Kekulé structure enumeration yields unique SMILES

Martin Mann[1] and Bernhard Thiel[2]

[1]Bioinformatics, Department for Computer Science, University of Freiburg,
George-Köhler-Allee 106, 79106 Freiburg, Germany,
[2]Institute for Theoretical Chemistry, University of Vienna, Währingerstrasse 17, 1090
Vienna, Austria
mmann@informatik.uni-freiburg.de

Abstract. A standard representation of molecules is based on graphs where atoms correspond to vertices and covalent bonds are represented by a number of edges according to the bond order. This depiction reaches its limitations for aromatic molecules where the aromatic ring can be encoded by different bond order layouts, i.e. Kekulé structures, since electrons are shared within the ring rather than fixed to a specific bond. Thus, several Kekulé structures are possible for aromatic molecules. Here, we propose a new constraint programming based approach to enumerate all Kekulé structures for a given molecule. Furthermore, the ambiguity information derived is used to enable a unique Kekulé-based SMILES encoding of the molecule independent of any aromaticity detection algorithm. This is of importance, since there is no generally accepted aromaticity definition available that covers all cases.

1 Introduction

Molecules are often depicted as undirected graphs representing atoms as vertices and covalent single, double, or triple bonds by an according number of edges as given in Fig. 1, known as structural formula. This works well as long as it is possible to specifically assign electron pairs shared between two atoms to individual bonds. In that case, a unique graph representation can be given. But the depiction fails as soon as electrons are not uniquely assignable, a phenomenon named mesomerism. A classic example is benzene shown in Fig. 1 a). Two different graph representations, i.e. bond assignments, can be given and these were first identified and introduced by August Kekulé in 1872 [6]. Since that time, such explicit structural formula for molecules with ambivalent rings (different single-double-bond assignments) are refered to as resonance or *Kekulé structures*. In the following, we will focus only on mesomerism of rings within molecules, other forms are discussed later. This ambiguity usually poses no problem for most applications but gets crucial as soon as a unique representation of a molecule is needed, e.g. within chemical compound databases [2, 5] or when atom mappings for reactions are to be identified [9].

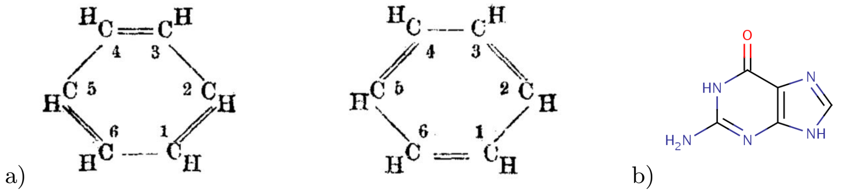


Fig. 1. a) The two isomeric structures of benzene identified by Kekulé (taken from [6]) and b) the Kekulé structure of guanine.

Whether or not a molecule gives rise to several Kekulé structures usually depends on the existence of (hetero) aromatic rings within the molecule. Within aromatic rings, bond electrons are shared within the ring and no unique single-double-bond assignment is possible, resulting in multiple Kekulé representations. The number of Kekulé structures is therefore combinatoric in the number of ambiguous aromatic rings part of the molecule. It was reasoned that the thermodynamic stability of a molecule is to some extent linked to its number of Kekulé structures [15, 3]. Most algorithms to enumerate Kekulé structures are based on graph theoretical studies and a lot of work was done in the early 80s. A thorough review of the early methods is given in [13] on pages 50-52. Therein, most algorithms were tailored to specific hydrocarbone molecule classes usually only covering benzene-like ring layouts and conjugations, e.g. [3, 1].

Within this contribution, we introduce a new constraint programming (CP) based method to enumerate all Kekulé structures for a given molecule. This encodes all possibibly ambiguouos edges and enumerates all valid bond assignments and thus all Kekulé structures. This is of importance for instance to parse molecules given in SMILES format [14] (later discussed in detail) or to provide all Kekulé variants where needed. For instance, we have recently introduced a CP-based approach for the computation of valid atom mappings for chemical reactions [8, 9]. Therein, the reaction's educt and product molecules are mapped onto each other revealing the bond breakings and formations occuring during the reaction. To this end, all Kekulé structures of all participating molecules have to be known and considered, since it is not known in advance, what specific Kekulé structure participates in the reaction. The approach is generic and not tailored to specific classes of molecules. As an input a single structural formula for each molecule has to be provided and all Kekulé structures are enumerated.

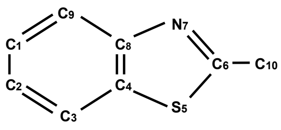
Beside the enumeration of Kekulé structures, we use the approach to enable the generation of unique SMILES strings without the need for aromaticity perception. SMILES is a standard format to represent molecules as strings [14]. The string is generated from a treelike-decomposition of the molecule, where ring closures are marked by according number pairs. For instance benzene depicted in Fig. 1 a) can be represented by `[H]C1=C([H])C([H])=C([H])C([H])=C([H])1` when hydrogens are explicitly encoded by `[H]`. Usually, hydrogens deducable from the structure are ommited leaving the SMILES `C1=CC=CC=C1`. Note, this encoding is the same for both benzene Kekulé structures, since a SMILES does not encode any node indexing.

The SMILES language copes with ring ambiguity by a special treatment of aromatic rings. Therein, bonds and atoms part of an aromatic ring are given a special lowercase label marking their ambiguity. It is left to the SMILES parser to pick one of the encoded Kekulé structures, to enumerate them all, etc. The benzene example from Fig. 1 would be encoded by `c1ccccc1` in contrast to the Kekulé structure encoding `C1=CC=CC=C1` discussed above.

The central problem for the standard SMILES approach is the lack of a decent definition of aromaticity that covers all cases of aromatic molecules [11]. Furthermore, aromaticity cannot be simply used interchangeably with mesomerism, i.e. the existence of several Kekulé structures. A simple example is guanine depicted in Fig. 1 b). Both rings of the molecule are usually assumed to be aromatic. Still guanine features only a single Kekulé structure and does not show the usual aromatic ambiguity. It is therefore generally hard to decide whether or not a ring within a molecule is aromatic or not and thus if it is to be treated ambiguous or not, which is central to generate unique SMILES [14]. Within our approach, we use a variant of the presented CP approach to identify all edges that enable ambiguity instead of annotating whole rings. Only these edges are treated differently in the SMILES generation, which results in a slightly different but aromaticity-independent SMILES encoding. The new SMILES encoding is only encoding ambiguity information where needed and results in a general, unique molecule string encoding.

2 Preliminaries

Given a structural formula of a molecule, it can be represented by an undirected graph (V, E) with vertex set V representing the atoms of the molecule and edge set $E = \{ \{v, v'\} \mid v, v' \in V\}$ representing the covalent bonds between these atoms. The bond order, i.e. the number of electron pairs shared within the bond, is given by the input adjacency matrix A where each entry $A_{v,v'} \in \{0, 1, 2, 3\}$ denotes the according bond order between v and v'. An example is given in Fig. 2.



A	1	2	3	4	5	6	7	8	9	10
1		1							2	
2	1		2							
3		2		1						
4			1		1			2		
5				1		1				
6					1		2			1
7						2		1		
8				2			1		1	
9	2							1		
10						1				

Fig. 2. An example molecular graph (without hydrogens) with $V = \{1 \ldots 10\}$ and the according adjacency matry A.

For such a graph (V, E), we identify the subgraph (V°, E°) covering only vertices and edges participating in rings since we want to enumerate Kekulé structures for ring ambiguity. To enumerate all rings, we apply the ring perception algorithm by Hanser [4], which first removes all vertices with degree one and successively decomposes the remaining ring structure into single rings in an iterative fashion. Since triple bonds form very strong and inflexible atom interactions, we ignore all triple bond containing rings. For the example in Fig. 2, the Hanser algorithm identifies the three rings $1{-}2{-}3{-}4{-}8{-}9{-}1$, $4{-}5{-}6{-}7{-}8{-}4$, and $1-2-3-4-5-6-7-8-9-1$, resulting in $V^\circ = \{1 \ldots 9\}$ (leaving out node 10).

Eventually, each vertex $v \in V^\circ$ participates in at least two edges and all edges $\{v, v'\} \in E^\circ$ are single or double bonds, i.e. $A_{v,v'} \in \{1, 2\}$, that might give rise to different Kekulé structures. All other "non-ring bonds" are assumed to be isomorphic between different Kekulé structures. Therefore, the problem of enumerating all Kekulé structures based on ring ambiguity reduces to the enumeration of all valid single-double bond assignments of the ring bonds in E°.

3 Enumerating all Kekulé structures

As introduced above, given the ring-covering subgraph (V°, E°) of a molecule's structure graph (V, E), it is sufficient to enumerate all valid single-double bond assignments for the bonds in E°. To this end, we formulate a constraint satisfaction problem as follows. For each edge $\{v, v'\} \in E^\circ$, we introduce a variable $X_{v,v'}$ with domain $D_{v,v'} = \{1, 2\}$. For each atom vertex $v \in V^\circ$, we add a linear constraint $\sum_{\{v,v'\} \in E^\circ} X_{v,v'} = \sum_{v' \in V^\circ} A_{v,v'}$, i.e. the sum over all bond orders for each atom has to be preserved by any assignment.

The example from Fig. 2 would result in 10 edge variables, e.g. $X_{1,2}, X_{2,3}, \ldots$ and 9 linear constraints, e.g. for vertex 4: $X_{3,4} + X_{4,5} + X_{4,8} = 4$.

Note, while given here in terms of integer domains that were implemented using the Gecode library v4.0 [12], an equivalent CSP can be formulated using Boolean variables and domains. In such a formulation, the boolean encoding would cover whether or not a bond is e.g. a double bond and the applied linear constraints would limit the number of double bonds to $\sum_{v' \in V^\circ} \max(0, A_{v,v'} - 1)$, i.e. the overall ring bond order minus the number of ring bonds.

Given such a CSP for a certain molecule, we can simply apply a standard first-fail depths-first-search to enumerate all valid single-double bond assignments and thus according Kekulé structures. This reveals two Kekulé structures for the discussed example molecule; both are presented in Fig. 3.

We have applied the procedure to molecules from the ChEBI database [2]. From the 15,944 molecules in the database, we derived a subset of 10,920 ring-containing molecules for which full atom information was available (68.5% of the database). For each molecule, we applied the given procedure to enumerate all Kekulé structures. In Tab. 1, we report the resulting statistics where the dataset was further clustered according to the number of rings present in a molecule.

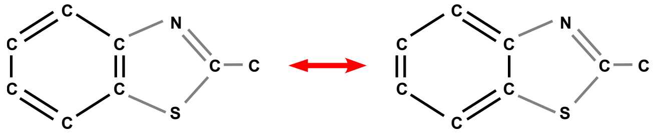


Fig. 3. The two Kekulé structures of the molecule depicted in Fig. 2 whereby both rings are aromatic.

		#ambiguous rings			#overlaps		#Kekulé structures		
#rings	#mols	median	mean	max	median	mean	median	mean	max
1	2871	0	0.4	1	0	0	1	1.4	2
2	2275	1	1	2	0	0.4	2	2.0	4
3	2261	2	1.6	3	1	0.9	2	2.5	8
4	1621	0	1.4	4	2	1.9	1	2.7	16
5	806	1	1.7	5	3	3.0	2	3.1	24
6-10	909	0	2.1	9	4	5.7	1	6.7	288
> 10	177	0	3.8	49	27	122.2	1	7.2	256

Table 1. Statistics on the number of ambivalent rings, the number of shared bonds between rings (#overlaps), and the number of distinct Kekulé structures within the ChEBI data set clustered by the number of rings per molecule.

When investigating the median of the number of ambiguous rings it becomes clear that most rings are non-ambiguous (median ~ 0) while there is on average at least one ring with ambiguity. Their average number only slightly increases with the number of rings a molecule features. The median of the number of bonds shared between rings is given in column #overlaps. Inspecting the numbers, most molecules in our data set seem to sport individual rings instead of ring fusions as e.g. for guanine in Fig. 1 b). Only for larger molecules with multiple rings, fused ring systems become more common.

5,816 molecules (53.3%) show multiple Kekulé structures, highlighting the need for appropriate ambiguity handling and enumeration. For 3,459 molecules, all present rings were ambiguous. Table 2 gives statistics on the number of ambiguous bonds for molecules with multiple Kekulé structures. On average, about half of all ring participating bonds are ambiguous. This is mainly due to ring fusions, where e.g. an ambiguous benzene ring is fused with a non-ambiguous ring. In such a case, all non-shared bonds of the second ring are non-ambiguous resulting in the presented statistics.

When averaging over the whole ChEBI data set, a “mean ring molecule” features about 3 rings where one is ambiguous with about 5 ambiguous bonds, which results in 2-3 Kekulé structures on average. This gives rise to the need for canonicalization to enable unique molecule representations within databases as discussed in the next section.

	#mols	#ringBonds
#rings	ambi/all	mean(ambi) / mean(all)
1	1276/2871	6.0 / 6.0 = 100%
2	1320/2275	6.0 / 11.0 = 54%
3	1506/2261	9.8 / 16.2 = 60%
4	798/1621	11.7 / 21.3 = 55%
5	405/806	12.7 / 25.3 = 50%
6-10	435/909	16.1 / 32.8 = 49%
> 10	76/177	17.2 / 35.4 = 49%

Table 2. Statistics on the number of ambivalent ring bonds (#ringBonds ambi) and the overall number of bonds participating in rings (#ringBonds all) for all molecules with at least two distinct Kekulé structures (first number in column #mols vs. overall number) within the ChEBI data set clustered by the number of rings per molecule.

4 Unique SMILES with ambiguous bond encoding

The previous study on the ambiguity when representing molecules with specific bond assignments highlights the need for a unique canonical molecule representation, e.g. for database lookups etc. As discussed in the introduction, the SMILES encoding was introduced for this purpose with according canonicalization algorithms [14]. Therein, atoms are represented by according standard abbreviations like “C”, “H”, “Br”, etc. (all starting upper case and enclosed in brackets if longer than one character), and bonds formed by more than one electron pair are encoded by the special characters “=“ and “#” for double or triple bonds. The ambiguity resulting from aromatic rings was handled using special character encodings for atoms participating in such rings, i.e. using lower case characters as “c”, “o”, “n”, . . . for the common aromatic-ring atoms “C”, “O”, “N”, . . . respectively, as discussed for benzene in the introduction. Furthermore, an aromatic bond label “:” was introduced, which encodes the uncertainty if the bond is a single or a double bond. This encoding works well for simple standard cases of aromatic compounds. But the central problem is the decision whether or not a ring is aromatic or not, a question still not successfully solved in chemistry [11].

For instance, given the example molecule from Fig. 2. Depending on the aromaticity annotation, there are various possibilities to encode the molecule:

both rings aromatic	: `c12ccccc2sc(C)n1`
large ring aromatic	: `c12ccccc2SC(C)=N1`
small ring aromatic	: `c12C=CC=Cc2sc(C)n1`
Kekulé: left Fig. 3	: `C12C=CC=CC=2SC(C)=N1`
Kekulé: right Fig. 3	: `C12=CC=CC=C2SC(C)=N1`

The SMILES notation thus mixes the problem of defining a unique and compact string represention for molecules with the even harder problem of aromaticity perception. Here, we will try to disentangle the two problems and to provide a solution for the first, namely the generation of unique canonical SMILES without the need for aromaticity perception.

To this end, we simply fall back to the previous problem of Kekulé structure ambiguity, which poses the true problem for canonicalization. Currently, such ambiguity is intrinsically connected with aromaticity in the SMILES encoding, but there exist many counter-examples as e.g. guanine in Fig. 1 b). In contrast, we want to encode only for variation where it occurs, i.e. the ambiguous bonds within rings.

Given the CSP formulation from above, we only perform a constraint propagation until arc-consistency is reached. *No search* is performed. The bond-encoding variables $X_{v,v'}$ that are still unassigned $|D_{v,v'}| > 1$ encode for bonds $\{v, v'\} \in E$ can either be single or double bond, i.e. ambiguous bonds.

Once this subset of ambiguous bonds is identified, we can apply a variant of the canonical SMILES algorithm from [14], where

1. all atoms are treated non-aromatic (since no aromaticity perception was done),
2. ambiguous ring-bonds (non-assigned variables) are represented by the label ":",
3. non-ambiguous ring-bonds are represented by by single ("-") or double bond label ("=") based on the according variable assignment, and
4. non-ring bonds are labeled according to the initial molecule representation encoded by the adjacency matrix A with "-", "=", or "#" for $A_{v,v'} = 1, 2$, or 3, respectively.

Given this special treatment, we can derive unique canonical SMILES without aromaticity perception using the standard SMILES canonicalization implementation as e.g. available in the Graph Grammar Library GGL [7].

In Figure 3, only the bonds of the 6-ring given in black are ambiguous leaving 4 of the 5 bonds of the smaller 5-ring (in gray) unambiguous. This results in the new non-aromatic but ambiguity-encoding SMILES representation `C12:C:C:C:C:C:1SC(C)=N2` instead of the SMILES with aromaticity annotation `c12ccccc2sc(C)n1`, which does not easily reveal the two Kekulé structures and the source of ambiguity.

5 Future work

The current approach is restricted to mesomerism of molecular ring systems based on the ambiguity of single-double bond assignments. Still, there are further sources of mesomerism that result in multiple resonance structures. A common form is the shift of unbound (valence) electrons of atoms, that define its charge, to neighbored atoms, which directly results in ambiguity. Another source for different representations of basically the same molecule is a phenomenon called tautomerism, where adjacent hydrogens are shifted to neighbored atoms resulting in a changed bond order pattern of the molecule. Furthermore, combination of both can occur. Finally, ionizations of some atoms are possible, depending on the physical conditions. Sayle gives a detailed overview of the problem in [10].

Thus, we are planning to extend the described approach to further cases of mesomerism to enable a full enumeration of resonance structures for a given molecule. When applied to the presented ChEBI data set, this might reveal even stronger abundance of ambiguity when representing molecules as structural formula.

6 Conclusions

We have introduced a constraint programming based approach to enumerate the possible Kekulé structures for a molecule that result from ring-mesomerism. The approach was used to assess the abundance of such ambiguity in the ChEBI data base, revealing that half of the data set shows at least two Kekulé structures. Furthermore, it became obvious that this ambiguity results only from a fraction of the involved ring-bonds.

Given that such ambiguity is problematic when deriving unique molecule representations, we have extended the approach to yield canonical SMILES without need for aromaticity perception. The latter was the base for the standard SMILES approach to identify and deal with ambiguity. Since aromaticity is hard to define, the fact that not all aromatic rings are ambiguous, and given our statistics on ambiguous bonds, we proposed an approach that is independent of aromaticity assignment. To this end, we identify ambiguous bonds using our CP approach. Only these bonds are treated special during the standard canonical SMILES generation. Thus, we derive unique graph-based molecule representations.

References

1. John L. Bergan, Sven J. Cyvin, and Bjorg N. Cyvin. Number of kekulé structures of single-chain corona-condensed benzenoids (cycloarenes). *Chemical Physics Letters*, 125(3):218–220, 1986.
2. Kirill Degtyarenko, Paula de Matos, Marcus Ennis, Janna Hastings, Martin Zbinden, Alan McNaught, Rafael Alcántara, Michael Darsow, Mickaël Guedj, and Michael Ashburner. ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Research*, 36(suppl 1):D344–D350, 2008.
3. B. Džonova-Jerman-Blažič and N. Trinajsticí. Computer-aided enumeration and generation of the Kekulé structures in conjugated hydrocarbons. *Computers Chemistry*, 6(3):121–132, 1982.
4. T. Hanser, P. Jauffret, and G. Kaufmann. A new algorithm for exhaustive ring perception in a molecular graph. *J. Chem. Inf. Comp. Sci.*, 36(6):1146–1152, 1996.
5. M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe. KEGG for integration and interpretation of large-scale molecular data sets. *Nuc. Acids Res.*, 40(Database issue):D109–14, 2012.
6. August Kekulé. Ueber einige Condensationsproducte des Aldehyds. *Justus Liebigs Annalen der Chemie*, 162(1):77–124, 1872.

7. M. Mann, H. Ekker, and C. Flamm. The graph grammar library - a generic framework for chemical graph rewrite systems. In Keith Duddy and Gerti Kappel, editors, *Theory and Practice of Model Transformations, Proc. of ICMT 2013*, volume 7909 of *LNCS*, pages 52–53. Springer, 2013. Extended abstract at ICMT, long version at arXiv http://arxiv.org/abs/1304.1356.
8. M. Mann, H. Ekker, P.F. Stadler, and C. Flamm. Atom mapping with constraint programming. In R. Backofen and S. Will, editors, *Proceedings of the Workshop on Constraint Based Methods for Bioinformatics WCB12*, pages 23–29, Freiburg, 2012. Uni Freiburg. http://www.bioinf.uni-freiburg.de/Events/WCB12/proceedings.pdf.
9. M. Mann, F. Nahar, H. Ekker, P.F. Stadler, and C. Flamm. Atom mapping with constraint programming. In *Proceedings of the 19th International Conference on Principles and Practice of Constraint Programming, CP'13*, LNCS. Springer, 2013. Accepted for publication.
10. Roger A. Sayle. So you think you understand tautomerism? *Journal of Computer-Aided Molecular Design*, 24(6-7):485–496, 2010.
11. Amnon Stanger. What is... aromaticity: a critique of the concept of aromaticity-can it really be defined? *Chem. Commun.*, 0:1939–1947, 2009.
12. Gecode Team. Gecode: Generic constraint development environment, 2013. Available as an open-source library from http://www.gecode.org.
13. N. Trinajstić. *Chemical Graph Theory*, volume 1. CRC Press, 1983.
14. D. Weininger. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comp. Sci.*, 28(1):31–36, 1988.
15. G. W. Wheland. The number of canonical structures of each degree of excitation for an unsaturated or aromatic hydrocarbon. *J. Chem. Phys.*, 3(6):356–361, 1935.