PLOS COMPUTATIONAL BIOLOGY

TOPIC PAGE

Chemical graph generators

Mehmet Aziz Yirik*, Christoph Steinbeck

Friedrich Schiller Universität Jena, Institute for Inorganic and Analytical Chemistry, Jena, Germany

* yirik.mehmetaziz@uni-jena.de

Abstract

Chemical graph generators are software packages to generate computer representations of chemical structures adhering to certain boundary conditions. Their development is a research topic of cheminformatics. Chemical graph generators are used in areas such as virtual library generation in drug design, in molecular design with specified properties, called inverse QSAR/QSPR, as well as in organic synthesis design, retrosynthesis or in systems for computer-assisted structure elucidation (CASE). CASE systems again have regained interest for the structure elucidation of unknowns in computational metabolomics, a current area of computational biology.

Citation: Yirik MA, Steinbeck C (2021) Chemical graph generators. PLoS Comput Biol 17(1): e1008504. https://doi.org/10.1371/journal.pcbi.1008504

Editor: Daniel Mietchen, Museum für Naturkunde Berlin, GERMANY

Published: January 5, 2021

Funding: The author(s) received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Wikipedia Version: https://en.wikipedia.org/wiki/chemical_graph_generators

History

Molecular structure generation is a branch of graph) generation problems. Molecular structures are graphs with chemical constraints such as valences), bond multiplicity and fragments. These generators are the core of CASE systems. In a generator, the molecular formula is the basic input. If fragments are obtained from the experimental data, they can also be used as inputs to accelerate structure generation. The first structure generators were versions of graph generators modified for chemical purposes. One of the first structure generators was CONGEN,[1] originally developed for the DENDRAL project, the first artificial intelligence project in organic chemistry.[2] CONGEN dealt well with overlaps in substructures (Fig 1). The overlaps among substructures rather than atoms were used as the building blocks. For the case of stereoisomers, symmetry group calculations were performed for duplicate detection.

After DENDRAL, another mathematical method, MASS[3], a tool for mathematical synthesis and analysis of molecular structures, was reported. As with CONGEN, the MASS algorithm worked as an adjacency matrix generator. Many mathematical generators are descendants of efficient branch-and-bound methods from Igor Faradjev[4] and Ronald C. Read's orderly generation method.[5] Although their reports are from the 1970s, these studies are still the fundamental references for structure generators. In the orderly generation method, specific order-check functions are performed on graph representatives, such as vectors. For example, MOLGEN[6] performs a descending order check while filling rows of adjacency matrices. This descending order check is based on an input valence distribution. The literature classifies generators into two major types: structure assembly and structure reduction. The algorithmic complexity and the run time) are the criteria used for comparison.

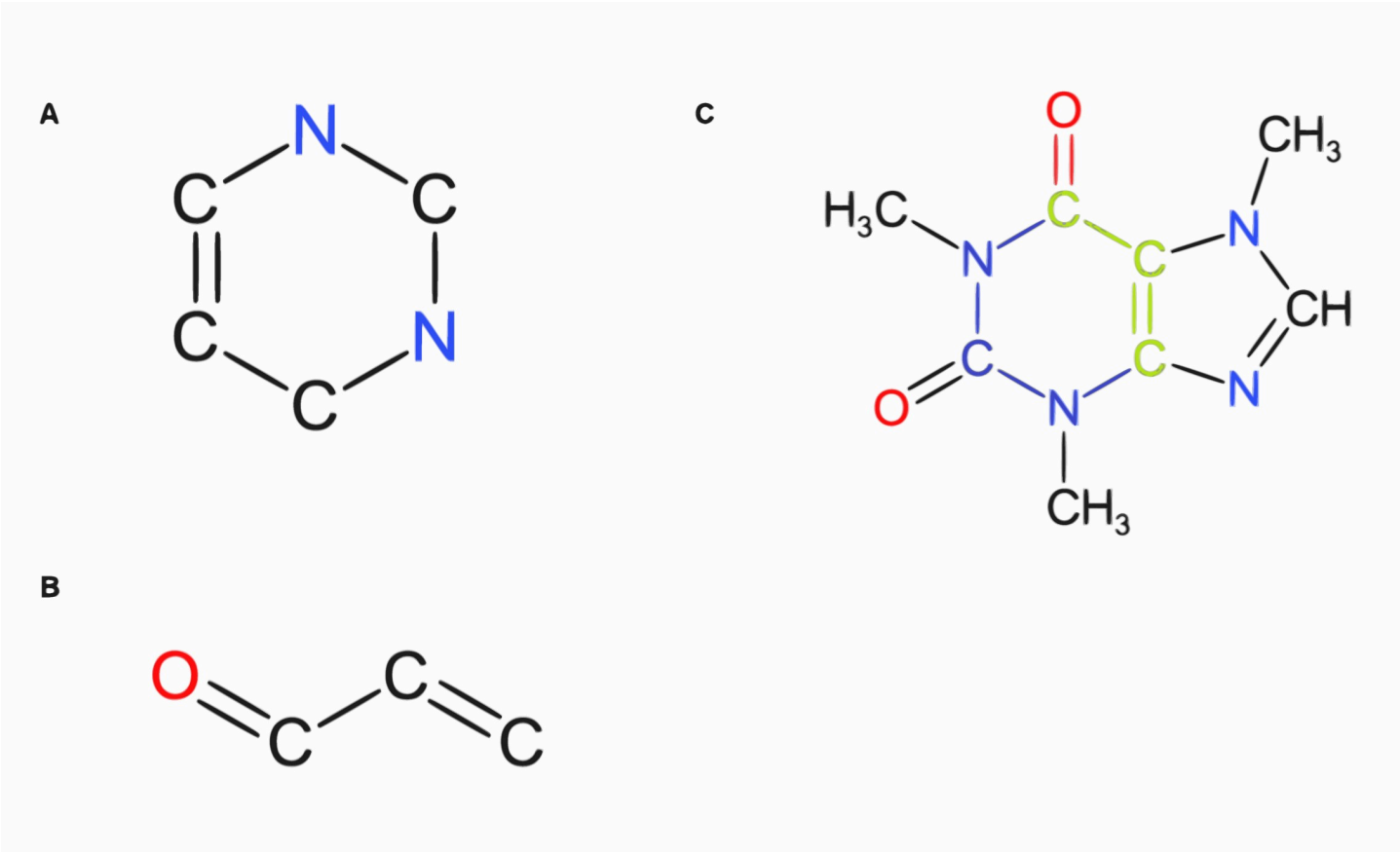

Fig 1. Overlapping substructures of caffeine. Two substructures of a caffeine molecule are given, **(A)** and **(B)**. The overlap of these substructures is highlighted in green in the caffeine structure **(C)**.

https://doi.org/10.1371/journal.pcbi.1008504.g001

Structure assembly

The generation process starts with a set of atoms from the molecular formula. In structure assembly, atoms are combinatorically connected to consider all possible extensions. If substructures are obtained from the experimental data, the generation starts with these substructures. These substructures provide known bonds in the molecule. One of the earliest attempts was made by Hidetsugu Abe in 1975 using a pattern recognition-based structure generator.[7] The algorithm had two steps: first, the prediction of the substructure from low-resolution spectral data; second, the assembly of these substructures based on a set of construction rules. Hidetsugu Abe and the other contributors published the first paper on CHEMICS,[8] which is a CASE tool comprising several structure generation methods. The program relies on a predefined non-overlapping fragment library. CHEMICS generates different types of component sets ranked from primary to tertiary based on component complexity. The primary set contains atoms, i.e., C, N, O and S, with their hybridization. The secondary and tertiary component sets are built layer-by-layer starting with these primary components. These component sets are represented as vectors and are used as building blocks in the process.

Substantial contributions were made by Craig Shelley and Morton Munk, who published a large number of CASE papers in this field. The first of these papers reported a structure generator, ASSEMBLE.[9] The algorithm is considered one of the earliest assembly methods in the field. As the name indicates, the algorithm assembles substructures with overlaps to construct

structures. ASSEMBLE overcomes overlapping by including a "neighbouring atom tag". The generator is purely mathematical and does not involve the interpretation of any spectral data. Spectral data are used for structure scoring and substructure information. Based on the molecular formula, the generator forms bonds between pairs of atoms, and all the extensions are checked against the given constraints. If the process is considered as a tree), the first node of the tree is an atom set with substructures if any are provided by the spectral data. By extending the molecule with a bond, an intermediate structure is built. Each intermediate structure can be represented by a node in the generation tree. ASSEMBLE was developed with a user-friendly interface to facilitate use. The second version of ASSEMBLE was released in 2000.[10] Another assembly method is GENOA.[11] Compared to ASSEMBLE and many other generators, GENOA is a constructive substructure search-based algorithm, and it assembles different substructures by also considering the overlaps.

The efficiency and exhaustivity of generators are also related to the data structures. Unlike previous methods, AEGIS[12] was a list-processing generator. Compared to adjacency matrices, list data requires less memory. As no spectral data was interpreted in this system, the user needed to provide substructures as inputs. Structure generators can also vary based on the type of data used, such as HMBC, HSQC and other NMR data. LUCY is an open-source structure elucidation method based on the HMBC data of unknown molecules[13], and involves an exhaustive 2-step structure generation process where first all combinations of interpretations of HMBC signals are implemented in a connectivity matrix, which is then completed by a deterministic generator filling in missing bond information. This platform could generate structures with any arbitrary size of molecules; however, molecular formulas with more than 30 heavy atoms are too time consuming for practical applications. This limitation highlighted the need for a new CASE system. SENECA was developed to eliminate the shortcomings of LUCY.[14] To overcome the limitations of the exhaustive approach, SENECA was developed as a stochastic method to find optimal solutions. The systems comprise two stochastic methods: simulated annealing and genetic algorithms. First, a random structure is generated; then, its energy is calculated to evaluate the structure and its spectral properties. By transforming this structure into another structure, the process continues until the optimum energy is reached. In the generation, this transformation relies on equations based on Jean-Loup Faulon's rules.[15] LSD (Logic for Structure Determination)[16] is an important contribution from French scientists. The tool uses spectral data information such as HMBC and COSY data to generate all possible structures. LSD is an open source structure generator released under the General Public License (GPL). A well-known commercial CASE system, StrucEluc,[17] also features a NMR based generator. This tool is from ACD Labs and, notably, one of the developers of MASS, Mikhail Elyashberg. COCON[18] is another NMR based structure generator, relying on theoretical data sets for structure generation. Except J-HMBC and J-COSY, all NMR types can be used as inputs.

In 1994, Chinese scientists reported an integer partitioning)-based structure generator.[19] The decomposition of the molecular formula into fragments, components and segments was performed as an application of integer partitioning. These fragments were then used as building blocks in the structure generator. This structure generator was part of a CASE system, ESESOC.[20]

A series of stochastic generators was reported by Jean-Loup Faulon. The software, MOLSIG,[21] was integrated into this stochastic generator for canonical labelling and duplicate checks.[22] As for many other generators, the tree approach is the skeleton of Jean-Loup Faulon's structure generators. However, considering all possible extensions leads to a combinatorial explosion. Orderly generation is performed to cope with this exhaustivity. Many assembly algorithms, such as OMG,[23] MOLGEN and Jean-Loup Faulon's structure

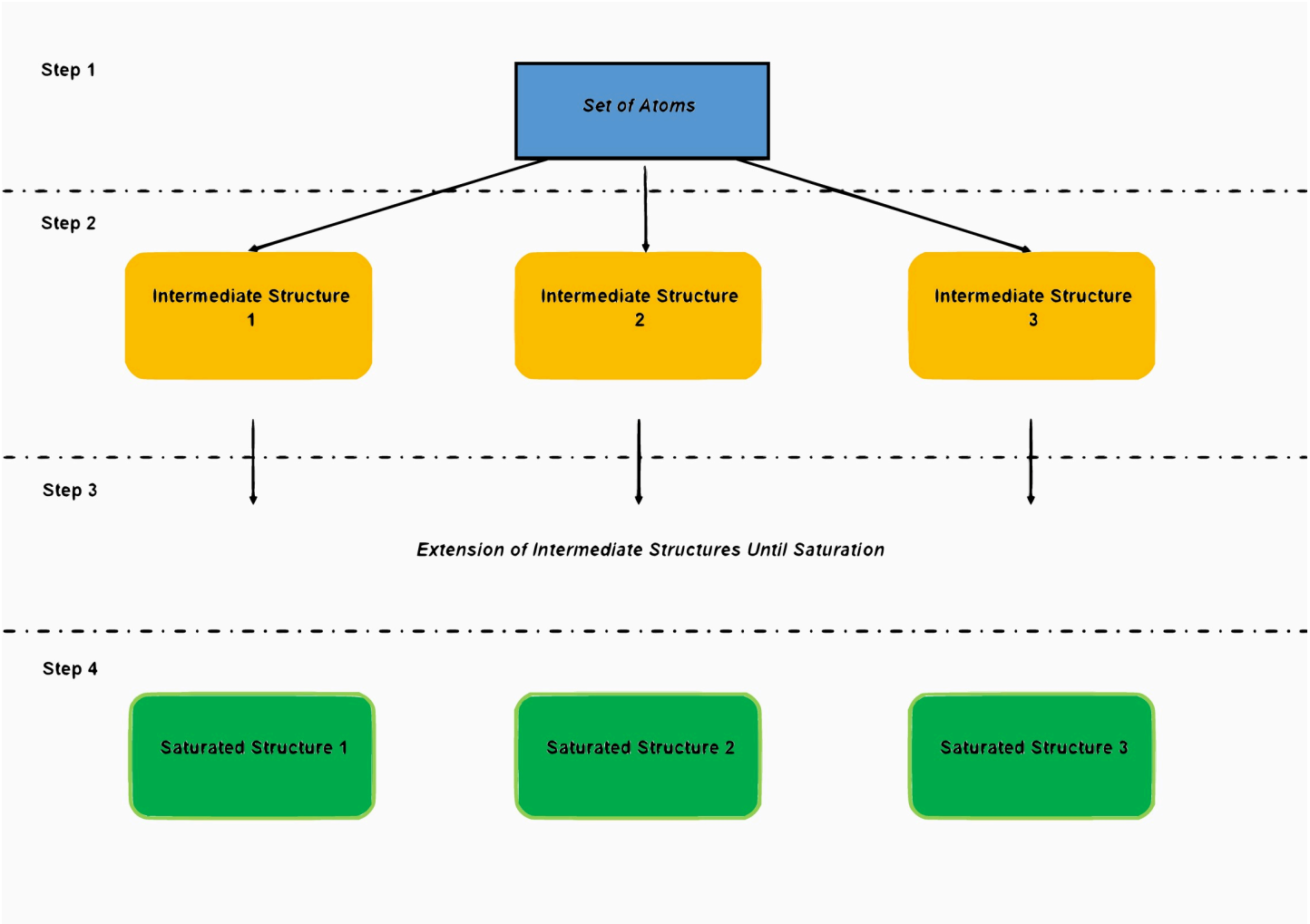

Fig 2. Breadth-first search generation. Molecular structure generation is explained step by step. Starting from a set of atoms, bonds are added between atom pairs until reaching saturated structures.

https://doi.org/10.1371/journal.pcbi.1008504.g002

generator[24], are orderly generation methods. Jean-Loup Faulon's structure generator relies on equivalence classes over atoms. Atoms with the same interaction type and element are grouped in the same equivalence class. Rather than extending all atoms in a molecule, one atom from each class is connected with other atoms. Similar to the former generator, Julio Peironcely's structure generator, OMG, takes atoms and substructures as inputs and extends the structures using a breadth-first search method (Fig 2). This tree extension terminates when all the branches reach saturated structures.

OMG generates structures based on the canonical augmentation method from Brendan McKay's NAUTY package. The algorithm calculates canonical labelling and then extends structures by adding one bond. To keep the extension canonical, canonical bonds are added. [25] Although NAUTY is an efficient tool for graph canonical labelling, OMG is approximately 2000 times slower than MOLGEN.[26] The problem is the storage of all the intermediate structures. OMG has since been parallelized, and the developers released PMG (Parallel Molecule Generator).[27] MOLGEN outperforms PMG using only 1 core; however, PMG outperforms MOLGEN by increasing the number of cores to 10.

A constructive search algorithm is a branch-and-bound method, such as Igor Faradjev's algorithm, and an additional solution to memory problems. Branch-and-bound methods are matrix) generation algorithms. In contrast to previous methods, these methods build all the connectivity matrices without building intermediate structures. In these algorithms, canonicity criteria and isomorphism checks are based on automorphism groups from mathematical group theory. MASS, SMOG[28] and Ivan Bangov's algorithm[29] are good examples in the literature. MASS is a method of mathematical synthesis. First, it builds all incidence matrices for a given molecular formula. The atom valences are then used as the input for matrix generation. The matrices are generated by considering all the possible interactions among atoms with respect to the constraints and valences. The benefit of constructive search algorithms is their low memory usage. SMOG is a successor of MASS.

Unlike previous methods, MOLGEN is the only maintained efficient generic structure generator, developed as a closed-source platform by a group of mathematicians as an application of computational group theory. MOLGEN is an orderly generation method. Many different versions of MOLGEN have been developed, and they provide various functions. Based on the users' needs, different types of inputs can be used. For example, MOLGEN-MS[30] allows users to input mass spectrometry data of an unknown molecule. Compared to many other generators, MOLGEN approaches the problem from different angles. The key feature of MOLGEN is generating structures without building all the intermediate structures and without generating duplicates.

In the field, the recent studies are from Kimito Funatsu's research group. As a type of assembly method, building blocks, such as ring systems and atom fragments, are used in the structure generation.[31] Every intermediate structure is extended by adding building blocks in all possible ways. To reduce the number of duplicates, Brendan McKay's canonical path augmentation method is used. To overcome the combinatorial explosion in the generation, applicability domain and ring systems are detected based on inverse QSPR/QSAR analysis. [32] The applicability domain, or target area, is described based on given biological as well as pharmaceutical activity information from QSPR/QSAR.[33] In that study, monotonically changed descriptors (MCD) are used to describe applicability domains. For every extension in intermediate structures, the MCDs are updated. The usage of MCDs reduces the search space in the generation process. In the QSPR/QSAR based structure generation, there is the lack of synthesizability of the generated structures. Usage of retrosynthesis paths in the generation makes the generation process more efficient. For example, a well-known tool called RetroPath [34] is used for molecular structure enumeration and virtual screening based on the given reaction rules.[35] Its core algorithm is a breadth-first method, generating structures by applying reaction rules to each source compound. Structure generation and enumeration are performed based on Brendan McKay's canonical augmentation method. RetroPath 2.0 provides a variety of workflows such as isomer transformation, enumeration, QSAR and metabolomics.

Besides these mathematical structure generation methods, the implementations of neural networks, such as generative autoencoder models,[36, 37] are the novel directions of the field.

Structure reduction

Unlike these assembly methods, reduction methods make all the bonds between atom pairs, generating a hypergraph. Then, the size of the graph is reduced with respect to the constraints. First, the existence of substructures in the hypergraph is checked. Unlike assembly methods, the generation tree starts with the hypergraph, and the structures decrease in size at each step. Bonds are deleted based on the substructures. If a substructure is no longer in the hypergraph, the substructure is removed from the constraints. Overlaps in the substructures were also

considered due to the hypergraphs. The earliest reduction-based structure generator is COCOA,[38] an exhaustive and recursive) bond-removal method. Generated fragments are described as atom-centred fragments to optimize storage, comparable to circular fingerprints [39] and atom signatures[40]. Rather than storing structures, only the list of first neighbours of each atom is stored. The main disadvantage of reduction methods is the massive size of the hypergraphs. Indeed, for molecules with unknown structures, the size of the hyper structure becomes extremely large, resulting in a proportional increase in the run time.

The structure generator GEN[41] by Simona Bohanec combines two tasks: structure assembly and structure reduction. Like COCOA, the initial state of the problem is a hyper structure. Both assembly and reduction methods have advantages and disadvantages, and the GEN tool avoids these disadvantages in the generation step. In other words, structure reduction is efficient when structural constraints are provided, and structure assembly is faster without constraints. First, the useless connections are eliminated, and then the substructures are assembled to build structures. Thus, GEN copes with the constraints in a more efficient way by combining these methods. GEN removes the connections creating the forbidden structures, and then the connection matrices are filled based on substructure information. The method does not accept overlaps among substructures. Once the structure is built in the matrix representation, the saturated molecule is stored in the output list. The COCOA method was further improved and a new generator was built, HOUDINI.[42] It relies on two data structures: a square matrix of compounds representing all bonds in a hyper structure is constructed, and second, substructure representation is used to list atom-centred fragments. In the structure generation, HOUDINI maps all the atom-centred fragments onto the hyper structure.

Mathematical basis

Chemical graphs

In a graph representing a chemical structure, the vertices) and edges) represent atoms and bonds, respectively (Fig 3). The bond order corresponds to the edge multiplicity, and as a result, chemical graphs are vertex and edge-labelled graphs. A vertex and edge-labelled graph is described as a chemical graph where is the set of vertices, i.e., atoms, and is the set of edges, which represents the bonds.

In graph theory, the degree) of a vertex is its number of connections. In a chemical graph, the maximum degree of an atom is its valence), and the maximum number of bonds a chemical element can make. For example, carbon's valence is 4. In a chemical graph, an atom is saturated if it reaches its valence. A graph is connected) if there is at least one path between each pair of vertices. Although chemical mixtures[43] are one of the main interests of many chemists, due to the computational explosion, many structure generators output only connected chemical graphs. Thus, the connectivity check is one of the mandatory intermediate steps in structure generation because the aim is to generate fully saturated molecules. A molecule is saturated if all its atoms are saturated.

Symmetry groups for molecular graphs

For a set of elements, a permutation is a rearrangement of these elements.[44] An example is given below (Table 1):

The second line of Table 1 shows a permutation of the first line. The multiplication of permutations, a and b, a function composition, as shown below.

$$(a \bullet b)\ (x) = a(b(x)) \tag{1}$$

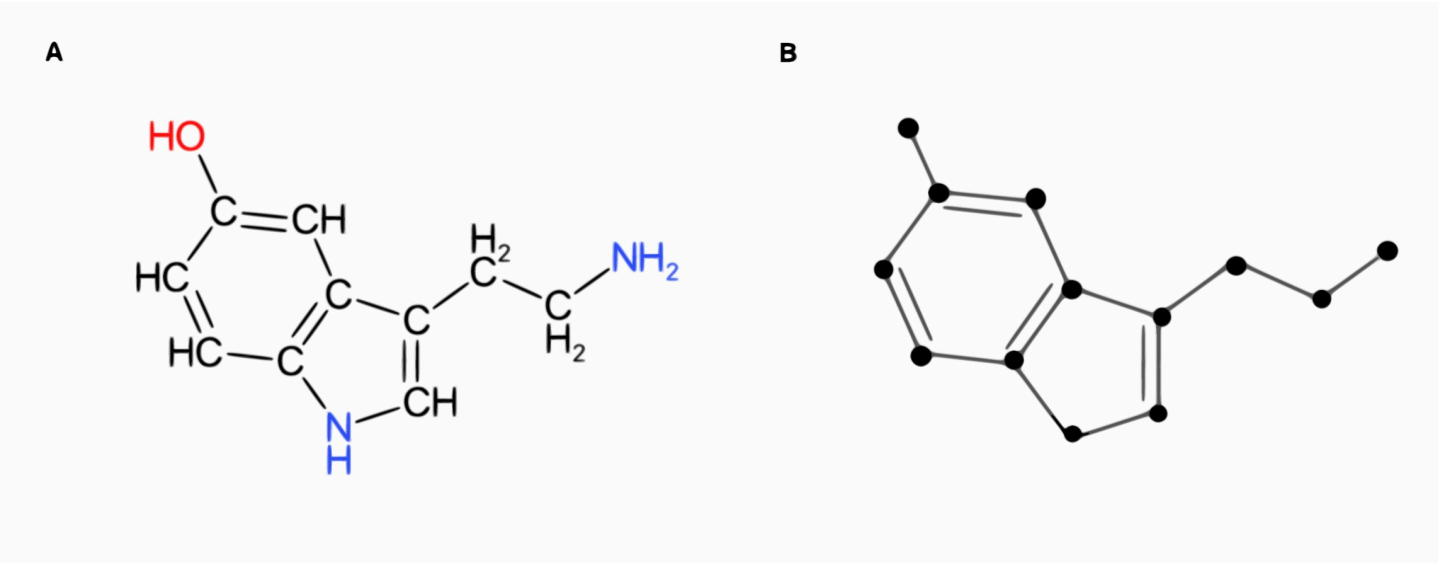

Fig 3. Graph representation of the serotonin molecule. (A) Molecular structure of serotonin. **(B)** Graph representation of the molecule.

https://doi.org/10.1371/journal.pcbi.1008504.g003

The combination of two permutations is also a permutation. A group, G, is a set of elements together with an associative binary operation • defined on G such that the following are true:

- There is an element I in G satisfying $g \bullet I = g$, for all elements g of G.
- For each element of G, there is an element g^{-1} such that $g \bullet g^{-1}$ is equal to the identity element.

The order) of a group is the number of elements in the group. Let us assume X is a set of integers. Under the function composition operation, $Sym(X)$ is a symmetry group, the set of all permutations over X. If the size of X is n, then the order of $Sym(X)$ is $n!$ Set) systems consist of a finite set X and its subsets, called blocks of the set. The set of permutations preserving the set system is used to build the automorphisms of the graph. An automorphism permutes the vertices of a graph; in other words, it maps a graph onto itself. This action is edge-vertex preserving. If (u,v) is an edge of the graph, $G = (V,E)$, and a is a permutation of V, then

$$a(u, v) = (a(u), a(v)) \tag{2}$$

A permutation a of V is an automorphism of the graph $G = (E, V)$, if $a(u, v)$ is an element of E, if (u, v) is an element of E.

The automorphism group of a graph G, denoted $Aut(G)$, is the set of all automorphisms on V. In molecular graphs, canonical labelling and molecular symmetry (Fig 4) detection are implementations of automorphism groups. Although there are well known canonical labelling methods in the field, such as InChI[45] and ALATIS[46], NAUTY is a commonly used software package for automorphism group calculations and canonical labelling.

Table 1. Permutation of set of integers.

x	1	2	3	4	5	6	7	8	9	10	11
$f(x)$	4	2	11	6	1	5	8	9	7	10	3

https://doi.org/10.1371/journal.pcbi.1008504.t001

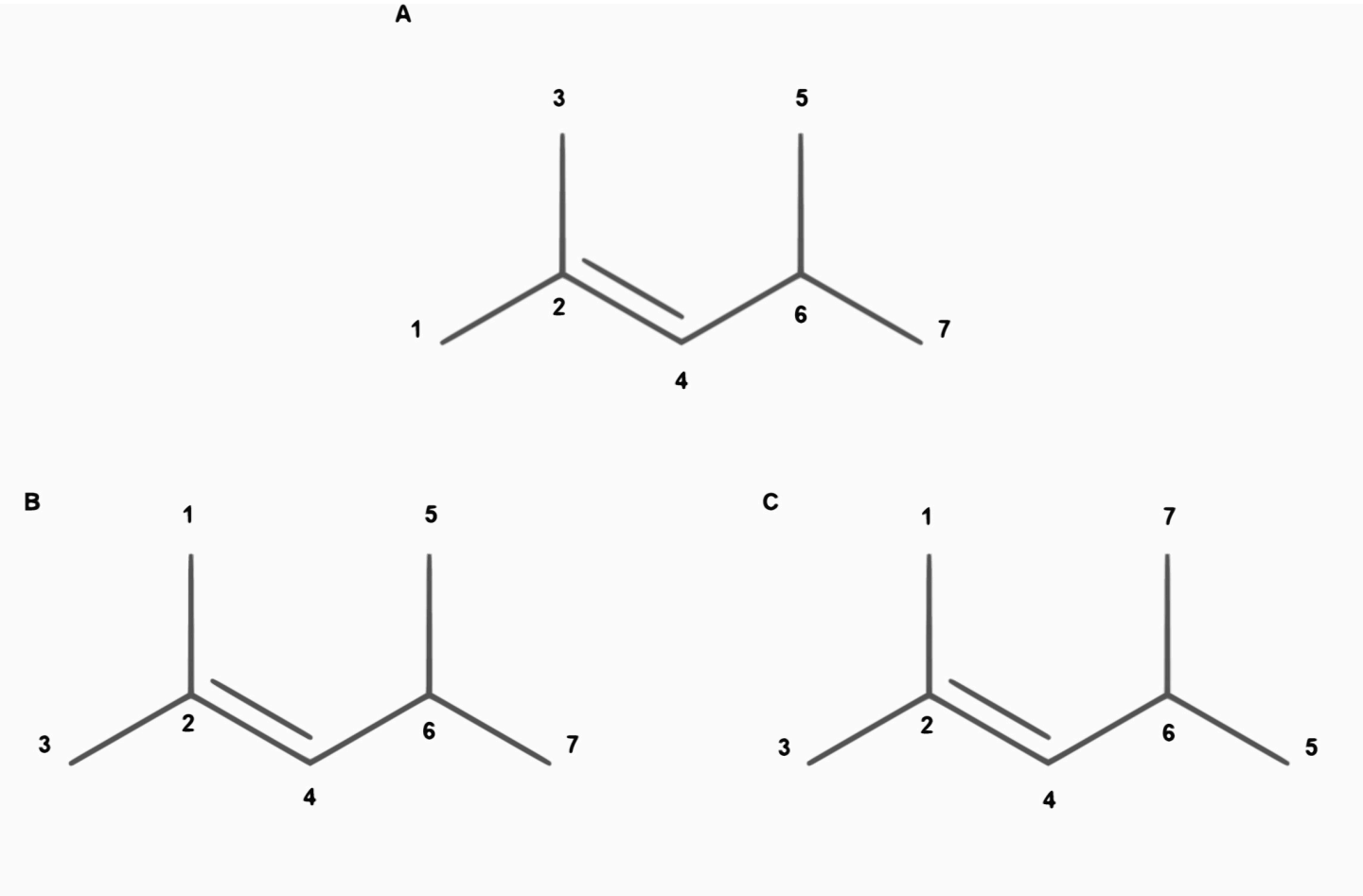

Fig 4. Molecular Symmetry. (A) The initial labelling of 2,4-Dimethyl-2-pentene. **(B)** and **(C)** are symmetries of the same molecule with different labels.

https://doi.org/10.1371/journal.pcbi.1008504.g004

Conclusion

The structural identification of unknown molecules is an interdisciplinary field involving mathematicians, chemists and computer scientists; moreover, it has led to the creation of the

Table 2. List of Available structure generators.

Name	Link
ASSEMBLE	http://www.upstream.ch/main.html?src=%2Findex.html
COCON	http://cocon.nmr.de
DENDRAL	http://www.softwarepreservation.org/projects/AI/DENDRAL/DENDRAL-CONGEN_GENOA.zip/view
LSD	http://eos.univ-reims.fr/LSD/index_ENG.html
MOLGEN	http://www.molgen.de/
MOLSIG	http://molsig.sourceforge.net
OMG	https://sourceforge.net/p/openmg/
PMG	https://sourceforge.net/projects/pmgcoordination/
SENECA	https://github.com/steinbeck/seneca
SMOG	http://ccl.net/cca/software/MS-DOS/SMOG/index.shtml

https://doi.org/10.1371/journal.pcbi.1008504.t002

field of mathematical chemistry and cheminformatics. The state-of-the-art methods comprise a variety of algorithms that can be classified into two groups; moreover, structure assembly has been the dominant approach in the field. Both assembly and reduction methods are incremental processes: all the intermediate structures are constructed based on previously generated structures, and duplicates are then excluded. The algorithms are generally breadth-first or depth-first search methods; and terminate once all the structures are saturated. The generation of too many intermediate structures and their storage make these algorithms inefficient. In the field, matrix generators have been attracting increasing interest from many scientists. According to the literature, there is still a lack of mathematical algorithms; more precisely, there is a lack of fast open-source structure generators.

List of available structure generators

The available software packages and their links are listed below (Table 2).

Supporting information

S1 Text. Version history of the text file.
(XML)

S2 Text. Peer reviews and response to reviews.
(XML)

References

1. Bruccoleri R, Karplus M. Prediction of the folding of short polypeptide segments by uniform conformational sampling. Biopolymers: Original Research on Biomolecules. 1987; 26(1):137–68. https://doi.org/10.1002/bip.360260114 PMID: 3801593
2. Sutherland G. DENDRAL-a computer program for generating and filtering chemical structures. STANFORD UNIV CALIF DEPT OF COMPUTER SCIENCE, 1967.
3. Serov V, Elyashberg M, Gribov L. Mathematical synthesis and analysis of molecular structures. Journal of Molecular Structure. 1976; 31(2):381–97.
4. Faradzev I. Constructive enumeration of combinatorial objects. 1978.
5. Colbourn C, Read R. Orderly algorithms for generating restricted classes of graphs. Journal of Graph Theory. 1979; 3(2):187–95.
6. Grüner T, Laue R, Meringer M. Algorithms for group actions applied to graph generation. Groups and Computation II; 1997: American Mathematical Soc.
7. Abe H, Jurs P. Automated chemical structure analysis of organic molecules with a molecular structure generator and pattern recognition techniques. Analytical Chemistry. 1975; 47(11):1829–35.
8. Sasaki S, Abe H, Hirota Y, Ishida Y, Kudo Y, Ochiai S, et al. CHEMICS-F: A Computer Program System for Structure Elucidation of Organic Compounds. Journal of Chemical Information and Computer Sciences. 1978; 18(4):211–22.
9. Shelley C, Munk M. CASE, a computer model of the structure elucidation process. Analytica Chimica Acta. 1981; 133(4):507–16.
10. Badertscher M, Korytko A, Schulz K, Madison M, Munk M, Portmann P, et al. Assemble 2.0: a structure generator. Chemometrics and Intelligent Laboratory Systems. 2000; 51(1):73–9.
11. Carhart R, Smith D, Gray N. GENOA: A computer program for structure elucidation utilizing overlapping and alternative substructures. 1981. https://doi.org/10.1007/BF00457449 PMID: 6975613
12. Luinge H, Van Der Maas J. AEGIS, an algorithm for the exhaustive generation of irredundant structures. Chemometrics and intelligent laboratory systems. 1990; 8(2):157–65.
13. Steinbeck C. LUCY—A program for structure elucidation from NMR correlation experiments. Angewandte Chemie International Edition in English. 1996; 35(17):1984–6.
14. Steinbeck C. SENECA: A platform-independent, distributed, and parallel system for computer-assisted structure elucidation in organic chemistry. Journal of chemical information and computer sciences. 2001; 41(6):1500–7. https://doi.org/10.1021/ci000407n PMID: 11749575

15. Faulon J. Stochastic generator of chemical structure. 2. Using simulated annealing to search the space of constitutional isomers. Journal of Chemical Information and Computer Sciences. 1996; 36(4):731–40.
16. Nuzillard J, Georges M. Logic for structure determination. Tetrahedron. 1991; 47(22):3655–64.
17. Blinov K, Elyashberg M, Molodtsov S, Williams A, Martirosian E. An expert system for automated structure elucidation utilizing 1H-1H, 13C-1H and 15N-1H 2D NMR correlations. Fresenius' journal of analytical chemistry. 2001; 369(7–8):709–14. https://doi.org/10.1007/s002160100757 PMID: 11371077
18. Junker J. Theoretical NMR correlations based structure discussion. Journal of cheminformatics. 2011; 3(1):1–4. https://doi.org/10.1186/1758-2946-3-1 PMID: 21214931
19. Hu C, Xu L. Principles for structure generation of organic isomers from molecular formula. Analytica chimica acta. 1994; 298(1):75–85.
20. Hao J, Xu L, Hu C. Expert system for elucidation of structures of organic compounds (ESESOC). Science in China Series B: Chemistry. 2000; 43(5):503–15.
21. Faulon J. Stochastic generator of chemical structure. 1. Application to the structure elucidation of large molecules. Journal of Chemical Information and Computer Sciences. 1994; 34(5):1204–18.
22. Faulon J, Churchwell C, Visco D. The signature molecular descriptor. 2. Enumerating molecules from their extended valence sequences. Journal of Chemical Information and Computer Sciences. 2003; 43 (3):721–34. https://doi.org/10.1021/ci020346o PMID: 12767130
23. Peironcely J, Rojas-Chertó M, Fichera D, Reijmers T, Coulier L, Faulon J, et al. OMG: open molecule generator. Journal of cheminformatics. 2012; 4(1):21. https://doi.org/10.1186/1758-2946-4-21 PMID: 22985496
24. Faulon J. On using graph-equivalent classes for the structure elucidation of large molecules. Journal of chemical information and computer sciences. 1992; 32(4):338–48.
25. McKay B, Piperno A. Practical graph isomorphism, II. Journal of Symbolic Computation. 2014; 60:94–112.
26. Yirik M. Blogger. 2020. [cited 2020]. Available from: https://mayphd.blogspot.com/2020/02/structure-generators-benchmark.html.
27. Jaghoori M, Jongmans S, de Boer F, Peironcely J, Faulon J, Reijmers T, et al. PMG: Multi-core Metabolite Identification. Electron Notes Theor Comput Sci. 2013; 299:53–60.
28. Molchanova M, Shcherbukhin V, Zefirov N. Computer generation of molecular structures by the SMOG program. Journal of chemical information and computer sciences. 1996; 36(4):888–99.
29. Bangov I, Kanev K. Computer-assisted structure generation from a gross formula: II. Multiple bond unsaturated and cyclic compounds. Employment of fragments. Journal of Mathematical Chemistry. 1988; 2(1):31–48.
30. Kerber A, Laue R, Meringer M, Varmuza K. MOLGEN-MS: Evaluation of low resolution electron impact mass spectra with MS classification and exhaustive structure generation. Adv Mass Spectrom. 2001; 15(939–940):22.
31. Miyao T, Kaneko H, Funatsu K. Ring system-based chemical graph generation for de novo molecular design. Journal of computer-aided molecular design. 2016; 30(5):425–46. https://doi.org/10.1007/s10822-016-9916-1 PMID: 27299746
32. Miyao T, Kaneko H, Funatsu K. Ring-System-Based Exhaustive Structure Generation for Inverse-QSPR/QSAR. Molecular informatics. 2014; 33(11–12):764–78. https://doi.org/10.1002/minf.201400072 PMID: 27485423
33. Miyao T, Arakawa M, Funatsu K. Exhaustive structure generation for inverse-QSPR/QSAR. Molecular informatics. 2010; 29(1–2):111–25. https://doi.org/10.1002/minf.200900038 PMID: 27463853
34. Delépine B, Duigou T, Carbonell P, Faulon J. RetroPath2. 0: A retrosynthesis workflow for metabolic engineers. Metabolic engineering. 2018; 45:158–70. https://doi.org/10.1016/j.ymben.2017.12.002 PMID: 29233745
35. Koch M, Duigou T, Carbonell P, Faulon J. Molecular structures enumeration and virtual screening in the chemical space with RetroPath2. 0. Journal of cheminformatics. 2017; 9(1):1–17. https://doi.org/10.1186/s13321-017-0252-9 PMID: 29260340
36. Kadurin A, Nikolenko S, Khrabrov K, Aliper A, Zhavoronkov A. druGAN: an advanced generative adversarial autoencoder model for de novo generation of new molecules with desired molecular properties in silico. Molecular pharmaceutics. 2017; 14(9):3098–104. https://doi.org/10.1021/acs.molpharmaceut.7b00346 PMID: 28703000
37. Blaschke T, Olivecrona M, Engkvist O, Bajorath J, Chen H. Application of generative autoencoder in de novo molecular design. Molecular informatics. 2018; 37(1–2):1700123. https://doi.org/10.1002/minf.201700123 PMID: 29235269

38. Christie B, Munk M. Structure generation by reduction: a new strategy for computer-assisted structure elucidation. Journal of Chemical Information and Computer Sciences. 1988; 28(2):87–93. https://doi.org/10.1021/ci00058a009 PMID: 3392122

39. Glen R, Bender A, Arnby C, Carlsson L, Boyer S, Smith J. Circular fingerprints: flexible molecular descriptors with applications from physical chemistry to ADME. IDrugs. 2006; 9(3):199. PMID: 16523386

40. Faulon J, Collins M, Carr R. The signature molecular descriptor. 4. Canonizing molecules using extended valence sequences. Journal of chemical information and computer sciences. 2004; 44 (2):427–36. https://doi.org/10.1021/ci0341823 PMID: 15032522

41. Bohanec S. Structure generation by the combination of structure reduction and structure assembly. Journal of chemical information and computer sciences. 1995; 35(3):494–503.

42. Korytko A, Schulz K, Madison M, Munk M. HOUDINI: a new approach to computer-based structure generation. Journal of chemical information and computer sciences. 2003; 43(5):1434–46. https://doi.org/10.1021/ci034057r PMID: 14502476

43. Massiot G, Nuzillard J. Computer-assisted elucidation of structures of natural products. Phytochemical analysis. 1992; 3(4):153–9.

44. Kreher D, Stinson D. Combinatorial algorithms: generation, enumeration, and search. ACM SIGACT News. 1999; 30(1):33–5.

45. Heller S, McNaught A, Pletnev I, Stein S, Tchekhovskoi D. InChI, the IUPAC international chemical identifier. Journal of cheminformatics. 2015; 7(1):23. https://doi.org/10.1186/s13321-015-0068-4 PMID: 26136848

46. Dashti H, Westler W, Markley J, Eghbalnia H. Unique identifiers for small molecules enable rigorous labeling of their atoms. Scientific data. 2017; 4:170073. https://doi.org/10.1038/sdata.2017.73 PMID: 28534867