Review on Chemical Graph Theory and Its Application in Computer-Assisted Structure Elucidation

Mehmet Aziz Yirik, Institute for Inorganic and Analytical Chemistry, Friedrich-Schiller University, Jena, Germany
yirik.mehmetaziz@uni-jena.de, ORCID ID: 0000-0001-7520-7215

Kumsal Ecem Çolpan, Institute of Plant Genetics, Heinrich-Heine Universität Düsseldorf, Düsseldorf, Germany
kumsal.colpan@hhu.de, ORCID ID: 0000-0002-8689-218X

Saskia Schmidt, Institute for Inorganic and Analytical Chemistry, Friedrich-Schiller University, Jena, Germany
saskia.schmidt@uni-jena.de, ORCID ID: 0000-0002-4802-228X

Maria Sorokina, Institute for Inorganic and Analytical Chemistry, Friedrich-Schiller University, Jena, Germany
maria.sorokina@uni-jena.de, ORCID ID: 0000-0001-9359-7149

Christoph Steinbeck, Institute for Inorganic and Analytical Chemistry, Friedrich-Schiller University, Jena, Germany
christoph.steinbeck@uni-jena.de, ORCID ID: 0000-0001-6966-0814

Corresponding author: yirik.mehmetaziz@uni-jena.de

Abstract: The chemical graph theory is a subfield of mathematical chemistry which applies classic graph theory to chemical entities and phenomena. Chemical graphs are main data structures to represent chemical structures in cheminformatics. Computable properties of graphs lay the foundation for (quantitative) structure activity and structure property predictions - a core discipline of cheminformatics.

It has a historic relevance for natural sciences, such as chemistry, biochemistry and biology, and is in the heart of modern disciplines, such as cheminformatics and bioinformatics. This review first covers the history of chemical graph theory, then provides an overview of its various techniques and applications for CASE, and finally summarises modern tools using chemical graph theory for CASE.

Keywords: Chemical Graph Theory; Computational Chemistry; CASE; Computer-Assisted Structure Elucidation

Introduction

The relationship between graph theory and chemistry has a long way from past to present. Different studies related to both disciplines have built very strong interactions in between and formed the scientific area known as chemical graph theory (Bonchev 1991). The first mentions of chemical graphs were in the late eighteenth century, where the perspective of chemistry was also affected by the ideas of Isaac Newton

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 November 2021 doi:10.20944/preprints202111.0546.v1

(Bonchev 1991). Although the studies about the interactions between the atoms gained speed during that century, the chemical bonds were not identified. Thus, the first usage of chemical graphs was for representing the hypothetical forces between the molecules and atoms (Bonchev 1991).

In 1805 John Dalton had built the first atomic model by representing atom types with specific circles, which could show only the chemical positions and number of atoms in a molecule (Cardwell 1969; Dalton 2010). However, August Kékule showed both physical positions and orientations of atoms in a molecule. In his "Tetrahedral Carbon Atom" model (Figure 1), he classified several organic molecules and visualized the bond orders between atoms, including the benzene ring (Hein 1966).

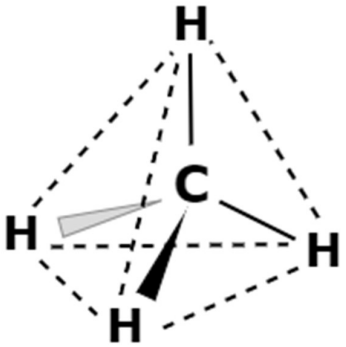

Figure 1. Tetrahedral carbon atom model.

This work led the three-dimensional thinking in chemical modelling to start and the tetrahedral carbon atom model affected not only organic but also inorganic chemical compounds' structure-modelling (Bonchev 1991).

Following Kékule's work, Alfred Werner developed the new scientific field, coordination chemistry by inventing the idea that the atoms have specific natural properties related to their location in a molecule (Werner 1893; Werner 1912). In addition, complex compounds were represented with octahedral models.

In 1861, Alexander Butlerov introduced the term "molecular structure" meaning that every chemical substance should have a fixed structure in molecular bases. This term explained several chemical properties of the substances (Butlerov 1861). To develop this concept, illustrations, analysis and formulations of different chemical compounds were proposed in years by scientists such as Johann Wolfgang Döbereiner, Alexander William Williamson and Archibald Scott Coupe (Alexanderson 2006). As another representation model, the line representation of bonds between atom pairs was first used in Bryan Higgins' chemical structure models (Higgins 1791). However, these lines were representing only the interatomic forces, not the specific bonds, therefore Couper's work has been counted as the first graphical edge representation of a chemical bond (Couper 1858). The molecular formula of acetic acid was defined and depicted as a chemical structure with straight lines between the atoms in a molecule to represent a chemical bond (Figure 2) (Couper 1858).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 November 2021 doi:10.20944/preprints202111.0546.v1

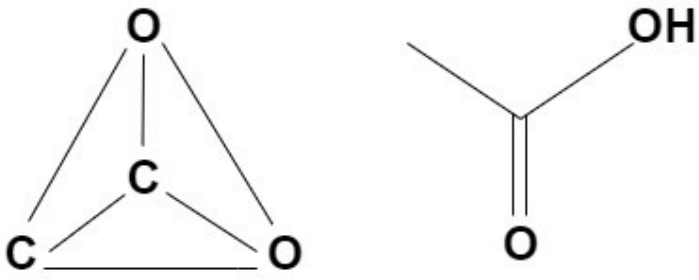

Figure 2. Acetic Acid Chemical and Structural Representation.

The next breakthrough idea leading to CGT is the one of fixed valence bonds of different atom types, and it was first published in the book of Edward Frankland: "*Lecture Notes for Chemical Students*" in 1866. In this book, several atom types were introduced, such as hydrogen, zinc, boron and carbon with their respective valence bonds (Frankland 1866). Following these studies, several other chemistry-based graph-theoretical analyses were made. Arthur Cayley and James Joseph Sylvester constructed chemical graphs with respect to the structural formulations of chemical substances (Cayley 1857; Sylvester 1878; Bonchev 1991). Cayley developed the tree representation of alkanes, kenograms, for structural isomer enumeration of alkanes (Figure 3). In addition, Sylvester labelled vertices with different letters to instruct chemical graphs with a variety of properties.

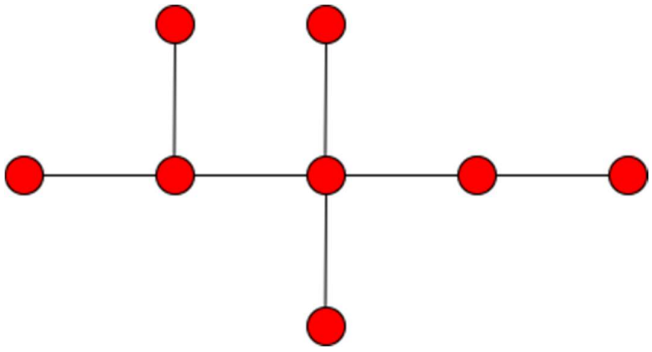

Figure 3. Alkane kenogram (2,3,3-trimethylpentane).

With all these fundamental theories being formulated, the need for pure mathematical analysis and applications in chemistry became explicit, in particular, for the mathematical analysis and discovery of unknown molecules. The evolution of chemical graph theory accelerated especially during the 20th century with the discovery and the synthesis of new molecules (Bonchev 1991). For the structure elucidation of unknown molecules, one of the possible preliminary steps used to be the identification of all isomers for a given or predicted molecular formula. Isomer enumeration became popular in the 1930s for the determination of the possible number of chemical structures for molecules. Besides an apparent simplicity of the task that is the isomer enumeration, the generation of all possible isomers became another key step towards structure elucidation. When formulated as a chemical graph theory problem, the structure generation of all possible molecular structures from a given molecular formula is a combinatorial generation of all possible graphs for given node degrees, i.e. the atom valences. In addition

 doi:10.20944/preprints202111.0546.v1

to molecular formulas as input, the inclusion of the substructure information during the generation process shrinks the search space of the chemical graph generation. In chemical graph theory, the earliest generator, CONGEN, came from a Stanford team in the 60s. Following that,many other structure generators have been developed since then.

On the other hand lies metabolomics, the study of the set of metabolites present within a given sample. The analyses of metabolomes are important for the discovery of new metabolites, the elucidation of disease processes in living organisms, and many other applications. Several methods for metabolome analyses are currently available, in particular mass spectrometry (MS) and nuclear magnetic resonance (NMR). The exact mass of a molecule almost always corresponds to one molecular formula. During MS analyses, in vitro fragmentation (i.e. breaking down the molecule in substructures), allows the obtention of a series of exact masses that correspond to a given set of substructures. Many CASE suites require such spectral information as input. The information provided in the spectral data, e.g. the present molecular substructures, helps the elucidation of new structures through the assembly of substructures in a single structure in a similar way one assembles pieces of a puzzle. In terms of the chemical graph theory, such information builds the substructures of the searched chemical structures. Therefore, the efficiency of the metabolomics data has a major impact on the accuracy and of CASE.

In this review, key applications of chemical graph theory for CASE are discussed: isomer enumeration, chemical graph generation, molecular fragmentation and molecular descriptors.

Isomer Enumeration

The first application of graph-theoretical techniques for solving a chemical problem was in the field of isomer enumeration. In 1811, Louis Joseph Gay-Lussac started a discussion of compounds with the same atom and bond sets but different arrangements. However, the definition of isomerism for chemical compounds was defined firstly as "possessing the same chemical constitution and molecular weight but differing properties" by Jöns Jakob Berzelius, in 1830 (Berzelius 1830). In other words, isomers consist of the same atom and bond sets with different arrangements. In 1862, Butlerov obtained different isomers of methane molecules substituted by chloride molecules (Butlerov 1862), which followed by a redefinition of the term “structural isomerism”, which was then explained as having the same chemical formula but different connections between the atoms of substances (Alen 2018). Isomers are classified into two groups: constitutional and steric. Constitutional isomers (Figure 4) provide only atom neighbourhood information.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 November 2021 doi:10.20944/preprints202111.0546.v1

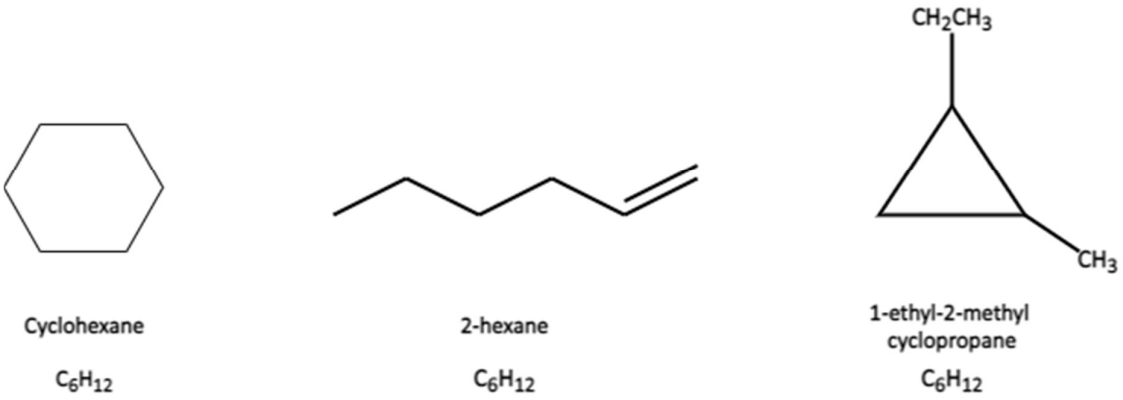

Figure 4. Constitutional Isomer of C_6H_{12}.

However, the configurational information, such as bond angle, bond length, and distance between disconnected atoms, are considered in stereoisomerism (Figure 5).

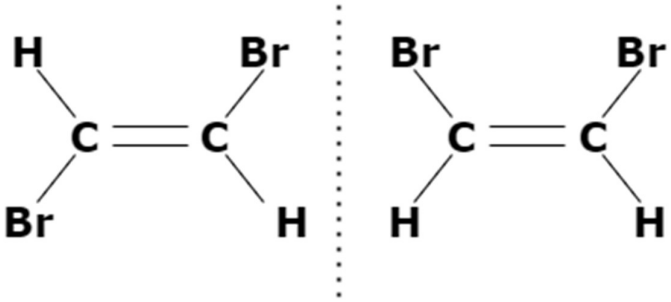

Figure 5. Stereoisomers of $C_2H_2Br_2$.

Louis Pasteur was the first scientist to introduce the term stereoisomerism, consisting of the same number and type of chemical bonds but in different orientations (Pasteur 1848; Alen 2018).

Chemical graphs are suitable for enumeration of structural isomeric compounds because of their ability to show only the topological positions of the atoms in a molecule. However, they are not convenient for stereoisomeric substances, because of their inability to give information about physical positions of atoms in molecules (Bonchev 1991).

The first contribution to the isomer enumeration problem came from a mathematician, James Joseph Slyvester. In his study, the enumeration of rooted trees was discussed. In chemistry, rooted trees correspond to alkyl radicals with i - non-hydrogen atoms (Figure 6).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 November 2021 doi:10.20944/preprints202111.0546.v1

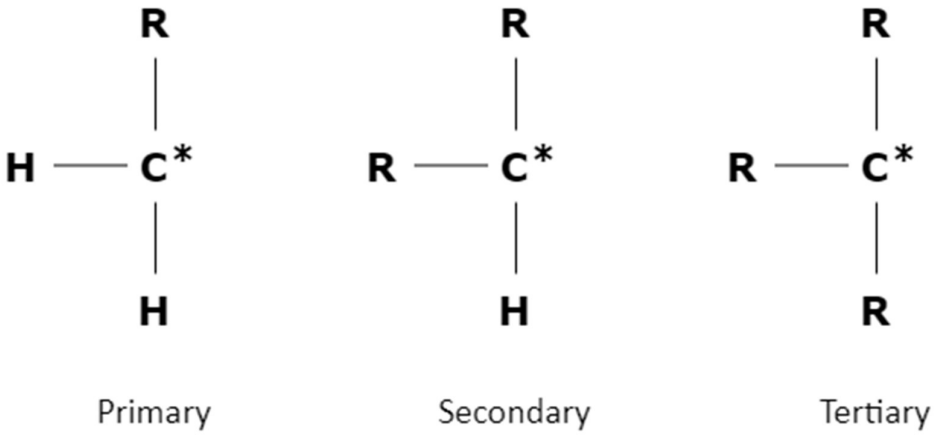

Figure 6. Alkyl Radicals.

This enumeration method relied on the polynomial representation of molecular connectivity. In other words, the coefficients of the polynomials represented the number of bonds between atoms. In 1874, Arthur Cayley extended the problem to the enumeration of unrooted trees. Alkanes are chemical examples of these unrooted trees. Cayley's contributions increased the interest in the problem. Following this study, Hugo Schiff, Hermann, Ferdinand Tiemann and many others worked on the enumeration of alkanes. However, including Cayley, none of them found the general formula for alkane isomer enumeration. In 1937, the general formula for isomer enumeration was found by George Polya (Beeler 2015). Polya's enumeration theory was a general counting method with implementation into a variety of fields. The method basically relied on symmetry operations. From mathematics, symmetry groups and their actions were used for the polynomial representations. For the calculation of polynomial coefficients, cycle indices of each symmetry were calculated. The calculation of the cycle indices was based on conjugacy classes of the acting symmetry group (Jiping 1992). The construction of Polya's polynomial was step by step described in the mathematical chemistry and cheminformatics book (Kerber et al. 2013). Another good example came from Pevac (Pevac and Crundwell 2000). In that study, Polya's enumeration method was implemented for enumeration of boat and chair cyclohexane isomers. For the molecule, first, its symmetry group was calculated. Then polynomial representation was constructed based on cycle indices. In the literature, isomer enumeration studies were mostly for special compound classes such as alkanes, aromatic hydrocarbons and polycyclic aromatic compounds (Bytautas and Klein 1998; Dias 1984; Balasubramanian 2018).

In CASE, the chemical space size of virtual compound libraries are detected by enumeration methods, however, the construction of the isomer sets are performed by the molecular structure generators.

Chemical Graph Generation

Molecular structure generation is a branch of graph generation problem. The earliest molecular structures were modified versions of graph generators. Structure generators generate computer representations of chemical structures adhering to certain boundary conditions. These generators are mostly BFS or DFS based combinatorial algorithms requiring these basic inputs: molecular formula and

 doi:10.20944/preprints202111.0546.v1

substructures (Holdsworth 1999; Putri, Tulus, and Napitupulu 2011). To elucidate the structure of an unknown molecule, all combinatorially possible molecular extensions should be taken into account. These algorithms extend intermediate structures in a recursive manner until the molecular saturation, however, the extension of intermediate structures causes a combinatorial explosion. Thus, many structure generators have been designed in line with mathematical theorems. In many efficient generators, group theory has been applied to accelerate the calculation of bond extension (Canals and Schober 2012). CONGEN was the earliest structure generator, a part of the first CASE system, DENDRAL (Sutherland 1967). CONGEN, first, built a tree based structure. Each node of the tree represented a substructure of the unknown molecule. These substructures were extended based on group theoretical lemmas. Besides group theory, many other mathematical theorems have been applied in the field. MASS, a tool for mathematical synthesis and analysis of molecular structures was a matrix based structure generator. This method was considered as an adjacency matrix generation algorithm (Serov, Elyashberg, and Gribov 1976). In the literature, structure generators are classified into two groups: structure assembly and structure reduction methods.

Assembly methods start the generation with a set of atoms from a molecular formula. Atoms are combinatorially assembled until the saturation. The earliest assembly method was ASSEMBLE from Munk and Shelley (Badertscher et al. 2000). This generator was a part of the CASE system, called CASE (Munk et al. 1982). ASSEMBLE was not able to deal with substructural overlaps. Contrary to ASSEMBLE, GENOA was a constructive substructure search-based assembly method, which well dealt with substructural overlaps (Carhart et al. 1981). In the 80s, a set of CASE papers, CHEMICS, was contributed to the field by Japanese scientists (Sasaki et al. 1978). The vector representation of components and their usage in the generation process were the novelties of the study. All component sets were ranked from primary to tertiary used in the extension process.

In assembly methods, molecular extensions usually end up with a combinatorial explosion. To cope with this problem, orderly structure generators have been developed. MOLGEN is a well known orderly structure generator (Gugisch et al. 2014). As descendants of DENDRAL and MASS methods, MOLGEN also generates structures first as connectivity matrices with respect to chemical constraints, then stored in an output file. In the matrix generation, rows (or columns) are built in descending order. MOLGEN is an efficient but commercial structure generator. Besides MOLGEN, another commercial structure generator is from one of the MASS developers, Michael Elyashberg, ACD Labs. The structure generator is a part of the commercial CASE system, StrucEluc (Blinov et al. 2001).

Unlike assembly methods, reduction methods construct a hypergraph with all possible bonds among atoms. First, the existence of substructures are checked in the hypergraph, then the irrelevant bonds are removed. If a substructure is not in the hypergraph anymore, it is removed from chemical constraints. In some assembly methods, substructural overlaps were not taken into account, however, all structure reduction methods dealt well with structural overlaps due to the hypergraph structure. COCOA was the earliest reduction method from Morton E. Munk and Craig A. Shelley, later integrated into the

CASE system, called SESAMI (Christie and Munk 1988; Madison et al. 1998). In COCOA algorithm, substructures were represented as atom centered fragments which were the lists of atoms' first neighbours. As successor of COCOA, HOUDINI was also released (Korytko et al. 2003). Although structural overlaps were taken into account, the massive size of hypergraphs was the main disadvantage of reduction approaches. Bohanec combined two methods to overcome the disadvantages of assembly and reduction methods. The algorithm, GEN, avoided irrelevant bonds and assembled atoms based on substructural information (Bohanec 1995).

In the field, MOLGEN is currently the fastest and an efficient structure generator, however unfortunately, it is proprietary. Alternative to a commercial tool, MAYGEN was developed as an open source molecular structure generator and it is the fastest open source structure generator in the field (Yirik, Sorokina, and Steinbeck 2021).

Molecular Fragmentation and Mass Spectrometry

In metabolomics, one of the major current challenges is the identification of unknown molecules. Liquid chromatography (LC) mass spectrometry(MS) and tandem mass spectrometry(MS/MS) are widely used techniques for molecular structure identification (Allard et al. 2016). Matching the experimentally obtained spectra against spectral libraries is a basic approach for such identification, also called dereplication. However, searching in the reference libraries also has limitations, such as the library sizes and reliability of the spectral data (Djoumbou-Feunang et al. 2019). Thus, successful structure identification through dereplication needs a massive increase in the number of accurate spectral libraries (Hufsky and Böcker 2017), however, the searched molecular formula can also be a new compound. Therefore, in most of the cases, the search in compound libraries does not yield any results. Besides these libraries, search in comprehensive molecular structure databases has also been used in structure identification. To overcome the limitations of spectral libraries and molecular databases, retrieval of structural information from molecular fragments has been the widely-used method in the field. To achieve this, a variety of in silico fragmentation methods have been developed in the last 20 years, in particular rule-based methods and combinatorial fragmenters.

The rule-based methods rely on self-determined rules, developed by experimental fragmentation patterns which provide particular structural properties (Allard et al. 2016). These methods grant consistent, fast and accurate results. The DENDRAL project recorded the first usage of fragmentation mass spectra and fragmentation rules. Nowadays, the state-of-the-art rule-based fragmenters are Mass Frontier, ACD/MS and MOLGEN/MS. The main disadvantage of rule-based methods is the need for expert-curated rules (Hufsky, Scheubert, and Böcker 2014; Ridder et al. 2012). Therefore, the usage of the combinatorial fragmentation approaches has been relatively increased (Bohanec 1995). Different from the rule-based methods, combinatorial algorithms generate fragments in an expert-free way, based on the cleavage of the chemical bonds in a given molecule. The cleaving bonds are first scored, then the bond disconnection is performed based on these penalty scores. Many combinatorial algorithms, such as FiD (Fragment iDentificator) and MetFrag, vary based on their scoring functions (Heinonen et al. 2008; Ruttkies et al. 2016). FiD is a software to identify the structure of the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 November 2021 doi:10.20944/preprints202111.0546.v1

resulting ions of the MS/MS spectrometry application to organic compounds with low molecular weight (Heinonen et al. 2008). It is advantageous for the analysis of compounds, for which the mechanisms of fragmentation are not perfectly known. In FiD, there are two different methods for fragmentation such as the single-step model and the multi-step model. The single-step model does not consider the intermediate fragments and with the multi-step model, the analysis for complex fragmentation pathways is possible. Another combinatorial fragmenter, MAGMA, arranges substructures of the input compound into their best form, explaining the fragmentation model of MS^n spectra tree (Ridder et al. 2012). In the different levels of MS^n data, there is a variety of fragment peaks and the algorithm constructs the hierarchical tree representation of these substructures in the MS^n data.

In chemical graph theory, understanding the generation of subgraphs of chemical structures is as crucial as the generation of molecular structures from a set of bonds and atoms. The coherent fragmentation accelerates the identification of the unknown molecule.

Molecular Descriptors

Molecular descriptors are defined by Todeschini and Consonni as:

"The molecular descriptor is the final result of a logical and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment."

Molecular descriptors are, in a broad sense, a way to describe and quantify a chemical structure with mathematical and cheminformatics tools. It needs to be clear that there is no molecular descriptor that fits all applications and that the same molecule can be meaningfully analyzed and described with different descriptors depending on the question to be answered and aims to be reached. Various types of molecular descriptors exist, and many involve chemical graph theory in their definition (Mauri, Consonni, and Todeschini 2017). Among those, there are chemical indices, topological indices, autocorrelation descriptors, geometrical descriptors and some of the molecular fingerprints. Most of them reveal themselves useful for CASE: to measure the topology and geometry between the query and target molecules, quickly identify identical features between a big number of chemical graphs or enable fast filtering of chemical libraries based on required molecular features.

Topological indices

Topological indices are two-dimensional molecular descriptors based on the topology of the molecular structure when represented as a graph. The molecular graph is the first topological index, as it is the 2D graph representation of a molecule. Therefore, the graph G=(V,E) represents the molecular structure, the set of vertices V represents the set of atoms where each vertex is an atom and the set of edges E represents the bonds between the atoms. The molecular graph is an undirected weighted sparse multigraph. The usage of a graph to depict a molecular structure allows applying well-known graph theory algorithms to it, in order to extract meaningful topological information. These molecular graphs are commonly represented as adjacency matrices or as collections of adjacency lists, where two atoms are adjacent if there is a chemical bond connecting them.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 November 2021 doi:10.20944/preprints202111.0546.v1

However, adjacency matrices are only one of the possible matrices that can be calculated from a molecular graph, and that can contain different information. These matrices, also called graph-theoretical matrices can be molecular descriptors by themselves or starting points for other molecular descriptors. There are three main types of graph-theoretical matrices: vertex matrices, edge matrices and incidence matrices. Vertex matrices are square matrices where rows and columns refer to graph vertices that are the atoms of the molecule, and each element of the matrix contains information related to the pair of atoms. Edge matrices are also square matrices where rows and columns refer to graph edges that are the chemical bonds of the molecule, and each element of the matrix contains information related to the pair of bonds. Incidence matrices are generally not square and contain information about different types of objects in their rows and columns, such as atoms versus edges, or even bigger molecular elements, such as cycles, molecular fragments or paths.

The adjacency matrix, which is a vertex matrix, allows calculating the Lovasz–Pelikan index, which is the largest eigenvalue of the matrix. The other most-used vertex matrix is the topological distance matrix, where each value d_{ij} of it represents the number of edges in the shortest path between the vertices i and j in the molecular graph. The most known and used topological index in chemistry, based on the topological distance matrix, is the Wiener index. This index is defined by summing the length of all shortest paths between non-hydrogen atom pairs (Trinajstic 2018). For a molecule, the half-sum of its distance matrix returns its Wiener index. Much more vertex, edge and incidence matrix-based molecular descriptors have been described across the years by numerous researchers and have then been compiled by Todeschini and Consonni (R Todeschini et al. 2009).

Using a matrix representation of molecular graphs allows applying a multitude of matrix operators and manipulators enabling calculation of numerous sets of molecular descriptors that highlight diverse information about the molecules.

Chemical Indices

The first usage of chemical indices came from the studies of Calingaert and Hladky (Calingaert and Hladky 1936). In the study, the proportion of molecular volume and the number of carbons in a hydrocarbon was described as a chemical index for the properties of hydrocarbons. The similar index was used also in the study of Kopp, summing the different atom types to describe the volume and densities of molecules (Kopp 1844). Besides the usage of atom numbers as indices, Wiener introduced one of the earliest graph invariants for the correlation between molecular properties and structural features (Randic 1975). In his study, the structural index was used for the determination of paraffin' boiling point (Wiener 1947). In 1971, Hosoya introduced a new index: the Hosoya index, which is the number of edge matchings in a graph and formulated as (Hosoya, Hosoi, and Gutman 1975):

$$H = \sum_{i=0}^{n/2} P(G, i) \quad (4)$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 November 2021 doi:10.20944/preprints202111.0546.v1

In this formula, P(G,i) is the number of matchings of i-mutually non-adjacent edges in the graph, in other words, i-covering of chemical graphs. The index has been used in a variety of applications such as the modelling of physicochemical properties of hydrocarbon atoms (Hosoya, Hosoi, and Gutman 1975). The usage of the index was described in Hosoya's review (Trinajstić 1986).In 1947, bonds were differentiated by valences of their end vertices with the study of Hartmann (Hartmann 1947). This study was extended by Randic's topological index in 1975 (Randic 1975). It was also called connectivity index since the formula was based on bond weights. In the formula, given below, d(i) and d(j) represent atom valences for vertices i and j. Bond weight is calculated with the formula $[d(i).d(j)]^{-½}$. Thus, the Randic index is calculated by summing all bond weights (Randic 1975):

$$\chi = \sum_{\text{all edges}} [d(i)\cdot d(j)]^{-1/2} .$$

(5)

This index is not a reliable descriptor for the characterization of molecules since non-isomorphic structures might have the same Randic index (Trinajstic 2018). Later, Balaban contributed to the field by slightly modifying the Randic Index (Balaban 1982). Different from the Randic index, the average of the bond weight sum was taken in the Balaban index. The index formula is :

$$J = [M/(\mu + 1)] \sum_{\text{all edges}} [D(i)\cdot D(j)]^{-1/2}$$

(6)

The average value is calculated by multiplying the Randic index by [M/(m+1)]. M is the number of edges and m is the number of cycles in the graph.

In the literature, there are more than 100 different chemical indices, often described as topological indices, which they are generally similar to.

Other theoretical molecular descriptors

Geometrical descriptors are based on three-dimensional molecular structures, where the position of the atoms in the 3D space is known and the connections between them are defined. Molecular 3D structures can be experimentally elucidated from crystallographic or NMR data or computed using molecular optimization algorithms. Geometrical descriptors have a higher information content than those based on 2D structures, such as topological descriptors and chemical indices, but have to be treated with the awareness that their values heavily depend on the molecule conformation, and can vary depending on the latter. Two of the most known classes of three-dimensional descriptors are the WHIM (Weighted Holistic Invariant Molecular) descriptors and the GETAWAY (Geometry, Topology, and AtomWeights Assembly) descriptors (Roberto Todeschini, Lasagni, and Marengo 1994; R Todeschini et al. 1996; Consonni et al. 2002).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 November 2021 doi:10.20944/preprints202111.0546.v1

Molecular fingerprints are structural keys that are a well-defined bit list of molecular features present or absent in a molecule. These features can have 2D and/or 3D structural properties and have been very useful for molecular searches, in particular the similarity search, due to their computational effectiveness. They are particularly suitable for Tanimoto and Tversky similarity indices. The molecular fingerprints are determined by fragmenting the molecule in all possible substructures following a set of rules. MACCS and PubChem fingerprints are the most widely used nowadays. The MACCS fingerprint has been developed by the MDL Information Systems (now BIOVIA) and the public version contains 166 pre-computed molecular features that might be present or absent in a molecule. The PubChem fingerprint has been developed to enable efficient and fast similarity search in the PubChem database and is composed of 881 pre-computed structural features. An alternative to the structural keys is hashed fingerprints. This type of fingerprint does not require a pre-computed set of molecular features, but rather breaks the molecular structure on a set of all possible substructures following a set of pre-defined rules. The path-centred Morgan fingerprints and the atom-centred circular fingerprints are the most-used hashed fingerprints, and they are generally used to find structural patterns and substructural similarities in molecules.

Toolkits

Most cheminformatics toolkits are mainly based on chemical graph theory (Table 1).

Table 1. List of Cheminformatics Toolkits.

Software Names	Web Links
3D-e-Chem	https://3d-e-chem.github.io/
Accord SDK	http://accelrys.com/products/datasheets/accord-software-development-kit.pdf
ACD/ StrucEluc	https://www.acdlabs.com/products/com_iden/elucidation/struc_eluc/
ADMET Predictor	http://www.simulations-plus.com
ASSEMBLE	www.upstream.ch/main.html
CACTVS	http://www.xemistry.com/academic
CDD Vault	https://www.collaborativedrug.com/cdd-vault
CDK	https://cdk.github.io/
ChemDoodle API	http://www.ichemlabs.com
chemf	https://github.com/stefan-hoeck/chemf
ChemmineR	http://manuals.bioinformatics.ucr.edu/home/chemminer
COCON	cocon.nmr.de
Daylight	http://www.daylight.com/products/toolkit.html
DENDRAL CONGEN+GENOA	www.softwarepreservation.org/projects/AI/DENDRAL/DENDRAL-CONGEN_GENOA.zip/view

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 November 2021 doi:10.20944/preprints202111.0546.v1

Enalos KNIME nodes	http://tech.knime.org/community/enalos-nodes
Enalos+ KNIME nodes	http://enalosplus.novamechanics.com/
frowns	http://frowns.sourceforge.net/
Helium	https://web.archive.org/web/20140407082845/http://www.moldb.net/helium.html
Indigo	http://lifescience.opensource.epam.com/indigo
LSD	eos.univ-reims.fr/LSD/index_ENG.html
Marvin, JChem	http://www.chemaxon.com
MAYGEN	https://github.com/MehmetAzizYirik/MAYGEN
MedChem Designer	http://www.simulations-plus.com
MedChem Studio	http://www.simulations-plus.com
Molecular Operating Environment (MOE)	https://web.archive.org/web/20160909172415/http://www.chemcomp.com/MOE-Cheminformatics_and_QSAR.htm
MolecularGraph.jl	https://github.com/mojaie/MolecularGraph.jl
MolEngine	https://www.scilligence.com/web/scilligence-regmol/
MOLGEN	www.molgen.de
MOLSIG	molsig.sourceforge.net
OEChem	http://eyesopen.com/
OMG	sourceforge.net/p/openmg
OpenBabel	http://openbabel.org/
OUCH	http://www.pharmash.com/posts/2010-08-02-ouch.html
PerlMol	https://web.archive.org/web/20120315121757/http://www.perlmol.org/
PMG	sourceforge.net/projects/pmgcoordination
Rcpi	https://bioconductor.org/packages/Rcpi
RDKit	http://www.rdkit.org/
SENECA	github.com/steinbeck/seneca
SMOG	ccl.net/cca/software/MS-DOS/SMOG
SMSD	http://www.ebi.ac.uk/thornton-srv/software/SMSD/

Conclusion

Chemical graph theory is a branch of mathematics combining graph theory and chemistry. The field was ensued due to the necessity of mathematical modelling. Molecular structures are represented as graphs and their mathematical analysis is performed based on the graph theorems. One of the key implementations of chemical graphs is the structure elucidation. For the identification of molecular structures, chemists need to build the structures with respect to its spectral data. The

methods for the elucidation process are part of chemical graph theory. In this review, we presented the key chemical graph theory methods for CASE. Starting from the spectral data of an unknown molecular structure, first the atom, bond and subgraph information are retrieved, then the enumeration as well as the generation of its isomers can be performed. There are further analyses to determine the best candidate structures, such as molecular descriptor based comparison. Although there are recently many effective contributions in the field, there is always the necessity of further enhancements. One of the modern directions of science is artificial intelligence, and the chemical graph theorems will have more and more impact on the improvements of AI methods such as machine learning, and deep neural networks.

References

Alen, R. 2018. *Carbohydrate Chemistry: Fundamentals And Applications*. World Scientific Publishing Company. https://books.google.de/books?id=tWdhDwAAQBAJ.

Alexanderson, Gerald L. 2006. "About the Cover: Euler and Königsberg's Bridges: A Historical View." *Bulletin of the American Mathematical Society* 43 (04): 567–74. https://doi.org/10.1090/S0273-0979-06-01130-X.

Allard, Pierre-Marie, Tiphaine Péresse, Jonathan Bisson, Katia Gindro, Laurence Marcourt, Van Cuong Pham, Fanny Roussi, Marc Litaudon, and Jean-Luc Wolfender. 2016. "Integration of Molecular Networking and In-Silico MS/MS Fragmentation for Natural Products Dereplication." *Analytical Chemistry* 88 (6): 3317–23. https://doi.org/10.1021/acs.analchem.5b04804.

Badertscher, Martin, Andrew Korytko, Klaus-Peter Schulz, Mark Madison, Morton E Munk, Pius Portmann, Martin Junghans, Patrick Fontana, and Ernö Pretsch. 2000. "Assemble 2.0: A Structure Generator." *Chemometrics and Intelligent Laboratory Systems* 51 (1): 73–79. https://doi.org/10.1016/S0169-7439(00)00056-3.

Balaban, Alexandru T. 1982. "Highly Discriminating Distance-Based Topological Index." *Chemical Physics Letters* 89 (5): 399–404. https://doi.org/10.1016/0009-2614(82)80009-2.

Balasubramanian, Krishnan. 2018. "Combinatorial Enumeration of Isomers of Superaromatic Polysubstituted Cycloarenes and Coronoid Hydrocarbons with Applications to NMR." *The Journal of Physical Chemistry A* 122 (41): 8243–57. https://doi.org/10.1021/acs.jpca.8b08784.

Beeler, Robert A. 2015. "Advanced Counting—Pólya Theory." In *How to Count*, edited by Robert A Beeler, 219–55. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-13844-2_8.

Berzelius, Jöns Jakob. 1830. "On the Composition of Tartaric Acid and Racemic Acid (John's Acid from the Vosges Mountains), on the Atomic Weight of Lead Oxide, Together with General Remarks on Those Substances Which Have the Same Composition but Different Properties." *Poggendorff's Annalen Der Physik Und Chemie* 19: 305.

Blinov, K A, M E Elyashberg, S G Molodtsov, A J Williams, and E R Martirosian. 2001. "An Expert System for Automated Structure Elucidation Utilizing 1 H- 1 H, 13 C- 1 H and 15 N- 1 H 2D NMR Correlations." *Fresenius' Journal of Analytical Chemistry* 369 (7–8): 709–14. https://doi.org/10.1007/s002160100757.

Bohanec, Simona. 1995. "Structure Generation by the Combination of Structure Reduction and Structure Assembly." *Journal of Chemical Information and Computer Sciences* 35: 494–503. https://doi.org/doi:10.1021/ci00025a017.

Bonchev, D. 1991. *Chemical Graph Theory: Introduction and Fundamentals*. Chemical Graph Theory. Taylor & Francis. https://books.google.de/books?id=X0AG7HhiccoC.

Butlerov, Alexander. 1861. *Zeitschrift Für Chemie Und Pharmacie: Correspondenzbl., Archiv u. Krit. Journal f. Chemie, Pharmacie u. d. Verwandten Disciplinen*. Zeitschrift Für Chemie Und Pharmacie: Correspondenzbl., Archiv u. Krit. Journal f. Chemie, Pharmacie u. d. Verwandten Disciplinen, 4. c. Bangel & Schmitt. https://books.google.de/books?id=R5RTAAAAcAAJ.

Butlerov, Alexander. 1862. "Ueber Die Verwandtschaft Der Mehraffinen Atome." *Zeitschrift Für*

Chemie 5: 297–304.

Bytautas, L, and D J Klein. 1998. "Chemical Combinatorics for Alkane-Isomer Enumeration and More." *Journal of Chemical Information and Computer Sciences* 38 (6): 1063–78. https://doi.org/10.1021/ci980095c.

Calingaert, George, and John W Hladky. 1936. "A Method of Comparison and Critical Analysis of the Physical Properties of Homologs and Isomers. The Molecular Volume of Alkanes *." *Journal of the American Chemical Society* 58 (1): 153–57. https://doi.org/10.1021/ja01292a044.

Canals, Benjamin, and H Schober. 2012. "Introduction to Group Theory." *EPJ Web of Conferences* 22 (March): 00004. https://doi.org/10.1051/epjconf/20122200004.

Cardwell, D S L. 1969. "John Dalton and the Progress of Science." *British Journal for the Philosophy of Science* 20 (2).

Carhart, Raymond E, Dennis H Smith, Neil A B Gray, James G Nourse, and Carl Djerassi. 1981. "Applications of Artificial Intelligence for Chemical Inference. 37. GENOA: A Computer Program for Structure Elucidation Utilizing Overlapping and Alternative Substructures." *The Journal of Organic Chemistry* 46 (8): 1708–18. https://doi.org/10.1021/jo00321a037.

Cayley, A. 1857. "XXVIII. On the Theory of the Analytical Forms Called Trees." *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 13 (85): 172–76. https://doi.org/10.1080/14786445708642275.

Christie, B D, and M E Munk. 1988. "Structure Generation by Reduction: A New Strategy for Computer-Assisted Structure Elucidation." *Journal of Chemical Information and Computer Sciences* 28 (2): 87–93. https://doi.org/10.1021/ci00058a009.

Consonni, Viviana, Roberto Todeschini, Manuela Pavan, and Paola Gramatica. 2002. "Structure/Response Correlations and Similarity/Diversity Analysis by GETAWAY Descriptors. 2. Application of the Novel 3D Molecular Descriptors to QSAR/QSPR Studies." *Journal of Chemical Information and Computer Sciences* 42 (3): 693–705. https://doi.org/10.1021/ci0155053.

Couper, Archibald Scott. 1858. "Sur Une Nouvelle Théorie Chimique." *Annales de Chimie et de Physique* 53: 488–89.

Dalton, John. 2010. *A New System of Chemical Philosophy*. *Cambridge Library Collection - Physical Sciences*. Vol. 1. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511736391.

Dias, Jerry Ray. 1984. "A Periodic Table for Polycyclic Aromatic Hydrocarbons. IV. Isomer Enumeration of Polycyclic Conjugated Hydrocarbons. 2." *Journal of Chemical Information and Computer Sciences* 24 (3): 124–35. https://doi.org/10.1021/ci00043a002.

Djoumbou-Feunang, Yannick, Allison Pon, Naama Karu, Jiamin Zheng, Carin Li, David Arndt, Maheswor Gautam, Felicity Allen, and David S Wishart. 2019. "CFM-ID 3.0: Significantly Improved ESI-MS/MS Prediction and Compound Identification." *Metabolites* 9 (4): 72. https://doi.org/10.3390/metabo9040072.

Frankland, E. 1866. *Lecture Notes for Chemical Students: Embracing Mineral and Organic Chemistry*. J. Van Voorst. https://books.google.de/books?id=YRtIAAAAIAAJ.

Gugisch, Ralf, Adalbert Kerber, Reinhard Laue, Axel Kohnert, Markus Meringer, Christoph Rücker, and Alfred Wassermann. 2014. "MOLGEN 5.0, A Molecular Structure Generator." In *Advances in Mathematical Chemistry and Applications*, edited by Ralf Gugisch, Adalbert Kerber, Axel Kohnert, Reinhard Laue, Markus Meringer, Christoph Rücker, and Alfred Wassermann, 113–38. BENTHAM SCIENCE PUBLISHERS. https://doi.org/10.2174/9781608059287114010010.

Hartmann, Hermann. 1947. "Eine Neue Quantenmechanische Behandlung von CH4 Und NH4+." *Zeitschrift Für Naturforschung A* 2: 489–92.

HEIN, GEORGE E. 1966. "Kekulé and the Architecture of Molecules." In , 1–12. https://doi.org/10.1021/ba-1966-0061.ch001.

Heinonen, Markus, Ari Rantanen, Taneli Mielikäinen, Juha Kokkonen, Jari Kiuru, Raimo A Ketola, and Juho Rousu. 2008. "FiD: A Software for Ab Initio Structural Identification of Product Ions from Tandem Mass Spectrometric Data." *Rapid Communications in Mass Spectrometry* 22 (19): 3043–52. https://doi.org/10.1002/rcm.3701.

Higgins, Bryan. 1791. *A Comparative View of the Phlogistic and Antiphlogistic Theories ... The Second*

Edition. J. Murray. https://books.google.de/books?id=1RXeDN5zzGIC.
Holdsworth, Jason. 1999. "The Nature of Breadth-First Search," February.
Hosoya, Haruo, Kikuko Hosoi, and Ivan Gutman. 1975. "A Topological Index for the Total?-Electron Energy." *Theoretica Chimica Acta* 38 (1): 37–47. https://doi.org/10.1007/BF01046555.
Hufsky, Franziska, and Sebastian Böcker. 2017. "Mining Molecular Structure Databases: Identification of Small Molecules Based on Fragmentation Mass Spectrometry Data." *Mass Spectrometry Reviews* 36 (5): 624–33. https://doi.org/10.1002/mas.21489.
Hufsky, Franziska, Kerstin Scheubert, and Sebastian Böcker. 2014. "Computational Mass Spectrometry for Small-Molecule Fragmentation." *TrAC Trends in Analytical Chemistry* 53: 41–48. https://doi.org/10.1016/j.trac.2013.09.008.
Jiping, Zhang. 1992. "On Finite Groups All of Whose Elements of the Same Order Are Conjugate in Their Automorphism Groups." *Journal of Algebra* 153 (1): 22–36. https://doi.org/10.1016/0021-8693(92)90146-D.
Kerber, Adalbert, Reinhard Laue, Markus Meringer, Christoph Rücker, and Emma Schymanski. 2013. *Mathematical Chemistry and Chemoinformatics*. DE GRUYTER. https://doi.org/10.1515/9783110254075.
Kopp, Hermann. 1844. "Ueber Den Zusammenhang Zwischen Der Chemischen Constitution Und Einigen Physikalischen Eigenschaften Bei Flüssigen Verbindungen." *Annalen Der Chemie Und Pharmacie* 50 (1): 71–144. https://doi.org/10.1002/jlac.18440500105.
Korytko, A, K-P Schulz, M S Madison, and M E Munk. 2003. "HOUDINI: A New Approach to Computer-Based Structure Generation." *Journal of Chemical Information and Computer Sciences* 43 (5): 1434–46. https://doi.org/10.1021/ci034057r.
Madison, Mark, Klaus-Peter Schulz, Andrew Korytko, and Morton E Munk. 1998. "SESAMI: An Integrated Desktop Structure Elucidation Tool." *The Internet Journal of Chemistry* 1: CP1-U22.
Mauri, Andrea, Viviana Consonni, and Roberto Todeschini. 2017. "Molecular Descriptors." In *Handbook of Computational Chemistry*, 2065–93. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-27282-5_51.
McEachran, Andrew D, Ilya Balabin, Tommy Cathey, Thomas R Transue, Hussein Al-Ghoul, Chris Grulke, Jon R Sobus, and Antony J Williams. 2019. "Linking in Silico MS/MS Spectra with Chemistry Data to Improve Identification of Unknowns." *Scientific Data* 6 (1): 141. https://doi.org/10.1038/s41597-019-0145-z.
Munk, Morton E, Craig A Shelley, Hugh B Woodruff, and Mark O Trulson. 1982. "Computer-Assisted Structure Elucidation." *Fresenius' Zeitschrift Für Analytische Chemie* 313 (6): 473–79. https://doi.org/10.1007/BF00483534.
Pasteur, Luis. 1848. *Comptes Rendus Hebdomadaires de l'Académie Des Sciences*.
Pevac, S, and G Crundwell. 2000. "Pólya's Isomer Enumeration Method: A Unique Exercise in Group Theory and Combinatorial Analysis for Undergraduates." *Journal of Chemical Education* 77 (10): 1358. https://doi.org/10.1021/ed077p1358.
Putri, Sheila Eka, Tulus Tulus, and Normalina Napitupulu. 2011. *Implementation and Analysis of Depth-First Search (DFS) Algorithm for Finding The Longest Path*. https://doi.org/10.13140/2.1.2878.2721.
Randic, Milan. 1975. "Characterization of Molecular Branching." *Journal of the American Chemical Society* 97 (23): 6609–15. https://doi.org/10.1021/ja00856a001.
Ridder, Lars, Justin J J van der Hooft, and Stefan Verhoeven. 2014. "Automatic Compound Annotation from Mass Spectrometry Data Using MAGMa." *Mass Spectrometry* 3 (Special_Issue_2): S0033–S0033. https://doi.org/10.5702/massspectrometry.S0033.
Ridder, Lars, Justin J J van der Hooft, Stefan Verhoeven, Ric C H de Vos, René van Schaik, and Jacques Vervoort. 2012. "Substructure-Based Annotation of High-Resolution Multistage MS n Spectral Trees." *Rapid Communications in Mass Spectrometry* 26 (20): 2461–71. https://doi.org/10.1002/rcm.6364.
Ruttkies, Christoph, Emma L Schymanski, Sebastian Wolf, Juliane Hollender, and Steffen Neumann. 2016. "MetFrag Relaunched: Incorporating Strategies beyond in Silico Fragmentation." *Journal of Cheminformatics* 8 (1): 3. https://doi.org/10.1186/s13321-016-0115-9.
Sabater, Carlos, Agustín Olano, Nieves Corzo, and Antonia Montilla. 2019. "GC–MS

Characterisation of Novel Artichoke (Cynara Scolymus) Pectic-Oligosaccharides Mixtures by the Application of Machine Learning Algorithms and Competitive Fragmentation Modelling." *Carbohydrate Polymers* 205 (February): 513–23. https://doi.org/10.1016/j.carbpol.2018.10.054.

Sasaki, Shin-ichi, Hidetsugu Abe, Yuji Hirota, Yoshiaki Ishida, Yoshihiro Kudo, Shukichi Ochiai, Keiji Saito, and Tohru Yamasaki. 1978. "CHEMICS-F: A Computer Program System for Structure Elucidation of Organic Compounds." *Journal of Chemical Information and Computer Sciences* 18 (4): 211–22. https://doi.org/10.1021/ci60016a007.

Serov, V.V., M.E. Elyashberg, and L.A. Gribov. 1976. "Mathematical Synthesis and Analysis of Molecular Structures." *Journal of Molecular Structure* 31 (2): 381–97. https://doi.org/10.1016/0022-2860(76)80018-X.

Sutherland Stanford University., Computer Science Department., United States., Advanced Research Projects Agency., Georgia L. 1967. *Dendral–a Computer Program for Generating and Filtering Chemical Structures*. [Stanford, Calif.]: [Computer Science Dept., Stanford University].

Sylvester, J. J. 1878. "On an Application of the New Atomic Theory to the Graphical Representation of the Invariants and Covariants of Binary Quantics, with Three Appendices." *American Journal of Mathematics* 1 (1): 64. https://doi.org/10.2307/2369436.

Todeschini, R, V Consonni, R Mannhold, H Kubinyi, and G Folkers. 2009. *Molecular Descriptors for Chemoinformatics: Volume I: Alphabetical Listing / Volume II: Appendices, References*. Methods & Principles in Medicinal Chemistry. Wiley. https://books.google.de/books?id=03iAsdAcXHcC.

Todeschini, R, M Vighi, R Provenzani, A Finizio, and P Gramatica. 1996. "Modeling and Prediction by Using Whim Descriptors in QSAR Studies: Toxicity of Heterogeneous Chemicals on Daphnia Magna." *Chemosphere* 32 (8): 1527–45. https://doi.org/10.1016/0045-6535(96)00060-4.

Todeschini, Roberto, Marina Lasagni, and Emilio Marengo. 1994. "New Molecular Descriptors for 2D and 3D Structures. Theory." *Journal of Chemometrics* 8 (4): 263–72. https://doi.org/10.1002/cem.1180080405.

Trinajstić, N. 1986. *Mathematical and Computational Concepts in Chemistry*. Ellis Horwood Series in Mathematics and Its Applications. Ellis Horwood. https://books.google.de/books?id=1C4jAQAAIAAJ.

Trinajstic, N. 2018. *Chemical Graph Theory*. CRC Press. https://books.google.de/books?id=33laDwAAQBAJ.

Werner, A. 1912. "Über Die Raumisomeren Kobaltverbindungen." *Justus Liebigs Annalen Der Chemie* 386 (1–2): 1–272. https://doi.org/10.1002/jlac.19123860102.

Werner, Alfred. 1893. "Beitrag Zur Konstitution Anorganischer Verbindungen." *Zeitschrift Für Anorganische Chemie* 3 (1): 267–330. https://doi.org/10.1002/zaac.18930030136.

Wiener, Harry. 1947. "Structural Determination of Paraffin Boiling Points." *Journal of the American Chemical Society* 69 (1): 17–20. https://doi.org/10.1021/ja01193a005.

Yirik, Mehmet Aziz, Maria Sorokina, and Christoph Steinbeck. 2021. "MAYGEN: An Open-Source Chemical Structure Generator for Constitutional Isomers Based on the Orderly Generation Principle." *Journal of Cheminformatics* 13 (1): 48. https://doi.org/10.1186/s13321-021-00529-9.