Discrete Mathematics for Combinatorial Chemistry

Adalbert Kerber*, Reinhard Laue†, Thomas Wieland‡

May 2, 1998

Abstract

The aim is a description of discrete mathematics used in a project[1] devoted to the implementation of a software package for the simulation of combinatorial chemistry.

1 Discrete mathematics and chemistry

In the March issue 1996 of the *Notices of the American Mathematical Society,* the following article was published:

> George A. Hagedorn: Crossing the Interface between Chemistry and Mathematics.

It was motivated by the report

> Mathematical Challenges from Theoretical/Computational Chemistry,

published by the National Academy Press and available via Internet under the address *http://www.nap.edu/readingroom/books/mctcc.* In fact, G. A. Hagedorn contributes to this "crossing the interface" by making helpful remarks on applications of quantum mechanical resonances, which play an important role in chemical reactions.

The report of the National Academy is very useful, since it brings a very important field of applications of all kinds of mathematics into focus, listing

*University of Bayreuth, Department of Mathematics, D-95440 Bayreuth, Germany, kerber@uni-bayreuth.de, http://www.mathe2.uni-bayreuth.de

†University of Bayreuth, Department of Mathematics, D-95440 Bayreuth, Germany, laue@uni-bayreuth.de, http://www.mathe2.uni-bayreuth.de

‡Deutsches Zentrum für Luft- und Raumfahrt e.V., Deutsches Fernerkundungsdatenzentrum, Oberpfaffenhofen, D-82234 Weßling, Germany, thomas.wieland@dlr.de, http://www.dfd.dlr.de

[1] financial support by the BMBF under contract 03-KE7BA 1-4 is gratefully acknowledged

a lot of examples of constructive cross-fertilization between mathematics and chemistry. This is by no means trivial since the mathematization of chemistry is several hundred years younger than the mathematization of physics and it is not yet well established.

As the present conference lays emphasis upon applications of *discrete* mathematics in chemistry, we should like here to point to the fact that there is (or at least was) a cross–fertilization in the other direction, too. Let us describe our personal experiences with that, during the last thirty years.

The crucial point is that still today, the *graph theoretic model of molecules* is dominating, a model that was motivated by a classical problem of chemistry, namely by the problem of *isomerism.* Let us briefly indicate the history of that problem, since it is a decisive part of the history of graph theory (the other sources of graph theory are better known: Euler's solution of the Königsberg bridge problem, and Kirchoff's description of electrical networks).

In 1797, Alexander von Humboldt, a German geographer, famous for the scientific results of his expeditions to South America, published a book [10] with the title "Versuche über den gereizten Muskel- und Nervenfaser nebst Vermuthungen über den chemischen Prozess des Lebens", in which he made (in volume II, page 128) the following surprising statement:

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

This expresses his opinion that substances should exist with the very same constituents (atomic constitutents, in today's language, like oxygen, hydrogen, carbon, nitrogen and metal) but with different properties. And he did this long before the concepts of molecular structure, chemical bond or valence of atoms were introduced (for a detailed history and a list of corresponding references see e.g. [16] and the literature cited there). A quarter of a century later, after the development of suitable analytic methods, Humboldt's thesis was shown to be true by the famous chemists Joseph-Louis Gay-Lussac, Justus von Liebig and Friedrich Wöhler. Here is a quotation from a footnote by Gay-Lussac to a paper by Wöhler which shows that he clearly foresaw the combinatorial reason for that phenomenon:

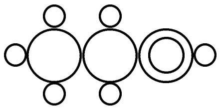
> ...comme ces deux acides sont très différents, il faudrait pour expliquer leur différence admettre entre leurs éléments un mode de combinaison différent. C'est un objet qui appelle un nouveau examen.

It is interesting to know that Gay-Lussac was a close friend, and von Liebig a protégé of von Humboldt, but there is no proof yet that von Humboldt had told them about this problem.

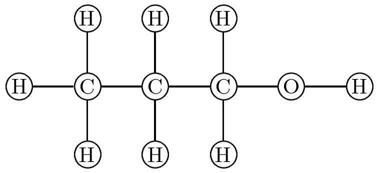
About 1830 Berzelius called this phenomenon *isomerism.* Chemists tried to understand it, and to find a solution by sketching molecules. Here are a few prominent ways of drawing the alcohol C_2H_5OH. The first one is the version by Couper:

$$\begin{array}{ll} \mathrm{C} & \left\{\begin{array}{l} \mathrm{O}\cdots\mathrm{OH} \\ \mathrm{H}_2 \end{array}\right. \\ \vdots & \\ \mathrm{C} & \qquad\cdots\mathrm{H}_3 \end{array}$$

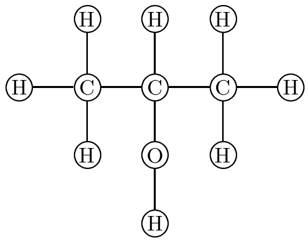
The next drawing is due to Loschmidt



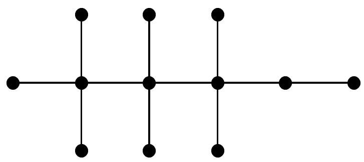
But the solution (which was in fact already in the hands of Loschmidt and Couper, as you can easily see, except that Couper had a different opinion about the weight of oxygen, and so he put two such atoms instead of one) was not clear until Alexander Crum Brown introduced the following notation which in modern terms is the *graph theoretical notation* for molecules. The given example demonstrates why there are in fact *two* molecular graphs corresponding to alcohol with the chemical formula C_3H_7OH:



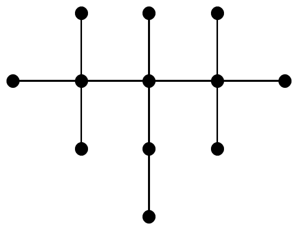
and



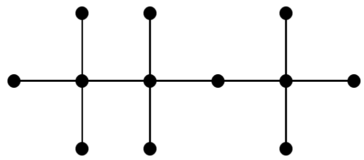
The corresponding alcohols differ, for example, in their boiling points which are 97.1 °C and 82.4 °C, respectively. If we leave out the names of the atoms then we immediately obtain the corresponding *graphs:*



and

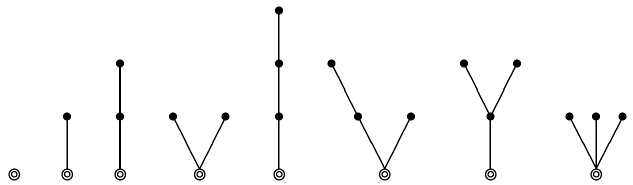


There are altogether *three* graphs with 3 carbon atoms, 8 hydrogen atoms and a single oxygen, but one of them does not represent an alcohol, since in this third case the oxygen is not connected to a hydrogen atom. For sake of completeness, here is the third graph as well:



2 Graphs and molecular graphs

In parallel with this development in chemistry, the mathematician A. Cayley considered so-called *rooted trees* ([4]). He saw the connection with chemistry, and he recognized that the number of *rooted* trees with root degree ≤ 3 is exactly the number of isomers of alcohols ([5]), since $C_nH_{2n+1}OH$, for natural numbers n, is the formula for alcohol, which always has a substructure of the form COH. If that substructure is identified with the root of the rooted tree, we get the desired and canonic bijection. The rooted tree corresponds to the skeleton of the alcohol, which is the molecule without the hydrogen atoms. Here are the smallest rooted trees, they all have root degree ≤ 3. It is easy to obtain the corresponding alcohols. The two rooted trees with exactly two vertices $\bullet$ correspond to the alcohols of formula C_3H_7OH:



At about the same time J. J. Sylvester published a note (in *Nature* **17** (1877-1878), 284) entitled *Chemistry and Algebra* [20], where he introduced the name *graph* for this kind of mathematical structure, *and he took this name from chemistry.* The paper deals with invariants and mentions the connection to chemistry by stating

> Every invariant and covariant thus becomes expressible by a *graph* precisely identical with a Kekuléan diagram or chemicograph.

On the side of chemistry a lot of activities was devoted to the *enumeration* of chemical isomers. In a paper of Lunn and Senior the first connections between this kind of a problem and group theoretic methods were mentioned. But the main clarification is due the famous paper

> G. Pólya: Anzahlbestimmungen für Gruppen, Graphen und Chemische Verbindungen. *Acta Mathematica* **68**, (1937), 145-254.

This paper, together with its long-time overlooked predecessor by J. H. Redfield [15], can be considered as the foundation of *graphical enumeration* or, more generally, of the enumeration *under group action* or even of *algebraic combinatorics.* (For further historical notes see e.g. [2].)

Molecular graphs are (usually connected) multigraphs, the points of which are colored by atom names. This will be described in more detail now since we need it to prepare the description of applications to combinatorial chemistry. It obviously is a challenge for mathematics to provide efficient algorithms and implementations for the *fast and redundancy free generation of the molecular graphs corresponding to a molecular formula and (optionally) further conditions like prescribed or forbidden substructures, ring sizes etc.*

In this paper we introduce *labeled multi-graphs* — like many other discrete structures and following the ideas of Pólya — as mappings. For example, a labeled multigraph on p points and with edge multiplicities at most equal to $m-1$, can be considered as a mapping

$$\gamma : \binom{p}{2} \to m,$$

where $p := \{0, 1, \ldots, p-1\}$ means the set of numbers of the points, while $\binom{p}{2} = \{\{0,1\}, \{0,2\}, \ldots, \{p-2, p-1\}\}$ denotes the set of *pairs of points.* The set $m := \{0, \ldots, m-1\}$ stands for the set of admissible bond multiplicities, while the value $\gamma(\{i,j\}) = k$ means that the points i and j of that particular graph are connected by a k-fold edge.

In order to describe *organic molecules* we take the usual model, identifying atoms with vertices and bonds with edges. The atomic types are defined by an additional mapping β from the set p of points of the multigraph to the set $\mathcal{E} := \{C, H, O, N, \ldots\}$ of admissible chemical elements. A molecular graph then is a pair (γ, β) of a graph and a coloring of the vertices with atomic types. (It is clear that not all such mappings are admissible since a coloring of a point by an atom name needs to be compatible with the valence of the atom in question.)

Furthermore we call

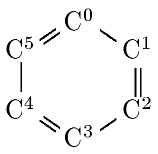
$$\eta : \binom{t}{2} \to \{0, \ldots, m-1\} \text{ with } t \subseteq p,\ \forall i,j \in t:\ \eta(\{i,j\}) = \gamma(\{i,j\})$$

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

Following Pólya, we denote by Y^X the set of all the mappings from X to Y, hence a labeled molecular graph turns out to be a pair

$$(\gamma, \beta), \text{ where } \gamma \in m^{\binom{p}{2}}, \text{ while } \beta \in \mathcal{E}^p.$$

For example, the labeled benzene ring, the *skeleton* of the benzene molecule with the hydrogen atoms left off is



The numbers of the points are indicated as upper indices of the carbon atoms (which are the atom types with which the atoms are colored by β). More formally, this molecular graph is the pair (γ, β), where γ and β are as follows:

pairs	$\{0,1\}$	$\{0,2\}$	$\{0,3\}$	$\{0,4\}$	$\{0,5\}$	$\{1,2\}$	$\{1,3\}$	$\{1,4\}$	$\ldots$
$\gamma(\{i,j\})$	1	0	0	0	2	2	0	0	$\ldots$

points	0	1	2	3	4	5
$\beta(i)$	C	C	C	C	C	C

We shall use these labeled molecular graphs later on in order to describe the generation of particular combinatorial libraries.

But let us mention here, that *a molecule is an unlabeled molecular graph,* since the numbering of the atoms does not really matter, which means an orbit of the symmetric group on the set of all labeled molecular graphs corresponding to a particular molecular formula.

In the past this graph theoretic model of molecule lead to very many publications using discrete mathematics in chemistry. The development of such applications of discrete mathematics, in particular of graph theory, in chemistry during the last twenty years, say, can be seen while reading the journal MATCH (communications in mathematical and in computer chemistry), founded in 1975 during a conference on "Graph Theory in Chemistry" by the organizer of that meeting, O. E. Polansky, Max-Planck-Institute, Mülheim, Germany. Together with him and André Dreiding, one of us (A.K.) was one of the organizers of that meeting, and with pleasure we mention that this journal still exists, it may be the worldwide oldest journal for (discrete) mathematical chemistry, and one of the organizers of the present conference, Patrick Fowler, is in its editorial board, too. When you read MATCH, then you can clearly see, that

- during the first ten years, say, the main emphasis was laid upon basic examination of the graph theoretic model of molecules and its applications, counting of isomers, eigenvalues of graphs, topological indices etc.

- then the first generators of molecular graphs corresponding to a given molecular formula and (optional) further conditions were described (see vol. 27), that were available on cheap and efficient computers. (The very first generator was in fact implemented in the USA by Lederberg and his co-workers in the famous DENDRAL project. It already used the main ingredients of today's generators, but at that time, computers were not yet efficient or cheap enough in order to make DENDRAL as successful as it should have been.) These generators serve for *molecular structure elucidation* by providing all the molecular graphs that correspond to a given set of data.

- and nowadays, since efficient generators are available, the next step is done already. 3D-placements are considered, conformations are classified, combinatorial chemistry will be simulated.

All in all I should like to say that the cross-fertilization between discrete mathematics and chemistry is in very good progress.

3 Combinatorial chemistry

Rapid technical screening methods have been developed recently which allow quickly to test thousands of chemical substances if they are helpful or not. Thus the results of mass synthesis (in contrast to classical synthesis of a single substance) can rapidly be searched for an efficient pharmaceutical drug, say. Correspondingly, automatic devices for controlled mass synthesis are available now.

This is a challenge for mathematicians to provide methods that allow a simulation of such experiments in order to see *in advance* if a planned mass synthesis can deliver an interesting substance, at least in principle. The history of mass synthesis is described in the review article

> I. Ugi: Fast and permanent changes in preparative and pharmaceutical chemistry through multicomponent reactions and their 'libraries'. *Proc. Estonian Acad. Sci. Chem.,* 1995, **44**,4,237-273.

Let us pick some of the crucial points of these developments from this article:

- 1850 Strecker discovers the first 3-component-reaction.

- 1929 Bergs and Bucherer find the first 4-component-reaction.

Then *isocyanide chemistry* was invented which completely differs from the rest of organic chemistry: Isocyanides are the only stable organic chemical compounds which contain divalent carbon C^{II}. This is always formed from a starting material with a functional group of C^{IV}. An important step in this branch of chemistry was made 1948 by Rothe who found antibiotic *Xanthocilline.* In 1957, Ugi et al. found the first widely applicable methods of isocyanide preparation and *isocyanide chemistry really began.*

- 1959 Ugi et al. find a four-component-reaction of isocyanides: U-4CR.

One of the first attempts to apply the U-4CRs was made by the Swedish company AB Astra. It was preparation of Xylocain, that is still one of the most widely used dental anaesthetics in the world.

- 1961 Ugi suggests first "liquid library" in book on Isonitrile Chemistry: Obtainable from U-4CR using 40 different compounds (so that, in principle, 2 560 000 different chemical products can be formed)

In 1966 Bodanszky and Ondetti recognized that U-4CRs would have great advantages in the synthesis of peptide derivatives, 5-CR's and 6-CR's were invented. The first 7-CR was reported in 1993 by Ugi and Dömling. It turned out that quite *often the MCRs proceed in better yields, if the components are not just mixed but are added sequentially!*

- 1982 Furka et al. introduce peptide libraries,

- 1993 Still/Yoon introduce a peptide library of 117 649 different members.

4 Combinatorial libraries

The aim of an experiment in combinatorial chemistry is the synthesis of a suitable *library* of molecules. There are single and multistep methods to do that, and they all use well defined *building blocks* and known *reactions* for the generation of such a library which is then *screened* in order to find synthesized molecules with the required medical or biological activity.

A mathematical model for a chemical reaction between two molecules (there are, of course, much more difficult cases, but for the purposes here, it suffices to consider two-component syntheses of the form $\mathrm{A}+\mathrm{B} \to \mathrm{C}$) is a *reaction scheme,* say

$$((\eta_1, \beta_1),\ (\eta_2, \beta_2),\ \rho),$$

which is a pair of labeled molecular graphs, together with a mapping ρ that describes the reaction in the form of a matrix[2]:

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

For example, peptides are built by joining amino acids via condensation of the carboxyl group (COOH) and the amid group ($\mathrm{NH_2}$). The building blocks (η_i, β_i) contain a substructure of the following form:

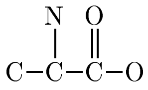
$$\begin{array}{ccc} & {}^{3}\mathrm{N} & \mathrm{O}^{4} \\ & | & \| \\ -{}^{1} & \mathrm{C}-\mathrm{C}^{2} & -\mathrm{O}^{5} \end{array}$$

[2] This definition is a simplification of the situation and is only used for a formalization of the construction problem discussed below. For more sophisticated purposes more comprehensive approaches like the algebra of *be*- &*r*-matrices of [6] are necessary.

The condensation is represented by the mapping

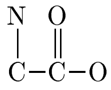
$$\rho = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ -\infty & -\infty & -\infty & -\infty & -\infty \end{pmatrix}$$

We consider, for example, the amino acids



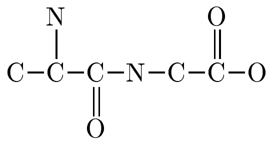
Alanin

and

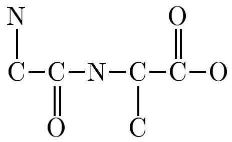


Glycin

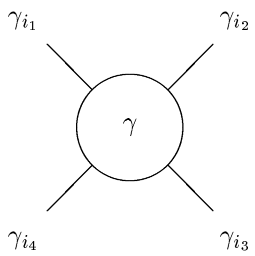
(For sake of simplicity we took of the numbering of the atoms which is supposed to be — in the common substructure — the very same as indicated above.) Taking Alanin first gives



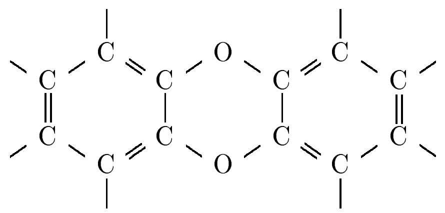
While Glycin as η_1 yields



This was a case of so-called *single-attachment.* The case of *multiple attachment* is, of course, more complicated. Consider the following core with several reaction sites and a number of ligand compounds, assuming, for sake of simplicity, that each of them contains exactly one substructure isomorphic to (η_2, χ_2) and that the sites are all equivalent. Here is a sketch of such a situation where there are exactly four equivalent sites:



From the mathematical point of view this situation is the very same situation as in the case when we want to obtain *permutational isomers,* say the 22 isomers of dioxin which has the skeleton



E. Ruch and co-workers have shown [17, 18], how permutational isomers are bijectivley related to double-cosets. The set of permutational isomers of the dioxin is Jr.bijective to the set of double cosets

$$V_4 \backslash S_8 / S_4 \oplus S_4.$$

(The Kleinian four group V_4 comes from the fact that this is the symmetry group of the skeleton, while the subgroup $S_4 \oplus S_4$ is due to the fact that we have to distribute 4 chlorine and 4 hydrogen atoms among the 8 free valences on which the symmetric group S_8 acts transitively.) This set of double cosets is of order 22, but even more: Using double cosets we can *construct* and not just count, a complete system of representatives of these orbits, i. e. we can construct the molecular graphs corresponding to these 22 permutational isomers of dioxin!

5 A mathematical model

Slightly more abstract than the consideration of permutational isomers is the approach used by Pólya [13] in order to consider isomers and many other situations. He considered sets

$$Y^X := \{f : X \to Y\}$$

of mappings together with actions of a group G on X and the corresponding action of G on Y^X :

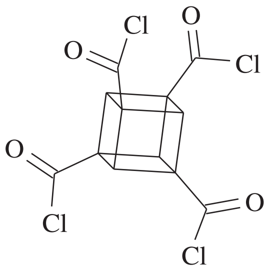
$$G \times Y^X \to Y^X : (g, f) \mapsto f \circ g^{-1}.$$

It is not difficult to see that also here the set of orbits of G on Y^X consisting of mappings of a fixed weight is bijectively related to a set of double cosets.

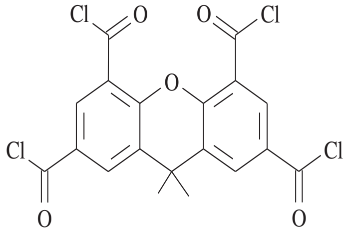
Let us apply this approach to some of the famous examples from combinatorial chemistry, taken from

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

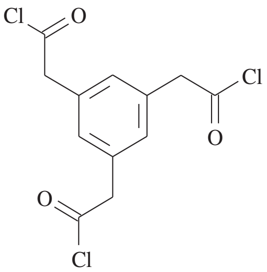
Here are the cores that are considered:



I

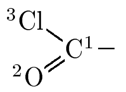


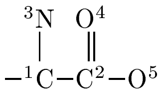
II



III

The first (upper left) is a cuban derivative, the second one a xanthene, and the third one is a benzene triacid chlorine. The reaction scheme consists of the substructures





and the matrix

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

6 The size of the library

The situation just described can easily be reformulated in terms of Pólya's approach. The four active sites of the cuban or of the xanthen, as well as the three active sites of the benzene triacid chlorine, respectively, form the set X while the set Y consists of a subset of the set of the twenty natural amino acids. Attaching amino acids at the active sites is the same as forming a mapping $f \in Y^X$, and the equivalence classes of attachments are the orbits of Y^X under the symmetry group of the core. The general formula for the number of such orbits with respect to the symmetry group G is obtained from the well-known Lemma of Cauchy-Frobenius. If applied to the Pólya situation, where a group

G acts on a set X and therefore also on Y^X, we obtain the following formula for the number of orbits of G on Y^X :

$$\frac{1}{|G|}\sum_{g\in G}|Y|^{c(g)},$$

where $c(g)$ means the number of cyclic factors of g on X. For example in the case of the cuban and if we use as symmetry group the full symmetric group S_4, then we obtain for the number of orbits the expression [23]

$$\frac{1}{24}\left(|Y|^4+6\cdot|Y|^3+11\cdot|Y|^2+6\cdot|Y|\right).$$

If we take Y to be the set of 20 amino acids, then we obtain 8 855 orbits of S_4 and 13 700 orbits of A_4. A table with all the values, for the alternating group, is given in Table 1.

In the xanthen case, the symmetry group is of order 2, and so the formula for the size of the library turns out to be

$$\frac{1}{2}(|Y|^4+|Y|^2).$$

The full library, i. e. $|Y|=20$ gives the number 80 200.

In the case of the benzene triacid chlorine, the symmetry group is the cyclic group of order 3 on the set of three active sites, and so the size of the library is

$$\frac{1}{3}(|Y|^3+2\cdot|Y|).$$

For $|Y|=20$, we obtain the number 2680.

7 The elements of the library by weight

The evaluation of the size of the library is, of course, only the very first step in order to get an idea how big the library in principle might be. Much more important is the *construction of the elements of it.* A first step towards this refinement can also be done using Pólya's theory of enumeration under finite group actions.

Pólya's Theorem gives the numbers of orbits of G on Y^X by weight (=multiplicities of amino acids) in terms of the so-called *cycle index polynomial:*

$$Cyc(G,X):=\frac{1}{|G|}\sum_{g\in G}\prod_{i=1}^{|X|}x_i^{a_i(g)},$$

if $a_i(g)$ denotes the number of orbits of $\langle g\rangle$ on X which are of length i, i.e. the number of i-cycles of g on X. The number of orbits of G on Y^X which consist of

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

Table 1: Sizes of libraries depending on the number of building-blocks used

mappings f that take the value $y \in Y$ with multiplicity b_y is then the coefficient of the monomial $\prod_{y \in Y} y^{b_y}$ in the polynomial

$$Cyc(G \mid \sum y^i)$$

which arises from the cycle index by replacing the indeterminate x_i by the polynomial $\sum_y y^i$.

Using SYMMETRICA[3], we obtain, for example, in the dioxin case the following generating function for the permutational isomers:

$$Cl_8 + 2H_1Cl_7 + 10H_2Cl_6 + 14H_3Cl_5 + 22H_4Cl_4 + 14H_5Cl_3 + 10H_6Cl_2 + 2H_7Cl_1 + H_8.$$

The summand $22H_4Cl_4$ means that there are exactly 22 permutational isomers containing both 4 hydrogen and 4 chlorine atoms.

There exists an interesting reformulation in terms of irreducible representations $[\alpha]$, α a partition of the number $|X|$, for short: $\alpha \vdash |X|$, of the symmetric group S_X and the corresponding Schur functions $\{\alpha\}$:

$$Cyc(G \mid \sum y^i) = \sum_{\alpha \vdash |X|} (IG \uparrow S_X, [\alpha]) \cdot \{\alpha\}.$$

($IG \uparrow S_X$ means the representation of S_X, induced by the identity representation IG of the subgroup $G \leq S_X$.)

The cuban case is trivial since the symmetry group is the symmetric group S_4, and so the resulting generating function is simply the Schur polynomial $\{4\}$.

A bit less trivial is the situation of the triacid chlorine, where the symmetry group G is the cyclic group of order 3, and so the resulting generating function for the elements of the library by weight is

$$Cyc(C_3) = \{3\} + \{1^3\}.$$

In the case when Y consists of 4 amino acids, say of types which we denote by a, b, c and d, then we obtain, using the computer algebra package SYMMETRICA:

$$d^3 + cd^2 + c^2d + c^3 + bd^2 + 2\,bcd + bc^2 + b^2d + b^2c + b^3 + ad^2$$
$$+2\,acd + ac^2 + 2\,abd + 2\,abc + ab^2 + a^2d + a^2c + a^2b + a^3$$

The summand $2acd$, for example, shows that there are two essentially different attachments of an amino acid of type a, an amino acid of type c and an amino acid of type d. This is, of course, obvious, but it shows that we can use Schur polynomials in order to get a refined count of the elements of the combinatorial library.

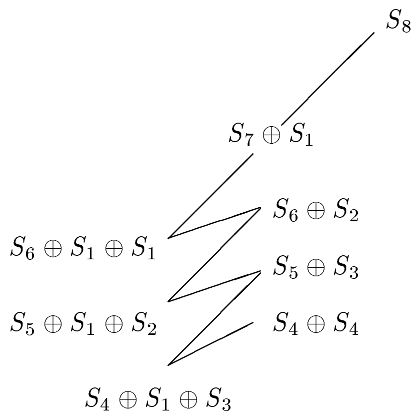
[3]SYMMETRICA can be obtained via Internet from http://www.mathe2.uni-bayreuth.de

8 The generation of the library

Here we can use — at least for our particular examples taken from [3] — the very same method as in the case of the construction of permutational isomers: *there exists a canonical bijection onto a set of double cosets.* The reason is that the symmetric group S_X on X acts transitively on the set of mappings $f \in Y^X$ which have a prescribed weight (which means prescribed multiplicities $f^{-1}(y)$ for the values $y \in Y$, for our chemical examples this means: prescribed multiplicities of amino acids). Hence the orbit of a fixed mapping (= labeled molecule) under that symmetric group is essentially the same as the set $S_X/(S_X)_{f_0}$ of left cosets of the stabilizer $(S_X)_{f_0}$. Therefore the set of orbits $G(f)$ of the symmetry group G (of the cubane derivative, say) on this orbit $S_X(f_0)$ can be mapped bijectively onto the corresponding set of double cosets

$$\varphi\colon \mathrm{Orb}(G, S_X(f_0)) \to G\backslash S_X/(S_X)_{f_0} : G(gf_0) \mapsto Gg(S_X)_{f_0}.$$

Thus, once we have obtained a complete system of representatives of these double cosets we obtain, by an application of the inverse mapping φ^{-1}, a complete system of molecules with prescribed multiplicities of the admissible amino acids. This can be carried out, as it is described in several papers in detail, by going up and down along a leporello of Young subgroups, from S_X to $(S_X)_{f_0}$, a method that applies in the same way to the construction of full sets of permutational isomers. This method was invented by B. Schmalz (see [19]), applied to the construction of combinatorial designs and called the *ladder game.* Here is a selfexplanatory example of such a folder which can be used in the dioxin case, in order to construct all the permutational isomers successively:



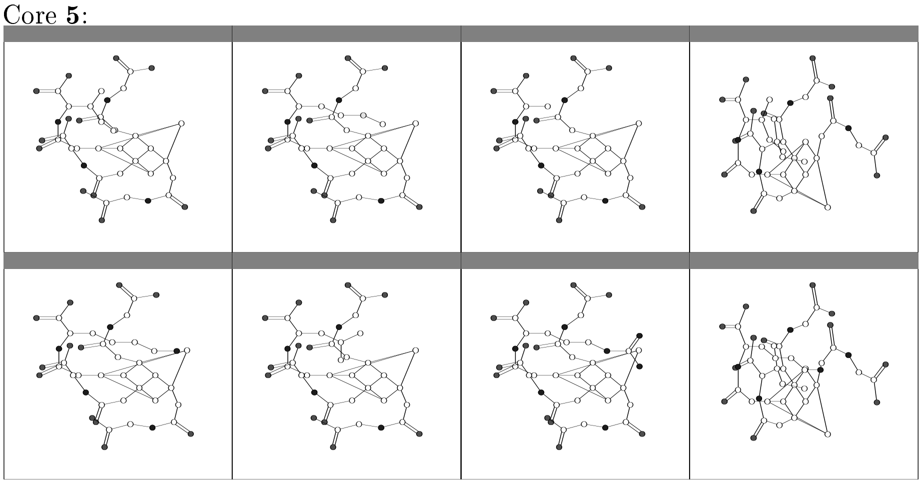
The method used is to start from above with the subgroup S_8 of S_8. It has exactly one left coset and so there is exactly one orbit of V_4 on this set of left cosets, and hence exactly one isomer containing 8 hydrogen atoms and no chlorine. In the second step this one-elment set of left cosets of S_8 is split up into the 8 left cosets of $S_7 \oplus S_1$. It turns out that V_4 has exactly two orbits on this set, representatives of which yield the two different permutational isomers containing 7 hydrogen atoms and 1 chlorine, ..., until we reach $S_4 \oplus S_4$, obtaining from the orbits of V_4 on the set of left cosets the corresponding permutational isomers with 4 hydrogen and 4 chlorine atoms. The "Homomorphism Principle", see [11],[12], shows how we can obtain a transversal of this bigger set of orbits from the smaller set we had before. This principle is combined with orderly generation which is popular in constructive combinatorics (see [14],[7]).

This rather general and efficient approach allows a very rapid generation of the combinatorial libraries. In the three cases from our example (taking the natural amino acids as building blocks), we examined a computing speed of 40 structures per second on a Pentium 90 MHz PC, writing all solutions to the hard disk. Figure 1 shows six molecules from each of the three libraries. More details are given in [22, 23, 24].

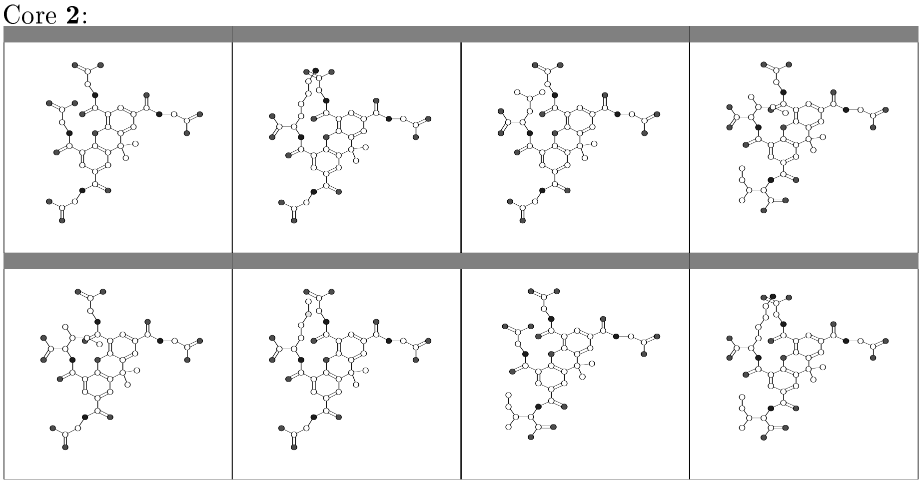
9 Summary

We have shown that enumeration under finite group action can cover famous examples of combinatorial chemistry. The methods used are Pólya's Ansatz to choose suitable sets X, Y and a finite group G acting on X. The examination of the induced action of G on the set of mappings Y^X allows to evaluate the total number of elements in a combinatorial library arising from a core with the set X as set of active sites by attaching elements of a set Y of building blocks, and with respect to the symmetry group G of X. This number is obtained by an easy application of the Cauchy-Frobenius Lemma. Moreover we can obtain (by an application of Pölya's Theorem) a generating function for the elements of the library, which enumerates these elements by weight. Finally we can even construct the library using the double coset reformulation of the orbits, the Homomrophism Principle and Orderly Generation. The corresponding software is under development.

Core **5**:



Core **2**:



Core **3**:

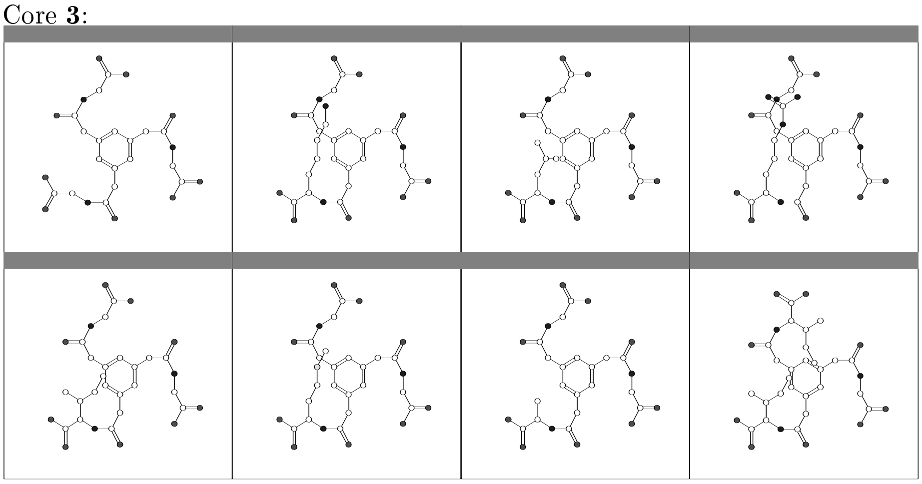


Figure 1: Extracts from the combinatorial libraries produced from the structures **5**, **2** and **3** and the natural amino acides

References

[1] C. Benecke, R. Grund, R. Hohberger, A. Kerber, R. Laue and T. Wieland. *Chemical Isomerism, a Challenge for Algebraic Combinatorics and for Computer Science. Proc. of IEEEC 1995,* LN in Computer Science, Springer-Verlag 1996.

[2] N.L. Biggs, K.E. Lloyd, and R.J. Wilson. *Graph theory 1736-1936.* Clarendon Press, 1977.

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

[4] A. Cayley. On the theory of the analytical forms called trees. *Phil. Mag.*, **13**, pp. 172–176, 1857.

[5] A. Cayley. On the analytical forms called trees, with application to the theory of chemical combinations. *Rep. Brit. Assoc. Adv. Sci.*, **45**, pp. 257–305, 1875.

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

[7] I. A. Faradžev. Constructive enumeration of combinatorial objects. *Problèmes combinatoires et théorie des graphes (Colloq. Internat. CNRS, Univ. Orsay, Orsay, 1976)*, **260**, 131–135, 1978.

[8] R. Grund, A. Kerber, and L. Laue. Construction of discrete structures, especially isomers. to appear in *Discr. Appl Math.*

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

[10] A. von Humboldt. *Versuche über den gereizten Muskel- und Nervenfaser nebst Vermuthungen über den chemischen Prozess des Lebens.* Decker & Compagnie, Leipzig, und Heinrich August Rottmann, Posen, 1797.

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

[12] R. Laue. Construction of combinatorial objects – a tutorial. *Bayreuther Mathematische Schriften*, **43**, pp. 53–96, 1993.

[13] G. Pólya. Kombinatorische Anzahlbestimmungen für Gruppen, Graphen und chemische Verbindungen. *Acta mathematica*, **68**, pp. 145–253, 1937.

[14] R.C. Read. Every one a winner. *Ann. Discr. Math.*, **2**, pp. 107–120, 1978.

[15] H.J. Redfield. The theory of group-reduced distributions. *Am. J. Math.*, **49**, pp. 433–455, 1927.

[16] D.H. Rouvray. The Origins of Chemical Graph Theory. In: *D. Bonchev/D. h. Rouvray (eds.): Chemical Graph Theory, 1-39.* Mathematical Chemistry Series, vol. 1, Abacus Press, 1991.

[17] E. Ruch, W. Hässelbarth, B. Richter. Doppelnebenklassen als Klassenbegriff und Nomenklaturprinzip für Isomere und ihre Abzählung. *Theoretica Chimica Acta*, **19** (1970), 288-300.

[18] E. Ruch, D. J. Klein. Double cosets in chemistry and physics. *Theoretica Chimica Acta*, **63** (1983), 447-472.

[19] B. Schmalz. Verwendung von Untergruppenleitern zur Bestimmung von Doppelnebenklassen. *Bayreuther Mathematische Schriften*, **31**, pp. 109–143, 1990.

[20] J.J. Sylvester. Chemistry and Algebra. *Nature*, **17**, 1878. a

[21] I. Ugi, A. Dömling, B. Gruber, M. Heilingbrunner, C. Heiss, W. Hörl. Formale Unterstützung bei Multikomponentenreaktionen – Automatisierung der Synthesechemie. In: *R. Moll (Hrsg.): Software-Entwicklung in der Chemie 9,* S. 114-128, GDCh, Frankfurt am Main, 1995.

[22] T. Wieland. Mathematical Simulations in Combinatorial Chemistry. *MATCH*, **34**, pp. 179–206, 1996.

[23] T. Wieland. Combinatorics of Combinatorial Chemistry. *J. Math. Chem.*, **21**, pp. 141–157, 1997.

[24] T. Wieland. Konstruktionsalgorithmen bei molekularen Graphen und deren Anwendung. *MATCH*, **36**, pp. 5–155, 1997.