DM840 Algorithms in Cheminformatics

Daniel Merkle

September 5, 2022

Subject overview

- Representation of Molecular Structures
- Combinatorial Structures (Counting, Generating Functions, ...)
- Structure Representation (Canonicalization Algorithms)
- Structure Invariant (e.g. Topological Indices)
- Graph Grammars ("Formal Languages" for Graphs)
- Synthesis Planning (e.g., Shortest Paths in Hypergraphs)
- Enzymatic Design (Discrete Optimization, ILP)
- Concurrency Theory and Causality (e.g. Petri Nets, Category Theory)
- Artificial Chemistries (e.g. "Lattices")
- Quantitative Structure Activity Relationship
  Principal Component Analysis
  Algorithms for Minimum Cycle Basis
- Organization Theory
- Stoichiometric Models
- Metabolic Networks and Metabolic Pathways
- (Flux Balance Analysis)
- ...

Representation of Molecular Structures

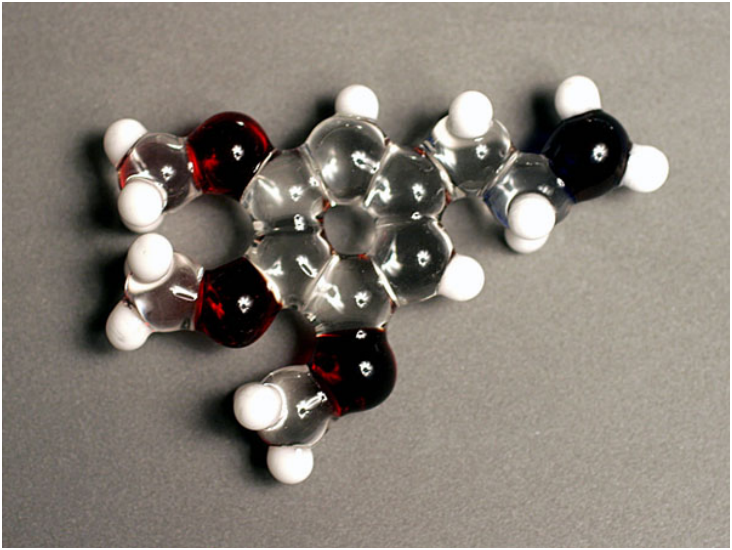


Figure: From the peyote cactus (Lophophora williamsii)

Representation of Molecular Structures

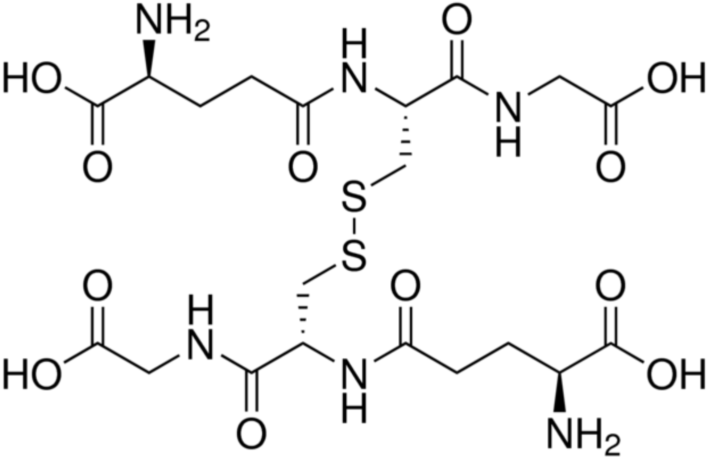


Figure: L-Glutathione oxidized

https://www.sigmaaldrich.com/catalog/product/sigma/g4376?lang=en

Representation of Molecular Structures

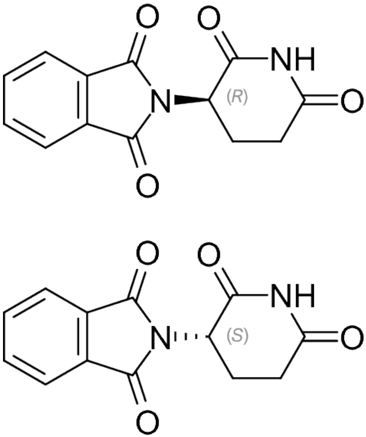
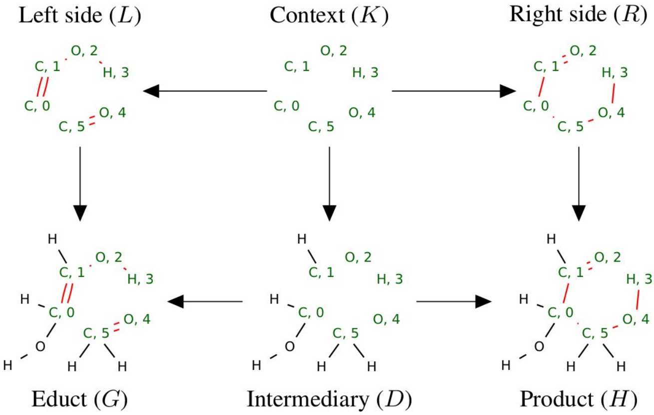
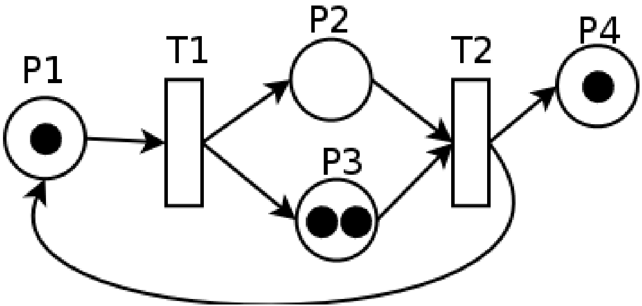


Figure: Thalidomide enantiomers (Contergan)

Graph Grammars



Petri Nets



Metabolic Networks and Metabolic Pathways

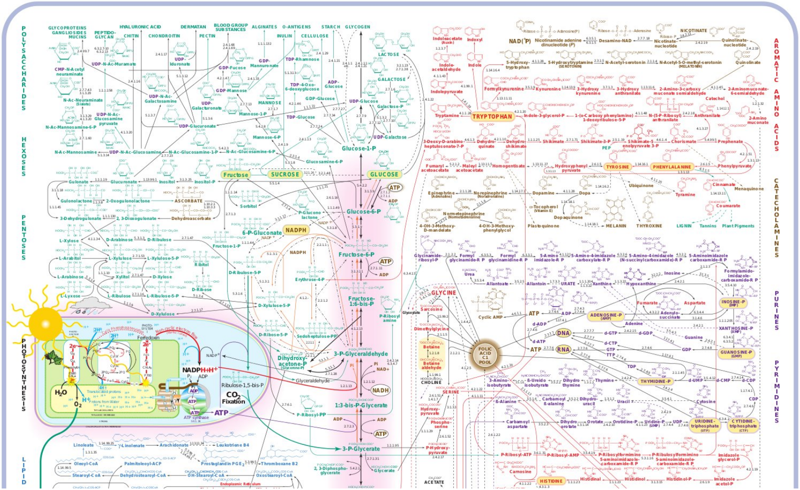
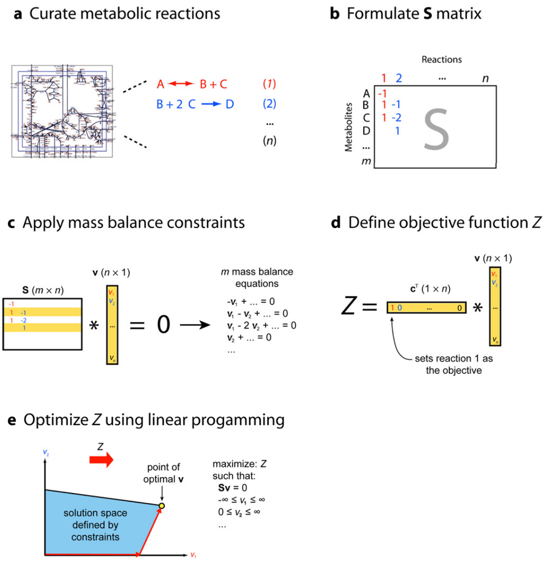


Figure: Typical metabolic network of a cell; (click here for the pdf)

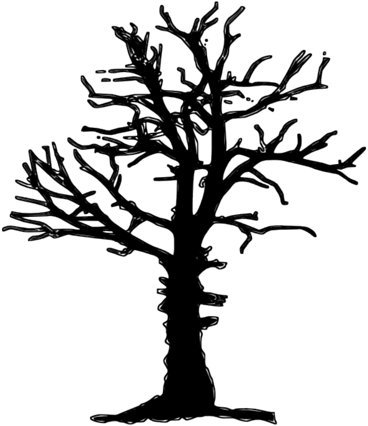
Flux Balance Analysis



Motivation

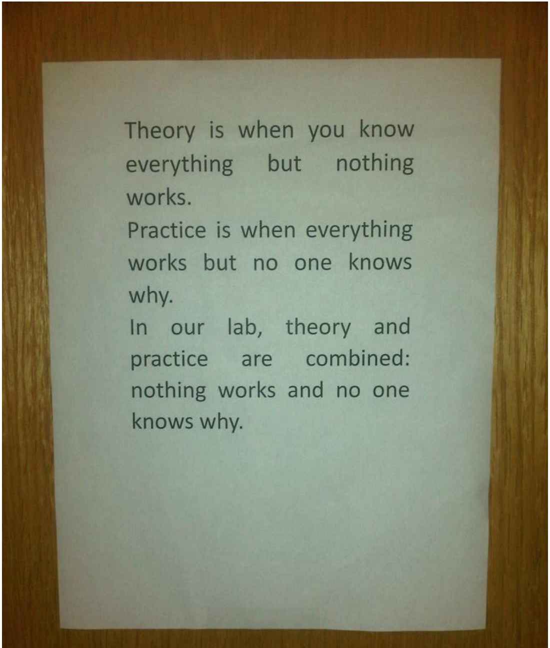
Why is the sky blue?

The research situation



Money

Theory - Practice



The End

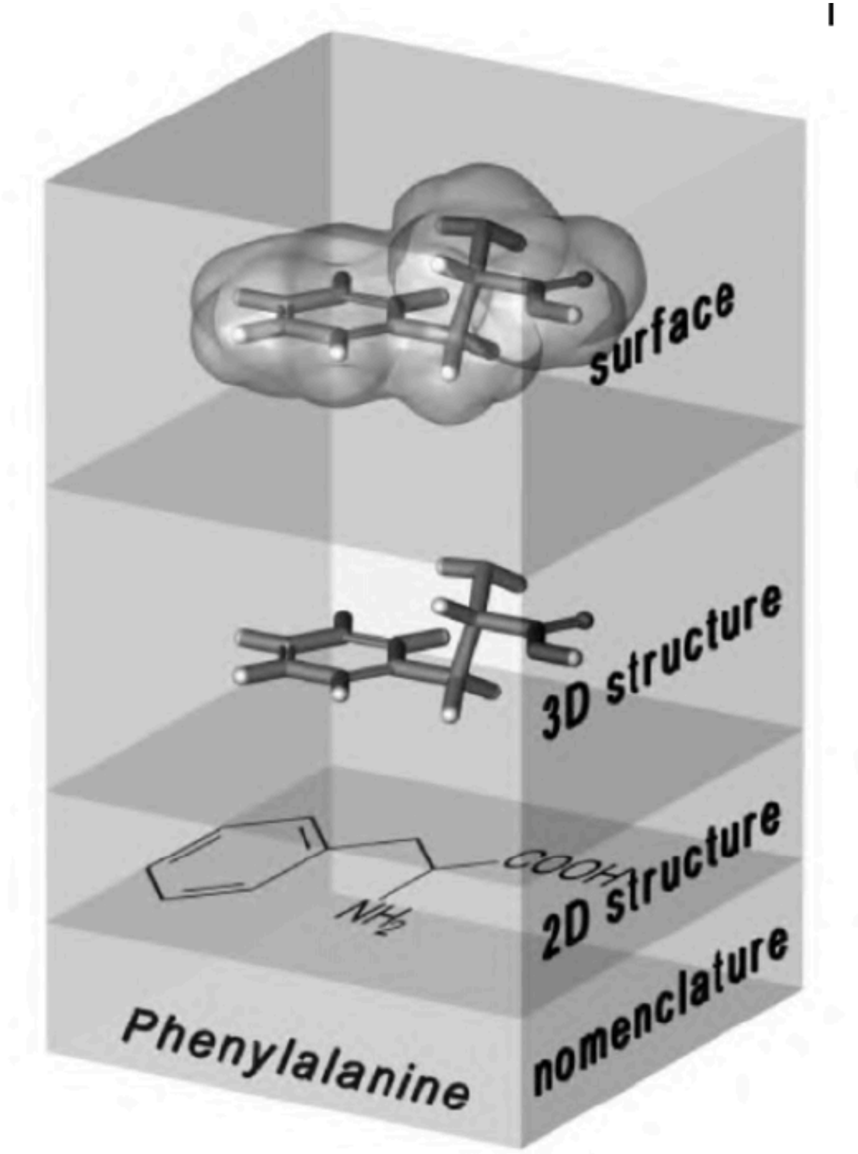


Figure 2-1. Hierarchical scheme for representations of a molecule with different contents of structural information.

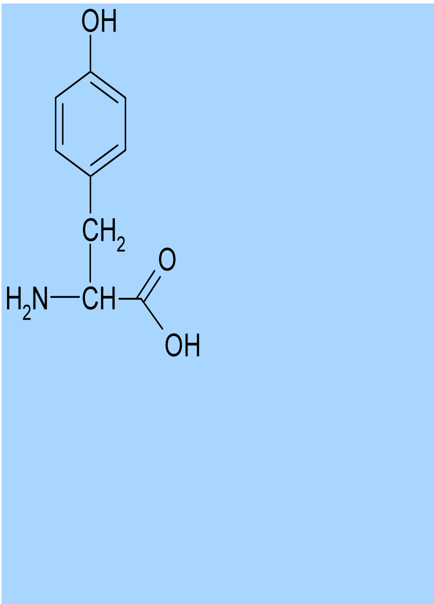
Representing a chemical structure

- How much information do you want to include?
  - atoms present
  - connections between atoms
    - bond types
  - stereochemical configuration
  - charges
  - isotopes
  - 3D-coordinates for atoms

$C_8H_9NO_3$

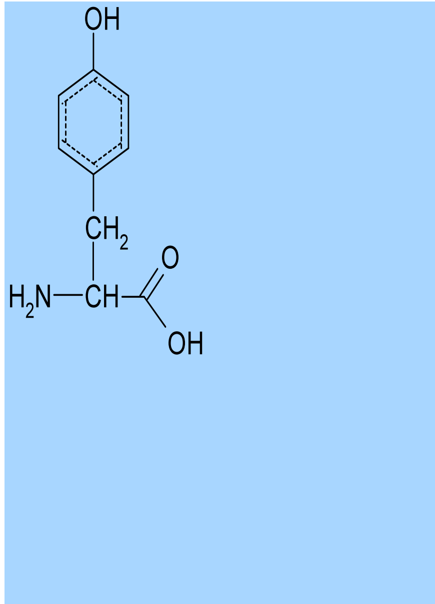
Representing a chemical structure

- How much information do you want to include?
  - atoms present
  - connections between atoms
    - bond types
  - stereochemical configuration
  - charges
  - isotopes
  - 3D-coordinates for atoms



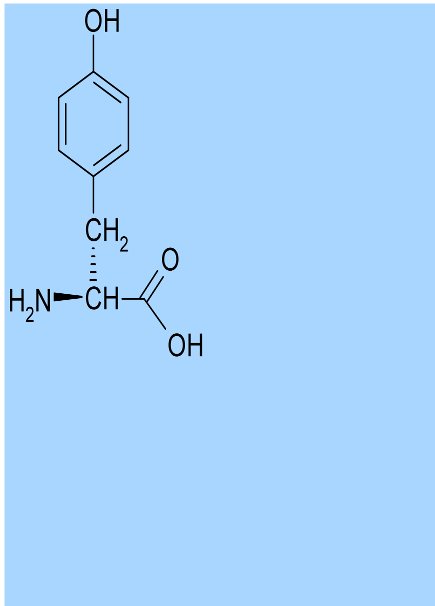
Representing a chemical structure

- How much information do you want to include?
  - atoms present
  - connections between atoms
    - bond types
      (aromatic ring identification)
  - stereochemical configuration
  - charges
  - isotopes
  - 3D-coordinates for atoms



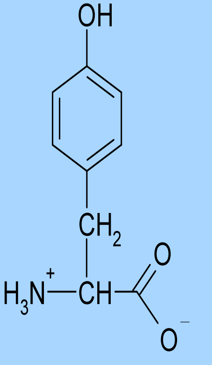
Representing a chemical structure

- How much information do you want to include?
  - atoms present
  - connections between atoms
    - bond types
  - stereochemical configuration
  - charges
  - isotopes
  - 3D-coordinates for atoms



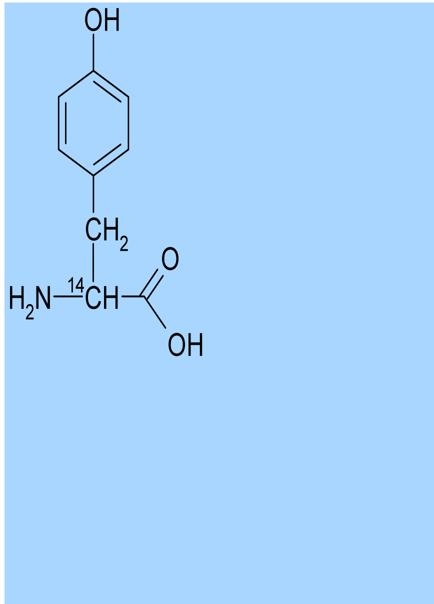
Representing a chemical structure

- How much information do you want to include?
  - atoms present
  - connections between atoms
    - bond types
  - stereochemical configuration
  - charges
  - isotopes
  - 3D-coordinates for atoms



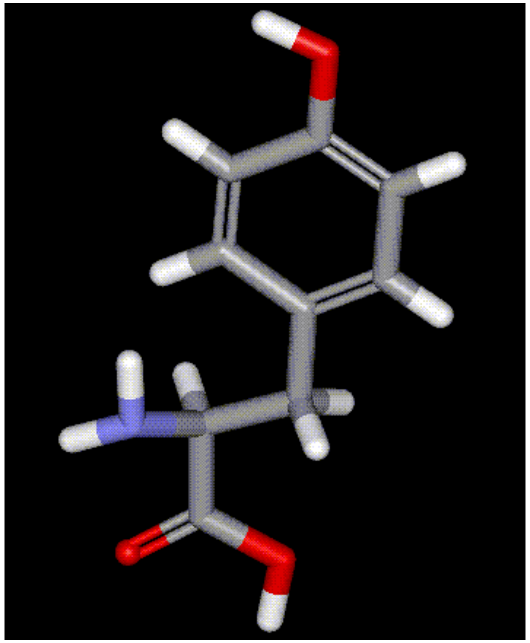
Representing a chemical structure

- How much information do you want to include?
  - atoms present
  - connections between atoms
    - bond types
  - stereochemical configuration
  - charges
  - isotopes
  - 3D-coordinates for atoms

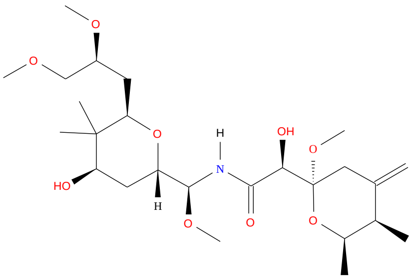


Representing a chemical structure

- How much information do you want to include?
  - atoms present
  - connections between atoms
    - bond types
  - stereochemical configuration
  - charges
  - isotopes
  - 3D-coordinates for atoms



Store Substance as Graphics or by Name



- Pederin
- **2H-Pyran-2-glycolamide**, N-((6-(2,3-dimethoxypropyl) tetrahydro-4-hydroxy-5,5-dimethyl-2H-pyran-2-yl)methoxymethyl) tetrahydro-2-methoxy-5,6-dimethyl-4- methylene-
- **D-manno-Nonitol**, 2,6-anhydro-3,5,7-trideoxy-1-C-(((2S)-hydroxy((2R,5R,6R)-tetrahydro-2-methoxy-5,6-dimethyl-4-methylene-2H-pyran-2-yl)acetyl)amino)-5,5-dimethyl-1,8,9-tri-O-methyl-,(1S)-

Elements of Formal Grammars

1. **Terminal Symbols** T (represented by lowercase letters).
2. **Nonterminals Symbols** N (represented by uppercase letters).
3. **Production Rules** with a left- and a right-hand side consisting of strings of these symbols.
4. **Start Symbol** (also called Axiom)

The example grammar defines the language of all strings of the form $\{ax_1x_2\ldots x_kb \mid k \geq 0 \wedge x_i \in \{a, b\}\}$. (A is the Axiom).

members: `ab, abab, aaabbb`
non-members: `a, b, ababa`

```
T =  {a,b}
N =  {A, B, C}

A -> aA | aB
B -> aB | bB | C
C -> bC | b
```

Chomsky Hierarchy

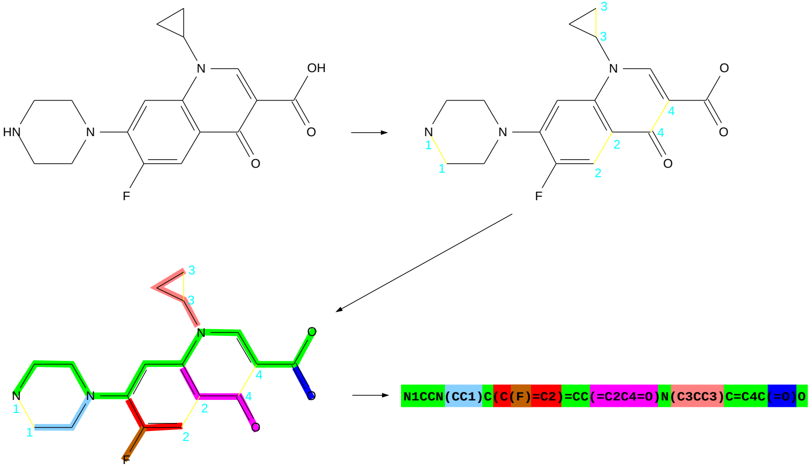
Type No	Type Name	Rule Pattern	
0	unrestricted	$x \longrightarrow y$,	$x \in (N \cup T)_+$ $y \in (N \cup T)*$
1	context sensitive	$x \longrightarrow y$,	$x \in (N \cup T)_+$ $y \in (N \cup T)_+$ $\lvert x \rvert \leq \lvert y \rvert$
2	context free	$x \longrightarrow y$,	$x \in N$ $y \in (N \cup T)*$
3	regular	$w \longrightarrow x \mid yz$	$w, x, z \in N$ $y \in T$

BNF grammar of Daylight's SMILES

```
smiles            ::= chain terminator
chain             ::= branched_atom | chain branched_atom
                    | chain bond branched_atom | chain '.' branched_atom
branched_atom     ::= atom ringbond* branch*
atom              ::= bracket_atom | aliphatic_organic | aromatic_organic
                    | '*'
ringbond          ::= bond? DIGIT | bond? '\%' DIGIT DIGIT
branch            ::= '(' chain ')' | '(' bond chain ')' | '(' '.' chain ')'
bracket_atom      ::= '[' isotope? symbol chiral? hcount? charge? class? ']'
isotope           ::= NUMBER
symbol            ::= element_symbols | aromatic_symbols | '*'
chiral            ::= '@' | '@@'
hcount            ::= 'H' DIGIT?
charge            ::= '-' DIGIT? | '+' DIGIT?
class             ::= ':' NUMBER
aliphatic_organic ::= 'B' | 'C' | 'N' | 'O' | 'S' | 'P'| 'F'| 'Cl'| 'Br'| 'I'
aromatic_organic  ::= 'b' | 'c' | 'n' | 'o' | 's' | 'p'
element_symbols   ::= 'H' | 'He' | 'Li' | 'Be' | 'B' | 'C' | 'N' | 'O' | etc
aromatic_symbols  ::= 'c' | 'n' | 'o' | 'p' | 's' | 'se' | 'as'
bond              ::= '-' | '=' | '#' | '\$' | ':' | '/' | '\'
terminator        ::= SPACE | TAB | '\n' | '\0'
```

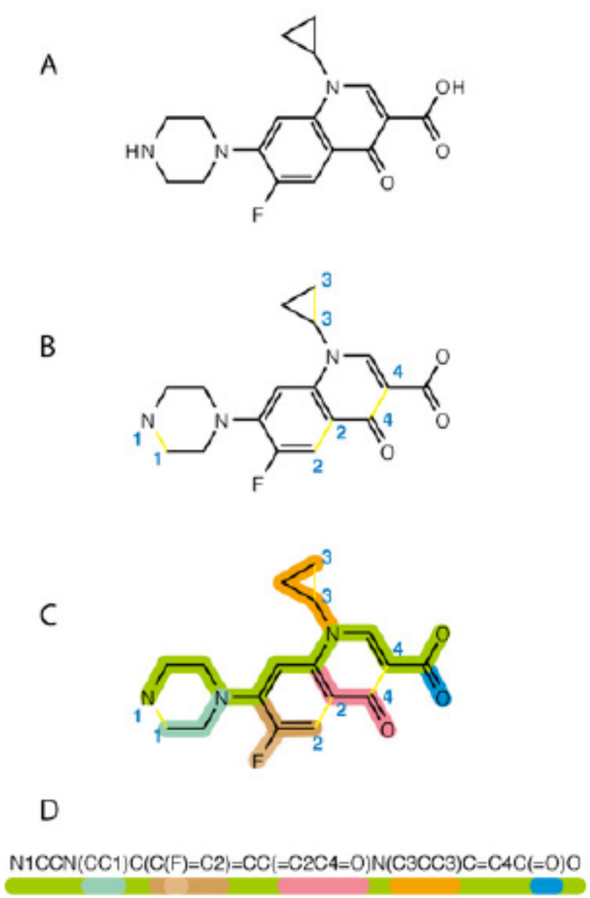
http://www.opensmiles.org/

Generation of SMILES



modified from `http://en.wikipedia.org/wiki/File:SMILES.png`

SMILES
(Simplified Molecular Input Line Entry System)



Six basic rules:

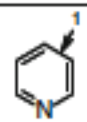
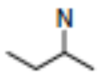
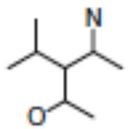
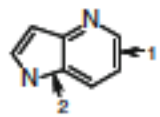
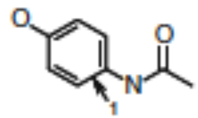
1) Atoms by atomic symbol
2) Hydrogen atoms added to saturate free valences
3) Neighboring atoms stand next to one another
4) Double and triple bonds by = and #
5) Branching shown by parentheses
6) Rings shown by digit at ring closures

Canonical SMILES: unique for each structure

Isomeric SMILES: describe isotopism, configuration around double bonds and tetrahedral centers, chirality

SMILES

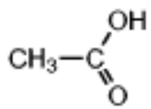
Illustrative SMILES: molecular structures and the corresponding SMILES strings are paired vertically. The numbered arrows on the three cyclic molecular structures are not part of the molecules. They are used to indicate the break points for deriving the corresponding SMILES strings (*see* text)

			
CCC	CC=C	CC#C	c1ccncc1
			
CCC(C)N	CC(C)C(C(C)N)C(C)O	c1cc2c(cc[nH]2)nc1	CC(=O)Nc1ccc(cc1)O

J.Z. Zhou (ed.), Chemical Library Design, Methods in Molecular Biology 685, 2011

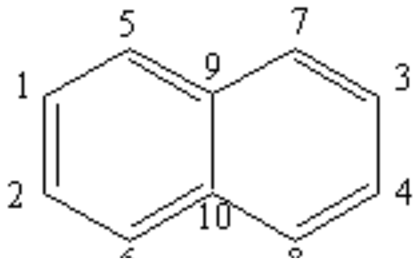
SLN
(SYBYL Line Notation)

Table 2-3. Basic SLN syntax without description of attributes and macro atoms.

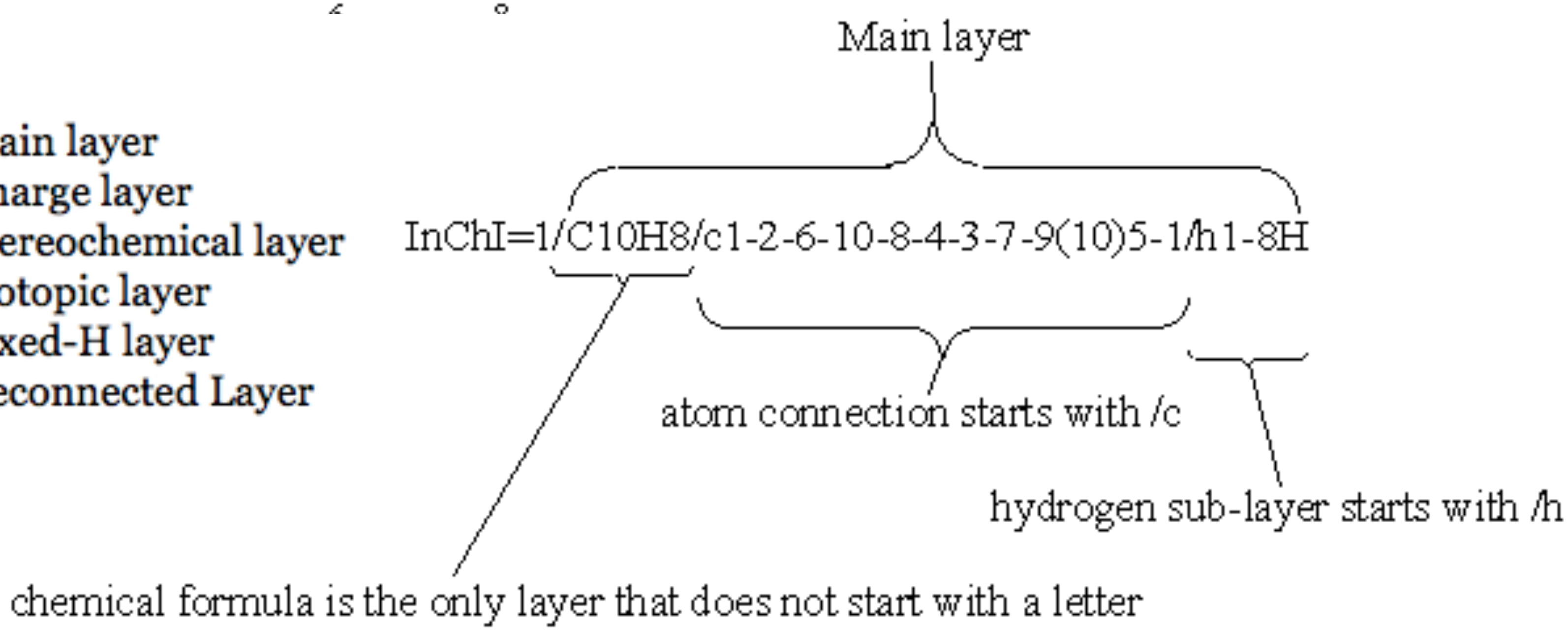
SLN	*Chemical structure*	*Compound name*
Atoms: Atoms are represented by their atomic symbols. The first letter is upper-case, and in two-letter symbols the second letter is lower-case. Hydrogen atoms must be specified.		
CH4	CH_4	methane
NH2	$-NH_2$	amine
Bonds: Single bonds are omitted; double, triple, and aromatic bonds are indicated by the symbols " = ", " # " and " : ", respectively. In contrast to SMILES, aromaticity is not an atomic property, but a property of bonds. A period indicates the start of a new part of the structure.		
HC(=O)OH	HCOOH	formic acid
Na.OH	NaOH	sodium hydroxide
Branches: Branches are indicated by parentheses.		
CH3C(=O)OH		acetic acid
Cyclic structures: Ring closures are described by a bond to a previously defined atom which is specified by a unique ID number. The ID is a positive integer placed in square brackets behind the atom. An " @ " indicates a ring closure.		
C[15]H2CH2CH2CH2CH2CH2@15		cyclohexane
O[6]:CH:CH:CH:CH:@6		furan

InChI
(IUPAC International Chemical Identifier)

InChI=1/C10H8/c1-2-6-10-8-4-3-7-9(10)5-1/h1-8H



1. Main layer
2. Charge layer
3. Stereochemical layer
4. Isotopic layer
5. Fixed-H layer
6. Reconnected Layer



Adjacency and Distance matrices

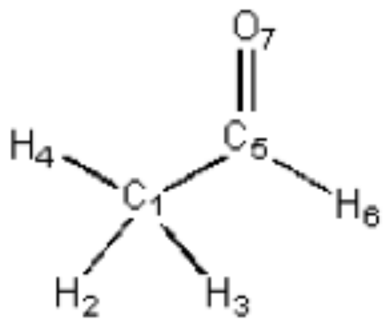
Acetaldehyde: CH3CH=O

Adjacency Matrix

C1	0	1	1	1	1	0	0
H2	1	0	0	0	0	0	0
H3	1	0	0	0	0	0	0
H4	1	0	0	0	0	0	0
C5	1	0	0	0	0	1	1
H6	0	0	0	0	1	0	0
O7	0	0	0	0	1	0	0

Distance Matrix

0	1	1	1	1	2	2
1	0	2	2	2	3	3
1	2	0	2	2	3	3
1	2	2	0	2	3	3
1	2	2	2	0	1	1
2	3	3	3	1	0	2
2	3	3	3	1	2	0

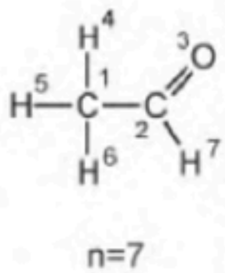


1: atoms i j are bonded

0: atoms i j are not bonded

Length of shortest path between atoms i j

Bond matrix



a)

	1	2	3	4	5	6	7
1		1		1	1	1	
2	1		2				1
3		2					
4	1						
5	1						
6	1						
7		1					

=

b)

	1	2	3	4	5	6	7
1		1		1	1	1	
2			2				1
3							
4							
5							
6							
7							

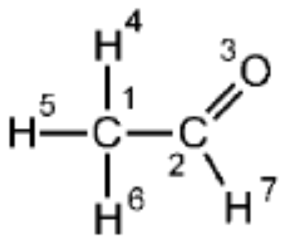
=

c)

	1	2	3
1		1	
2			2
3			

Figure 2-17. a) The redundant bond matrix of ethanal with the zero values omitted. b) It can be compressed by reduction to the top right triangle. c) Omitting the hydrogen atoms provides the simplest non-redundant matrix representation.

Connection Table



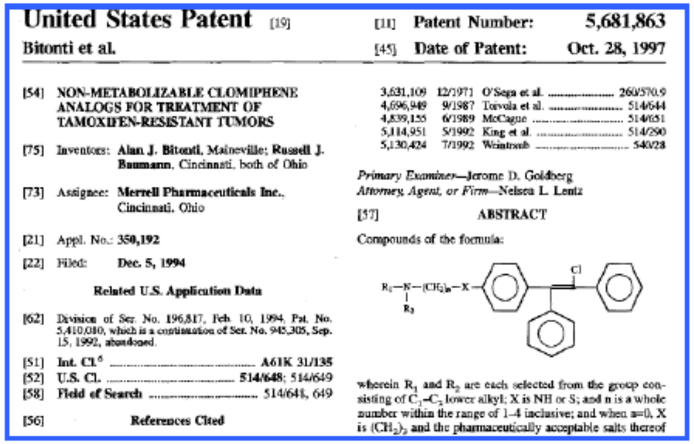
Atom list	
1	C
2	C
3	O
4	H
5	H
6	H
7	H

Bond list		
1st atom	2nd atom	bond order
1	2	1
2	3	2
2	7	1
1	4	1
1	5	1
1	6	1

Figure 2-20. A connection table: the structure diagram of ethanal, with the atoms arbitrarily labeled, is defined by a list of atoms and a list of bonds.

Special notations of chemical structures

- Markush structures



United States Patent [19]
Bitonti et al.

[11] Patent Number: 5,681,863
[45] Date of Patent: Oct. 28, 1997

[54] NON-METABOLIZABLE CLOMIPHENE ANALOGS FOR TREATMENT OF TAMOXIFEN-RESISTANT TUMORS

[75] Inventors: **Alan J. Bitonti**, Maineville; **Russell J. Baumann**, Cincinnati, both of Ohio

[73] Assignee: **Merrell Pharmaceuticals Inc.**, Cincinnati, Ohio

[21] Appl. No.: **350,192**

[22] Filed: **Dec. 5, 1994**

Related U.S. Application Data

[62] Division of Ser. No. 196,817, Feb. 10, 1994, Pat. No. 5,410,080, which is a continuation of Ser. No. 945,305, Sep. 15, 1992, abandoned.

[51] **Int. Cl.**[6] **A61K 31/135**
[52] **U.S. Cl.** **514/648**; 514/649
[58] **Field of Search** 514/648, 649

[56] **References Cited**

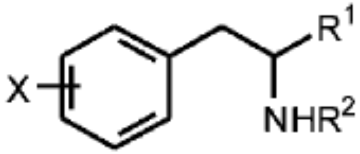
3,631,109	12/1971	O'Sega et al.	260/570.9
4,696,949	9/1987	Toivola et al.	514/644
4,839,155	6/1989	McCague	514/651
5,114,951	5/1992	King et al.	514/290
5,130,424	7/1992	Weintraub	540/28

Primary Examiner—Jerome D. Goldberg
Attorney, Agent, or Firm—Nelsen L. Lentz

[57] ABSTRACT

Compounds of the formula:

wherein R_1 and R_2 are each selected from the group consisting of C_1–C_2 lower alkyl; X is NH or S; and n is a whole number within the range of 1–4 inclusive; and when n=0, X is $(CH_2)_2$ and the pharmaceutically acceptable salts thereof



R^1 = H or small alkyl, halogen, OH, COOH

R^2 = H, CH_3

X = H, $(CH_2)_nCH_3$

Figure 2-62. The substituted phenyl derivative is an example of a typical Markush structure. Herein, a number of compounds are described in one structure diagram by fill-ins. Phenylalanine is one of these structures when R^1 is COOH, R^2 is H, and X is H.

Special notations of chemical structures

- Fingerprints

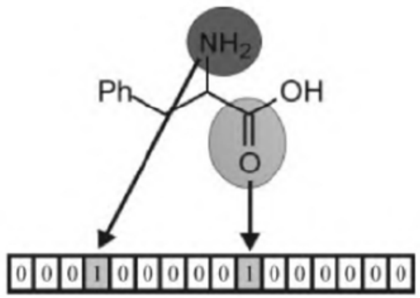


Figure 2-64. How an excerpt from a binary code could appear, if only $-NH_2$ and C=O are available in the fragment library.

MACCS fingerprints: 166 structural keys that answer questions of the type:

- Is there a ring of size 4?
- Is at least one F, Br, Cl, or I present?

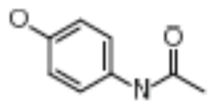
where the answer is either TRUE (1) or FALSE (0)

SMILES .smi file

```
N12CCC36C1CC(C(C2)=CCOC4CC5=O)C4C3N5c7ccccc76 Strychnine
c1ccccc1C(=O)OC2CC(N3C)CCC3C2C(=O)OC cocaine
COc1cc2c(ccnc2cc1)C(O)C4CC(CC3)C(C=C)CN34 quinine
OC(=O)C1CN(C)C2CC3=CCNc(ccc4)c3c4C2=C1 lyseric acid
CCN(CC)C(=O)C1CN(C)C2CC3=CNc(ccc4)c3c4C2=C1 LSD
C123C5C(O)C=CC2C(N(C)CC1)Cc(ccc4O)c3c4O5 morphine
C123C5C(OC(=O)C)C=CC2C(N(C)CC1)Cc(ccc4OC(=O)C)c3c4O5 heroin
c1ncccc1C1CCCN1C nicotine
CN1C(=O)N(C)C(=O)C(N(C)C=N2)=C12 caffeine
C1C(C)=C(C=CC(C)=CC=CC(C)=CCO)C(C)(C)C1 vitamin a
```

MOLfile

(a)



(b)

```
 SMMXDraw12120917342D

11 11  0  0  0  0  0  0  0  0999 V2000
   12.3082   -7.1882    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   13.0242   -7.6016    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   13.7402   -7.1882    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   14.4562   -7.6016    0.0000 N   0  0  3  0  0  0  0  0  0  0  0  0
   15.1722   -7.1882    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   15.1722   -6.3615    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
   15.8882   -7.6016    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   13.7402   -6.3615    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   13.0242   -5.9481    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   12.3082   -6.3615    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   11.5922   -5.9481    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
  2  1  2  0  0  0  0
  3  2  1  0  0  0  0
  3  4  1  0  0  0  0
  4  5  1  0  0  0  0
  5  6  2  0  0  0  0
  5  7  1  0  0  0  0
  8  3  2  0  0  0  0
  9  8  1  0  0  0  0
 10  9  2  0  0  0  0
 10 11  1  0  0  0  0
  1 10  1  0  0  0  0
M  END
```

Header block

Counts line

Atom block

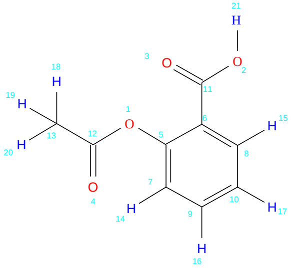
Bond block

Connection Table (Ctab)

Fig. 2.1. Illustrative example of a MOLFile for acetaminophen (also known as paracetamol). (a) Molecular structure of acetaminophen, commonly known as Tylenol. Tylenol is a widely used medicine for reducing fever and pain. (b) MOLFile for acetaminophen.

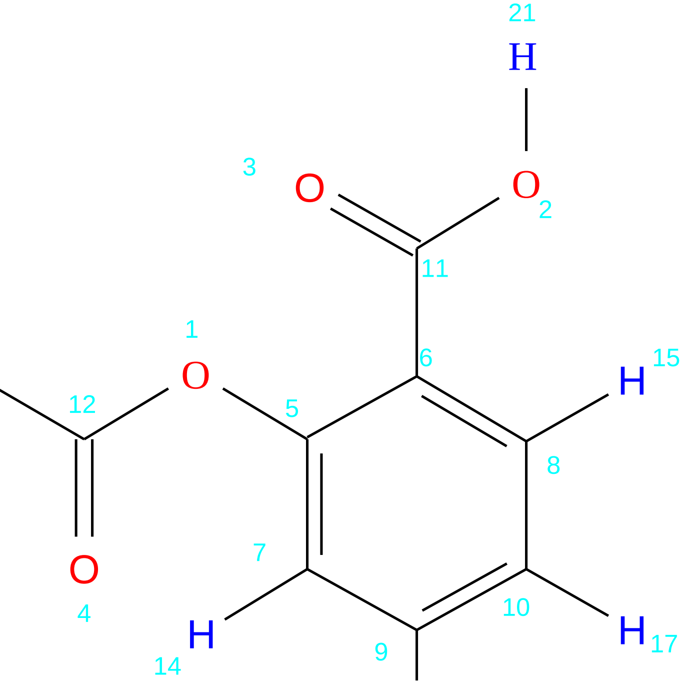
Aspirin in SDF Format

```
2244
  -OEChem-03090904423D

 21 21  0     0  0  0  0  0  0999 V2000
    1.8152   -0.9382    4.0419 O   0  0  0  0  0  0  0  0  0  0  0  0
    5.1920   -2.1043    2.0467 O   0  0  0  0  0  0  0  0  0  0  0  0
    3.9623   -2.6855    3.8563 O   0  0  0  0  0  0  0  0  0  0  0  0
    2.9441    1.1113    3.9712 O   0  0  0  0  0  0  0  0  0  0  0  0
    1.8509   -0.9767    2.6799 C   0  0  0  0  0  0  0  0  0  0  0  0
    2.9180   -1.5734    2.0082 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.8008   -0.4105    1.9570 C   0  0  0  0  0  0  0  0  0  0  0  0
    2.9348   -1.6038    0.6137 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.8176   -0.4410    0.5626 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.8846   -1.0376   -0.1090 C   0  0  0  0  0  0  0  0  0  0  0  0
    4.0236   -2.1714    2.7435 C   0  0  0  0  0  0  0  0  0  0  0  0
    2.4171    0.2017    4.5978 C   0  0  0  0  0  0  0  0  0  0  0  0
    2.3265    0.1503    6.0942 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.0345    0.0550    2.4732 H   0  0  0  0  0  0  0  0  0  0  0  0
    3.7445   -2.0729    0.0609 H   0  0  0  0  0  0  0  0  0  0  0  0
   -0.0005   -0.0011   -0.0002 H   0  0  0  0  0  0  0  0  0  0  0  0
    1.8958   -1.0640   -1.1947 H   0  0  0  0  0  0  0  0  0  0  0  0
    1.2777    0.1264    6.3998 H   0  0  0  0  0  0  0  0  0  0  0  0
    2.7936    1.0443    6.5170 H   0  0  0  0  0  0  0  0  0  0  0  0
    2.8566   -0.7302    6.4654 H   0  0  0  0  0  0  0  0  0  0  0  0
    5.9361   -2.5075    2.5428 H   0  0  0  0  0  0  0  0  0  0  0  0
  1  5  1  0  0  0  0
  1 12  1  0  0  0  0
  2 11  1  0  0  0  0
  2 21  1  0  0  0  0
  3 11  2  0  0  0  0
  4 12  2  0  0  0  0
  5  6  1  0  0  0  0
  5  7  2  0  0  0  0
  6  8  2  0  0  0  0
  6 11  1  0  0  0  0
  7  9  1  0  0  0  0
  7 14  1  0  0  0  0
  8 10  1  0  0  0  0
  8 15  1  0  0  0  0
  9 10  2  0  0  0  0
  9 16  1  0  0  0  0
 10 17  1  0  0  0  0
 12 13  1  0  0  0  0
 13 18  1  0  0  0  0
 13 19  1  0  0  0  0
 13 20  1  0  0  0  0
M  END
```

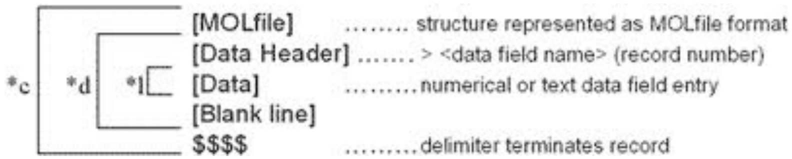


```
2244
  -OEChem-03090904423D

 21 21  0     0  0  0  0  0  0999 V2000
    1.8152   -0.9382    4.0419 O   0  0  0
    5.1920   -2.1043    2.0467 O   0  0  0
    3.9623   -2.6855    3.8563 O   0  0  0
    2.9441    1.1113    3.9712 O   0  0  0
    1.8509   -0.9767    2.6799 C   0  0  0
    2.9180   -1.5734    2.0082 C   0  0  0
    0.8008   -0.4105    1.9570 C   0  0  0
    2.9348   -1.6038    0.6137 C   0  0  0
    0.8176   -0.4410    0.5626 C   0  0  0
    1.8846   -1.0376   -0.1090 C   0  0  0
    4.0236   -2.1714    2.7435 C   0  0  0
    2.4171    0.2017    4.5978 C   0  0  0
    2.3265    0.1503    6.0942 C   0  0  0
```

```
    2.8566   -0.7302    6.4654
    5.9361   -2.5075    2.5428
  1  5  1  0  0  0  0
  1 12  1  0  0  0  0
  2 11  1  0  0  0  0
  2 21  1  0  0  0  0
  3 11  2  0  0  0  0
  4 12  2  0  0  0  0
  5  6  1  0  0  0  0
  5  7  2  0  0  0  0
  6  8  2  0  0  0  0
  6 11  1  0  0  0  0
```


21
H
3
O
O
2
11
1
6
H
15
O
12
5
8
7
O
10
4
H
H
17
14
9

SDfile

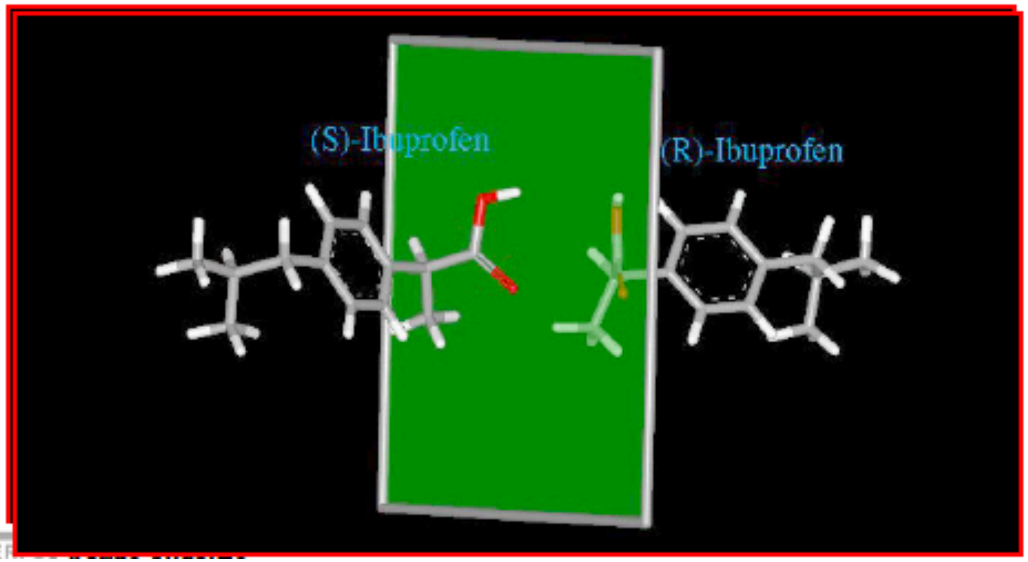


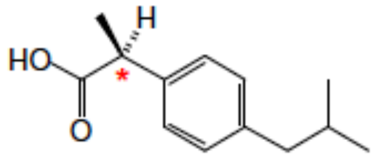

```
name
jscorina  12209406473DS
LongName
  7  6
   -0.0187    1.5258    0.0104 C   0  0  0  0  0
    0.0021   -0.0041    0.0020 C   0  0  0  0  0
    1.6831    2.1537   -0.0024 S   0  0  0  0  0
   -1.4333   -0.5336    0.0129 C   0  0  0  0  0
    2.0692    1.9811   -1.7665 C   0  0  0  0  0
   -1.4126   -2.0635    0.0045 C   0  0  0  0  0
    1.4620    3.1542   -2.5386 C   0  0  0  0  0
  2  1  1  0  0  0
  3  1  1  0  0  0
  4  2  1  0  0  0
  5  3  1  0  0  0
  6  4  1  0  0  0
  7  5  1  0  0  0

> <NSC>
19

> <CAS_RN>
638-46-0

$$$$
```

Importance of stereochemistry

Enantiomers (mirror image molecules) have

- ☆ identical physical and chemical properties
- ☆ different optical rotation
- ☆ often different biological activities

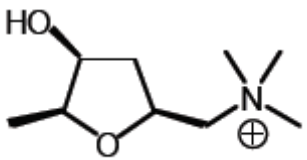




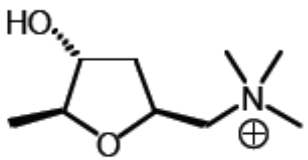
3D link

Importance of stereochemistry

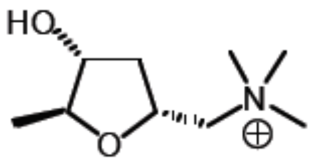
- Muscarine (muscarinic agonist): Natural product most potent
- 7 other stereoisomers possible: Considerable less potent



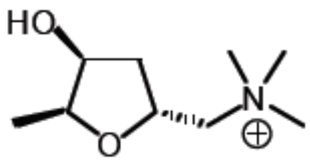
(+)-(2R,4R,5S)
1



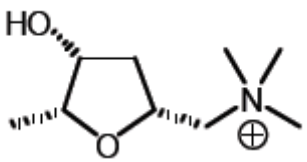
(-)-(2R,4S,5R)
2754



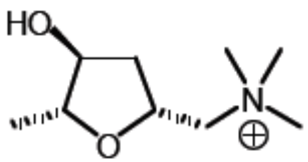
(+)-(2S,4S,5S)
1259



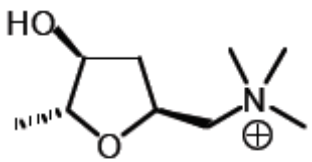
(-)-(2S,4R,5S)
263



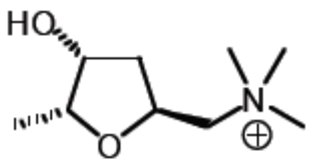
(-)-(2S,4S,5R)
331



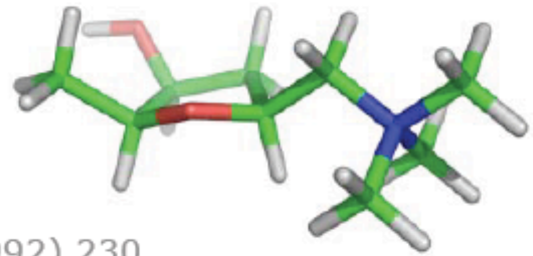
(+)-(2S,4R,5S)
1479



(-)-(2R,4R,5R)
1072



(+)-(2R,4S,5R)
1738



De Amici et al., Chirality 4 (1992) 230

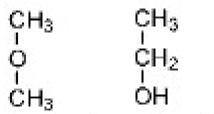
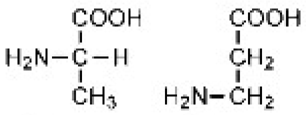
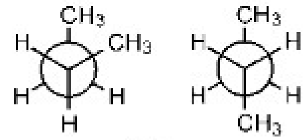
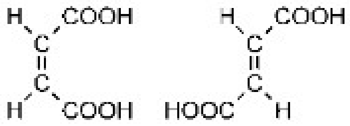
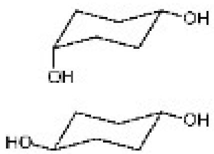
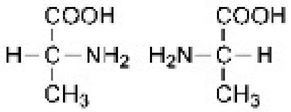
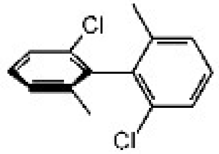
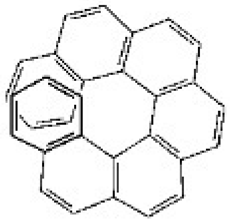
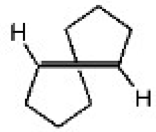
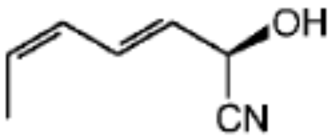
<table>
<tr><td colspan="7">isomers
compounds have an identical empirical formula, but different molecular structures</td></tr>
<tr><td colspan="2">constitutional isomers</td><td colspan="5">stereoisomers</td></tr>
<tr><td rowspan="3">structural isomers
different functional groups

dimethyl ether   ethanol</td><td rowspan="3">positional isomers
identical functional groups at different places

α-alanine   β-alanine</td><td colspan="4">configurational isomers</td><td rowspan="3">conformational isomers
different conformers by rotation around a single bond

gauche   n-butane   anti</td></tr>
<tr><td rowspan="2">cis/trans isomers
neighboring groups of double bonds can have two directions.

maleic acid (cis = Z)   fumaric acid (trans = E)
Ring systems can also be differentiated into cis/trans arrangements of substituents.

cis-
1,4-cyclohexanediol
trans-</td><td colspan="3">chirality</td></tr>
<tr><td>chiral center
if a molecule has one chiral atom, the isomers are like an image and mirror image (enantiomers); molecules with more than one chiral atom exist as enantiomers and diastereomers

D-alanine   L-alanine</td><td>axial chirality (atropisomerism, helicene)
if a molecule has four ligands which are placed pairwise along an axis and are not in one plane (atropisomer)

2,2′-dichloro-6,6′-dimethyl-1,1′-biphenyl
a special case of axial chirality is the arrangement of the molecule as a right- or left-handed helix (helicene)

heptahelicene</td><td>planar chirality
if the arrangement of a molecule can be distinguished into different sides

E-cyclooctene</td></tr>
</table>

Figure 2-67. Classification of isomeric structures of organic compounds.

Stereochemistry in SMILES



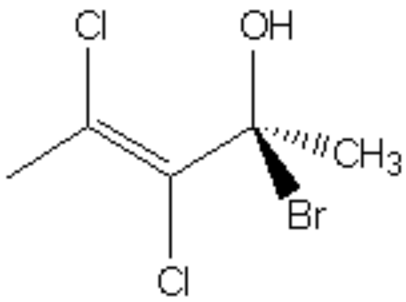
SMILES: C/C=C\C=\[C@@H](O)C#N

Figure 2-78. The stereochemistry of (2*R*,3*E*,5*E*)-2-hydroxy-3,5-heptadiene nitrile can be expressed in the SMILES notation with @ or (back)slashes.

@@: when viewed from atom along the bond to the chiral center, the sequence of atoms (H), (O) and (C#N) appear clockwise

"/" and "\”: “cis” and “trans” configuration

Stereochemistry in InChI



InChI=1/C5H7BrCl2O/c1-3(7)4(8)5(2,6)9/h9H,1-2H3/b4-3+/t5-/m1/s1

Stereochemical layer

Double bond stereo sub-layer

sp^3 stereo sub-layer

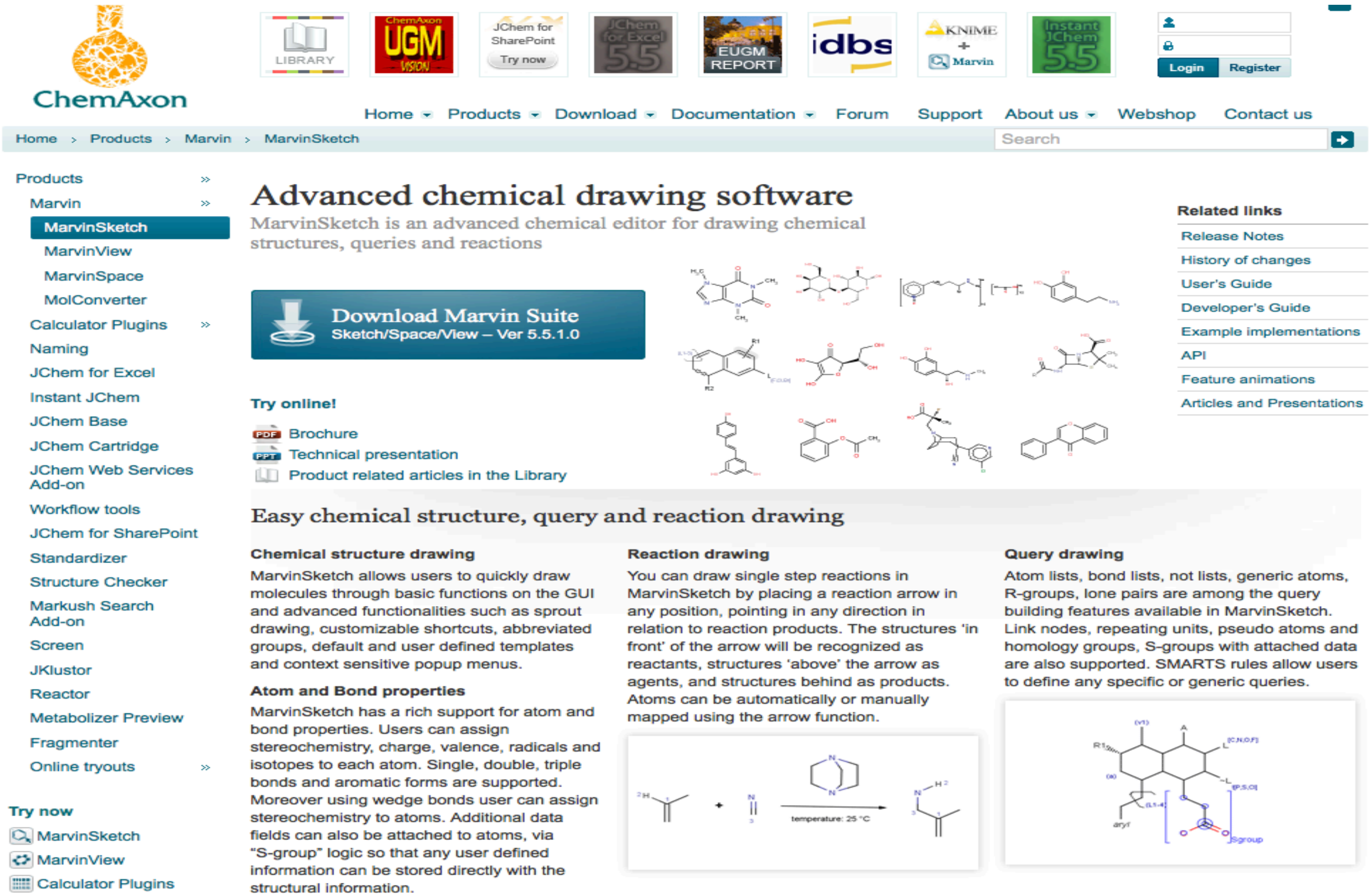
File formats

Table 2-5. The most important file formats for exchange of chemical structure information.

File format	*Suffix*	*Comments*	*Support*	*Ref.*
MDL Molfile	*.mol	Molfile; the most widely used connection table format	*www.mdli.com*	50
SDfile	*.sdf	Structure-Data file; extension of the MDL Molfile containing one or more compounds	*www.mdli.com*	50
RDfile	*.rdf	Reaction-Data file; extension of the MDL Molfile containing one or more sets of reactions	*www.mdli.com*	50
SMILES	*.smi	SMILES; the most widely used linear code and file format	*www.daylight.com*	20, 21
PDB file	*.pdb	Protein Data Bank file; format for 3D structure information on proteins and polynucleotides	*www.rcsb.org*	53
CIF	*.cif	Crystallographic Information File format; for 3D structure information on organic molecules	*www.iucr.org/iucr-top/cif/*	55
JCAMP	*.jdx, *.dx, *.cs	Joint Committee on Atomic and Molecular Physical Data; structure and spectroscopic format	*www.jcamp.org/*	56
CML	*.cml	Chemical Markup Language; extension of XML with specialization in chemistry	*www.xml-cml.org*	57–59

Molecular editors and viewers

http://www.chemaxon.com/products/marvin/



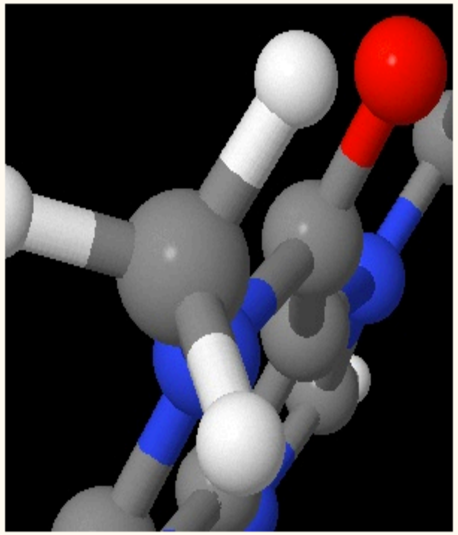
Molecular editors and viewers

http://jmol.sourceforge.net/

Jmol

Jmol: an open-source Java viewer for chemical structures in 3D

with features for chemicals, crystals, materials and biomolecules



Jmol is an interactive web browser applet.

This is a still image, but you can get an animated display of Jmol abilities by clicking here.

(The applet may take some seconds to load. Please, wait and do not reload the page in the meantime.)

The latest stable version is Jmol 12.0

Structure models

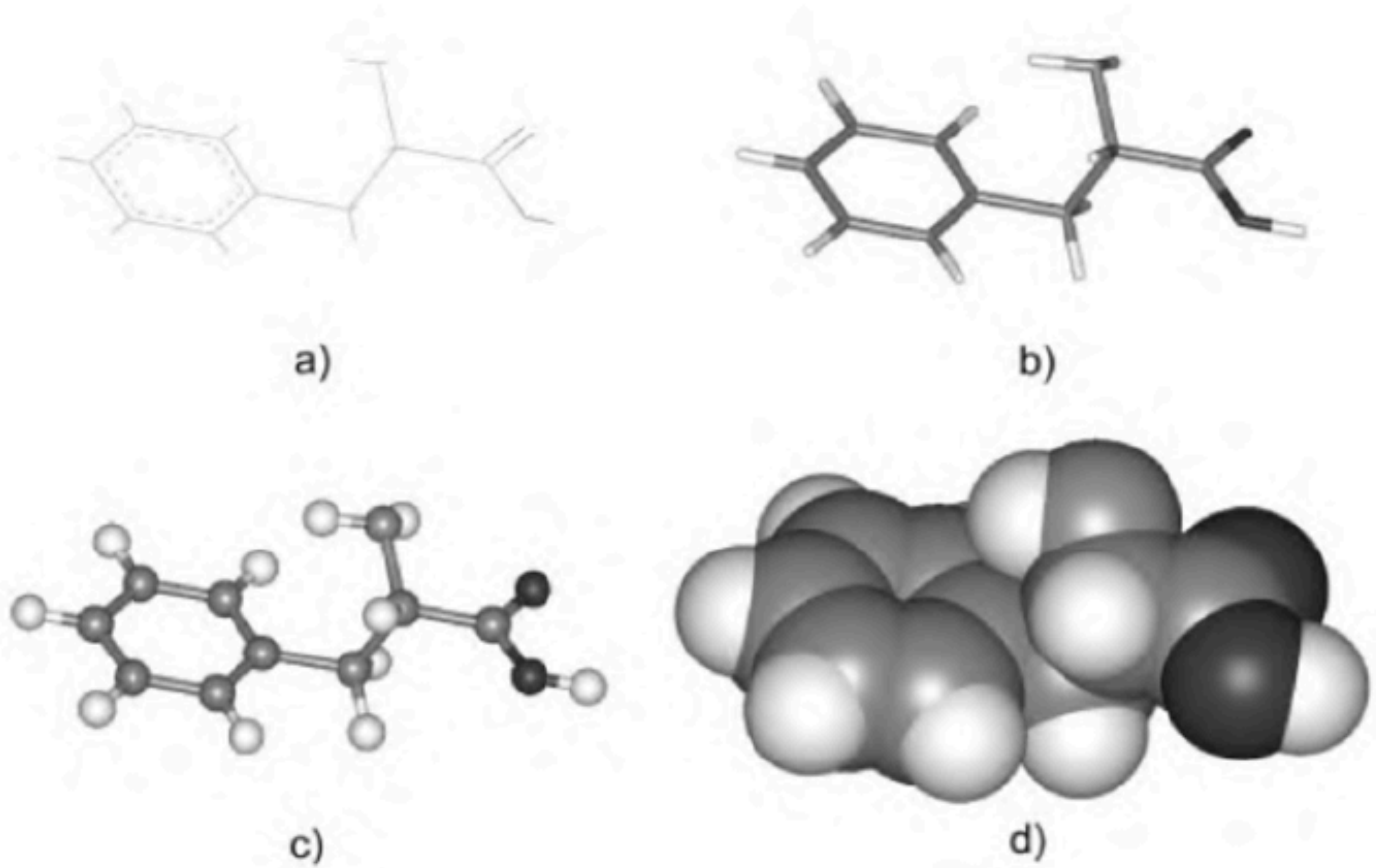


Figure 2-123. The most common molecular graphics representations of phenylalanine a) wire frame; b) capped sticks; c) balls and sticks; d) space-filling.

Format conversion

http://cactus.nci.nih.gov/translate/

Home | About | Contact | Disclaimer | Privacy

NCI/CADD Group

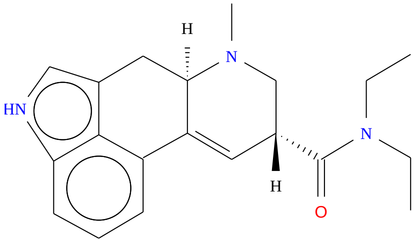
Online SMILES Translator and Structure File Generator

Form | News | Help | Acknowledgments

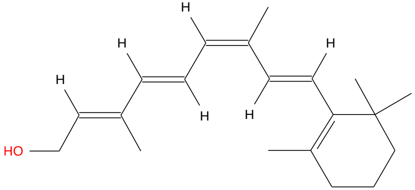
Input Format	Unique SMILES Output Format (Unique SMILES)	
C12C3C4C1C5C4C3C25 Start Structure Editor Please choose this field if you want to submit your own SMILES strings or create a SMILES string using the Structure Editor. A submitted file has precedence, so delete any entry below if you want to submit a new SMILES string.	○ Display on screen ○ SMILES TXT file ○ SDF ○ PDB ◉ MOL (only single structure generated)	Use ◉ Kekule or ○ Aromatic SMILES representation (choose "Aromatic" for closer approximation to Daylight USMILES) SD, PDB or MOL files should contain ◉ 2D ○ 3D coordinates
Browse... Please choose this field if you want to translate your own files. The service will automatically recognize SD files (single and multiple structure), text files with multiple SMILES fields, MOL files and PDB files (and in fact any other format CACTVS recognizes).	If the input file contains a single structure, the output will also be single structure. Multiple structure input formats will generate multiple structure output for those formats that support this. Otherwise, only the first structure will be used. SD files will contain a UNIQUE_SMILES field for unique SMILES and an USER_SUPPLIED_SMILES field for the user-supplied SMILES (if avaliable)	

Reset | Translate

Examples of SMILES



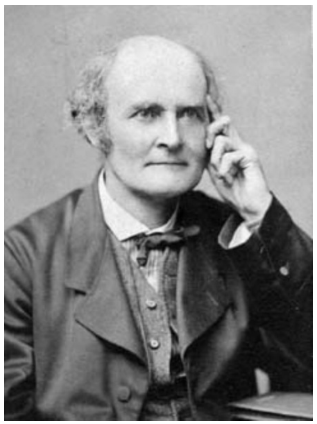
CCN(CC)C(=O)[C@H]1CN(C)[C@@H]2Cc3c[nH]c4cccc(C2=C1)c34



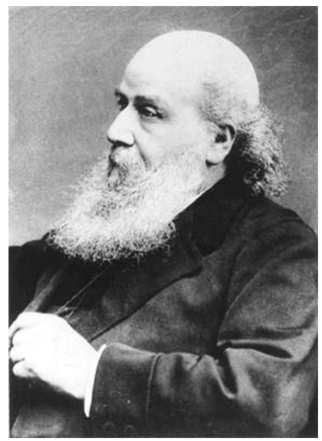
C/C\(=C\CO)/C=C/C=C(/C)\C=CC1=C(C)CCCC1(C)C

Graph Theory / Algebra / Chemistry

Enumeration of Chemical Isomers

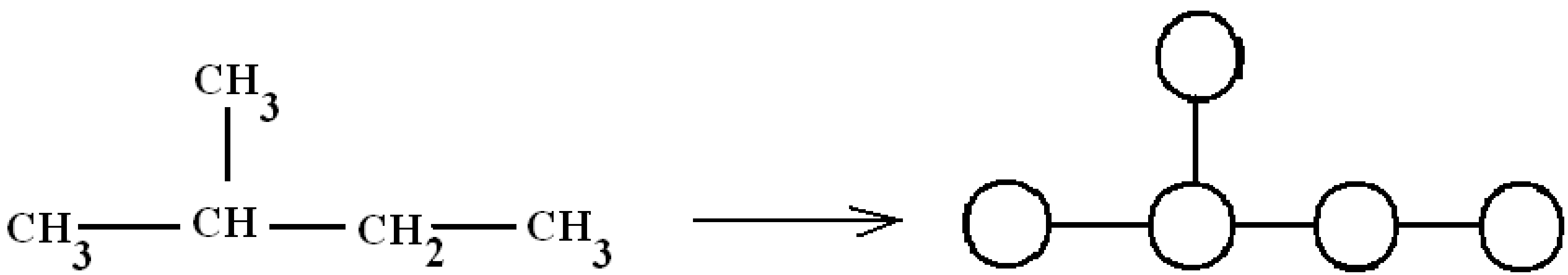


Arthur Cayley

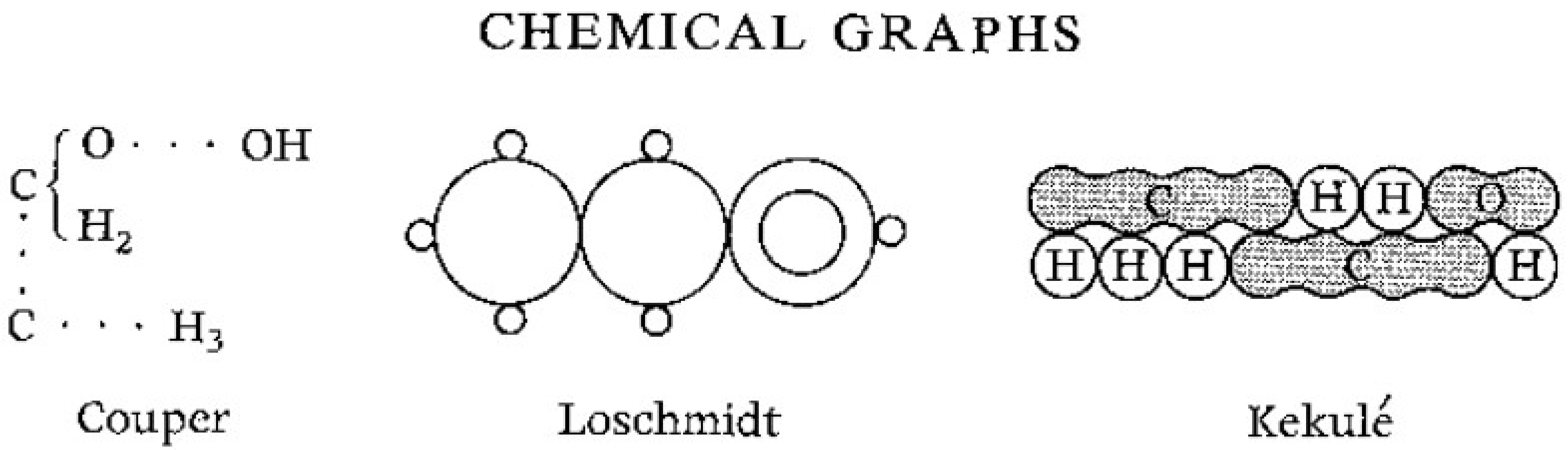


James J. Sylvester

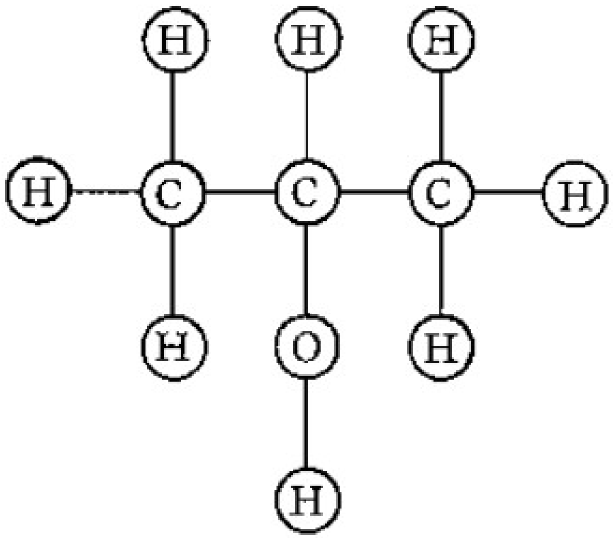
George Polya

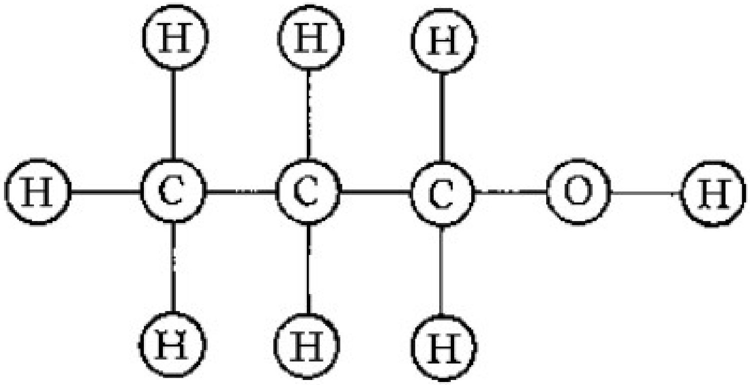


Couper (1858) / Loschmidt (1861) / Kekulé (1861)



Crum Brown (1864) and Frankland (1866)





8 cm. from the primary. *Reverse the wires in the secondary circuit, reverse the wires in the primary circuit, how you please, the mercury always moves towards the point of the capillary.*

8. Shouting or singing (excepting the above-mentioned note) produces no visible effect under the conditions mentioned in Experiments 5, 6, and 7.

9. If the secondary coil be now moved close up, so as to cover as completely as possible the primary, talking to the telephone with the ordinary voice, *i.e.* with moderate strength and at any pitch, produces a definite movement of the mercury column for each word, some sounds of course giving more movement than others, *but the movement is always towards the end of the capillary.* Singing the note mentioned in Experiments 5, 6, and 7 loudly, produces a movement too large to be measured with the electrometer.

Reversing the poles of the magnet in the telephone does not alter the results of Experiments 5, 6, 7, and 9.

On mentioning the above results to Dr. Burdon Sanderson, he suggested that the apparently anomalous behaviour of the electrometer might be accounted for, by supposing that the mercury moved *quicker* when a current passed towards the point of the capillary than when it flowed in the opposite direction; so that if a succession of rapidly alternating currents be passed through the instrument, the mercury will always move towards the point of the capillary, the movement away from the point being masked by the sluggishness of the instrument in that direction. That this explanation is the correct one is proved by the following experiment:—The current from two Grove's cells is sent through a metal reed vibrating 100 times a second, the contact being made and broken at each vibration, the primary wire of a Du Bois Reymond's induction-coil is also included in the circuit; on connecting the electrometer with the secondary coil placed at an appropriate distance the mercury always moves to the point of the tube whatever be the direction of the current. F. J. M. PAGE

Physiological Laboratory, University College, London, February 2

NOTE.—On February 4 Prof. Graham Bell kindly placed at my disposal a telephone much more powerful than any of those I had previously used. On speaking to this instrument, the electrometer being in the circuit, movements of the mercury column as considerable as those in Experiment 9 were observed.—F. J. M. P.

CHEMISTRY AND ALGEBRA

IT may not be wholly without interest to some of the readers of NATURE to be made acquainted with an analogy that has recently forcibly impressed me between branches of human knowledge apparently so dissimilar as modern chemistry and modern algebra. I have found it of great utility in explaining to non-mathematicians the nature of the investigations which algebraists are at present busily at work upon to make out the so-called *Grundformen* or irreducible forms appurtenant to binary quantics taken singly or in systems, and I have also found that it may be used as an instrument of investigation in purely algebraical inquiries. So much is this the case that I hardly ever take up Dr. Frankland's exceedingly valuable "Notes for Chemical Students," which are drawn up exclusively on the basis of Kekulé's exquisite conception of *valence*, without deriving suggestions for new researches in the theory of algebraical forms. I will confine myself to a statement of the grounds of the analogy, referring those who may feel an interest in the subject and are desirous for further information about it to a memoir which I have written upon it for the new *American Journal of Pure and Applied Mathematics*, the first number of which will appear early in February.

The analogy is between atoms and *binary* quantics exclusively.

I compare every binary quantic with a chemical atom. The number of factors (or rays, as they may be regarded by an obvious geometrical interpretation) in a binary quantic is the analogue of the number of *bonds*, or the *valence*, as it is termed, of a chemical atom.

Thus a linear form may be regarded as a monad atom, a quadratic form as a duad, a cubic form as a triad, and so on.

An invariant of a system of binary quantics of various degrees is the analogue of a chemical substance composed of atoms of corresponding *valences*. The order of such invariant in each set of coefficients is the same as the number of atoms of the corresponding *valence* in the chemical compound.

A co-variant is the analogue of an (organic or inorganic) compound radical. The orders in the several sets of coefficients corresponding, as for invariants, to the respective valences of the atoms, the free valence of the compound radical then becomes identical with the degree of the co-variant in the variables.

The weight of an invariant is identical with the number of the bonds in the chemicograph of the analogous chemical substance, and the weight of the leading term (or basic differentiant) of a co-variant is the same as the number of bonds in the chemicograph of the analogous compound radical. Every invariant and covariant thus becomes expressible by a *graph* precisely identical with a Kekuléan diagram or chemicograph. But not every chemicograph is an algebraical one. I show that by an application of the algebraical law of reciprocity every algebraical graph of a given invariant will represent the constitution in terms of the roots of a quantic of a type reciprocal to that of the given invariant of an invariant belonging to that reciprocal type. I give a rule for the geometrical multiplication of graphs, *i.e.* for constructing a *graph* to the product of in- or co-variants whose separate graphs are given. I have also ventured upon a hypothesis which, whilst in nowise interfering with existing chemicographical constructions, accounts for the seeming anomaly of the isolated existence as "monad molecules" of mercury, zinc, and arsenic—and gives a rational explanation of the "mutual saturation of bonds."

I have thus been led to see more clearly than ever I did before the existence of a common ground to the new mechanism, the new chemistry, and the new algebra. Underlying all these is the theory of pure colligation, which applies undistinguishably to the three great theories, all initiated within the last third of a century or thereabouts by Eisenstein, Kekulé, and Peaucellier.

Baltimore, January 1 J. J. SYLVESTER

PALMEN ON THE MORPHOLOGY OF THE TRACHEAL SYSTEM

DR. PALMEN, of Helsingfors, has recently published an interesting memoir on the tracheal system of insects. He observes that although the gills of certain aquatic larvæ are attached to the skin very near to the points at which the spiracles open in the mature insects, and though spiracles and gills do not co-exist in the same segment, yet the point of attachment of the gills never exactly coincides with the position of the future spiracle. Moreover, he shows that even during the larval condition, although the spiracles are not open, the structure of the stigmatic duct is present, and indeed that it opens temporarily at each moult, to permit the inner tracheal membrane to be cast, after which it closes again. In fact, then, he urges, the gills and spiracles do not correspond exactly, either in number or in position, and there can therefore be between them no genetic connection. He concludes that the insects with open tracheæ are not derived from ancestors provided with gills,

Some Polya Enumeration

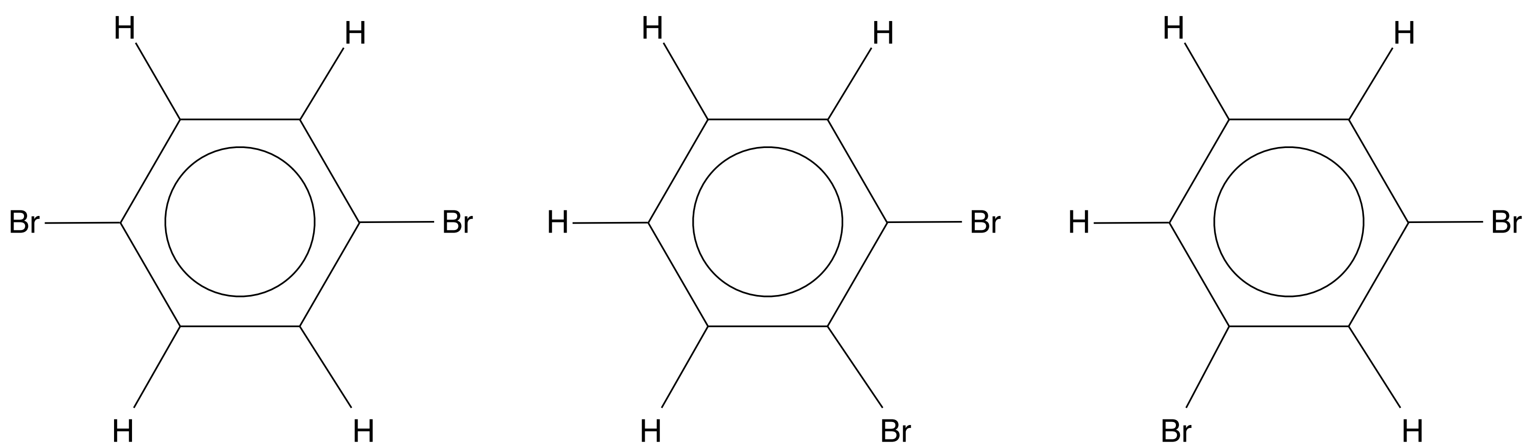
Pólya Theory

Formally: Pólya Theory counts equivalence classes, where the equivalence classes are induced by group actions. Since groups describe symmetry, Pólya Theory is counting the number of distinct objects in the presence of symmetry.

Informally: Pólya Theory does "common sense" counting.

Chemical Isomers

Pólya's original objective was to determine the number of distinct compounds given a chemical formula.



There are three distinct compounds with the formula $C_6H_4Br_2$.

Example

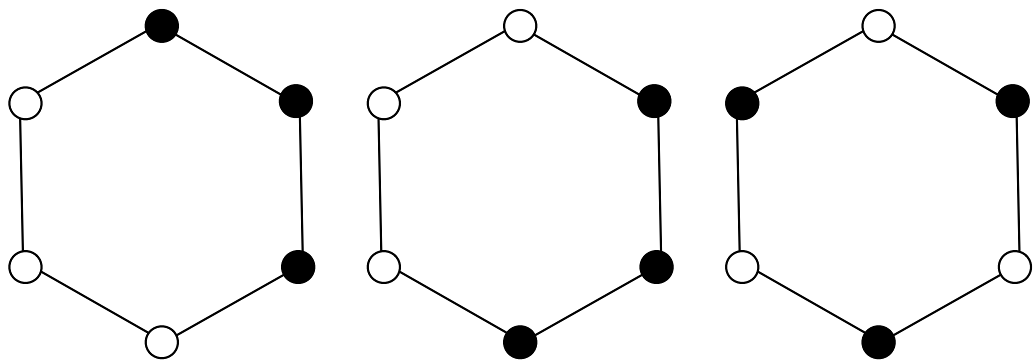


Figure 1: Common sense says that these the two bracelets on the left are the “same”, the third bracelet is “different”

Another Example

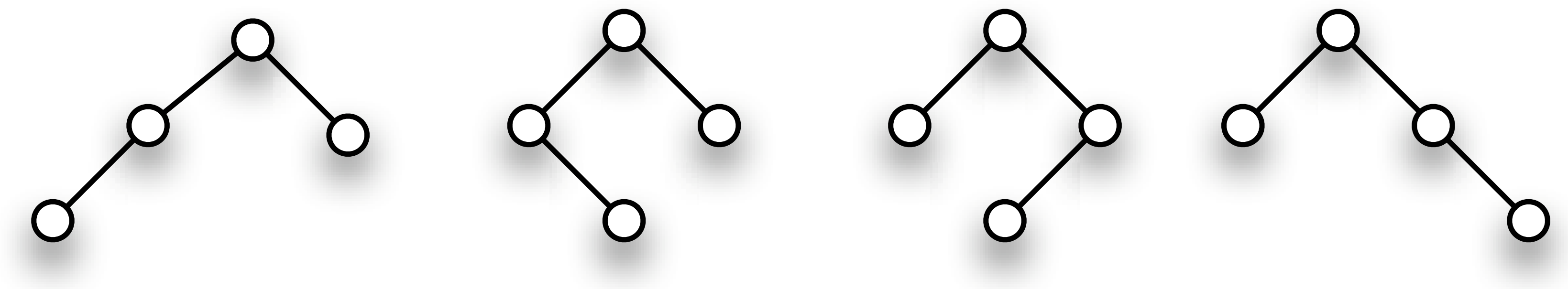


Figure 2: These 4 binary trees are equivalent if left and right are considered indistinguishable

Yet Another Example

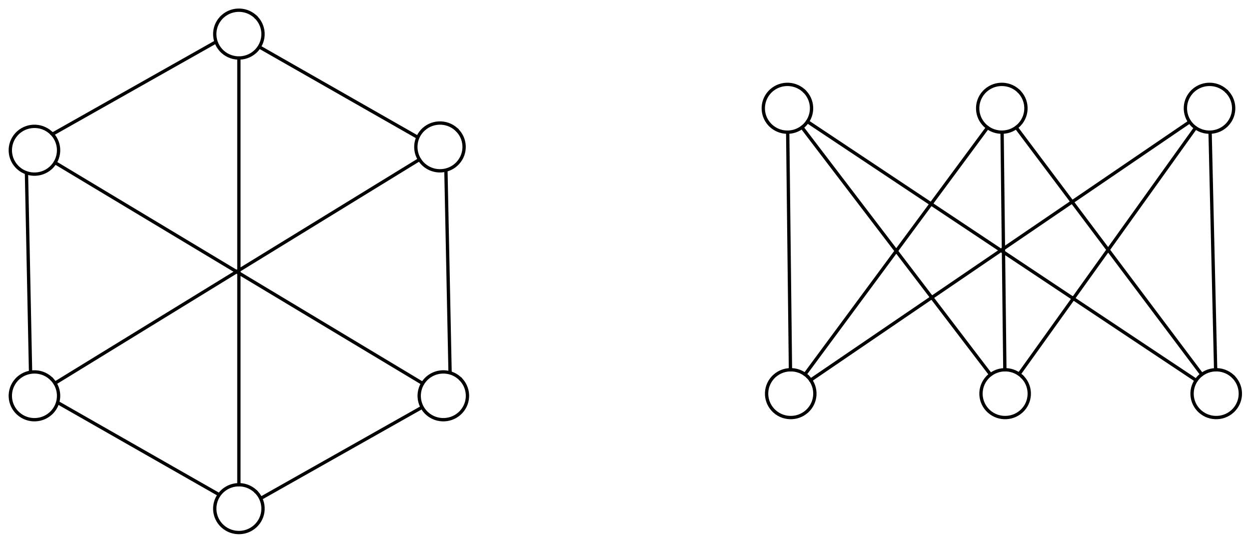


Figure 3: These two graphs are isomorphic, although it is not visually obvious

And Another

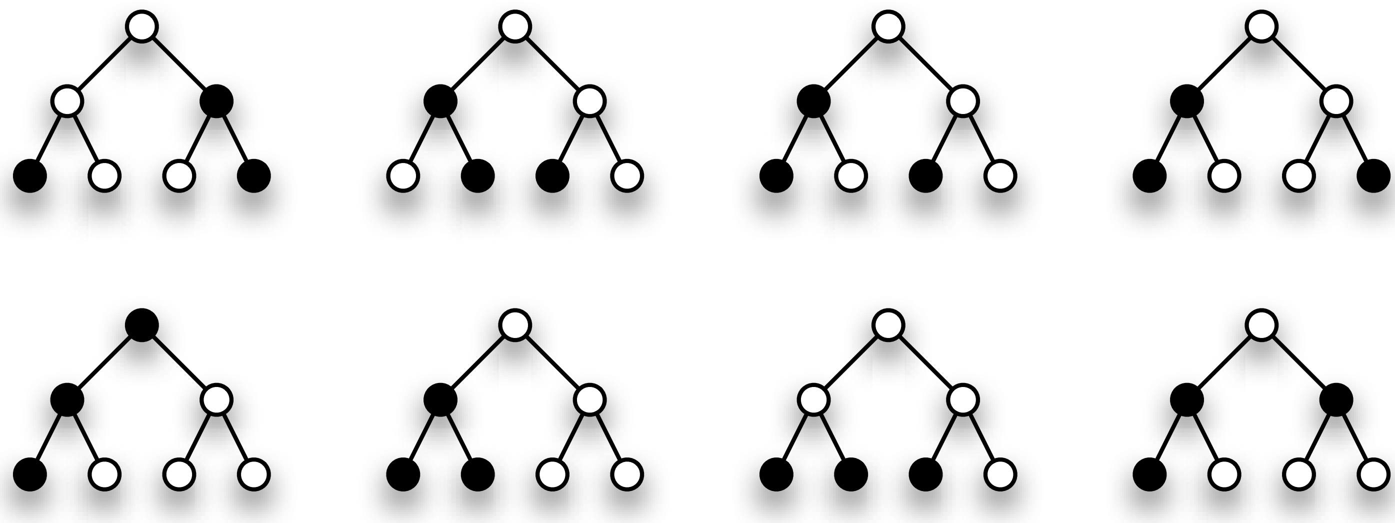
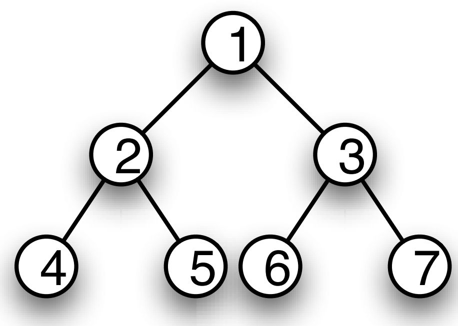


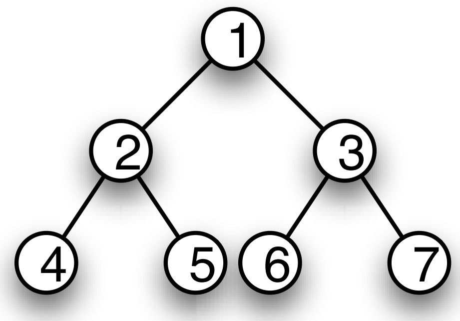
Figure 4: If left and right are indistinguishable then the top row of graph colorings are equivalent, and the bottom row are different.

Example Calculation

Problem: Count the number of distinct black and white colorings of the tree in Figure 7 if left and right are indistinguishable.



Step 1



Write down the permutations that leave the tree invariant if left and right are indistinguishable. This is a group G generated by:

$$\pi_1 = (1)(2\ 3)(4\ 6)(5\ 7)$$

$$\pi_2 = (1)(2)(3)(4\ 5)(6)(7)$$

$$\pi_3 = (1)(2)(3)(4)(5)(6\ 7)$$

Step 2

List the group elements, calculate the monomials.

Element	Cycle Representation	Monomial
I	(1)(2)(3)(4)(5)(6)(7)	x_1^7
π_1	(1)(2 3)(4 6)(5 7)	$x_1 x_2^3$
π_2	(1)(2)(3)(4 5)(6)(7)	$x_1^5 x_2$
π_3	(1)(2)(3)(4)(5)(6 7)	$x_1^5 x_2$
$\pi_2\pi_3 = \pi_3\pi_2$	(1)(2)(3)(4 5)(6 7)	$x_1^3 x_2^2$
$\pi_1\pi_2 = \pi_3\pi_1$	(1)(2 3)(4 6 5 7)	$x_1 x_2 x_4$
$\pi_1\pi_3 = \pi_2\pi_1$	(1)(2 3)(4 7 5 6)	$x_1 x_2 x_4$
$\pi_1\pi_2\pi_3$	(1)(2 3)(4 7)(5 6)	$x_1 x_2^3$

Table 1: The elements of G

Step 3

Add the monomials to get the *cycle index* for the group:

$$P_G(x_1, x_2, x_4) = \frac{1}{8}(x_1^7 + 2x_1^5 x_2 + 2x_1 x_2^3 + 2x_1 x_2 x_4 + x_1^3 x_2^2)$$

Formally the cycle index is defined:

$$P_G(x_1, x_2, \ldots, x_{|D|}) = \frac{1}{|G|} \sum_{\pi \in G} x_1^{l_1(\pi)} x_2^{l_2(\pi)} \ldots x_{|D|}^{l_{|D|}(\pi)}$$

where D is the the set acted on by elements of G, $|D|$ is the size of the set, $l_k(\pi)$ is the number of cycles of length k in π.

Step 4

Pólya's Enumeration Theorem says that the number of distinct k-colorings is

$$
\begin{aligned}
P_G(k,k,k,k) &= \frac{1}{8}(k^7 + 2k^6 + 2k^4 + 2k^3 + k^5) \\
&= \frac{k^3}{8}(k^4 + 2k^3 + k^2 + 2k + 2) \\
&= \frac{k^3}{8}(k+1)(k^3 + k^2 + 2)
\end{aligned}
$$

This must be an integer for all integer values of k, so $k^3(k+1)(k^3+k^2+2)$ must be divisible by 8.

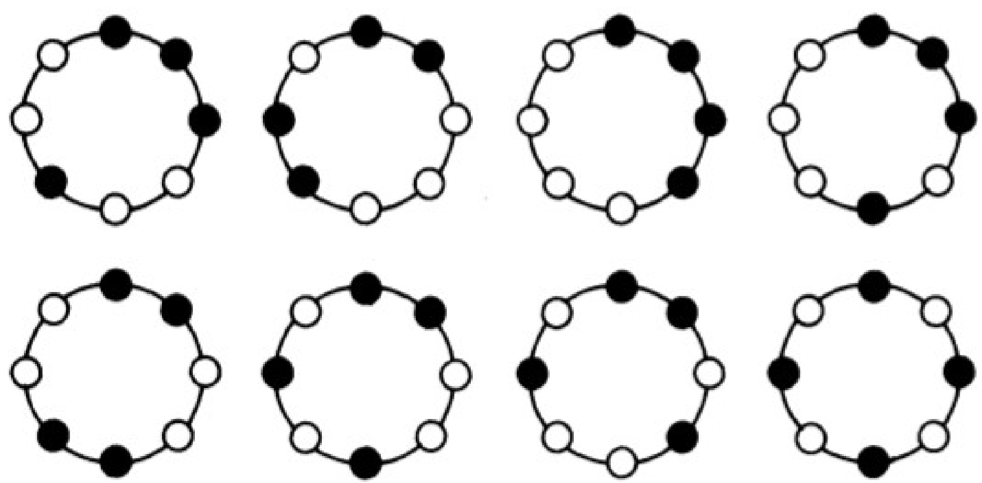
The Solution

The number of 2-colorings of the binary tree, with left and right indistinguishable, is

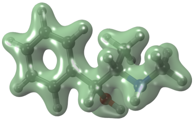
$$P_G(2,2,2,2) = \frac{1}{8}(2^7 + 2^5 + 2^7 + 2^4 + 2^5)$$
$$= 42$$

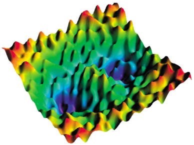
$$1 + x + 4x^2 + 5x^3 + 8x^4 + 5x^5 + 4x^6 + x^7 + x^8.$$

Thus, for example, there are five inequivalent necklaces having three black beads, and eight with equal numbers of black and white beads. The latter are shown in FIGURE 1.

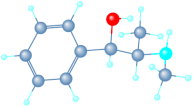


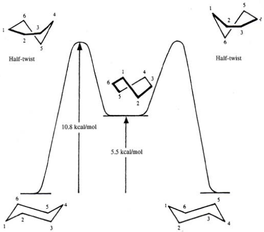
Levels of Abstraction in Computational Chemistry



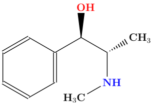


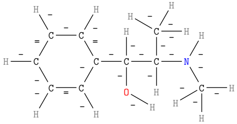
Potential energy surface





Reaction coordinate



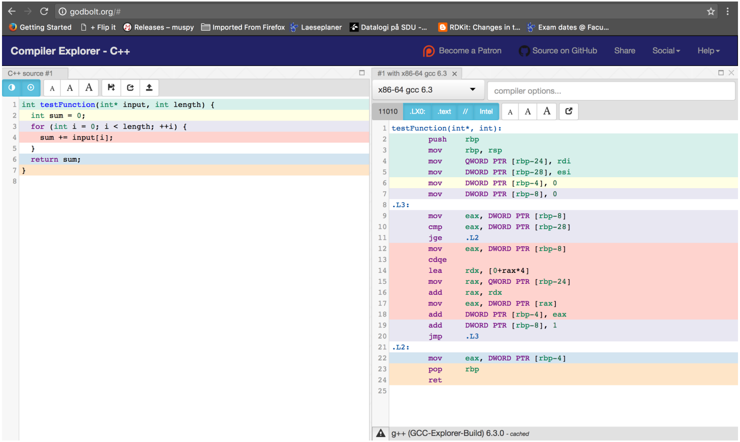


$L \leftarrow K \rightarrow R$

Graph grammar

[Andersen et al., Proceedings of the Royal Society A, 2017]

Levels of Abstraction in Programming



Declarative Description ↔ DSL ↔ C++ ↔ Assembler

Levels of Abstraction in Computer Science

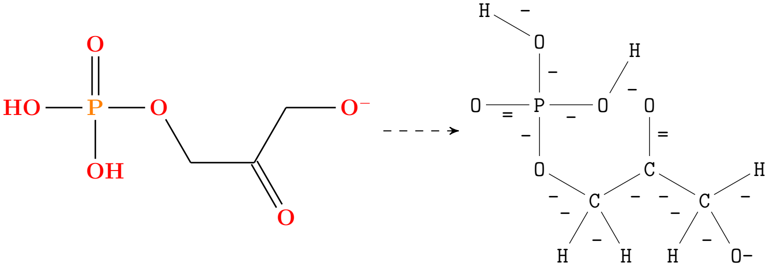
"The psychological profiling [of a Computer Scientist] is mostly the ability to shift levels of abstraction, from low level to high level. To see something in the small and to see something in the large."

Donald Knuth

Modelling and Analysis of Chemical Systems

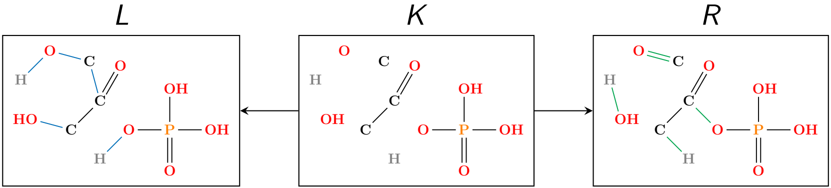
Modelling and Analysis of Chemical Systems

1. Model molecules as labelled graphs.
   - An old idea: [J. J. Sylvester, *Chemistry and Algebra*, Nature 1878]
   - Molecule: simple, connected, labelled graph.
   - Vertex labels: atom type, charge.
   - Edge labels: bond type.



Modelling and Analysis of Chemical Systems

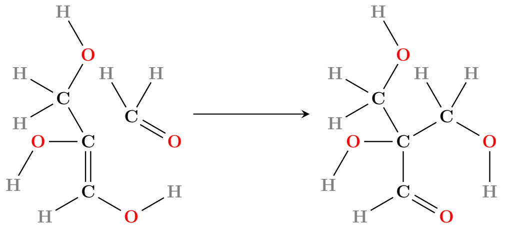
2. Model reaction types and graph transformation rules.

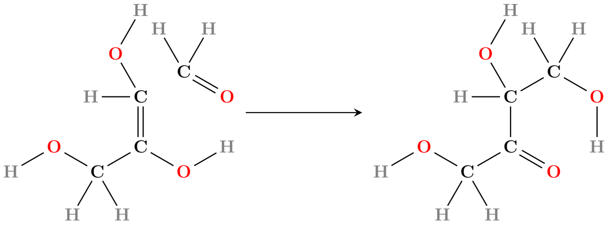


Example: Carbon rearrangement

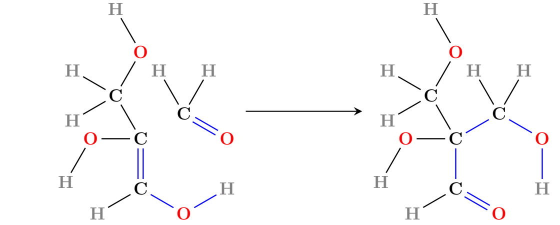
- Aldolase: ketone + aldehyde $\longrightarrow$ ketone
- Aldose-Ketose: aldehyde $\longrightarrow$ ketone
- Ketose-Aldose: ketone $\longrightarrow$ aldehyde
- Phosphohydrolase: H_2O+CnP $\longrightarrow$ Cn+Pi
- Phosphoketolase Pi+ketone $\longrightarrow$ carbonyl + CnP+water
- Transaldolase: Cn+Cm$\longrightarrow$ C(n+3)+C(m-3)
- Transketolase: Cn+Cm$\longrightarrow$ C(n+2)+C(m-2)

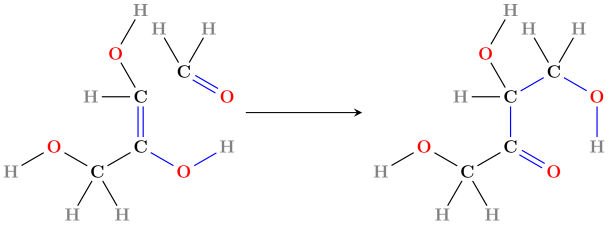
Chemical Reactions (Educts $\rightarrow$ Products)





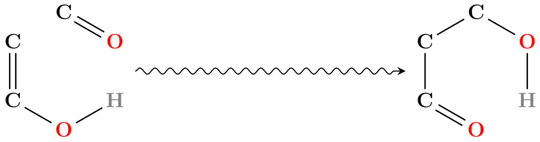
Chemical Reactions (of the Same Type)

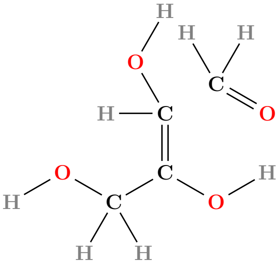




Chemical Reaction Patterns

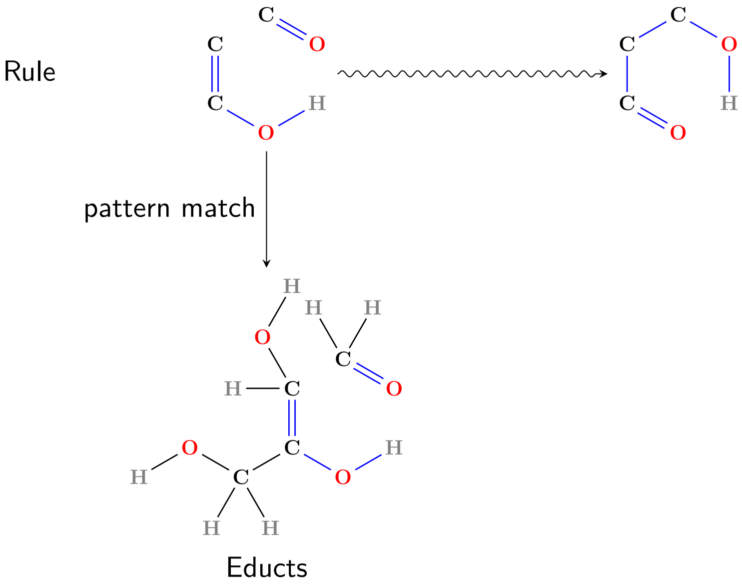
Rule



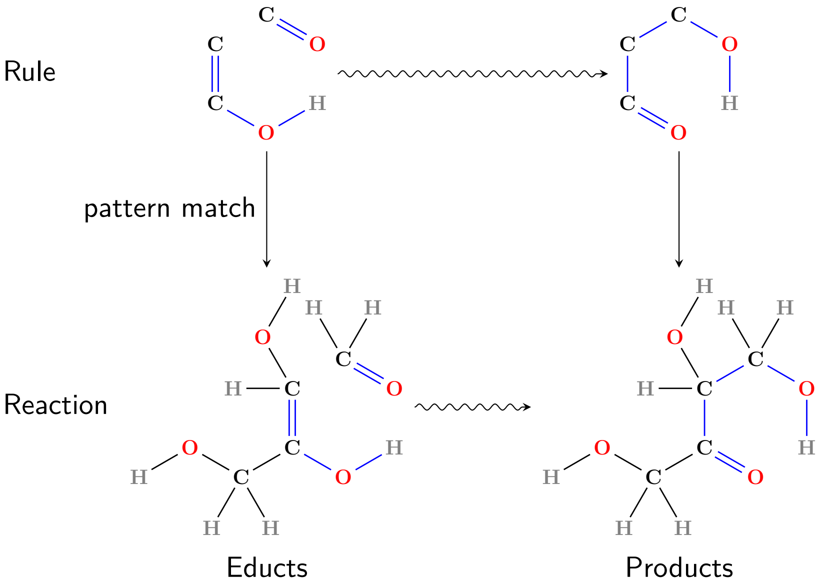


Educts

Chemical Reaction Patterns

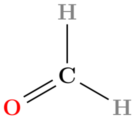


Chemical Reaction Patterns

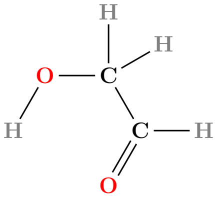


Grammar Example: The Formose Chemistry

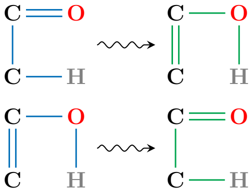
Formaldehyde:



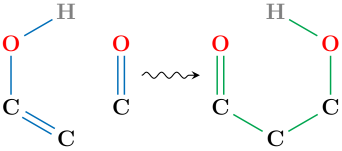
Glycolaldehyde:



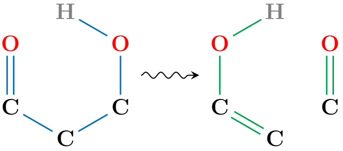
Keto-enol tautomerism:



Aldol addition:



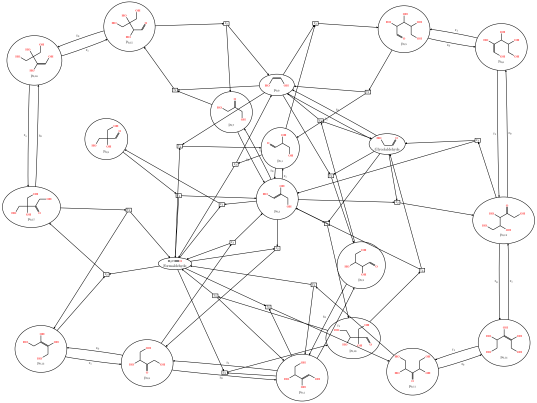
Retro aldol addition:



Modelling and Analysis of Chemical Systems

3. Generate a reaction network.

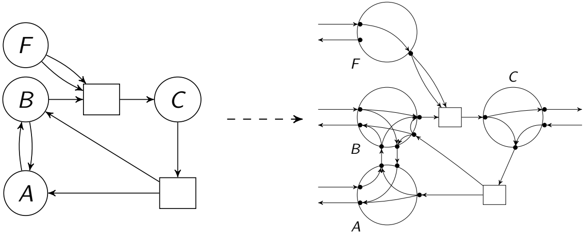
```
dg = dgRuleComp(inputGraphs,
    addSubset(inputGraphs) >> rightPredicate[
        lambda d: all(countCarbon(a) <= 5 for a in d.right)
    ](        repeat(inputRules)     )
)
dg.calc()
```



Modelling and Analysis of Chemical Systems

4. Set up pathway model.

Network → Extended network → Expanded network → ILP



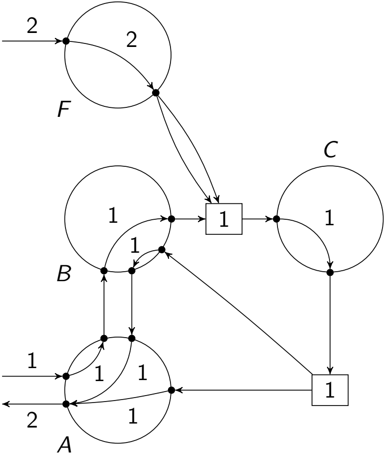
Conservation constraints:

$$\sum_{e\in\delta^+_{\widetilde{E}}(v)} m_v(e^+)f(e) - \sum_{e\in\delta^-_{\widetilde{E}}(v)} m_v(e^-)f(e) = 0 \qquad \forall v \in \widetilde{V}$$

Modelling and Analysis of Chemical Systems

5. Formulate pathway question.

Example: Given 2 formaldehyde and 1 glycolaldehyde, how can 2 glycolaldehyde be produced through autocatalysis.

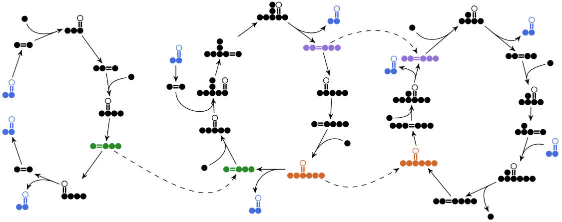


Modelling and Analysis of Chemical Systems

6. Enumerate many alternate pathways.

Example (Formose):
Network: all molecules with at most 9 carbon atoms.

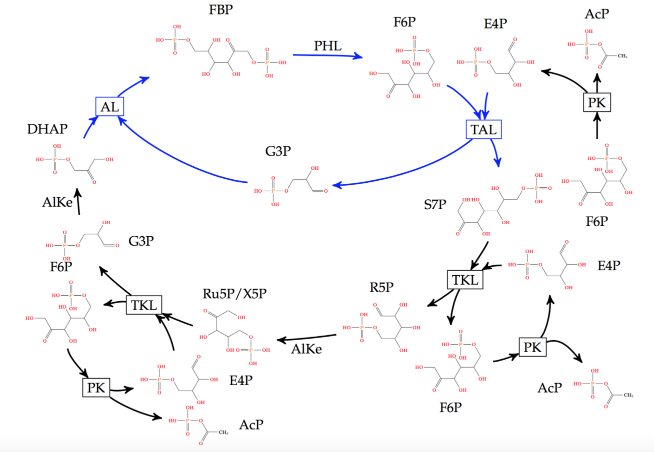
	Maximum #C						
Reactions used	4	5	6	7	8	9	Sum
6	0	0	1	1	1	2	5
7	0	0	0	0	0	2	2
8	1	5	7	17	37	68	135
9	0	0	12	12	37	69	130
10	0	12	50	274	849	—	$\geq$ 1185
11	0	5	41	190	738	—	$\geq$ 974
							$\geq$ 2431



Another Example: Non-oxidative Glycolysis

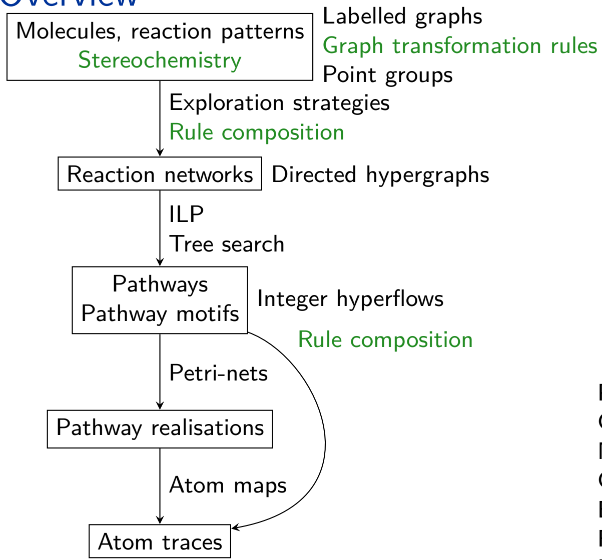
You specify: F6P + 2 $P_i \rightarrow$ 3 AcP + $2H_2O$

You get (for example):



Many alternatives for suggestion in [Bogorad, Lin, and Liao, Nature, 2013]

Overview



Category theory
Double Pushout
Rule composition
Monomorphisms
Isomorphisms
Canonicalisation
Automorphisms

Software package: MØD

C++, Python, Bash, LaTeX

Pentose phosphate pathway
Glycolysis (EMP and ED)
Non-oxidative glycolysis
Citric acid cycle
Enzyme mechanisms
Formose
Prebiotic chemistry (HCN)
Eschenmoser's GLX scenario
DNA templated computing

http://cheminf.imada.sdu.dk