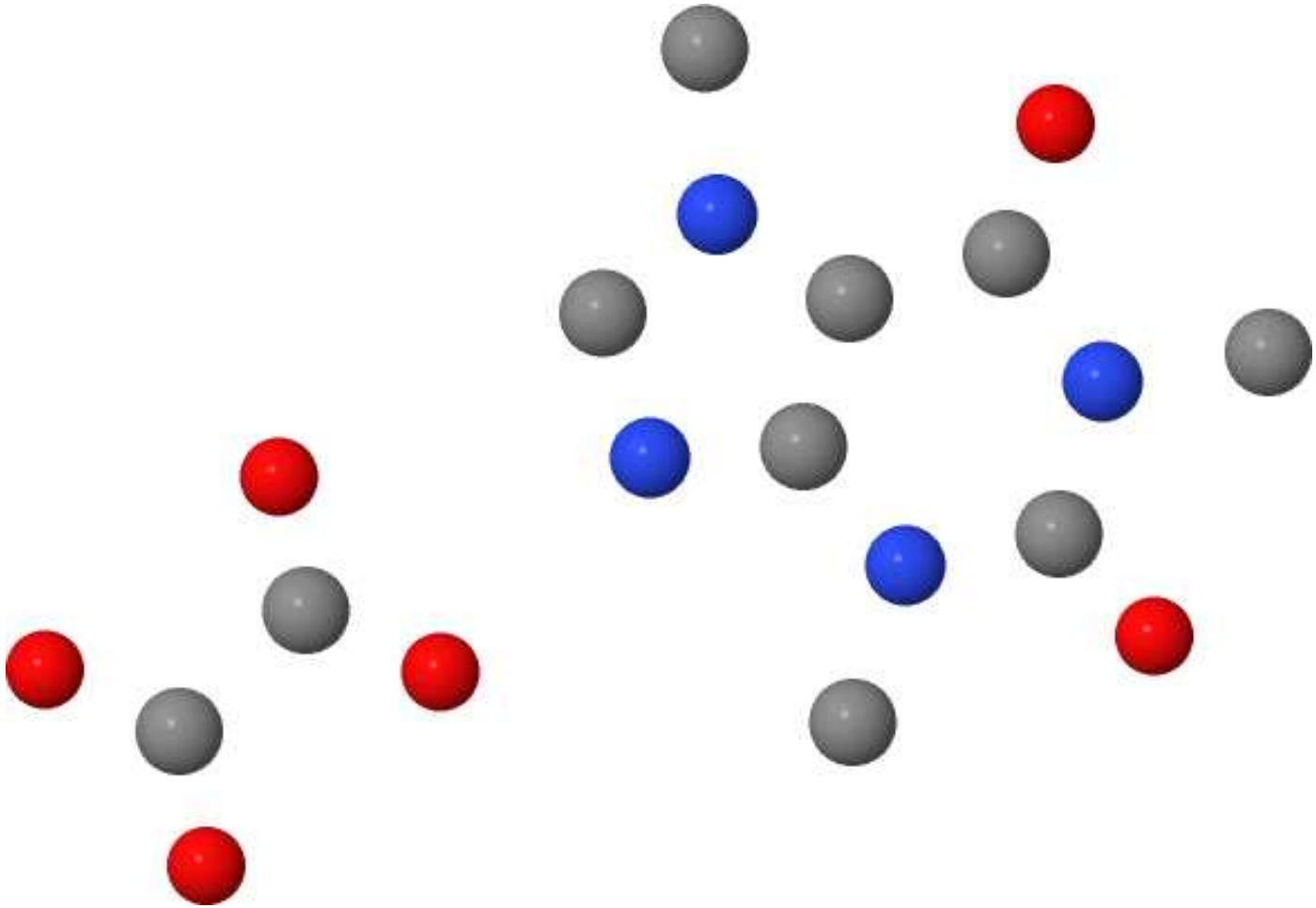
Graph algorithms in crystallography and cheminformatics

Andrius Merkys

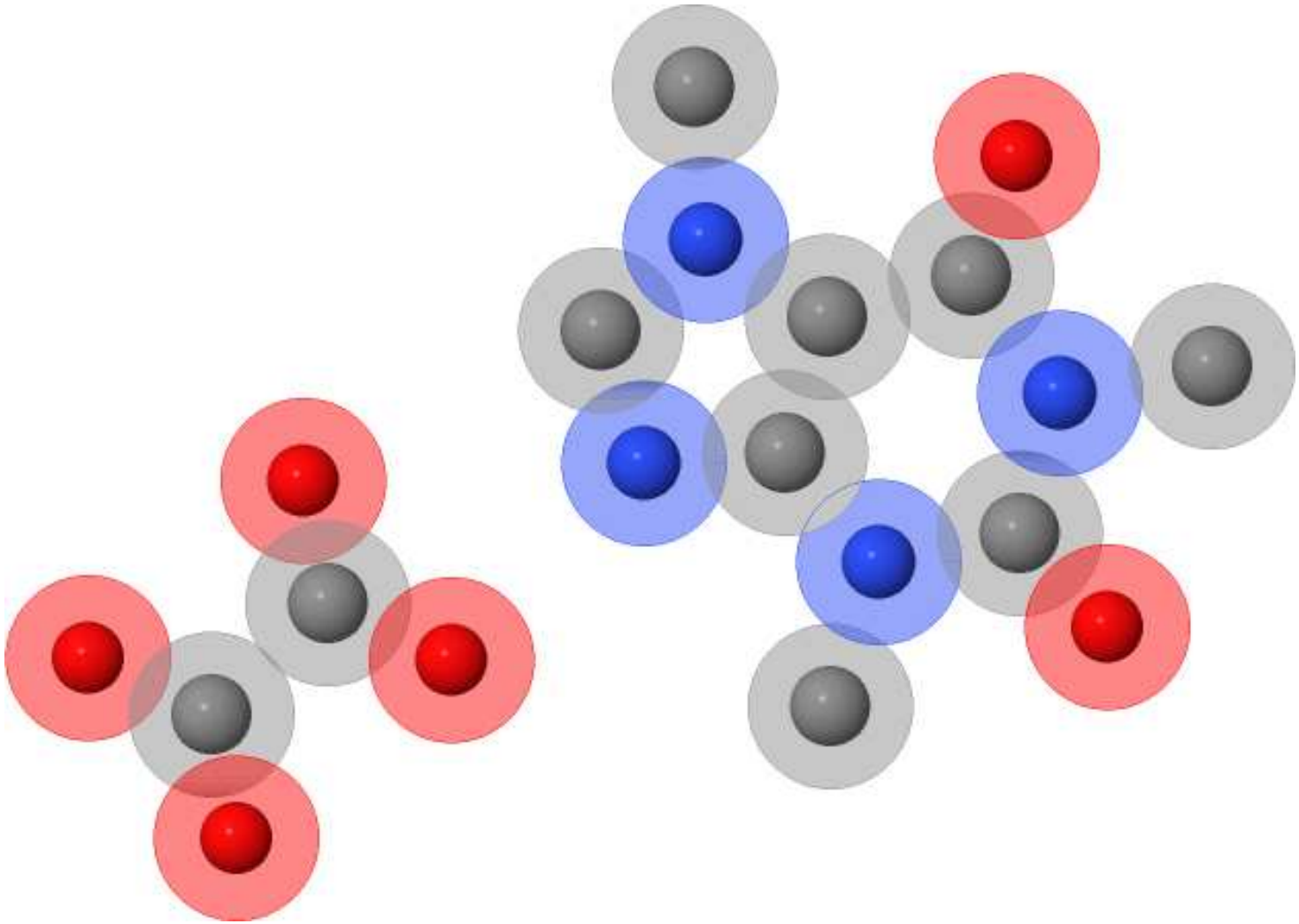
Grenoble, 2024

Vilnius University Institute of Biotechnology

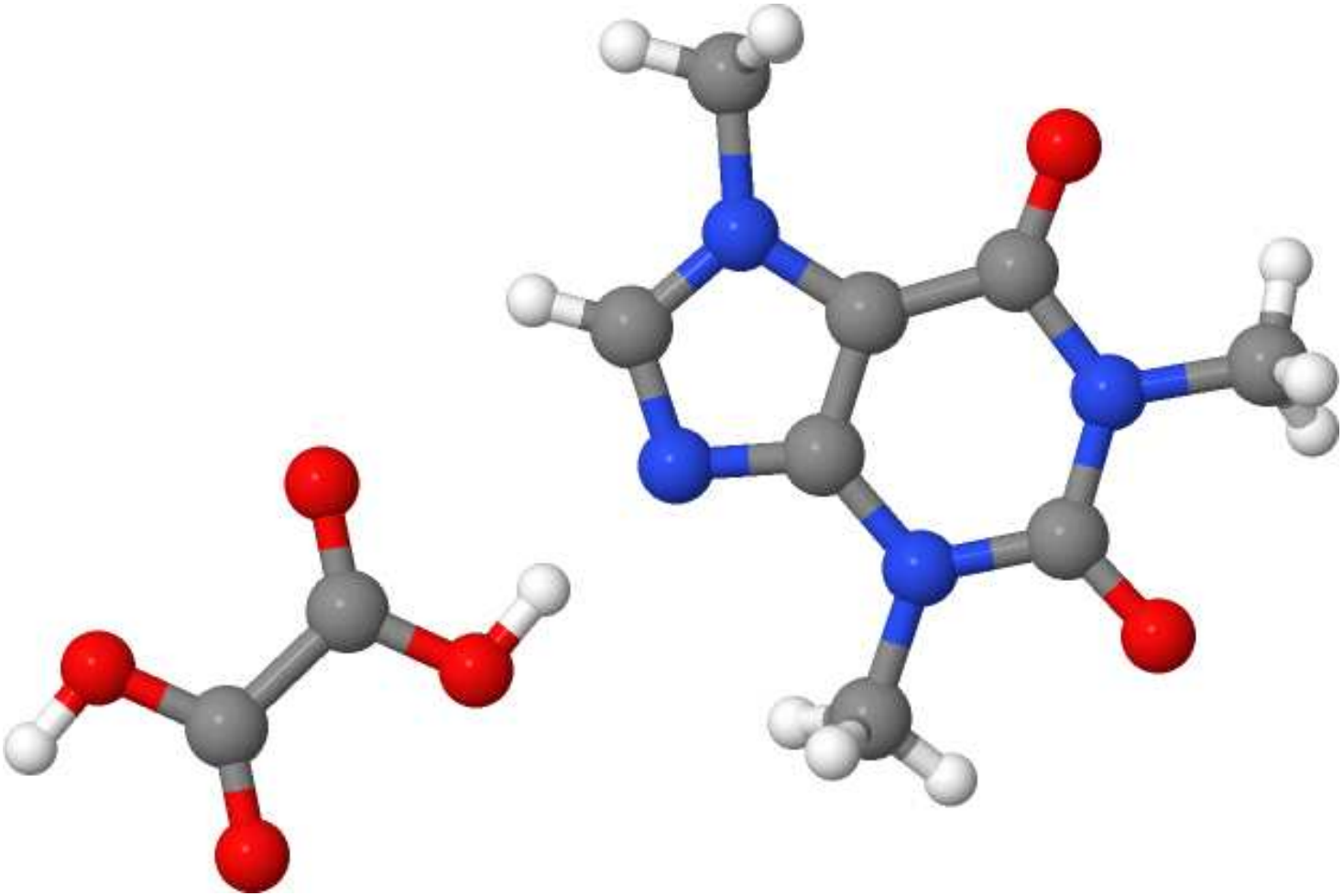
Output of X-ray crystallography



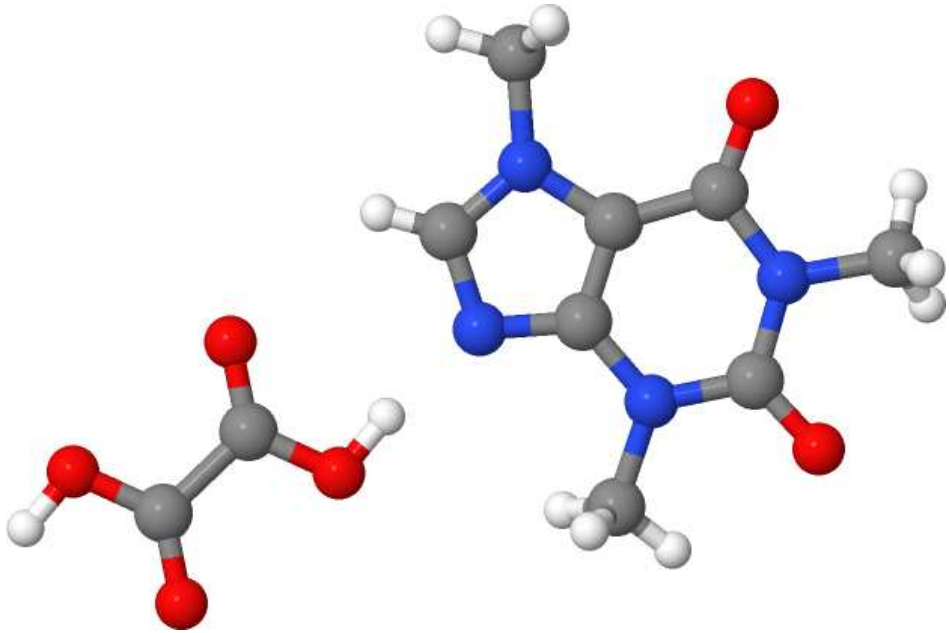
Detection of connectivity

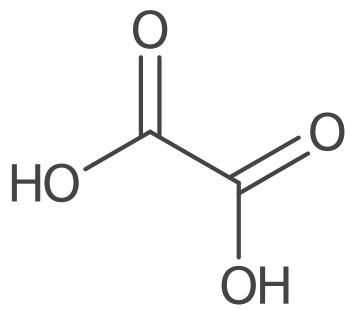


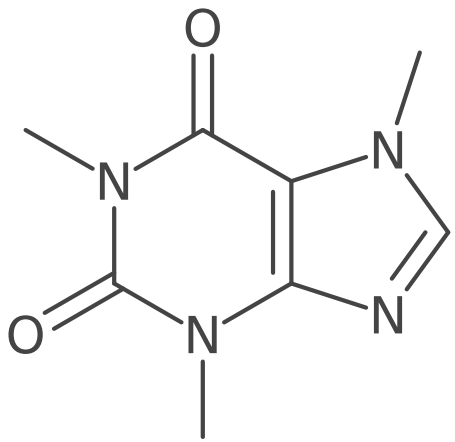
Molecular graph + 3D coordinates



Structural formula







Properties of molecular graphs

Properties

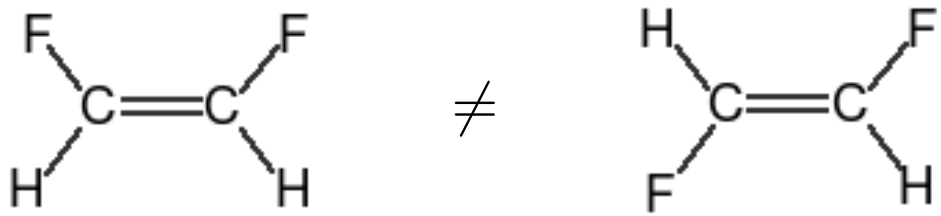
- undirected
- no self-loops
- non-multigraphs (no parallel edges)
- usually planar (Simmons & Maggio, 1981)

Colors

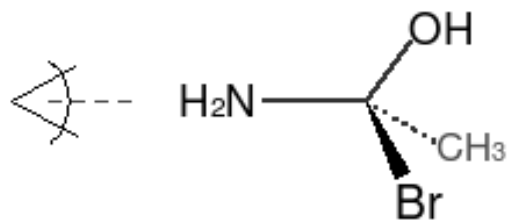
- vertices – chemical element, sometimes: charge, isotope
- edges – bond order (1, 2, 3, 4)

Stereochemistry

- Cis/trans configuration around double bonds



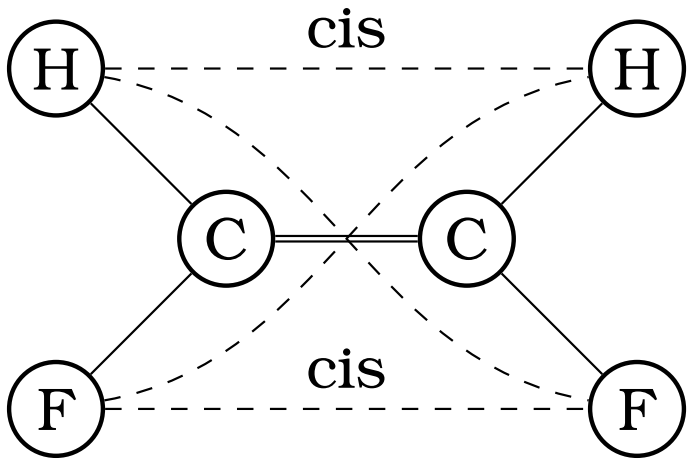
- Tetrahedral chirality



http://opensmiles.org/opensmiles.html
OpenSMILES specification

Cis/trans configuration in molecular graphs

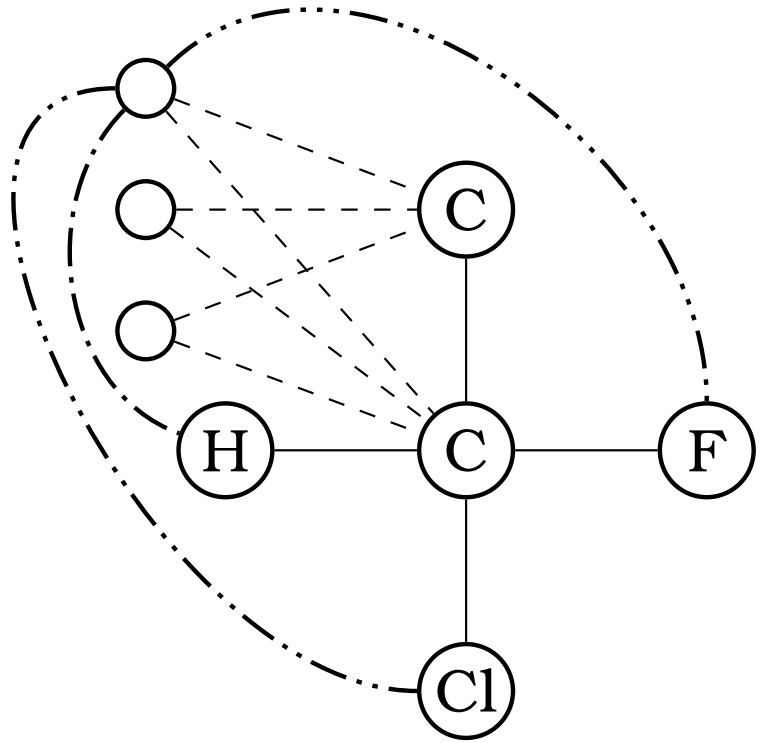
- Spatial relations can be represented by additional colored edges



Merkys et al., 2023

Tetrahedral chirality in molecular graphs

- Some relations need additional vertices



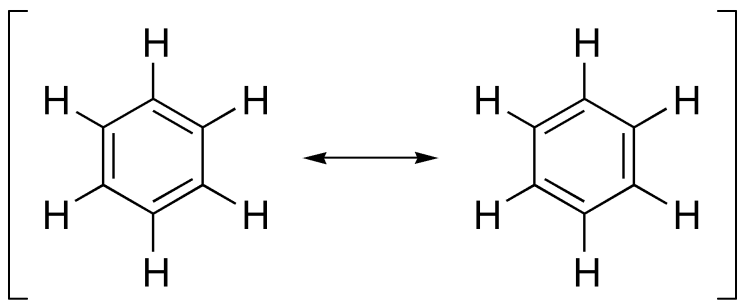
Merkys et al., 2023

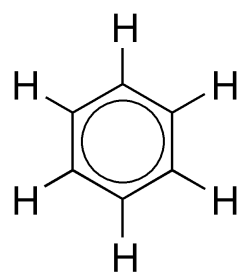
Cycles

Importance

- hold the structure (limit the conformational space)
- base for aromaticity

Aromatic cycles





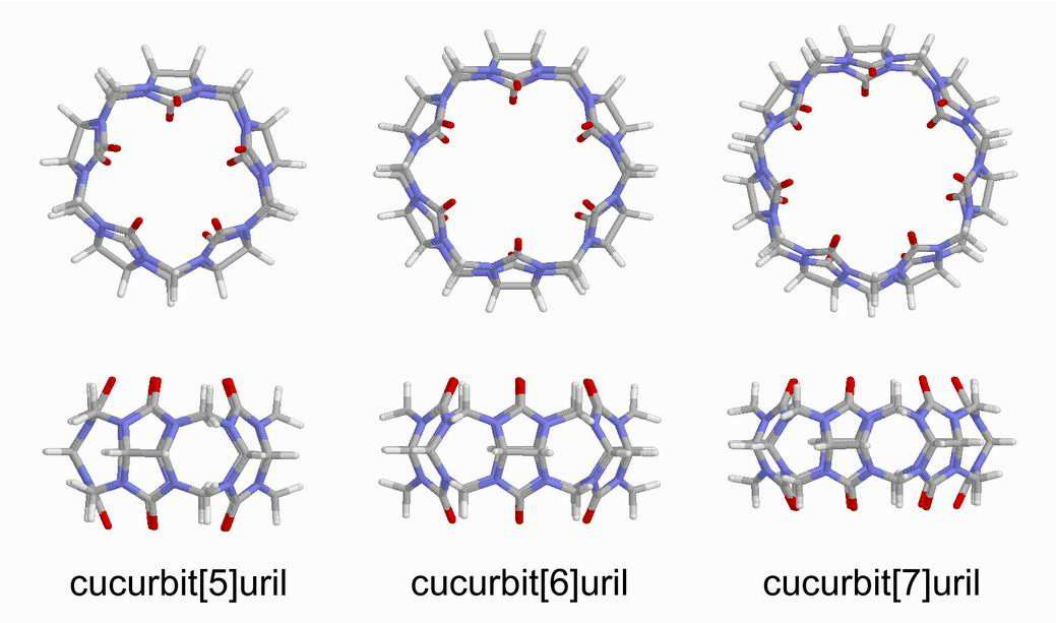
Smallest set of smallest rings (SSSR)

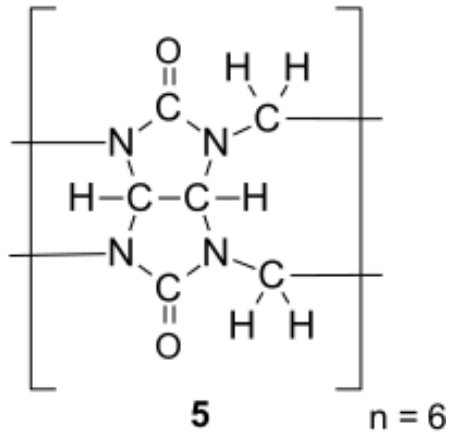
Closed path $\in$ SSSR iff there are no chords

Downs et al. 1989 algorithm:

1. choose L – length of the largest cycle
2. find all paths from all vertices (depth-first search)
3. search is terminated when L is reached
4. search is terminated when other criteria for $SSSR$ are violated

Cycles. Example: cucurbituril





- $n \times 2$ five-membered cycles
- n eight-membered cycles
- two $n \times 4$-membered cycles

https://commons.wikimedia.org/wiki/File:Models_of_cucurbiturils.jpg
M stone at English Wikipedia, CC-BY-SA 3.0
https://commons.wikimedia.org/wiki/File:CucurbiturilSynthesis.svg
V8rik at English Wikipedia, CC-BY-SA 3.0

All cycles in a graph

All closed paths without same vertex appearing more than once

Hanser, Jauffret, Kaufmann 1996 algorithm:

1. choose vertex v and remove it from graph
2. connect v neighbours with edges representing paths (through v) between them
3. every self-loop found is a cycle in the original graph
4. loop to step 1 until graph becomes empty

All cycles in a graph (2)

Examples

- cube – 28 cycles
- fullerene (a.k.a. C_{60}) – 8018 cycles

Problems

- computationally expensive
- no known low-level (C/C++) implementations

Search and comparison

Search – look for a matching subgraph

- boils down to subgraph isomorphism problem
- property vectors (a.k.a. fingerprints) may be used

Comparison – check if two graphs match

- identity – graph isomorphism problem
- property vectors (a.k.a. fingerprints) may be used

Property vectors a.k.a. fingerprints

Tanimoto similarity index

$$T_s(A, B) = \frac{\sum_i A_i \wedge B_i}{\sum_i A_i \vee B_i}$$

Example:

$$T_s(\texttt{1001}, \texttt{1010}) = \frac{1}{3}$$

Open Babel v3.1.1 uses:

- FP3 – 55 properties
- FP4 – 309 properties

Molecular graph representation formats

- IUPAC preferred name (*Pyridine*)
  - uniquely defines a compound
  - difficult to read/write by computer
- SMILES (`c1ccncc1`)
  - easy to read/write by computer
  - could be read/write by human
  - many competing standards
  - not a unique representation
- InChI (`InChI=1S/C5H5N/c1-2-4-6-5-3-1/h1-5H`)
  - uniquely defines a compound
  - difficult to read/write by human
  - sole I/O library is non-free software*

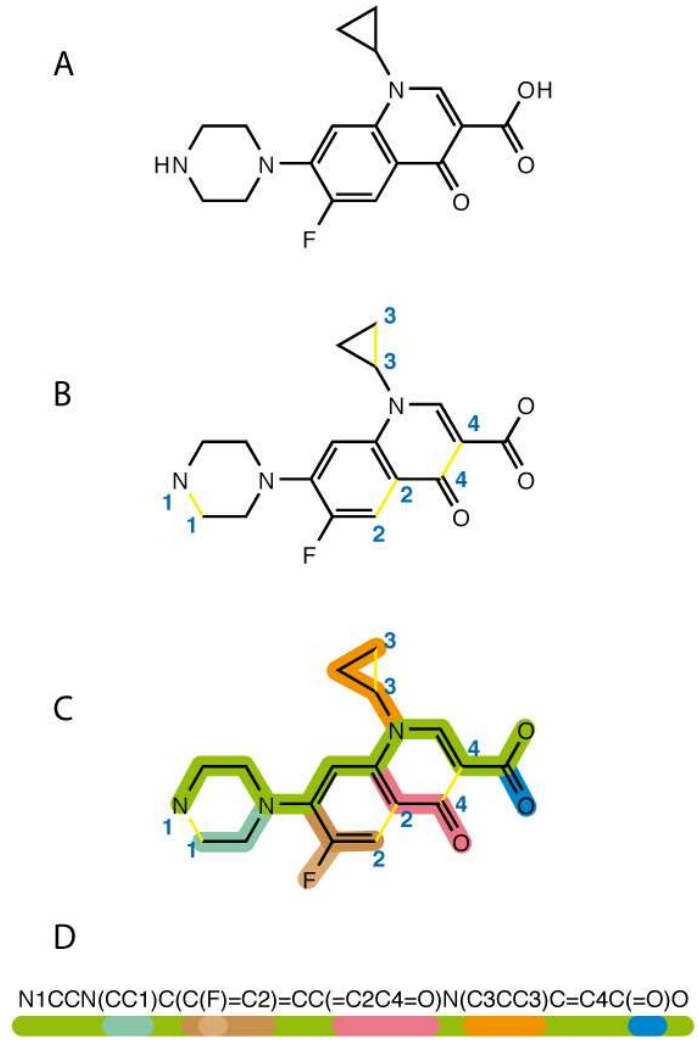
IUPAC names

- Reading
  - **OPSIN**
- Writing
  - ChemAxon
  - Lexichem
  - Nomenclator
  - STOUT
  - ChemOnomatopist

	ChemOnomatopist v0.9.0	STOUT v2.0
Correct PIN	1262	1132
Alternative name	752	1874
Incorrect	1074	690
Refused	608	0
Time	$\approx$ 4 min.	$\approx$ 230 min.

SMILES format

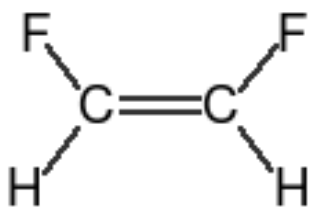
- depicts spanning tree of a graph
- vertices are named after chemical elements
- branches from main chain are written in parentheses
- edges not in spanning tree are marked by numbers
- bonds of order 2 or more are depicted by `=`, `#`, `$`



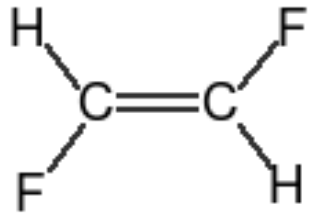
https://commons.wikimedia.org/wiki/File:SMILES.png
Fdardel & DMacks, CC-BY-SA 3.0

SMILES – representation of a molecular graph

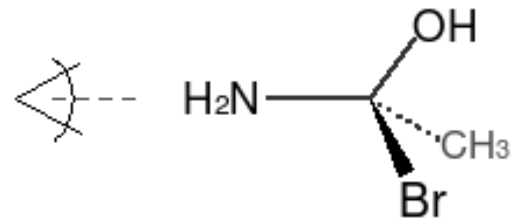
- aromatic atoms are written in lowercase
- cis/trans configuration is marked with up/down edges
- tetrahedral chirality is marked with `@` or `@@` and establishing atom enumeration order



`F\C=C/F`



`F/C=C/F`



`N[C@]C(Br)(O)`

- support for chirality for higher coordination numbers

http://opensmiles.org/opensmiles.html
OpenSMILES specification

Morgan's algorithm

Steps

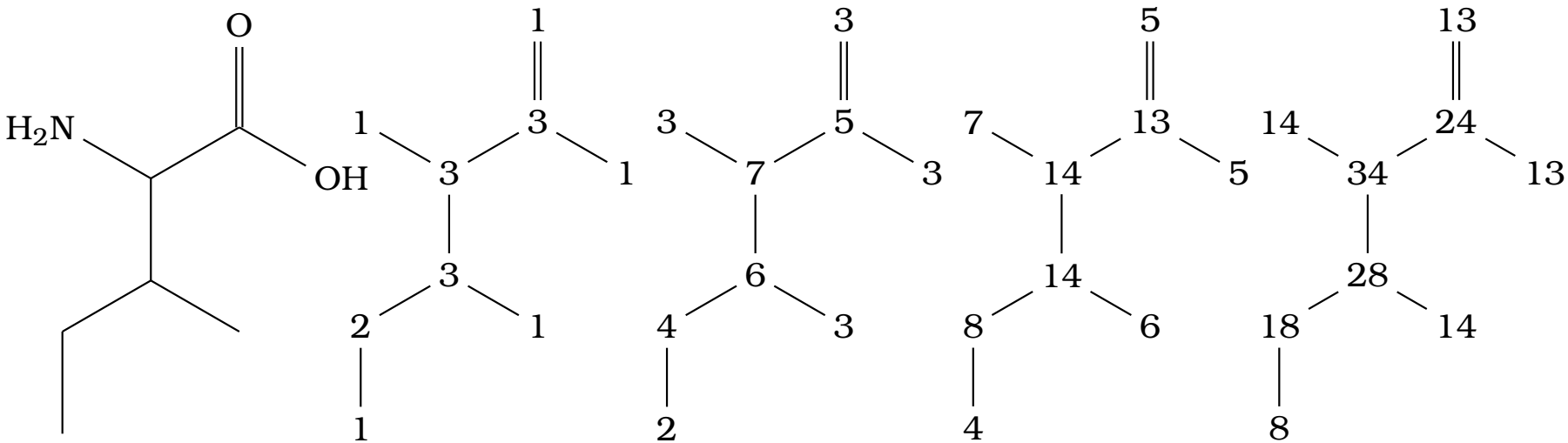
1. $S_0(v) = \deg(v), \forall v \in V$
2. $S_i(v) = \sum S_{i-1}(n)$, sum over all neighbours of v
3. ...

Loop is stopped when the number of distinct $S_i(v)$ stops increasing.

Equivalence classes are defined by $S_i(v)$ values.

Morgan, 1965

Morgan's algorithm. Example



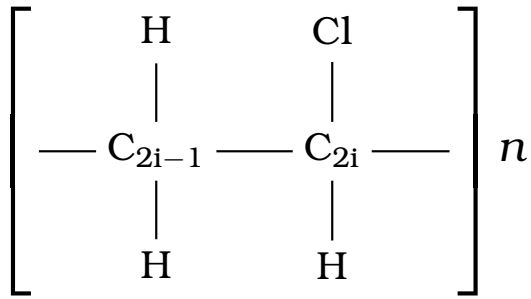
$\{1, 2, 3\}$ $\{2, 3, 4, 5, 6, 7\}$ $\{4, 5, 6, 7, 8, 13, 14\}$ $\{8, 13, 14, 18, 24, 28, 34\}$

Molecular graph isomorphism

- Faulon's algorithm (Faulon, 1998)
  - input – simple graph
  - $O(N^2)$, observed on carbon nanotubes

- *Nauty* (McKay & Piperno, 2014)
  - supports vertices with attributes (color)
  - employed in InChI
  - $O(N)$ (Faulon, Collins & Carr, 2004)

Polymeric molecules



Construction of molecular graph

1. choose a representing “monomer” (how?)
2. redirect boundary-crossing edges back to the “monomer”
   - parallel edges or self-loops may appear
   - multigraphs may be represented by line graphs

Representing polymer molecules with quotient graphs

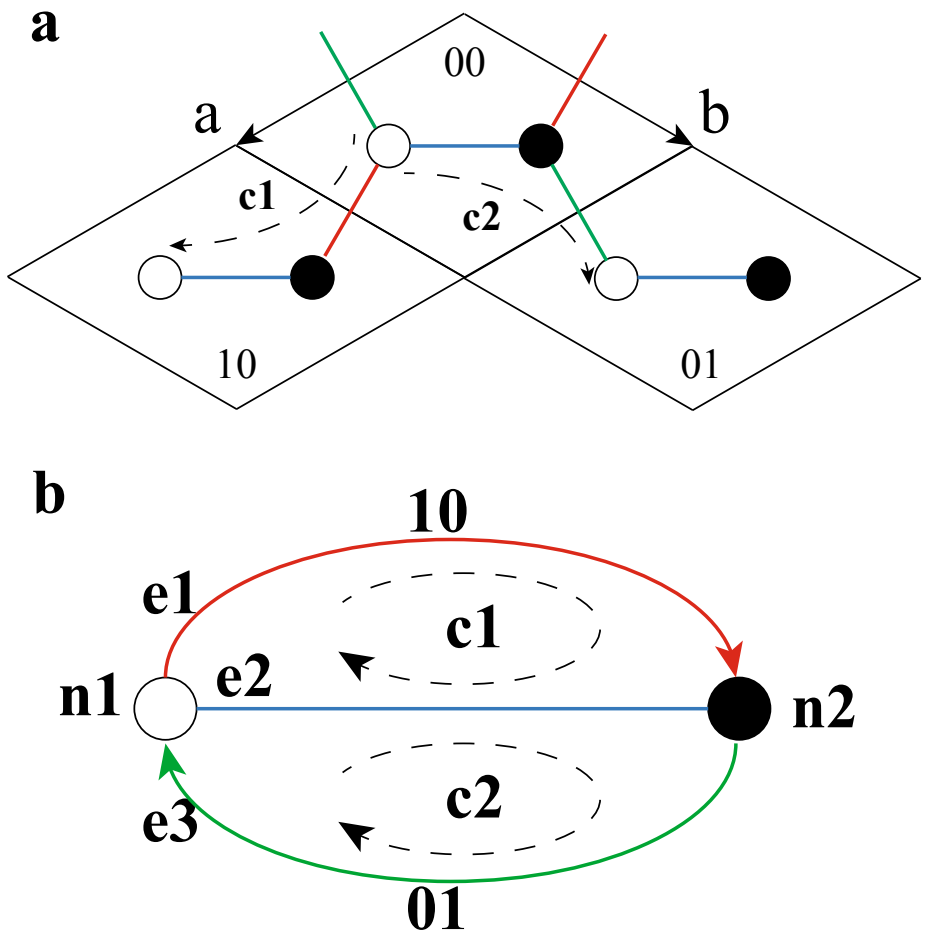
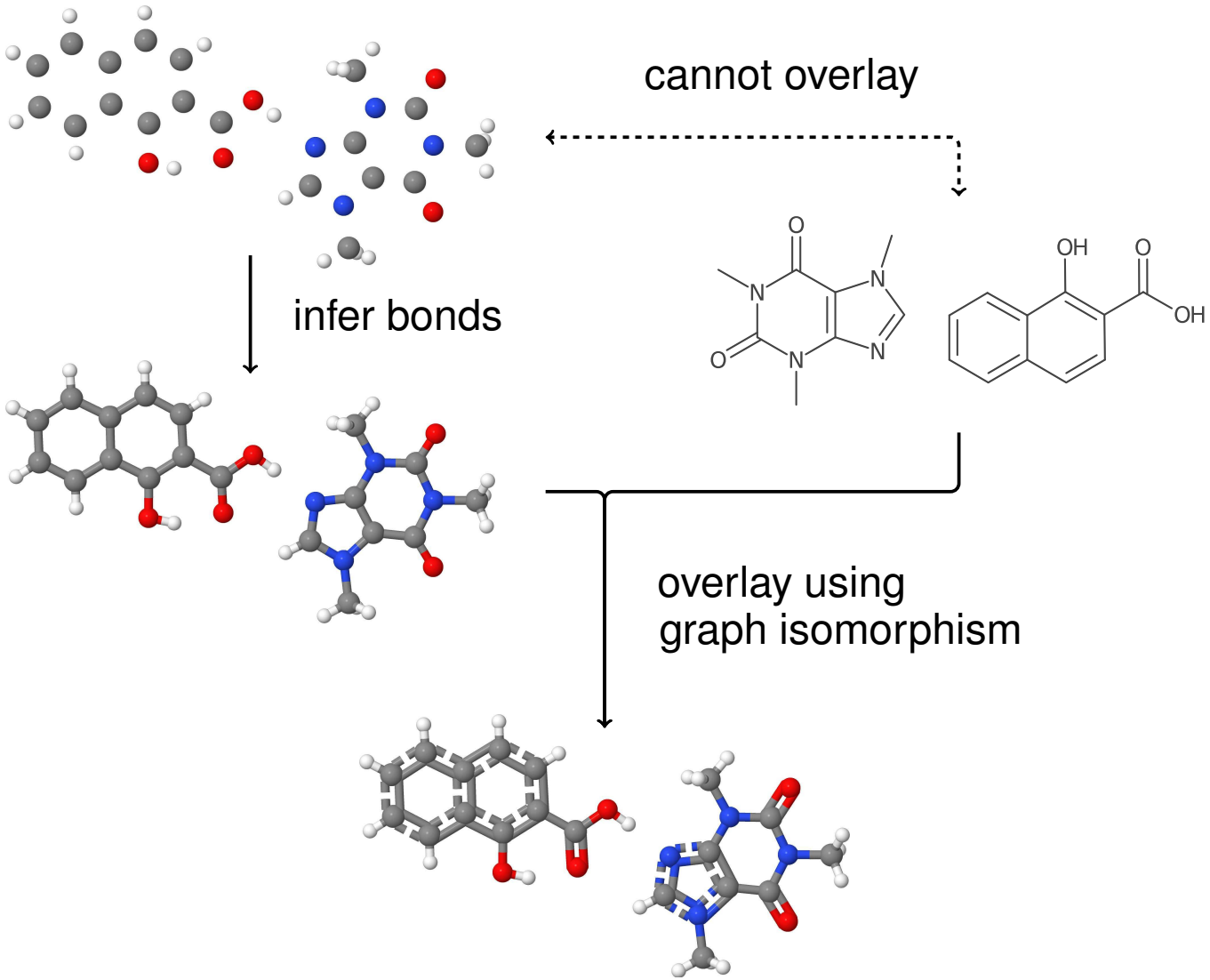


Fig. 1 Graphene net and its quotient graph. a Graphene net in 2D space; b The quotient graph of graphene net. n_1, n_2 are nodes. e_1, e_2, e_3 are edges. c_1, c_2 are the two basic cycles of graphene net.

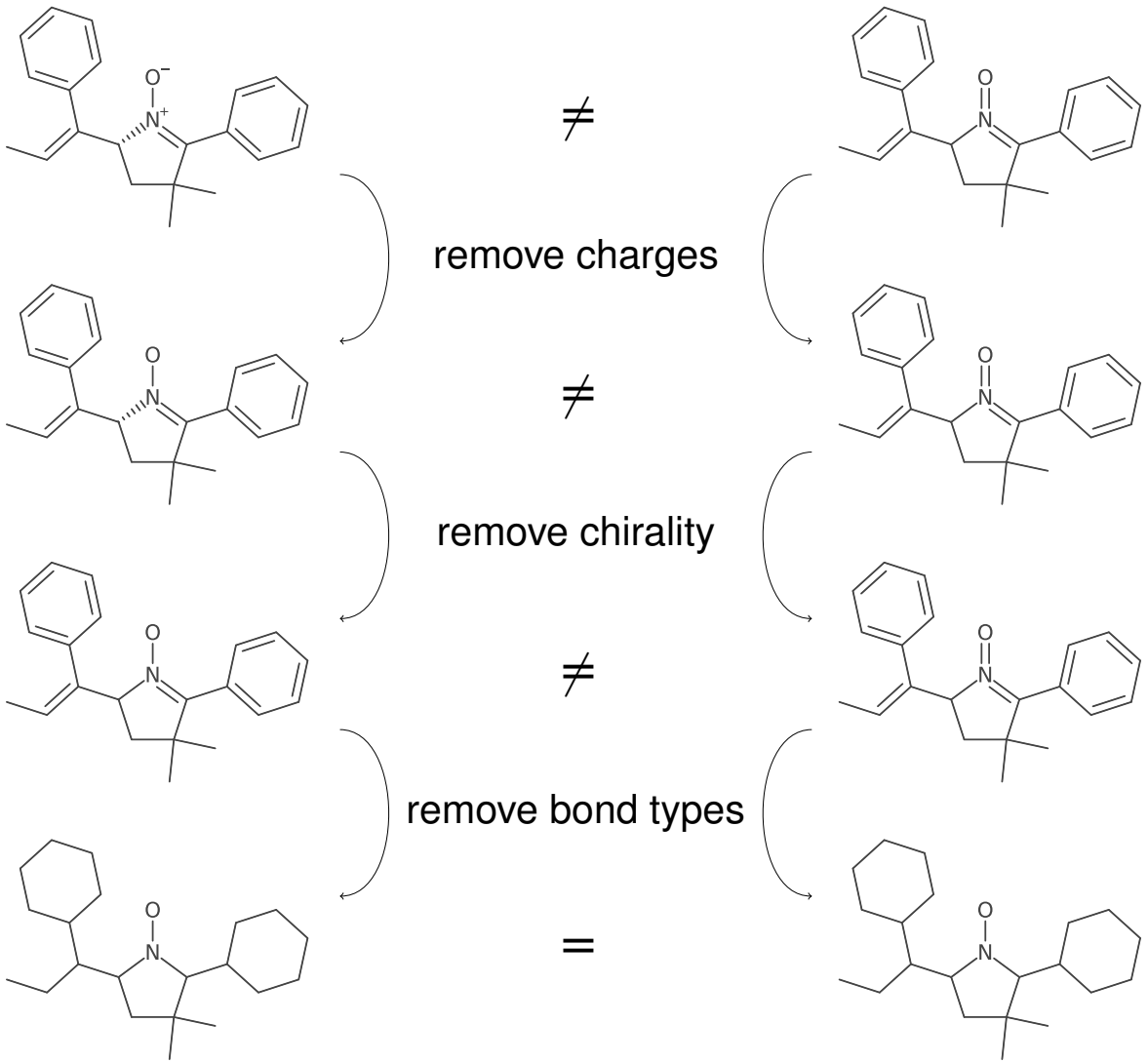
Gao et al., 2020

Chemical annotation of X-ray structures



https://www.crystallography.net/archives/2021/posters/IUCr-XXV/poster-1476.pdf
Merkys et al., poster presentation at IUCr25

Comparison of molecular graphs



https://www.crystallography.net/archives/2021/posters/IUCr-XXV/poster-1476.pdf
Merkys et al., poster presentation at IUCr25

Author-provided vs. curated annotations in COD

#	H atoms	aromaticity	atom types	charge	chirality	cis/trans	order	extra moieties
18424		×					×	
2990		×				×	×	
2803		×		×			×	
2191								
1878		×			×		×	
754					×			
467	×	×		×			×	
464		×		×		×	×	
381	×			×				
270		×					×	×
250	×	×					×	
192		×		×	×		×	
164		×			×	×	×	
145						×		
116				×			×	
85	×		×				×	
78								×
75	×	×				×	×	
56			×				×	
45				×				
37	×						×	
36	×							
35		×		×			×	×
34	×			×			×	
33	×	×		×			×	×
32474	total *(471 not shown here for brevity)*							
2196	unknown							

Problems

- What is a bond?
- How to describe interactions between more than two atoms?
- How to uniquely represent aromatic cycles?