall\ i \in k : \varepsilon(\phi(i)) = \varepsilon'(i) \text{ and } \zeta(\phi(i)) = \zeta'(i),$$

- a *closed substructure* of M, if

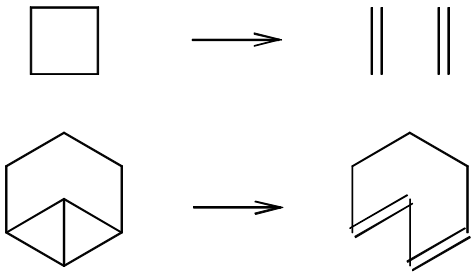
$$\exists\ \phi \in \text{Emb}_{\subseteq^c}(\gamma',\gamma)\ :\ \forall\ i \in k : \varepsilon(\phi(i)) = \varepsilon'(i) \text{ and } \zeta(\phi(i)) = \zeta'(i),$$

- an *induced substructure* of M, if

$$\exists\ \phi \in \text{Emb}_{\subseteq^i}(\gamma',\gamma)\ :\ \forall\ i \in k : \varepsilon(\phi(i)) = \varepsilon'(i) \text{ and } \zeta(\phi(i)) = \zeta'(i).$$

◇

When a chemist looks for a substructure he usually is interested in what we call a closed substructure, rather than what we call simply a substructure or an induced substructure. Thus, ethane is not usually considered a substructure of ethene, at least in the context of possible reactions. On the other hand, cyclobutane is considered a substructure of tricyclo[4.1.0.$0^{2,7}$]heptane, with the cyclobutane graph being a closed, not an induced subgraph in the tricyclo[4.1.0.$0^{2,7}$]heptane molecular graph. This is demonstrated in the following specific reactions both being considered possible.



In order to allow a more flexible description of structural properties, in our computer programs further concepts are integrated:

- In molecular substructures *alternatives* for chemical elements, atomic states and bond multiplicities are allowed.
- Substructure restrictions are introduced in order to describe graph–theoretical properties such as distances between atoms, bond patterns and neighbourhoods of atoms, or prescribed and forbidden ring sizes.

For detailed descriptions of ambiguous molecular substructures and substructure restrictions see [15, 18].

4 Chemical Reactions

To every chemist it is obvious that a specific reaction is characterised by its reactant(s) and its product(s). At the same time, the essential aspect of a reaction is what happens to the starting material(s), so that a reaction may be described by detailing its reactant(s) and the changes that occur. Whenever corresponding changes occur to different reactants, such reactions belong to the same class. In the following these ideas are expressed in mathematical terms. The basic definition is

4.1 Definition (chemical reaction) Assume a positive integer n and a set $\mathcal{E}$ of chemical elements together with $\mathcal{Z}_{\mathcal{E}} = \bigcup_{X \in \mathcal{E}} \mathcal{Z}_X$, the set of admissible atomic states of the elements in $\mathcal{E}$. An ordered pair

$$C := (M, M') \in \mathcal{M}_n \times \mathcal{M}_n$$

consisting of two molecular graphs $M = (\varepsilon, \zeta, \gamma)$ and $M' = (\varepsilon', \zeta', \gamma')$ is called a *chemical reaction* if $\varepsilon = \varepsilon'$. M is the *reactant graph* and M' the *graph of the product.*

The set $\mathrm{Con}(M)$ of connectivity components of M is the set of *starting materials*, the set $\mathrm{Con}(M')$ of connectivity components of M' the set of *products*. By $\mathcal{C}_n$ we shall indicate in the following the set of chemical reactions with n atoms involved.

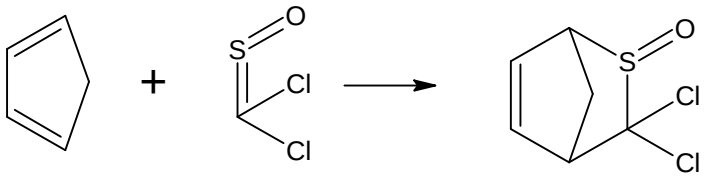
◇

Instead of the mathematical pair notation (M, M') chemists use the mapping notation, writing

$$M \longrightarrow M'.$$

Chemists usually link the components of the reactant or product graph by '+' signs.

4.2 Example The following figure shows the Diels–Alder reaction of cyclopentadiene and thiophosgene S–oxide[6]. The reactant graph (left) and the product graph (right) consist of two components and a single component, respectively:



◇

It is obvious that the changes of atomic states and bonds caused by chemical reactions are of particular interest. Concepts similar to the following were earlier introduced by Dugundji and Ugi [1], Fujita [12], and Temkin, Zeigarnik and Bonchev [13].

4.3 Definition (reaction graph of changes, bond change graph) The *reaction graph of changes* is defined to be the pair

$$\Delta C := (\Delta\zeta, \Delta\gamma) \in \Delta\mathcal{Z}^n \times \mathcal{G}_{n,[-3,3]} =: \Delta\mathcal{C}_n,$$

the components $\Delta\zeta$ and $\Delta\gamma$ of which are defined as follows:

i) $\Delta\zeta$ is a sequence

$$(\Delta\zeta(0), \ldots, \Delta\zeta(n-1)),$$

the i-th component of which is

$$\Delta\zeta(i) := (\Delta v_i, \Delta p_i, \Delta q_i, \Delta r_i) \in \mathbb{Z} \times \mathbb{Z} \times \mathbb{Z} \times \mathbb{B} =: \Delta\mathcal{Z}$$

describing the change of the state of atom i:

- $\Delta v_i := v_{\zeta'(i)} - v_{\zeta(i)}$ means the change of valence of atom i,
- $\Delta p_i := p_{\zeta'(i)} - p_{\zeta(i)}$ denotes the change of the number of free electron pairs of atom i,

[6]If desired, the bond between S and O in sulfoxides etc. may be described as a single bond between $\mathsf{S}(3,0,1,0)$ and $\mathsf{O}(1,3,-1,0)$ rather than as the simplistic double bond between $\mathsf{S}(4,1,0,0)$ and $\mathsf{O}(2,2,0,0)$ used here.

- $\Delta q_i := q_{\zeta'(i)} - q_{\zeta(i)}$ is the change of charge on atom i, while
- $\Delta r_i := r_{\zeta'(i)} \dot{\vee} r_{\zeta(i)}$ indicates the change of the radical character[7] at atom i.

For short: $\Delta\zeta$ is the *distribution of the changes of atomic states* caused by reaction C.

ii) The second component of the reaction graph of changes is the graph

$$\Delta\gamma$$

whose vertices are the atoms $0, ..., n-1$, and atom i is connected to atom j if and only if the bond between these atoms is changed during the reaction. I.e., vertex i is connected to vertex j if and only if $\gamma'(\{i,j\}) - \gamma(\{i,j\}) \neq 0$. Moreover, this edge is labeled by the difference

$$\Delta\gamma(\{i,j\}) := \gamma'(\{i,j\}) - \gamma(\{i,j\}).$$

Since the bond multiplicities in γ and in γ' are at most 3, we obtain

$$\Delta\gamma(\{i,j\}) \in [-3,3],$$

for short:

$$\Delta\gamma \in \mathcal{G}_{n,[-3,3]}.$$

This label $\Delta\gamma(\{i,j\})$ describes the change of bond multiplicity between atoms i and j. $\Delta\gamma$ is therefore called the *bond change graph* of C. $\diamond$

We are now able to formulate an important definition of the present paper, a mathematical model for reactions in organic chemistry on the level of integral chemistry [1].

4.4 Definition (reaction graph) A chemical reaction C is completely described by its reactant graph M together with ΔC. Therefore we call the quintuple

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

the reaction graph of C. $\diamond$

Using the above notation of a chemical reaction we write:

$$M' = \Delta C \circ M,$$

where ΔC is applied to M in the following way:

$$\Delta C \circ M = (\Delta\zeta, \Delta\gamma) \circ (\varepsilon, \zeta, \gamma) := (\varepsilon, \Delta\zeta \circ \zeta, \Delta\gamma \circ \gamma)$$

and for $i, j \in n$, $i \neq j$,

$$\begin{aligned} (\Delta\zeta \circ \zeta)(i) &:= \Delta\zeta(i) \circ \zeta(i) \\ (\Delta\gamma \circ \gamma)(\{i,j\}) &:= \gamma(\{i,j\}) + \Delta\gamma(\{i,j\}). \end{aligned}$$

The distribution of atomic states in the product is

$$\Delta\zeta(i) \circ \zeta(i) := (v_{\zeta(i)} + \Delta v_i,\ p_{\zeta(i)} + \Delta p_i,\ q_{\zeta(i)} + \Delta q_i,\ r_{\zeta(i)} \dot{\vee} \Delta r_i).$$

[7] An atom's radical character may or may not be changed in a reaction. It is changed if the atom bears an unpaired electron either before or after the reaction, but not both. The 'exclusive or' ($\dot{\vee}$) linking of two logical variables behaves analogously and is therefore used here.

4.5 Definition (reaction center) Assume a chemical reaction

$$C = ((\varepsilon, \zeta, \gamma), (\varepsilon, \zeta', \gamma')) \in \mathcal{C}_n.$$

Then

$$\mathrm{Cen}(C) := \{0 \leq i \leq n-1 \mid \zeta(i) \neq \zeta'(i) \vee \exists\, j : \gamma(\{i,j\}) \neq \gamma'(\{i,j\})\}$$

is called the *reaction center* of C. ◇

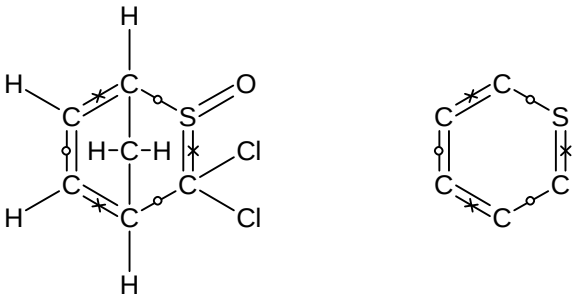
Hence, by definition, a reaction center consists of those atoms whose atomic states or bonds are changed in the reaction. Thus, the reaction may alternatively be described by its reactant graph, the reaction center, and the changes of states and bonds.

4.6 Definition (reaction center graph) Let C denote a chemical reaction. The subgraph induced by the reaction center

$$\mathrm{RCG}(C) := (\varepsilon|_{\mathrm{Cen}(C)}, \zeta|_{\mathrm{Cen}(C)}, \gamma|_{\mathrm{Cen}(C)}, \Delta\zeta|_{\mathrm{Cen}(C)}, \Delta\gamma|_{\mathrm{Cen}(C)})$$

is called the *reaction center graph* of C. ◇

4.7 Example The following picture shows the reaction graph (left) and the reaction center graph (right) of the above Diels–Alder reaction:



Bonds that are formed in the reaction are indicated by small circles '∘', bonds that are broken are indicated by crosses '×'. Thus, for the sake of simplicity, these symbols replace bond labels $+1$ and -1.
Finally the reaction center graph can be split into the reaction substructure (left) and the bond change graph (right):

◇

It is often observed that two nonidentical reactions are essentially the same. In such cases the reaction center graphs are identical or similar. Compare the above reaction with the analogous reaction between 1,3–butadiene and thiophosgene S–oxide, or the

reactions of cyclopentadiene or 1,3–butadiene with the dibromo analog of thiophosgene S–oxide: All the reaction center graphs are identical. In the cycloaddition of a 1,3–diene and maleic acid anhydride, the reaction center graph is as before except that the S is replaced by a C, and this reaction center graph obviously describes the essence of the Diels–Alder class of reactions.

4.8 Definition (reaction scheme) Assume a natural number k. A *reaction scheme* is a triple

$$R := (S, \Delta\zeta, \Delta\gamma) \in \mathcal{S}_k \times \Delta\mathcal{Z}^k \times \mathcal{G}_{k,[-3,3]} =: \mathcal{R}_k$$

consisting of a substructure S, the distribution of the *change of states* $\Delta\zeta$ and the *bond change graph* $\Delta\gamma$. In this context S is also called the *reaction substructure*. $\diamond$

This definition allows to describe one-reactant reactions such as cleavages and rearrangements as well as synthesis reactions with arbitrary numbers of reactants. Example 4.7 shows a canonical way to obtain a suitable reaction substructure from a specific reaction. However, the reaction substructure may be defined in a more or less restrictive manner in order to allow fewer or more reaction products to result from application of the reaction scheme, respectively.
The application of a reaction scheme $R = (S, \Delta\zeta, \Delta\gamma) \in \mathcal{R}_k$ to a molecular graph $M = (\varepsilon, \zeta, \gamma) \in \mathcal{M}_n$ is done in two steps. First we search for an embedding of the reaction substructure S in M. If we find such an embedding $\phi \in \mathrm{Emb}_{\subseteq^c}(S, M)$, then we apply both the distribution of change of states and the bond change graph to M in the following way: ϕ induces the mapping

$$-^{\phi} : \Delta\mathcal{C}_k \longrightarrow \Delta\mathcal{C}_n, \quad (\Delta\zeta, \Delta\gamma) \longmapsto (\Delta\zeta, \Delta\gamma)^{\phi} := (\Delta\zeta^{\phi}, \Delta\gamma^{\phi}),$$

where, for all $i \in n$,

$$\Delta\zeta^{\phi} := \begin{cases} \Delta\zeta\left(\phi^{-1}(i)\right) & \text{if } i \in \phi(k), \\ (0, 0, 0, false) & \text{else,} \end{cases}$$

and, for $i, j \in n$, $i \neq j$,

$$\Delta\gamma^{\phi} := \begin{cases} \Delta\gamma\left(\{\phi^{-1}(i), \phi^{-1}(j)\}\right) & \text{if } i, j \in \phi(k), \\ 0 & \text{else.} \end{cases}$$

An application of R to M with respect to ϕ can be defined as

$$R \circ_{\phi} M := (\Delta\zeta, \Delta\gamma) \circ_{\phi} M := (\Delta\zeta, \Delta\gamma)^{\phi} \circ M.$$

But we should note that $R \circ_{\phi} M$ need not be a molecular graph.

4.9 Definition (set of product graphs) Assume positive natural numbers k, n, where $k \leq n$, $R = (S, \Delta\zeta, \Delta\gamma) \in \mathcal{R}_k$ a reaction scheme and $M \in \mathcal{M}_n$ a molecular graph. The *set of product graphs* obtained by application of R to M is

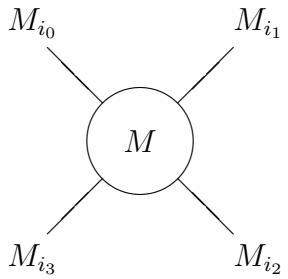
$$\mathrm{Prod}_R(M) := \{R \circ_{\phi} M \in \mathcal{M}_n \mid \phi \in \mathrm{Emb}_{\subseteq^c}(S, M)\}.$$

$\diamond$

This mathematical model allows to simulate chemical reactions, in the sense to generate *all products that may arise from a given set of reactants and a given reaction scheme*.
Quite a different question is that for the quantitative result of chemical reactions, i.e. for the amounts (concentrations) of products in the case of competing reactions. The outcome depends on reactivities, i.e. on free enthalpies of activation of the competing reactions and on temperature, and thus does not seem amenable to graph theoretic modeling, at least at present.

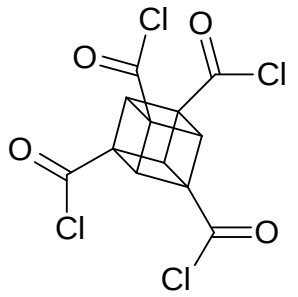
5 Libraries obtained from a Central Molecule

Assume a *central molecule* M and *reaction partners* M_i, $i \in a$, together with a reaction scheme $R = (S, \Delta\zeta, \Delta\gamma)$. We suppose that R means a two-component synthesis, i.e. the graph underlying S consists of two connectivity components A and B. We assume that A is embedded in M by k nonoverlapping ϕ_j, $j \in k$. For each j, the atoms defined in M by ϕ_j are a reactive site. Each reaction partner M_i, on the other hand, is assumed to contain a single embedding of B, the other component of the reaction substructure.
By applying the reaction scheme we attach the reaction partners in various combinations to the reactive sites of the central molecule. For $k = 4$ we may sketch the situation as follows:

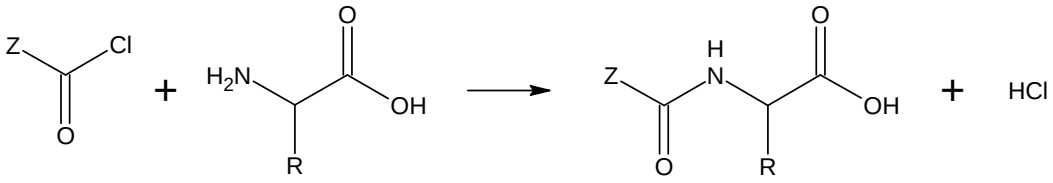


Here the index $i_j \in a$ in the substituent symbols M_{i_j} refers to the identity of the substituent, the subindex $j \in k$ refers to the numbering of the reactive sites in the central molecule M.
The essentially different attachments of substituents to the central molecule can be generated using the symmetry group of M [19]. The automorphism group $\mathrm{Aut}(M)$ acts on the reactive sites of M and induces a subgroup G of the symmetric group S_k acting on these sites. The essentially different attachments are the orbits of the symmetry group G on the set of all the a^k attachments.

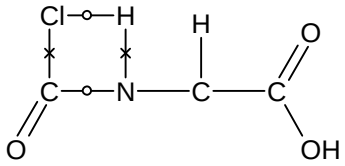
5.1 Example As an example we consider the exhaustive amidation of a particular cubanetetracarboxylic acid tetrachloride as a central molecule [20, 21]:



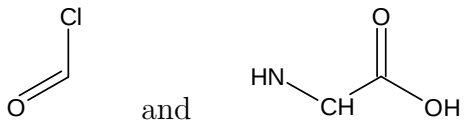
$a = 20$ natural amino acids are attached to this central molecule as shown below. An acyl chloride group reacts with an amino group in α position to the carboxyl group:



Proline contains only one hydrogen bound to the N–atom. To include proline, we define the reaction scheme in the following way:



The two connectivity components of the reaction substructure are



A B

There are $k = 4$ embeddings of A found in the cubanetetracarboxylic acid tetrachloride, while there exists exactly one embedding B in each of the 20 amino acids. Altogether there are $20^4 = 160000$ possible attachments of four amino acid molecules to the central molecule. But with respect to the central molecule's automorphism group the essentially different attachments are obtained as different orbits of the operation of the symmetric group S_4 applied to the set of 20^4 mappings. The result is a combinatorial library of 8855 different structures:

$$|20^4 /\!/ S_4| = 8855.$$

This example shows the importance of a canonizer in the generation process. Without it, using the combinatorial approach, we would need to carry out $\binom{160000}{2}$ isomorphism tests in order to get rid of the duplicates. $\diamond$

6 Construction of Reaction Networks

In the following we describe how compound libraries are generated by *successive* application of reaction schemes. Such libraries are of particular importance in structure elucidation, as well as in combinatorial chemistry.
Most chemical processes can be described as *chemical reaction networks*. Such a network is a *bipartite directed graph*. Its vertex set is partitioned into compounds and reactions, i.e. the vertices represent either molecular graphs or reaction schemes. Its edges are directed, from reactants to their reactions, and from reactions to their products. While the same reaction scheme may occur in a chemical reaction network several times, a molecular graph can occur at most once, with its vertices (atoms) canonically labeled.
We shall generate for a given set of reactants and a given set of reaction schemes all molecular graphs that can occur. We shall run through a (partial) reaction network using a *breadth–first* strategy.
Initially we have to generalize a few notions. Above we introduced the application of a reaction scheme $R = (S, \Delta\zeta, \Delta\gamma)$ to a single molecular graph $M \in \mathcal{M}$ in order to obtain the set of product graphs:

$$\mathrm{Prod}_R\,(M) = \{R \circ_\phi M \in \mathcal{M}_n \mid \phi \in \mathrm{Emb}_{\subseteq^c}\,(S, M)\}.$$

We extend this definition to *sets of molecular graphs* and *sets of reaction schemes*, starting from a set

$$\mathcal{L} = \{M_i \mid i \in l\} \subseteq \mathcal{M}^C$$

of connected molecular graphs. In order to evaluate the set of all possible products arising by an application of R to $\mathcal{L}$, we have to examine the connectivity components of the reaction substructure S:

$$\mathrm{Con}(R) := \mathrm{Con}(S).$$

I.e. $|\mathrm{Con}(R)|$ is the maximum number of starting materials involved in a reaction of the present kind. For the Diels–Alder reaction, this number is 2. In the particular example reaction above, there are in fact two starting materials, though in an intramolecular Diels-Alder reaction there is but one. Thus for the Diels–Alder and many other typical synthetic reactions all *combinations (with repetition)* of two species in the set of starting materials have to be considered as potentially reactive. 'With repetition' means that e.g. a combination of two reactants may be made of two copies of the same species.
For the set of combinations with repetition of n objects out of a set of m objects we introduce the notation $m^n_{\leq}$. This set is a subset of the set of distributions of n out of m objects to n positions (of which there are m^n). The subset condition is that the positions occupied do not matter, i.e. all those distributions are considered equivalent that lead to the same result after the objects are arranged in increasing order. (We assume that a natural order is defined for the objects, as e.g. for the natural numbers, or some initial or canonical numbering for molecular graphs.) Thus the combinations with repetition are equivalent to weakly monotonously increasing mappings from n to m:

$$m^n_{\leq} := \{f \in m^n \mid \forall\, i : f(i) \leq f(i+1)\}.$$

Using this we can introduce the product graphs arising from an application of R to the library $\mathcal{L}$ as

$$\mathrm{Prod}_R(\mathcal{L}) := \bigcup_{k\in|\mathrm{Con}(R)|} \bigcup_{f\in l^k_{\leq}} \mathrm{Prod}_R\left(\bigoplus_{i\in k} M_{f(i)}\right),$$

$\bigoplus_{i\in k} M_{f(i)}$ is built by putting $M_{f(0)}, ..., M_{f(k-1)}$ together into one big (disconnected) molecular graph.
For a set $\mathcal{R}$ of reaction schemes we can define

$$\mathrm{Prod}_{\mathcal{R}}(\mathcal{L}) := \bigcup_{R\in\mathcal{R}} \mathrm{Prod}_R(\mathcal{L}).$$

Finally we have to decompose the product graphs into connectivity components and to eliminate duplicates that may occur. For this reason we define, for an arbitrary set $\mathcal{L}$ of molecular graphs

$$\mathrm{Con}(\mathcal{L}) := \bigcup_{M\in\mathcal{L}} \mathrm{Con}(M)$$

and

$$\kappa(\mathcal{L}) := \{\kappa(M) \mid M \in \mathcal{L}\}.$$

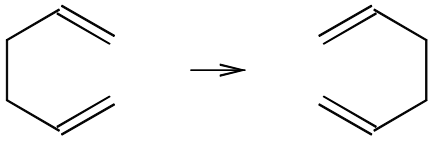
Here $\kappa(M)$ denotes the canonical labeling of M. An algorithm that computes a canonical labeling is described in [22].
We are now able to formulate an algorithm for the construction of the library of products that can arise by application of the set of reaction schemes $\mathcal{R}$ to a given set of molecular graphs $\mathcal{L}$:

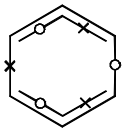
6.1 Algorithm $MolLib(\mathcal{L}, \mathcal{R})$

(1) $\mathcal{L}_0 \leftarrow \kappa(\mathcal{L})$, $k \leftarrow 0$
(2) **while** $\mathcal{L}_k \neq \emptyset$ **do**
(3) $\quad k \leftarrow k + 1$
(4) $\quad \mathcal{L}_k \leftarrow \kappa\left(\mathrm{Con}\left(\mathrm{Prod}_{\mathcal{R}}\left(\bigcup_{i\in k} \mathcal{L}_i\right)\right)\right) \setminus \bigcup_{i\in k} \mathcal{L}_i$
(5) $\quad Output(\mathcal{L}_k)$
(6) **end**

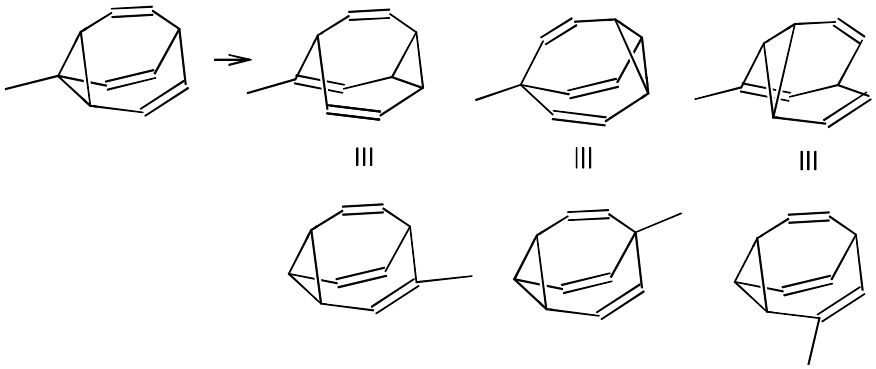
This means that in row (1) reactants are transformed into *canonical forms*, and after elimination of duplicates are assigned to $\mathcal{L}_0$. Of central importance is line (4). There we construct from the already obtained libraries $\mathcal{L}_i$, $i \in k$, *new* structures, collected in the library $\mathcal{L}_k$. The generation process stops as soon as no further structures are produced. This is checked in row (2).

6.2 Example The Cope rearrangement and its reaction center graph may be depicted as follows:





Starting from a particular methylbullvalene and this reaction scheme MOLGEN correctly generates all four methylbullvalenes, e. g.



The same result is of course obtained starting with any of the methylbullvalenes. By the corresponding reaction sequence in the parent bullvalene all ten CH units are rendered equivalent. ◇

In the following we shall modify algorithm 6.1 in order to apply it to specific problems.

7 Application: The Generation of MS Fragments

Our primary motivation to develop a structure generator working on a reaction network was the request to generate all fragments that occur in a compound's electron impact mass spectrum. Initially, we list a few characteristics of the chemistry that occurs in a mass spectrometer.

i) The set of reactants consists of a single molecule: $\mathcal{L} = \{M\}$.

ii) All reactions have one reactant only.

iii) The set of reaction schemes consists of two subsets, the set of ionisation schemes and the set of fragmentation schemes, $\mathcal{R} = \mathcal{R}_I \dot{\cup} \mathcal{R}_F$.

iv) In the first step an ionisation is applied to M, resulting in a radical cation or (in some cases) a cation plus a radical.

v) For the further steps only the cations are of interest.

vi) After ionisation we can apply an arbitrary number of fragmentations.

For i) and ii) no action is required, with respect to the other items we introduce the following modifications:

- To each reaction scheme we associate its *depth* which says on which level it is applicable. We specify for each reaction scheme an interval of non–negative integers:

$$\mathrm{depth}_{\mathcal{R}} : \mathcal{R} \longrightarrow \mathcal{I}(\mathbb{N}),$$

where

$$\mathrm{depth}_{\mathcal{R}}(R) = \begin{cases} [1,1] & \text{if } R \in \mathcal{R}_I, \\ [2,\infty[& \text{else.} \end{cases}$$

$\mathcal{I}(\mathbb{N})$ denotes the set of intervals on the natural numbers. This takes into account items iii), iv), and vi).

- With respect to item v), instead of Con() we introduce $\mathrm{Con}^+()$,

$$\mathrm{Con}^+(\mathcal{L}) := \{M \in \mathrm{Con}(\mathcal{L}) \mid \mathrm{cha}(M) = 1\},$$

for decomposition and selection of connectivity components of product graphs. $\mathrm{cha}(M)$ denotes the sum of charges of the atoms in M.

Since there is a single reactant for each reaction, we can restrict attention in line (4) of algorithm 6.1 to $\mathcal{L}_{k-1}$, otherwise $\mathrm{Prod}_{\mathcal{R}}\left(\bigcup_{i\in k}\mathcal{L}_i\right)$ would produce duplicates only. The modified algorithm, in addition, uses the notion of depth of the reaction schemes that we are going to apply:

7.1 Algorithm $MolLibMS(\mathcal{L}, \mathcal{R}, \mathrm{depth}_{\mathcal{R}}())$

(1) $\mathcal{L}_0 \leftarrow \kappa(\mathcal{L})$, $k \leftarrow 0$
(2) **while** $\mathcal{L}_k \neq \emptyset$ **do**
(3) $k \leftarrow k + 1$
(4) $\mathcal{R}' \leftarrow \{R \in \mathcal{R} \mid k \in \mathrm{depth}_{\mathcal{R}}(R)\}$
(5) $\mathcal{L}_k \leftarrow \kappa\left(\mathrm{Con}^+\left(\mathrm{Prod}_{\mathcal{R}'}\left(\mathcal{L}_{k-1}\right)\right)\right) \setminus \bigcup_{i\in k}\mathcal{L}_i$
(6) $Output(\mathcal{L}_k)$
(7) **end**

7.2 Example
In order to keep the number of ionisation and fragmentation schemes small we introduce several generic element symbols:

A: any element

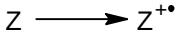
Y: heavy atom (i.e. any element except H)

Z: any element bearing a free electron pair (N, O, P, S, halogens)

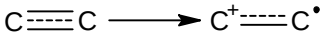
Alternatives for bond multiplicities will be coded graphically as follows:

We consider three ionisation reactions

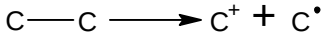
- n–ionisation



- π–ionisation

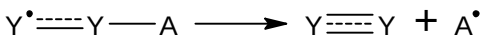


- σ–ionisation

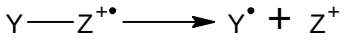


and the following fragmentation reactions

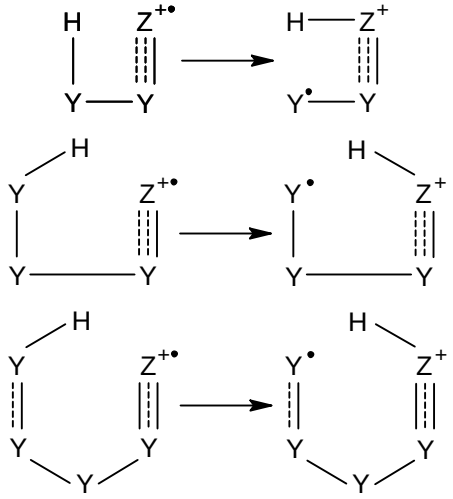
- α–cleavage



- σ–cleavage



- H–rearrangements



Of course several further reactions can occur in an MS. However, this minimalistic set of reaction schemes can already explain many peaks, as seen for the example methyl pentanoate.
Figure 1 shows that part of the MS reaction network initiated by n–ionisation (n–I). H–rearrangements on four and six atoms (H–R, McLafferty rearrangement) and α–cleavages are considered, resulting in fragment ions of mass 116, 115, 87, 85, 74, and 43. H–rearrangements on five atoms and fragments derived from σ–ionisation are left out for clarity. The former are irrelevant in our example, whereas the latter (not detailed here) are very important, resulting in fragments of mass 101, 87, 73, 59, 57, 43, 29, and 15. Figure 2, top, is the experimental spectrum of methyl pentanoate, Figure 2, middle, is the part of the spectrum explained by the above reactions, and the residual peaks are shown in Figure 2, bottom.
Figure 3 lists all 32 fragment ions that are generated from methyl pentanoate by the above reaction schemes. Structures are ordered by increasing mass. A structure's mass is given on the right hand side of its header.

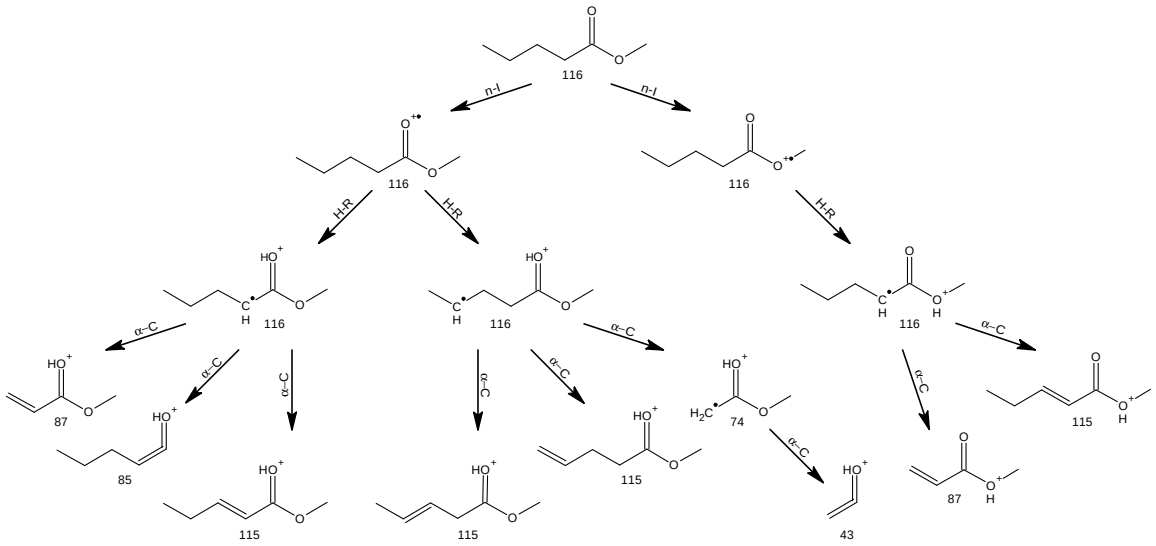


Figure 1: MS reactions of methyl pentanoate initiated by n–ionisation

Comparison of the fragments obtained by corresponding reactions from competing structure candidates (e.g. structures isomeric to methyl pentanoate) is an approach to automated structure elucidation via MS [15].
For early attempts to quantitatively model the reactions occurring in a mass spectrometer see references [23, 24]. ◇

8 Application: Generation of Combinatorial Libraries

Another important application of reaction based generation of molecular libraries is the simulation of combinatorial chemistry. We should be able to generate a library from building blocks and given reactions in order to examine libraries in advance (before or in lieu of synthesis). Often we find the following situation:

i) The set of reactants consists of two subsets, a set of central molecules and a set of ligands: $\mathcal{L} = \mathcal{L}_C \dot{\cup} \mathcal{L}_L$.

ii) Each central molecule can be used just once during the generation procedure, that is at the very beginning.

iii) Each reaction product contains at least one central molecule.

iv) All reactions have one or two reactants.

v) Reactions between two or more intermediates have to be neglected.

vi) Stoichiometric side products such as $\mathsf{H_2O}$, HCl, etc. are to be ignored.

In order to fulfil these conditions, we introduce the following restrictions:

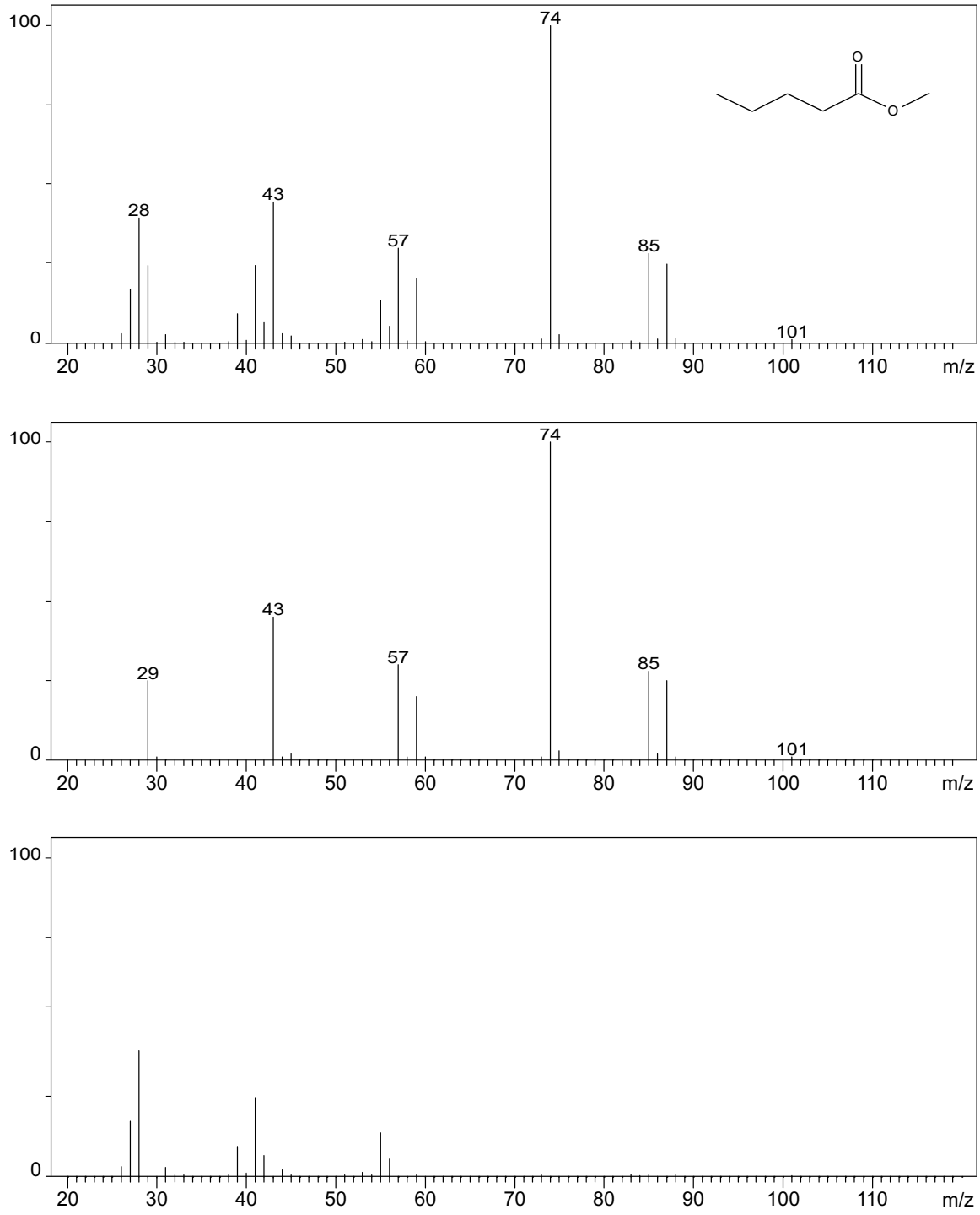


Figure 2: Experimental mass spectrum of methyl methyl pentanoate (top), and the parts of the spectrum explained (middle) and unexplained (bottom) by the reactions considered

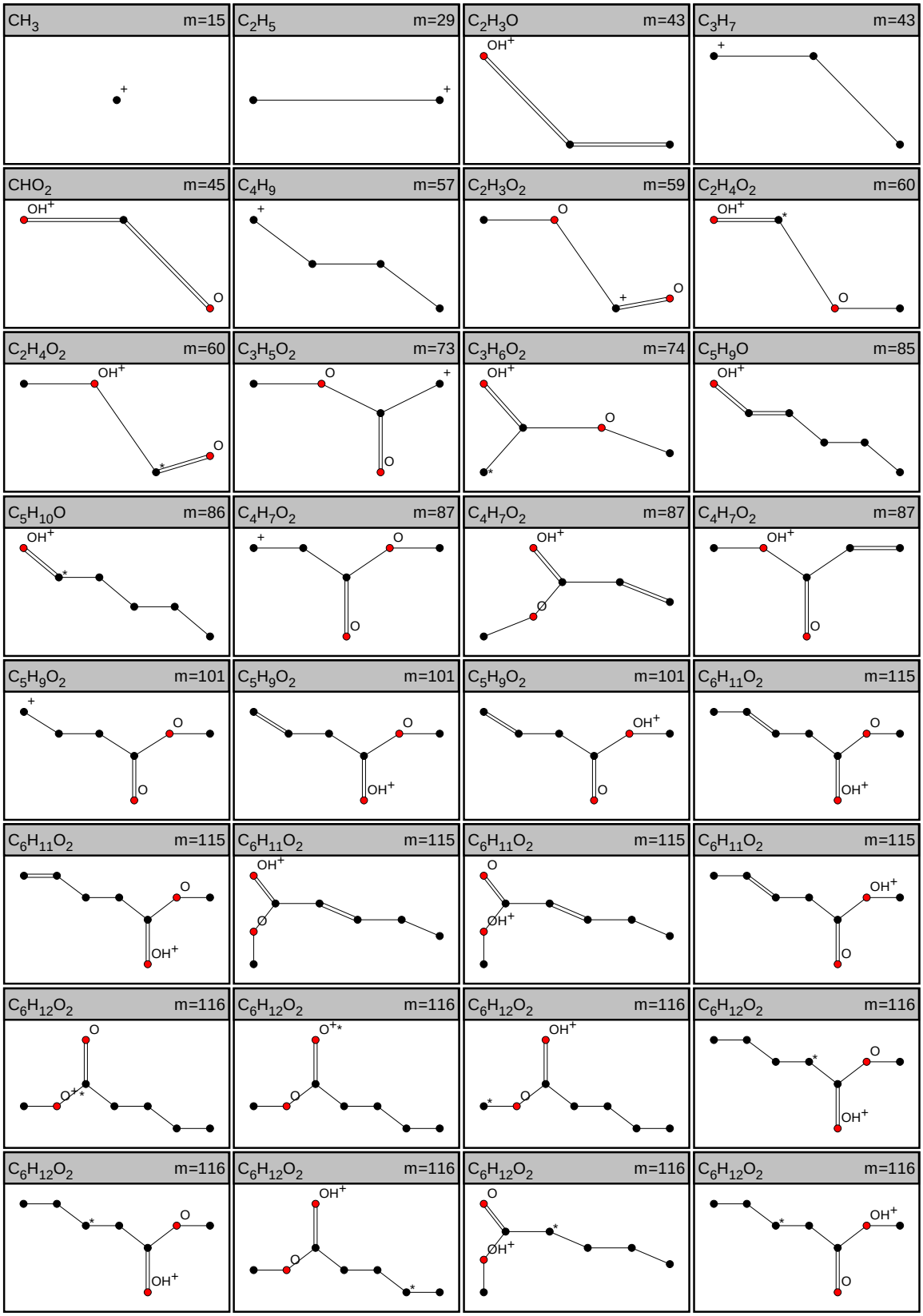


Figure 3: Fragment ions generated from methyl pentanoate

- To each reactant we associate a *depth*, in which it can occur during the reaction processes. The depth is given as an interval:

$$\mathrm{depth}_{\mathcal{L}} : \mathcal{L} \longrightarrow \mathcal{I}(\mathbb{N}),$$

where

$$\mathrm{depth}_{\mathcal{L}}(M) = \begin{cases} [0,0] & \text{if } M \in \mathcal{L}_C, \\ [1,\infty[& \text{else.} \end{cases}$$

This covers conditions i) – iii).

- In order to satisfy condition vi) we use, for the selection of connectivity components in the product graphs

$$\mathrm{Con}^{\geq}(\mathcal{L}) := \bigcup_{M \in \mathcal{L}} \mathrm{Con}^{\geq}(M),$$

where

$$\mathrm{Con}^{\geq}(M) := \{M' \in \mathrm{Con}(M) \mid \mathrm{size}(M') \geq \frac{1}{2}\mathrm{size}(M)\}.$$

Here $\mathrm{size}(M)$ means the number of atoms in M.

In order to make the algorithm applicable to different functions on the connectivity components, we introduce $\mathrm{Con}^*()$ as an additional argument. For combinatorial libraries we shall mostly choose

$$\mathrm{Con}^*() = \mathrm{Con}^{\geq}().$$

Conditions iv) and v) are considered in row (6) of algorithm 8.1.

8.1 Algorithm MolLibCC($\mathcal{L}, \mathcal{R}, \mathrm{depth}_{\mathcal{R}}(), \mathrm{depth}_{\mathcal{L}}(), \mathrm{Con}^*()$)

(1) $\mathcal{L}_0 \leftarrow \kappa(\{M \in \mathcal{L} \mid 0 \in \mathrm{depth}_{\mathcal{L}}(M)\})$, $k \leftarrow 0$
(2) **while** $\mathcal{L}_k \neq \emptyset$ **do**
(3) $k \leftarrow k + 1$
(4) $\mathcal{L}' \leftarrow \{M \in \mathcal{L} \mid k \in \mathrm{depth}_{\mathcal{L}}(M)\}$
(5) $\mathcal{R}' \leftarrow \{R \in \mathcal{R} \mid k \in \mathrm{depth}_{\mathcal{R}}(R)\}$
(6) $\mathcal{L}_k \leftarrow \kappa\left(\mathrm{Con}^*\left(\mathrm{Prod}_{\mathcal{R}'}\left(\mathcal{L}_{k-1} \cup \mathcal{L}'\right)\right)\right) \setminus \bigcup_{i \in k} \mathcal{L}_i$
(7) *Output*($\mathcal{L}_k$)
(8) **end**

In certain cases it is useful to have further tools at hand for the generation of molecular libraries:

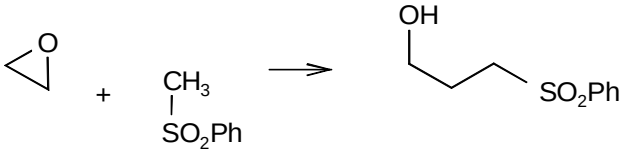
- Sometimes we just want to output the final products while intermediate ones are not of interest.

- It may also happen that reactants or reaction schemes should occur with prescribed multiplicities.

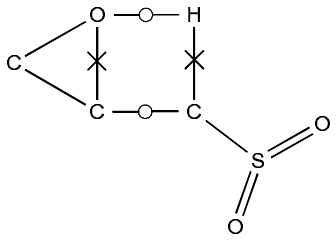
These features are also provided in the MOLGEN reaction module.

8.2 Examples

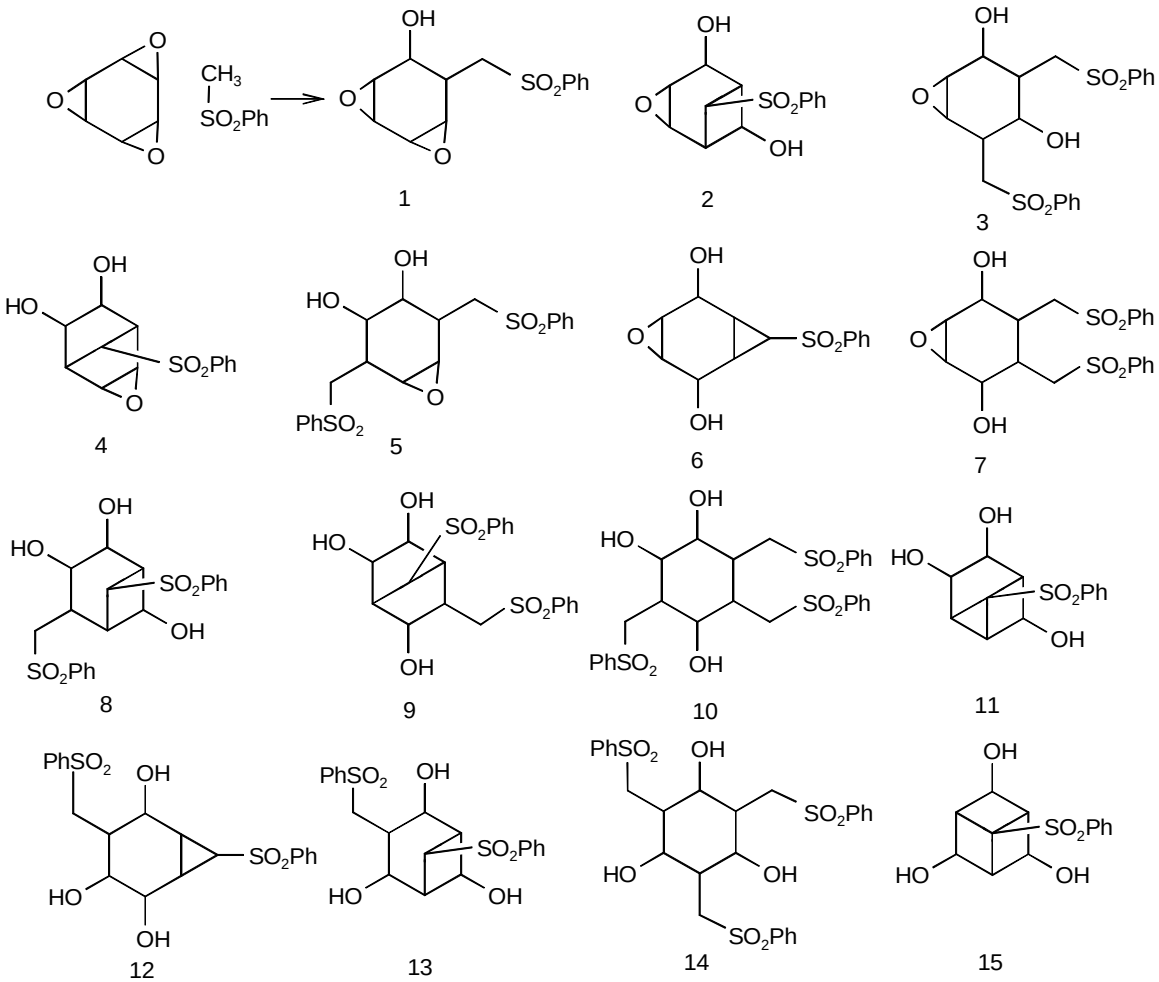
a) A sulfone bearing a proton geminal to the sulfonyl group is able, in the presence of a sufficiently strong base, to open an epoxide in a nucleophilic substitution reaction, forming a new $C-C$ bond. The epoxide oxygen thereby ends up as an alcohol oxygen, e. g.



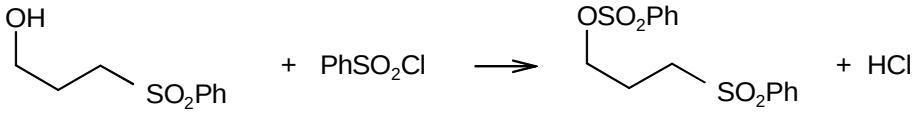
The corresponding reaction scheme may be written as follows:



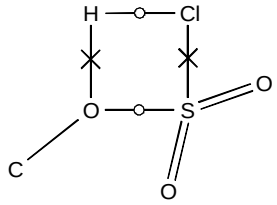
When this reaction scheme is applied to benzene trioxide, by intramolecular reaction steps carbocyclic rings may be formed. In fact, given this input, MOLGEN generates 15 products as follows. Of these, **8** – **15** are final products.



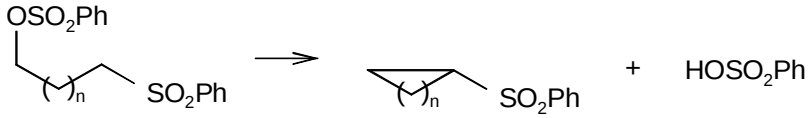
Experimentally many of these types of products were observed to result from reactions of *cis*–benzene trioxide with methyl phenyl sulfone or other acidified methanes [25, 26, 27].
b) Chemically, for a further ring closure a good leaving group is required, therefore we esterify the alcohols formed using benzenesulfonyl chloride, e. g.



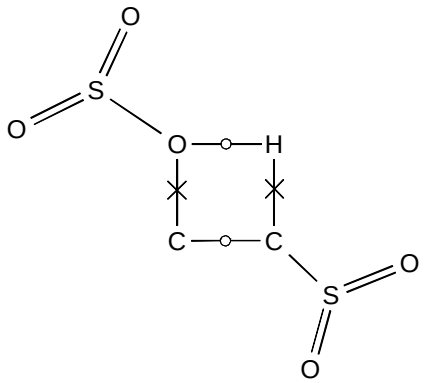
Applying the corresponding reaction scheme



to the above triols **8** – **15** MOLGEN generates the corresponding tris–benzenesulfonates.
c) Finally, in the presence of base a sulfone bearing an acidic H and a benzenesulfonate leaving group may be cyclized, e. g.

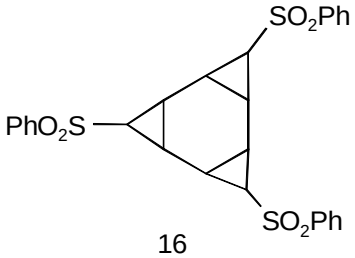


Given the tris-benzenesulfonates of **8** – **15** above and the corresponding reaction scheme

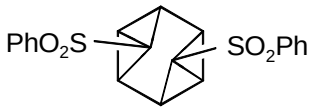


MOLGEN generates carbocyclic tris–σ–homobenzene **16**, disubstituted octabisvalene **17**, disubstituted cubane **18** and disubstituted cuneane **19** as possible reaction prod-

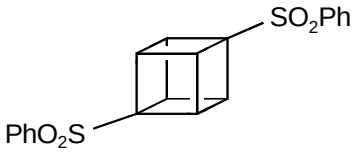
ucts, as well as some other interesting polycyclics.



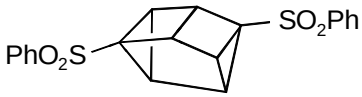
16



17



18

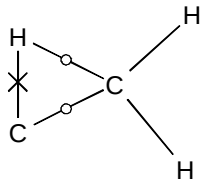


19

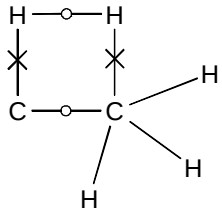
In fact, compound **17** was synthesized by this sequence of reactions [25, 26], while analogs of **16** were obtained by a corresponding sequence [27]. Compounds of types **18** and **19**, however, are not observed in vitro nor expected by a chemist. In silico generation of **18** and **19** results from MOLGEN at present not being able to consider stereochemistry of starting materials and reactions.

d) Since MOLGEN is a formal system, we are free to formulate reactions that seem unrealistic, and reactions may be formulated in a very flexible manner. Thus, an equivalent of the above ring closure reaction may be formulated directly using an alcohol as reactant instead of a benzenesulfonate. In that manner products **16**–**19** are obtained without the need to formulate the above sulfonylation reaction.

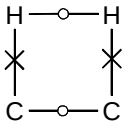
e) To demonstrate flexibility, we consider reaction schemes I - III:



I



II

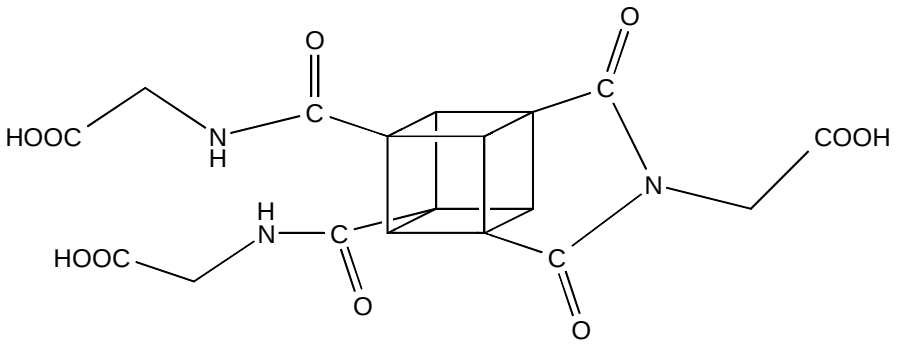


III

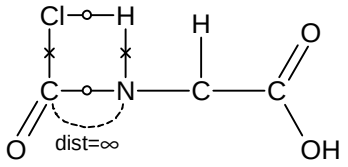
Starting with methane and methylene, repetitive use of I generates all alkanes via carbene insertions into $\mathsf{C}-\mathsf{H}$ bonds. The same result is obtained from methane alone and II, the formal condensation of an alkane and a methane molecule with release of molecular hydrogen. Finally, the more general reaction scheme III produces from methane all hydrocarbons (alkanes, alkenes, alkynes, cyclic and polycyclic hydrocarbons, arenes, etc.).

f) The problem in example 5.1 may also be solved by the network approach. However, the reaction scheme formulated in section 5.1 allows intramolecular reactions, i.e. ring

closures resulting in products of the following type:



To exclude this kind of products, a distance restriction is used, i.e. the distance between the carbon atom in A and the nitrogen atom in B is set to ∞, which means that these atoms must not be found in the same connectivity component:



◇

References

[1] J. Dugundji and I. Ugi. *An Algebraic Model of Constitutional Chemistry as a Basis for Chemical Computer Programs*. Top. Curr. Chem., 39:19–64, 1973.

[2] W.-D. Ihlenfeldt and J. Gasteiger. *Computer–Assisted Planning of Organic Syntheses: the Second Generation of Programs*. Angew. Chem. Int. Ed. Engl., 34:2613–2633, 1995.

[3] E. Fontain and K. Reitsam. *The Generation of Reaction Networks with RAIN. 1. The Reaction Generator*. J. Chem. Inf. Comput. Sci., 31:96–101, 1991.

[4] R. Höllering, J. Gasteiger, L. Steinhauer, K.-P. Schulz, and Herwig A. *Simulation of Organic Reactions: From the Degradation of Chemicals to Combinatorial Synthesis*. J. Chem. Inf. Comput. Sci., 40:482–494, 2000.

[5] J.-L. Faulon and A. G. Sault. *Stochastic Generator of Chemical Structure. 3. Reaction Network Generation*. J. Chem. Inf. Comput. Sci., 41:894–908, 2001.

[6] G. Benkö, C. Flamm, and P. F. Stadler. *A Graph–Based Toy Model of Chemistry*. J. Chem. Inf. Comput. Sci., 43:1085–1093, 2003.

[7] R. Herges. *Ordering Principle of Complex Reactions and Theory of Contracted Transition States*. Angew. Chem. Int. Ed. Engl., 33:255–276, 1994.

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

[9] 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.

[10] 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.

[11] 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.

[12] S. Fujita. *Computer–Oriented Representation of Organic Reactions.* Yoshioka Shoten Publishing Company, Kyoto, 2001.

[13] O. N. Temkin, A. V. Zeigarnik, and D. Bonchev. *Chemical Reaction Networks: A Graph–Theoretical Approach.* CRC Press, Boca Raton, FL, 1996.

[14] 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.

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

[16] C. Benecke, T. Grüner, A. Kerber, R. Laue, and T. Wieland. *Molecular Structure Generation with MOLGEN, new Features and Future Developments.* Fresenius J. Anal. Chem., 358:23–32, 1997.

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

[18] T. Grüner. *Strategien zur Konstruktion diskreter Strukturen und ihre Anwendung auf molekulare Graphen.* MATCH Commun. Math. Comput. Chem., 39:39–126, 1999.

[19] T. Wieland. *Konstruktionsalgorithmen bei molekularen Graphen und deren Anwendung.* MATCH Commun. Math. Comput. Chem., 39:7–155, 1997.

[20] T. Carell, E. A. Wintner, A. Bashir-Hashemi, and J. Rebek Jr. *Novel Method for Preparation of Libraries of Small Organic Molecules.* Angew. Chem. Int. Ed. Engl., 33:2059–2061, 1994.

[21] 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.

[22] 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.

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

[24] J. Gasteiger, W. Hanebeck, K.-P. Schulz, S. Bauerschmidt, and R. Höllering. *Automatic Analysis and Simulation of Mass Spectra.* Volume 4 of *Computer–Enhanced Analytical Spectroscopy*, 97–133. Kluwer Academic Publishers, 1993.

[25] C. Rücker. *Phenylsulfonylsubstituierte Octabisvalene, Synthesen und Reaktionen.* Chem. Ber., 120:1629–1644, 1987.

[26] C. Rücker and B. Trupp. *Pentacyclo[5.1.0.0^{2,4}.0^{3,5}.0^{6,8}]octane (Octabisvalene).* J. Am. Chem. Soc., 110:4828–4829, 1988.

[27] C. Rücker, H. Müller-Bötticher, W.-D. Braschwitz, H. Prinzbach, U. Reifenstahl, and H. Irngartinger. *Carbocyclic cis–[1.1.1]–Tris–σ–homobenzenes — Syntheses by Triple Epoxide → Cyclopropane Conversions, Structural Data, [σ2s+σ2s+σ2s]–Cycloreversions.* Liebigs Ann./Recueil, 967–989, 1997.