MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **80** (2018) 733-744

ISSN 0340 - 6253

MOLGEN, a Generator for Structural Formulas

Adalbert Kerber

Department of Mathematics, University of Bayreuth,
kerber@uni-bayreuth.de

(Received March 10, 2017)

Abstract

MOLGEN, a generator of all structural formulas corresponding to a given molecular formula, is described. Its history is mentioned, applications are touched, and further research projects are suggested.

From the history of Mathematical Chemistry

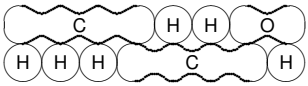
The first step towards Mathematical Chemistry was the attempt to develop a mathematical model of the concept of molecule using an *arithmetic description* of a molecule, a *molecular formula,* e.g. C_6H_6. But this does not suffice to distinguish molecules uniquely, as Alexander von Humboldt (1769-1859) *stated*, cf. [25], in vol. I, page 128, of [16], published in 1797:

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

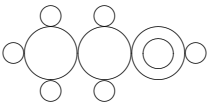
Thus, Humboldt states that chemical compounds (*Körper*) should exist that contain the same quantities of oxygen, hydrogen, carbon, nitrogen or metal while they are essentially different. On page 127 he even uses the word 'Bindung' (bond). This statement turned out to be true a quarter of a century later. Humboldt had met in Gay-Lussac's laboratory in Paris a young student, Justus Liebig. Humboldt saw immediately his talent and recommended him for a professorship in chemistry at the University of Gießen which Liebig received when he was just 21 years old. Liebig (1803-1873) proved Humboldt's statement in 1823/1824 by an examination

of the silver salt of fulminic acid, comparing it with silver isocyanate which F. Wöhler had examined.

This led to a description of molecules as *interaction models*, a *topological description* of molecule, a *structural formula*. Sketches of molecules were used, *chemicographs*, as they were called by Kekulé (1829-1869). Here is his sketch of C_2H_5OH:



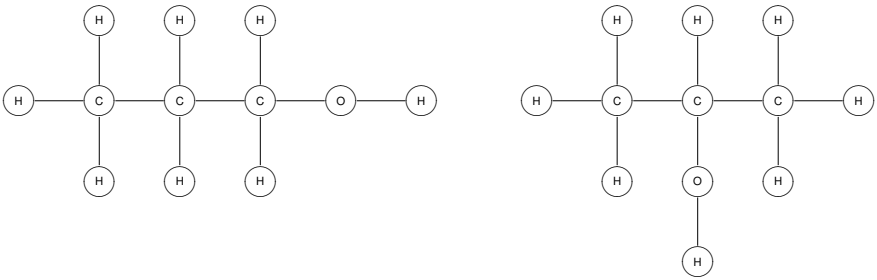
Loschmidt (1821-1895) used the following drawing for that molecule:



It contains touching points of circles, and if we replace these by lines and the circles by bullets, we get what mathematicians call a *graph*. Alexander Crum Brown (1838-1922), who used such lines instead of touching points, motivated the use of the resulting graphs in his thesis (1861) 'On the theory of chemical combination' as follows:

> It does not seem to me improbable that we may be able to form a mathematical theory of chemistry, applicable to all cases of composition and recomposition.

In 1864 he published an important paper [4] 'On the Theory of Isomeric Compounds', in which he, using his graphical formulas, discussed various types of isomerism. For example, he sketched C_3H_7OH in form of a graph, the vertices labeled by atom names, and he noticed that there are two essentially different sketches possible (there is in fact a third constitutional isomer of the molecular formula C_3H_8O, but it does not contain a hydroxyl group):



The mathematician J.J. Sylvester (1814-1897) connects in his paper 'Chemistry and Algebra' [36] algebraic invariants and chemical molecules, and he writes:

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

Thus, sketches of molecules, considered as interaction models (atoms interact by sharing electrons), gave rise to graph theory and Mathematical Chemistry. And, conversely, sketching molecules by graphs immediately made clear that there can exist several essentially different interaction models for a molecule with a given molecular formula, in terms of chemistry: that the phenomenon *isomerism* (a name that apparently J.J. Berzelius (1779-1848) introduced about 1830, cf. [1] or the reprint in German [2]) exists. Hence, the connection between mathematics and chemistry is not a one-way-street: Considering molecules as interaction models, gave rise to graph theory and, vice versa, graph theory made immediately clear that isomerism exists and, in mathematical terms, a *topological description* of molecules is necessary, the *second* step of Mathematical Chemistry.

For example, the molecular graphs of the historically first isomers, discovered by v. Liebig and Wöhler, look as follows:

> The molecular graph $\mathsf{Ag}-\mathsf{O}-\mathsf{N}^{+}\equiv\mathsf{C}^{-}$ describes the silver salt of fulminic acid, while silver isocyanate is represented by $\mathsf{Ag}-\mathsf{N}=\mathsf{C}=\mathsf{O}$.

Moreover, this discovery of a model of molecule as multigraph with given valences of its nodes stimulated mathematical research on the generation of graphs with given properties and, more generally, the enumeration (counting and generation) of finite structures, e.g., of all the structural formulas that correspond to a given molecular formula. The question arises how one can generate all of them since $\mathsf{C}_6\mathsf{H}_6$, for example, has 217, while $\mathsf{C}_{10}\mathsf{H}_{10}$ has already $369{,}067$ constitutional isomers.

A seminal paper in this direction of constructive theory, or, better say, classification of finite structures, including Mathematical Chemistry, is [32] by G. Pólya. And it should be mentioned that the classification of molecules by constitutional isomers led in 1975 during a meeting at the MPI at Mülheim to the birth of MATCH, due to O.E. Polansky.

MOLGEN and the generation of isomers

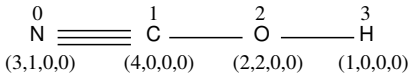
Here are occasions where chemists need the generation of constitutional isomers of molecules:

1. *Molecular structure elucidation* uses the generation of all structural formulas corresponding to a given molecular formula. For this purpose the generator DENDRAL was developed at Stanford University by E. Feigenbaum, B.G. Buchanan, J. Lederberg, C. Djerassi and their team for structure elucidation of analytes to be taken on the planet Mars. This was a pioneering project in artificial intelligence of the 1960s, although at that time it would have been impossible to do such complex computations on board of a spacecraft. They underestimated what generation of isomers needs on the side of the computers, e.g., MOLGEN 4, cf. [9], consists of about 160,000 lines of code. During discussions with Harald Brown, a member of that team, it became apparent that a lot of computer algebra (in particular group theory) is involved. We, working at J. Neubüser's chair, on group theory, representation theory and discrete structures at the RWTH Aachen in the seventies, decided therefore to develop a generator that could run on a PC.

2. *Comparing chemistry patents containing Markush structures with respect to overlap* is another field where generation of isomers and congeners is necessary. This requires the generation of the molecules for both patents, even more: a structural formula of the same isomer has to be obtained in both cases by the same data, i.e., *in canonical form*, a difficult problem in classification of discrete structures, but MOLGEN *does that automatically*!

3. There is also a reaction–based generation, see [37, 38, 21]. MOLGEN–COMB [10, 11] and MOLGEN–QSPR [20, 39, 40] can help to optimize an experiment of combinatorial chemistry *in advance*.

Molecular graphs are stored by MOLGEN in their labeled form, i.e., the atoms are numbered, for example, cyanic acid is described by a 'labeled multigraph, the nodes of which are colored by an atom state', by



.

The quadruple (3,1,0,0) describes the *atom state* of the nitrogen atom. The entry 3 means the

valence, 1 is the number of free electron pairs, the first entry 0 indicates the charge while the second 0 means that there is no unpaired electron present.

The first steps in our group towards generation of constitutional isomers were made by D. Moser and lead to our first generator, cf. his master's thesis [31], 1987, Bayreuth. Various further steps are described in master's and doctoral theses, and many papers were published. The summary, including a long list of publications and descriptions of the methods used and of their applications was recently published in the book [22].

We mentioned the application of the generation of constitutional isomers in molecular structure elucidation, cf. [41, 42, 26, 17, 27, 23, 34, 35, 33, 30]. Another one is the generation of amino acids and other biomolecule analogs in order to study questions arising from Astrobiology, see e.g. [29, 18, 3, 28], and the search for molecules of life on solar system bodies, which also approaches the aim of the original Dendral project, cf. the following link to a corresponding presentation by M. Meringer:

http://elib.dlr.de/102783/1/GRC160120ChemSpaceExp.pdf

MOLGEN exists in several versions

- MOLGEN 3.5 is a fast generator of molecular structures, a scientific software for chemists. Its basic version runs under Windows 7+. There is a handbook available, written by R. Gugisch, A. Kerber, A. Kohnert, R. Laue, M. Meringer, C. Rücker, A. Wassermann:

  http://molgen.de/documents/molgen_manual_35.pdf

- MOLGEN 4 is more flexible. It has interfaces to MOLGEN-MS and MOLGEN-QSPR for molecular structure elucidation, the results can immediately be used for QSPR/QSAR studies. Experiments of combinatorial chemistry can be optimized via mathematical simulations. Virtual combinatorial libraries can be constructed. The user can prescribe reactions and partners of reactions, see [37, 38, 21]. It is possible to check if a given library of molecules is contained in a virtual library. In order to predict physico-chemical or biological properties of a virtual library, there are at present 708 molecular descriptors implemented. Further information can be found in the manual by J. Braun, A. Kerber, R. Laue, M. Meringer, C. Rücker, see

  http://molgen.de/download/MOLGEN-QSPR_User_ Guide.pdf

Regression analysis allows to correlate molecular descriptors of the molecules in a generated (virtual) library with measured properties of a real library. The statistics package R can easily be applied. Multilinear classification and regression, regression trees, artificial neural networks and support vector machines are available for a search for candidates with prescribed target-properties.

- MOLGEN 5.0 combines efficiency (MOLGEN 3.5) with flexibility (MOLGEN 4). To achieve this, the software was reimplemented based on a new concept. It is available for Linux as well as for Windows systems and was developed by R. Gugisch, A. Kerber, A. Kohnert, R. Laue, M. Meringer, C. Rücker, A. Wassermann, see

  http://molgen.de/documents/manual_molgen50.pdf

  It allows to use a *fuzzy molecular formula*, for example, you may enter the following formula, containing atom types together with intervals of their multiplicities,

  C1-8H0-16N0-6O0-4

  In addition you can enter the condition 100 for the molecular mass and you will obtain in about 10 seconds the corresponding *molecular library* containing 33,537 structural formulas that were *constructed* and can be displayed and stored, cf. [15]. These structural formulas contain all the connectivity isomers that correspond to 16 molecular formulas covered by the fuzzy formula. A table of further sizes of molecular libraries, obtained from the fuzzy formula together with other conditions on the molecular mass is

mass	MF	MG	MGNAD	BS	MS
30	2	2	2	2	2
40	3	5	5	5	1
50	1	7	7	1	1
60	6	47	47	25	12
70	6	380	380	84	31
80	6	1,645	1,644	100	23
90	11	5,849	5,818	107	28
100	16	33,627	33,537	710	154

  where the abbreviations mean

  MF: number of molecular formulas,

  MG: number of corresponding molecular graphs,

  MGNAD: number of structural formulas without aromatic duplicates,

BS: number of isomers contained in the Beilstein Reaxys database,

MS: number of spectra in the NIST mass spectral library.

Here you can find links to all the manuals:

http://molgen.de/products.html

- A reduced version can be freely used online via

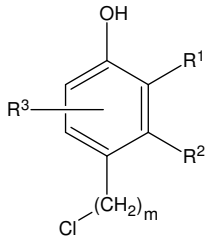
  http://molgen.de/online.html

  The computation time is limited to 5 minutes per job and the possibilities to prescribe and to forbid the occurence of substructures in the generated structures are not available in this internet version. The filter for aromatic duplicates is turned on by default.

Applications

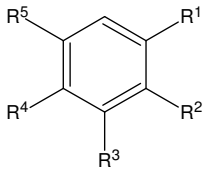
- Molecular structure generation is crucial whenever the unknown chemical compound may not be contained in the available databases. For a detailed description of an application of MOLGEN–MS to structure elucidation cf. [41]. The respective molecules were contaminants in groundwater of Bitterfeld, Germany: 150 spectra, 42 identified using NIST alone. 32 of these were confirmed using structure generation techniques. In addition, 20 further peaks were tentatively identified using structure generation techniques alone, resulting in a total of 62 tentative identifications.

- Here is a small artificial example how an application to the examination of patents might look like. Consider the Markush structure claiming compounds



  where R^1: CH_3 or C_2H_5, R^2: Alkyl (1–6 C atoms), R^3: NH_2, m : 1–3.

This formula covers a library $\mathcal{L}_1$ of altogether 396 structural formulas. We want to compare this with the Markush formula



where

R^1: CH_3, C_2H_5, OH,

R^2: Alkyl (1–6 C atoms),

R^3: OH, OCH_3, OC_2H_5, CH_3, C_2H_5,

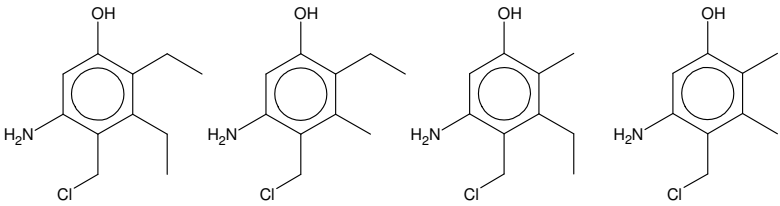
R^4: OH, CH_2Cl, NH_2,

R^5: H, CH_3, C_2H_5, NH_2.

An application of MOLGEN, cf. [19], shows that the corresponding patent library $\mathcal{L}_2$ consists of $|\mathcal{L}_2| = 5{,}939$ molecules. Moreover, we find that there is some overlap: Because of *canonic generation* we find in a fraction of a second

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

i.e., there are exactly four compounds claimed in both patents. The corresponding structural formulas are



An open problem

A generator for all stereoisomers corresponding to a given molecular graph is missing, although the theory is available, due to Dreiding, Dress, Gugisch, Rücker, cf. [5, 6, 7, 12, 13, 14]. Here are a few hints to the basic ideas and to relevant literature. Dreiding and Dress used *chirotopes* (strongly related to oriented matroids). Similar considerations are due to M. Klin, S. Tratch, N. Zefirov, cf. [24, 43, 44].

References

[1] J. J. Berzelius, On the composition of tartaric acid and racemic acid (John's acid of the Vosges, on the molecular weight of lead oxide, together with general observations on those bodies that have the same composition but different properties), *Kongliga Svenska Vetenskaps Academiens Handling (Transactions of the Royal Swedish Science Academy)* **49** (1830) 49–80.

[2] J. J. Berzelius, Über die Zusammensetzung der Weinsäure und Traubensäure (John's Säure aus den Voghesen), über das Atomengewicht des Bleioxyds, nebst allgemeinen Bemerkungen über solche Körper, die gleiche Zusammensetzung, aber ungleiche Eigenschaften besitzen, *Annalen der Physik und Chemie* **19** (1831) 305–335.

[3] H. J. Cleaves, M. Meringer, J. Goodwin, 227 Views of RNA: Is RNA unique in its chemical isomer space? *Astrobiology* **15** (2015) 538–558.

[4] A. Crum Brown, On the theory of isomeric compounds, *T. RSE* **23** (1864) 707–719.

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

[6] A. Dress, Chirotopes and oriented matroids, *Bayreuther Math. Schriften* **21** (1986) 14–68.

[7] A. Dress, A. Dreiding, H. Haegi, Classification of mobile molecules by category theory, *Stud. Phys. Theo. Chem.* **23** (1983) 39–58.

[8] M. E. Elyashberg, A. J. Williams, *Computer–Based Structure Elucidation from Spectral Data*, Springer, Berlin, 2015.

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

[10] T. Grüner, A. Kerber, R. Laue, M. Meringer, Mathematics for combinatorial chemistry, in: F. Keil, W. Mackens, H. Vob, J. Werther (Eds.), *Scientific Computing in Chemical Engineering II*, Springer, New York, 1999, pp. 74–81.

[11] R. Gugisch, A. Kerber, R. Laue, M. Meringer, J. Weidinger, MOLGEN-COMB, a software package for combinatorial chemistry, *MATCH Commun. Math. Comput. Chem.* **41** (2000) 189–203.

[12] R. Gugisch, Konstruktion von Isomorphieklassen orientierter Matroide, *Bayreuther Math. Schriften* **72** (2005) 1–129.

[13] R. Gugisch, A construction of isomorphism classes of oriented matroids, in: M. Klin, G. A. Jones, A. Jurisic, M. Muzychuk, I. Ponomarenko (Eds.), *Algorithmic Algebraic Combinatorics and Gröbner Bases*, Springer, Berlin, 2009.

[14] R. Gugisch, C. Rücker, Unified generation of conformations, conformers, and stereoisomers: A discrete mathematics-based approach, *MATCH Commun. Math. Comput. Chem.* **61** (2009) 117–148.

[15] R. Gugisch, A. Kerber, A. Kohnert, R. Laue, M. Meringer, C. Rücker, A. Wassermann, MOLGEN 5.0, a molecular structure generator, in: S. C. Basak, G. Restrepo, J. L. Villaveces (Eds.), *Advances in Mathematical Chemistry and Applications*, Elsevier, Sharjah, 2016, pp. 113–138.

[16] A. von Humboldt, *Versuche über die gereizte Muskel- und Nervenfaser, nebst Vermuthungen über den chemischen Prozeß des Lebens in der Thier- und Pflanzenwelt*, 2 volumes, Rottmann, Leipzig, 1797.

[17] S. Huntscha, T. B. Hofstetter, E. L. Schymanski, S. Spahr, J. Hollender, Biotransformation of benzotriazoles: Insights from transformation product identification and compound-specific isotope analysis, *Environ. Sci. Technol.* **48** (2014) 4435–4443.

[18] M. Ilardo, M. Meringer, S. Freeland, B. Rasulev, H. J. Cleaves, Extraordinarily adaptive properties of the genetically encoded amino acids, *Sci. Rep. UK* **5** (2015) 9414.

[19] A. Kerber, R. Laue, M. Meringer, An application of the structure generator MOLGEN to patents in chemistry, *MATCH Commun. Math. Comput. Chem.* **47** (2003) 169–172.

[20] A. Kerber, R. Laue, M. Meringer, C. Rücker, MOLGEN-QSPR, a software package for the study of quantitative structure property relationships, *MATCH Commun. Math. Comput. Chem.* **54** (2004) 187–204.

[21] A. Kerber, R. Laue, M. Meringer, C. Rücker, Molecules in silico: A graph description of chemical reactions, *J. Chem. Inf. Model.* **47** (2007) 805–817.

[22] A. Kerber, R. Laue, M. Meringer, C. Rücker, E. Schymanski, *Mathematical Chemistry and Chemoinformatics – Structure Generation, Elucidation and Quantitative Structure-Property Relationships*, de Gruyter, Berlin, 2014.

[23] A. Kerber, M. Meringer, C. Rücker, CASE via MS: Ranking structure candidates by mass spectra, *Croat. Chem. Acta* **79** (2004) 449–464.

[24] M. Klin, S. Tratch, N. Zefirov, 2D-configurations and clique-cyclic orientations of the graphs $L(K_p)$, *Rep. Mol. Theory* **1** (1990) 149–163.

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

[26] C. Meinert, E. Schymanski, E. Küster, R. Kühne, G. Schüürmann, W. Brack, Application of preparative capillary gas chromatography (pcGC), automated structure generation and mutagenicity prediction to improve effect-directed analysis of genotoxicants in a contaminated groundwater, *Environ. Sci. Pollut. Res.* **17** (2010) 885–897.

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

[28] M. Meringer, H. J. Cleaves, Exploring astrobiology using in silico molecular structure generation, *Phil. Trans. R. Soc. A* **375** (2017).

[29] M. Meringer, H. J. Cleaves, S. J. Freeland, Beyond terrestrial biology: Charting the chemical universe of α-amino acid structures, *J. Chem. Inf. Model.* **53** (2013) 2851–2862.

[30] M. Meringer, E. L. Schymanski, Small molecule identification with MOLGEN and mass spectrometry, *Metabolites* **3** (2013) 440–462.

[31] D. Moser, *Die computerunterstützte Konstruktion molekularer Graphen*, Diplom-arbeit, Universität Bayreuth, 1987.

[32] G. Pólya, Kombinatorische Anzahlbestimmungen für Gruppen, Graphen und chemische Verbindungen, *Acta Math.* **68** (1937) 145–254.

[33] E. L. Schymanski, C. M. J. Gallampois, M. Krauss, M. Meringer, S. Neumann, T. Schulze, S. Wolf, W. Brack, Consensus structure elucidation combining GC/EI-MS, structure generation, and calculated properties, *Anal. Chem.* **84** (2012) 3287–3295.

[34] E. L. Schymanski, M. Meringer, W. Brack, Matching structures to mass spectra using fragmentation patterns: are the results as good as they look?, *Anal. Chem.* **81** (2009) 3608–3617.

[35] E. L. Schymanski, M. Meringer, W. Brack, Automated strategies to identify compounds on the basis of GC/EI-MS and calculated properties, *Anal. Chem.* **83** (2011) 903–912.

[36] J. J. Sylvester, Chemistry and algebra, *Nature* **17** (1877-1878) 284.

[37] T. Wieland, Konstruktionsalgorithmen bei molekularen Graphen und deren Anwendung, *MATCH Commun. Math. Comput. Chem.* **36** (1997) 7–157.

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

[39] C. Rücker, M. Meringer, A. Kerber, QSPR using MOLGEN-QSPR: the example of haloalkane boiling points, *J. Chem. Inf. Comput. Sci.* **44** (2004) 2070–2076.

[40] C. Rücker, M. Meringer, A. Kerber, QSPR using MOLGEN-QSPR: the challenge of fluoroalkane boiling points, *J. Chem. Inf. Model.* **45** (2005) 74–80.

[41] E. Schymanski, *Integrated analytical and computer tools for toxicant identification in effect directed analysis*, PhD thesis 07/2011, Helmholtz Centre for Environmental Research, UFZ Leipzig.

[42] E. Schymanski, C. Meinert, M. Meringer, W. Brack, The use of MS classifiers and structure generation to assist in the identification of unknowns in effect-directed analysis, *Analyt. Chim. Acta* **615** (2008) 136–147.

[43] S. Tratch, Mathematical models in stereochemistry. I. Combinatorial characteristics of composition, connection, and configuration of organic molecules, *Russ. J. Org. Chem.* **31**(1995) 1189–1217.

[44] S. Tratch, M. Molchanova, N. Zefirov, A unified approach to characterization of molecular composition, connectivity and configuration: Symmetry, chirality, and generation problems for the corresponding combinatorial objects, *MATCH Commun. Math. Comput. Chem.* **61** (2009) 217–266.