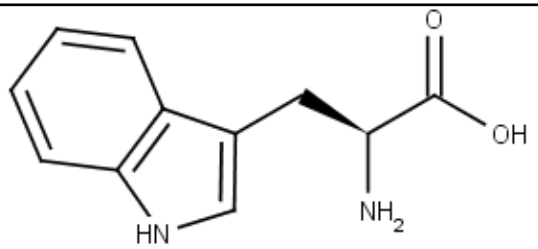
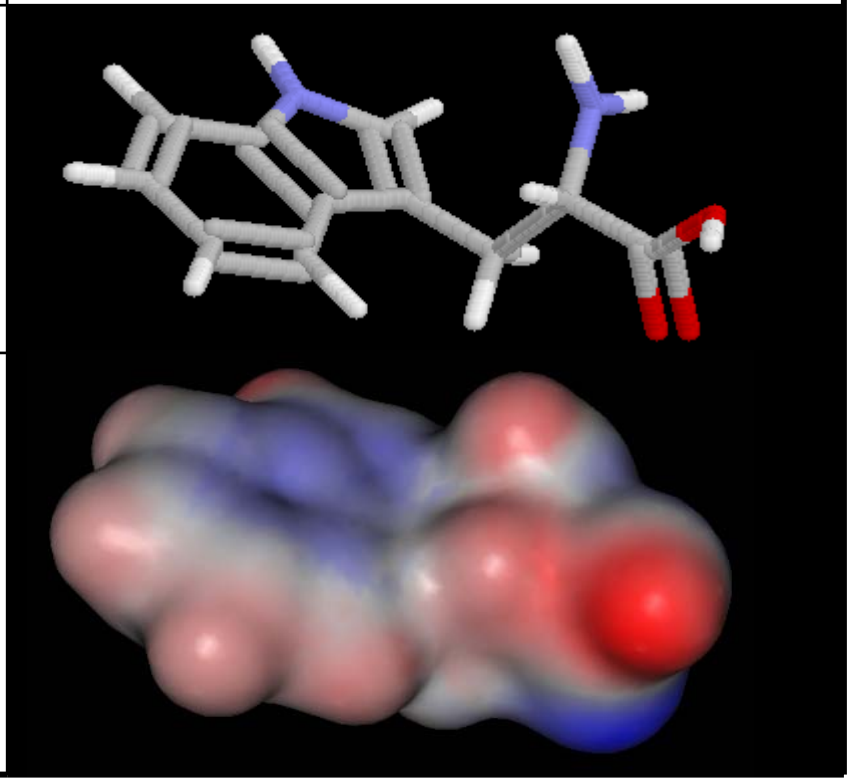
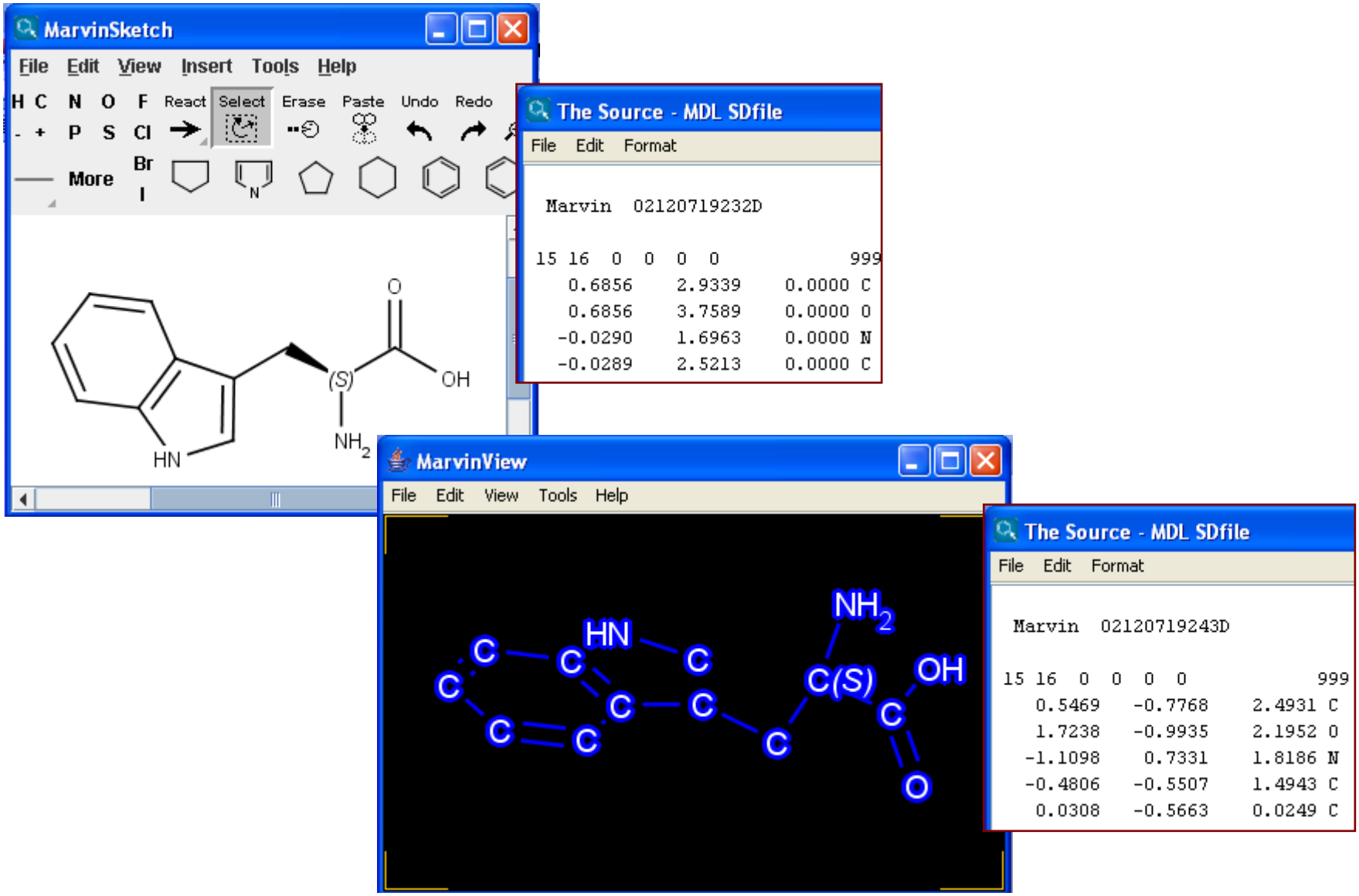
Representation of molecular structures

Coutersy of Prof. João Aires-de-Sousa, University of Lisbon, Portugal

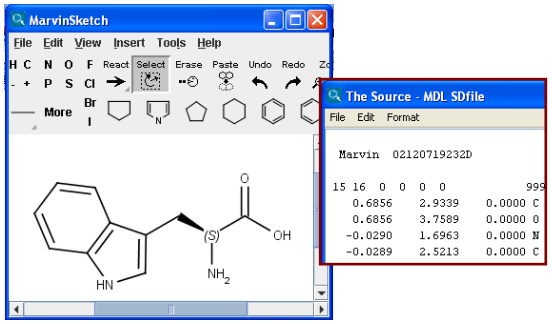
A hierarchy of structure representations

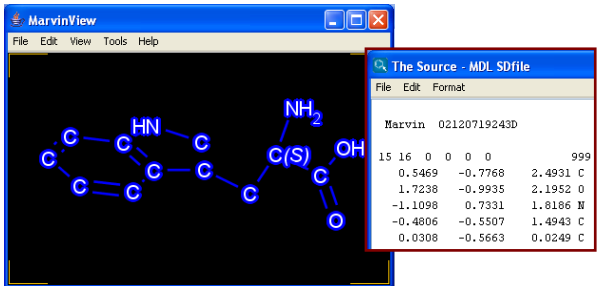
Name	(*S*)-Tryptophan
2D Structure	
3D Structure	
Molecular surface	

Storing molecular structures in a computer



Storing molecular structures in a computer





- Information must be **coded** into interconvertible formats that can be read by software applications.

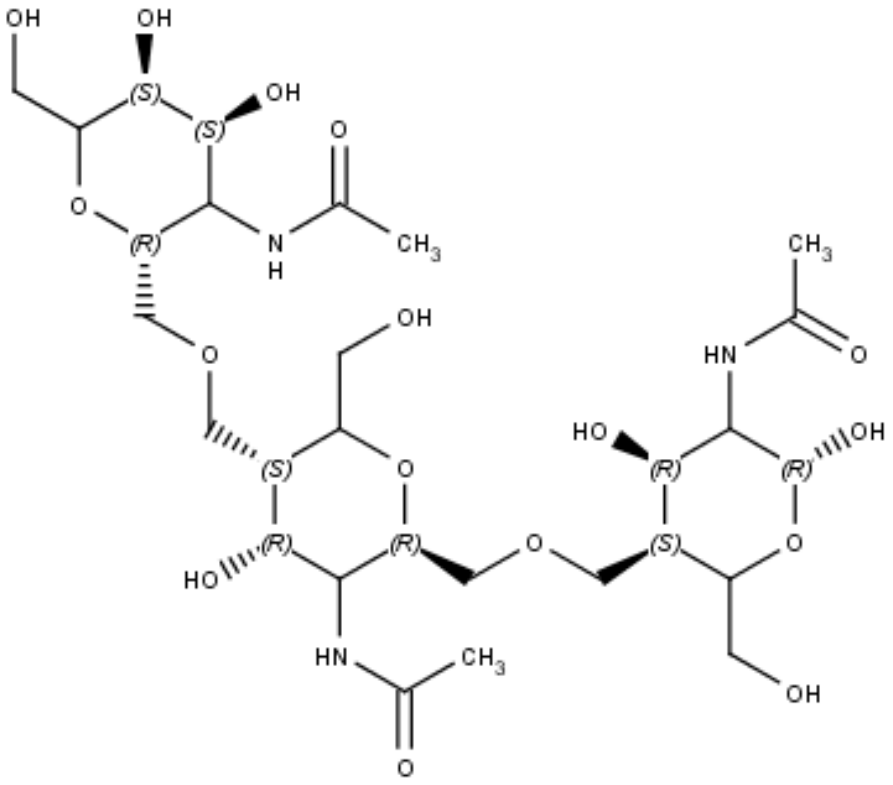
- Applications: visualization, communication, database searching / management, establishment of structure-property relationships, estimation of properties, …

Coding molecular structures

• A **non-ambiguous** representation identifies a single possible structure, e.g. the name ‘o-xylene’ represents one and only one possible structure.

• A representation is **unique** if any structure has only one possible representation (some nomenclature isn’t, e.g. ‘1,2-dimethylbenzene’ and ‘o-xylene’ represent the same structure).

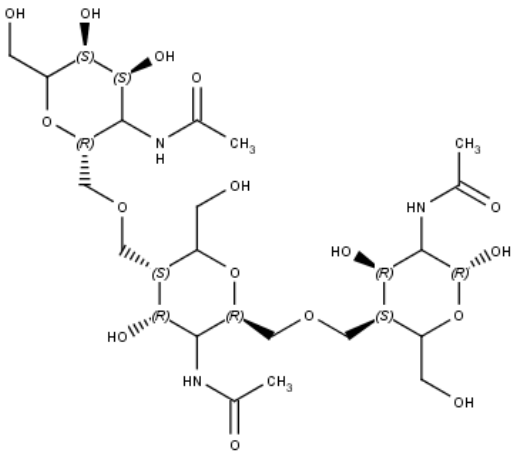
IUPAC Nomenclature



IUPAC name : N-[(2R,4R,5S)-5-[[(2S,4R,5S)-3-acetamido-5-[[(2S,4S,5S)-3-acetamido-4,5-dihydroxy-6-(hydroxymethyl)oxan-2-yl]methoxymethyl]-4-hydroxy-6-(hydroxymethyl)oxan-2-yl]methoxymethyl]-2,4-dihydroxy-6-(hydroxymethyl)oxan-3-yl]acetamide

IUPAC Nomenclature

- **Advantages**:
  - standardized systematic classification
  - stereochemistry is included
  - widespread
  - unambiguous
  - allows reconstruction from the name



- **Disadvantages**:
  - extensive rules
  - alternative names are allowed (non-unique)
  - long complicated names

IUPAC name : N-[(2R,4R,5S)-5-[[(2S,4R,5S)-3-acetamido-5-[[(2S,4S,5S)-3-acetamido-4,5-dihydroxy-6-(hydroxymethyl)oxan-2-yl]methoxymethyl]-4-hydroxy-6-(hydroxymethyl)oxan-2-yl]methoxymethyl]-2,4-dihydroxy-6-(hydroxymethyl)oxan-3-yl]acetamide

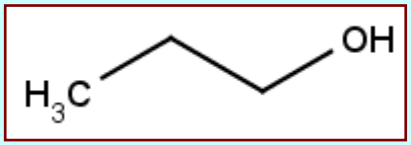
Linear notations

Represent structures by linear sequences of letters and numbers, e.g. IUPAC nomenclature.

Linear notations can be extremely compact, which is an advantage for the storage of structures in a computer (particularly when disk space is limited).

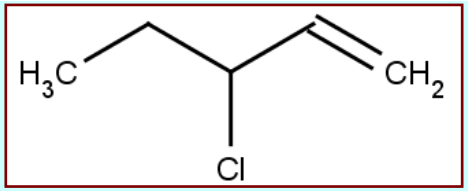
Linear notations allow for an easy transmission of structures, e.g. in a Google-type search, or in an email.

The SMILES notation

Example: 

SMILES representation : **CCCO**

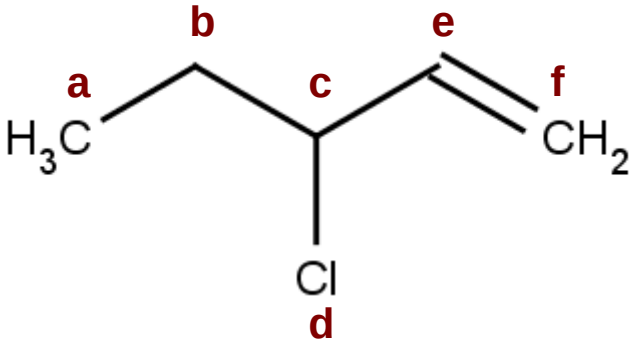
1. Atoms are represented by their atomic symbols.
2. Hydrogen atoms are omitted (are implicit).
3. Neighboring atoms are represented next to each other.
4. Double bonds are represented by ‘=‘, triple bonds by ‘#’.
5. Branches are represented by parentheses.
6. Rings are represented by allocating digits to the two connecting ring atoms.

Example : 

SMILES: **CCC(Cl)C=C**

The SMILES notation

1. Atoms are represented by their atomic symbols.
2. Hydrogen atoms are omitted (are implicit).
3. Neighboring atoms are represented next to each other.
4. Double bonds are represented by ‘=‘, triple bonds by ‘#’.
5. Branches are represented by parentheses.
6. Rings are represented by allocating digits to the two connecting ring atoms.

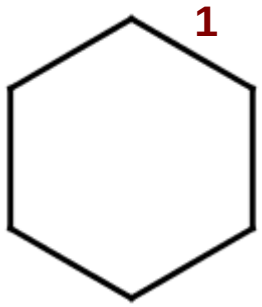


a b c d e f

SMILES: **CCC(Cl)C=C**

The SMILES notation

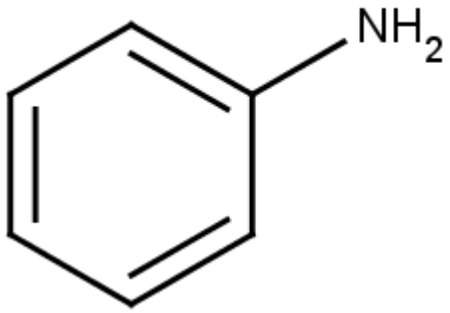
1. Atoms are represented by their atomic symbols.
2. Hydrogen atoms are omitted (are implicit).
3. Neighboring atoms are represented next to each other.
4. Double bonds are represented by ‘=‘, triple bonds by ‘#’.
5. Branches are represented by parentheses.
6. Rings are represented by allocating digits to the two connecting ring atoms.



SMILES: **C1CCCCC1**

The SMILES notation

1. Atoms are represented by their atomic symbols.
2. Hydrogen atoms are omitted (are implicit).
3. Neighboring atoms are represented next to each other.
4. Double bonds are represented by '=', triple bonds by '#'.
5. Branches are represented by parentheses.
6. Rings are represented by allocating digits to the two connecting ring atoms.
7. Aromatic rings are indicated by lower-case letters.

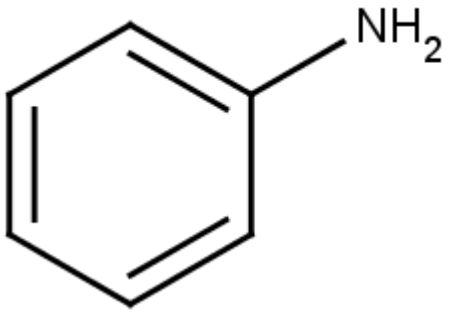


SMILES: **Nc1ccccc1**

The SMILES notation

- Is unambiguous (a SMILES string unequivocally represents a single structure).

- Is it unique ??



SMILES: **Nc1ccccc1**

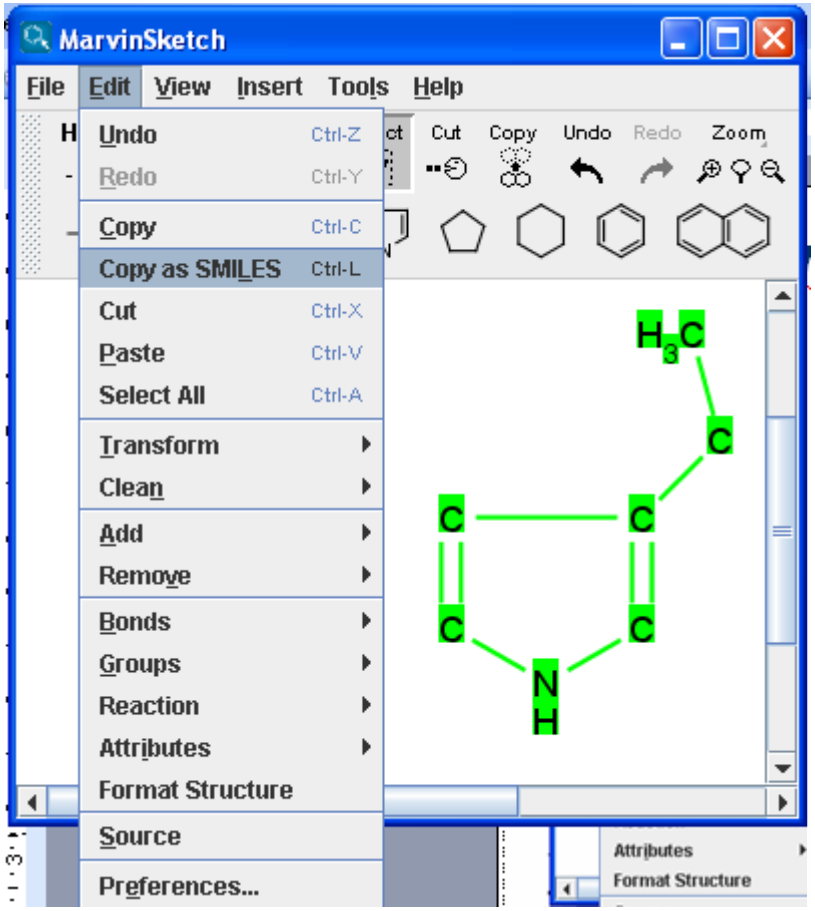
but also **c1ccccc1N**

or **c1cc(N)ccc1**

- **Solution**: algorithm that guarantees a canonical representation (each structure is always represented by the same SMILES string)

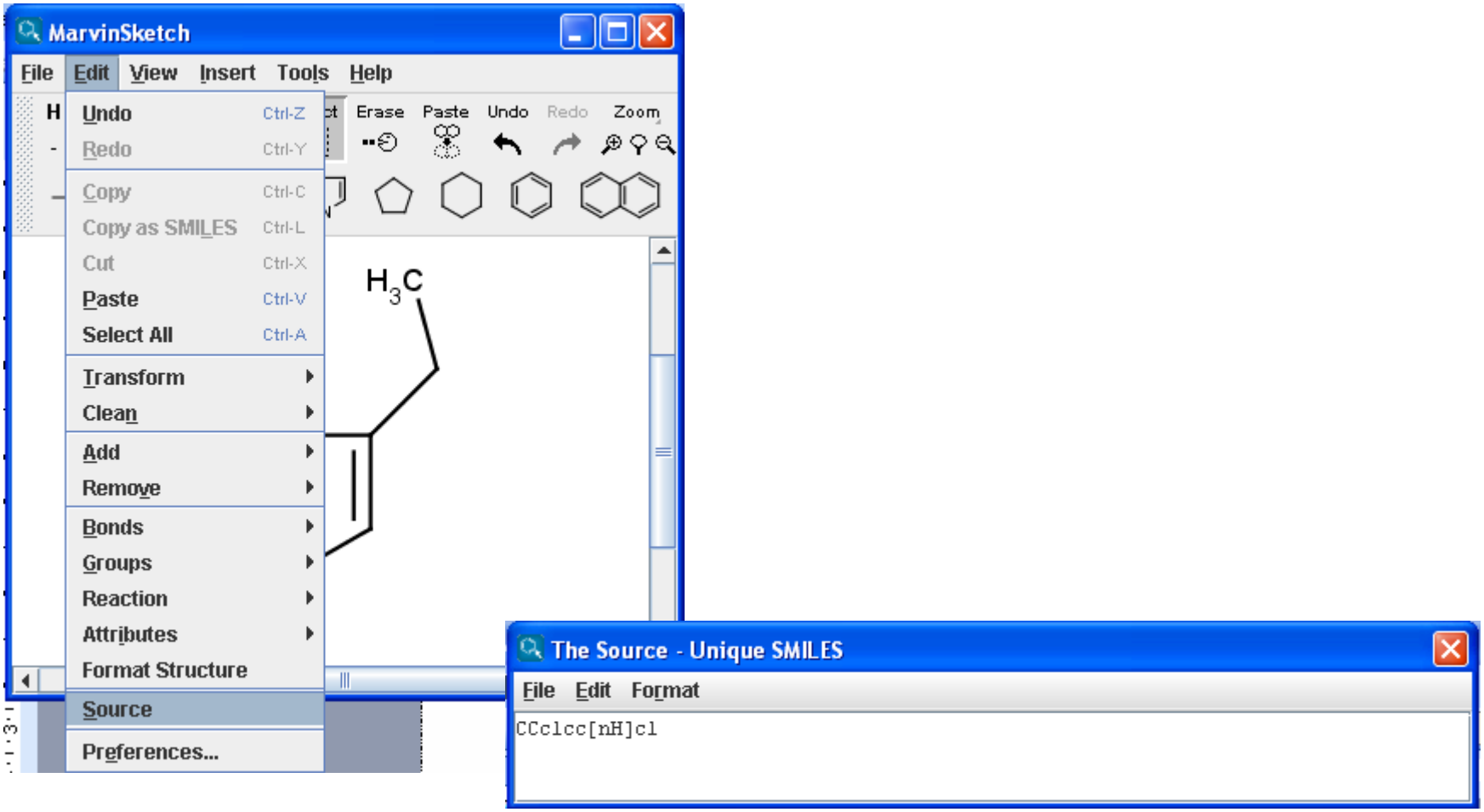
- More at: http://www.daylight.com/dayhtml_tutorials/index.html

SMILES notation in MarvinSketch



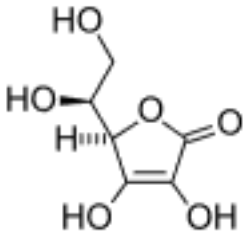
Paste

SMILES notation in MarvinSketch



The InChI notation (IUPAC International Chemical Identifier)

Example:



L-ascorbic acid

InChI=1/C6H8O6/c7-1-2(8)5-3(9)4(10)6(11)12-5/h2,5,7-10H,1H2/t2-,5+/m0/s1

A digital equivalent to the IUPAC name for a compound.

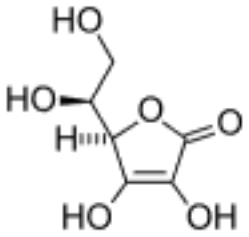
Five layers of information: connectivity, tautomerism, isotopes, stereochemistry, and charge.

An algorithm generates an unambiguous unique notation.

Official web site : http://www.iupac.org/inchi/

The InChI notation (IUPAC International Chemical Identifier)

Example:



L-ascorbic acid

InChI=1/C6H8O6/c7-1-2(8)5-3(9)4(10)6(11)12-5/h2,5,7-10H,1H2/t2-,5+/m0/s1

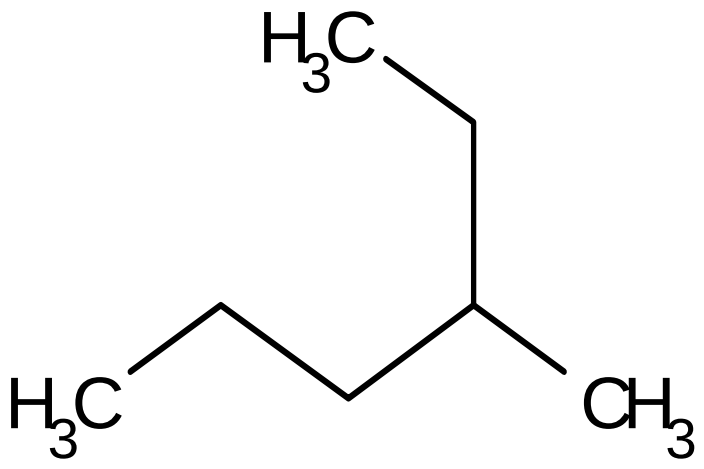
Each layer in an InChI string contains a specific class of structural information. This format is designed for compactness, not readability, but can be interpreted manually.

The length of an identifier is roughly proportional to the number of atoms in the substance. Numbers inside a layer usually represent the canonical numbering of the atoms from the first layer (chemical formula) except H.

Graph theory

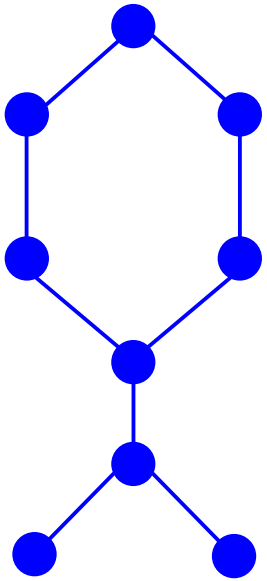
A molecular structure can be interpreted as a mathematical graph where each atom is a node, and each bond is an edge.

Such a representation allows for the mathematical processing of molecular structures using the graph theory.



Topological Graph Theory

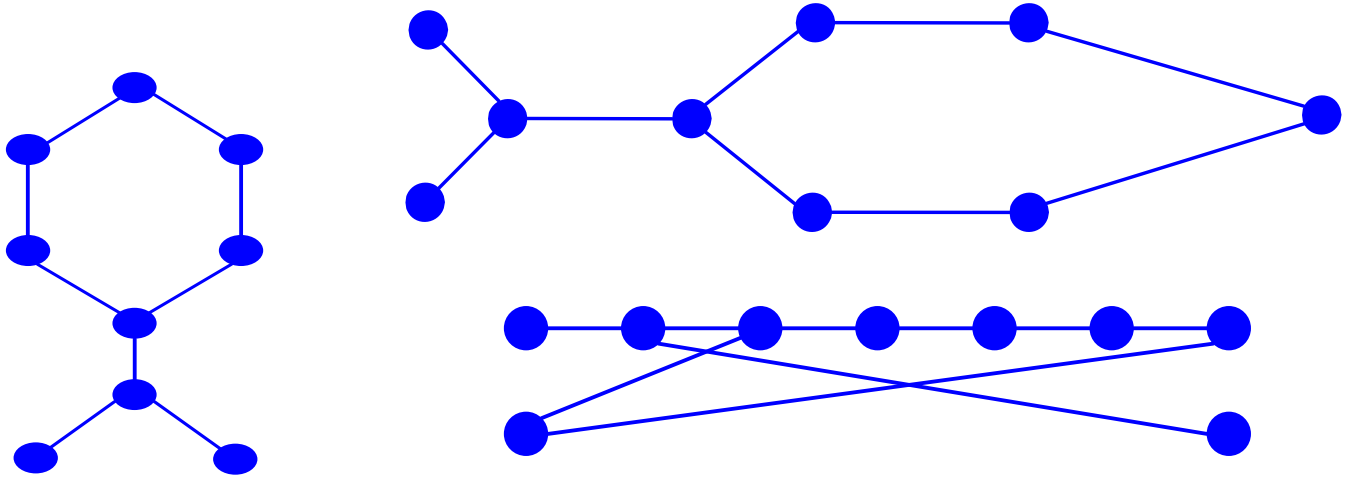
- branch of mathematics
  - particularly useful in chemical informatics and in computer science generally
- study of “graphs” which consist of
  - a set of “nodes”
  - a set of “edges” joining pairs of nodes



Properties of graphs

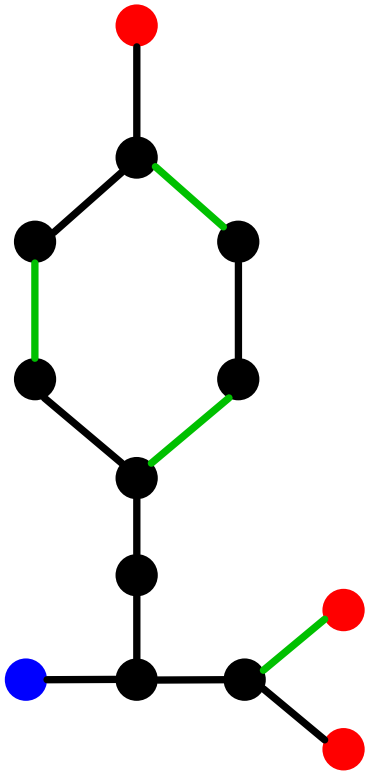
graphs are only about connectivity

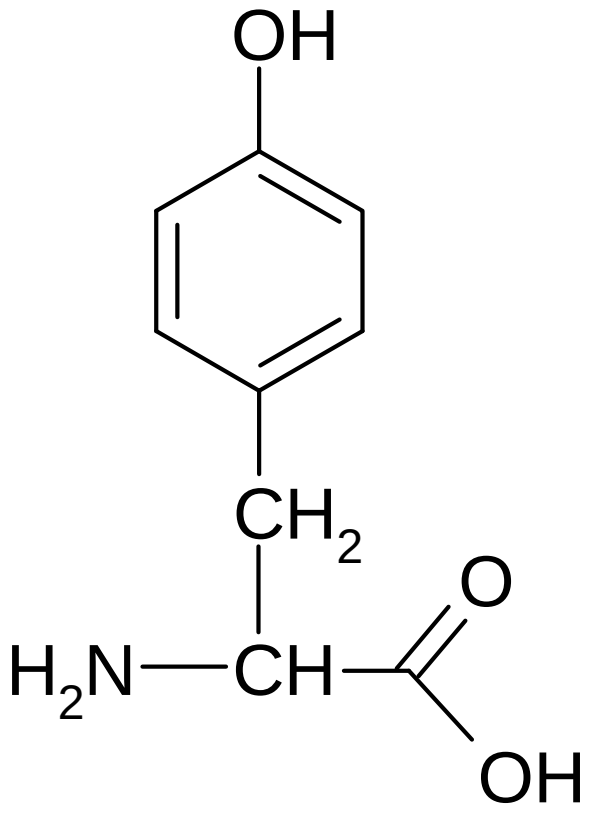
- spatial position of nodes is irrelevant
- length of edges are irrelevant
- crossing edges are irrelevant



Properties of Graphs

nodes and edges can be “coloured” to distinguish them





Structure Diagrams as Graphs

2D structure diagrams very like topological graphs

- atoms ↔ nodes
- bonds ↔ edges

terminal hydrogen atoms are not normally shown as separate nodes (“implicit” hydrogens)

- reduces number of nodes by ~50%
- “hydrogen count” information used to colour neighbouring “heavy atom” atom
- separate nodes sometimes used for “special” hydrogens

Advantages of using graphs

mathematical theory is well understood

graphs can be easily represented in computers

- many useful algorithms are known

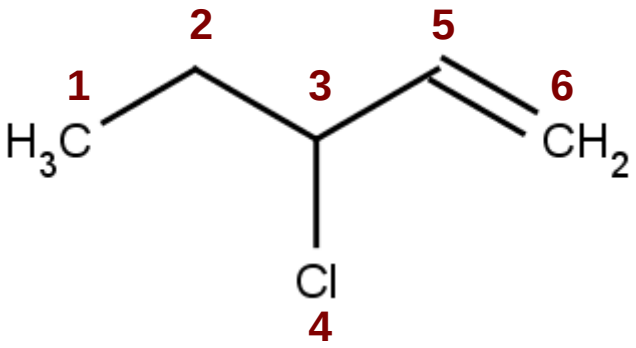
identical graphs $\Leftrightarrow$ identical molecules

different graphs $\Leftrightarrow$ different molecules

Matrix representations

A molecular structure with *n* atoms may be represented by an n × n matrix (H-atoms are often omitted).

Adjacency matrix : indicates which atoms are bonded.

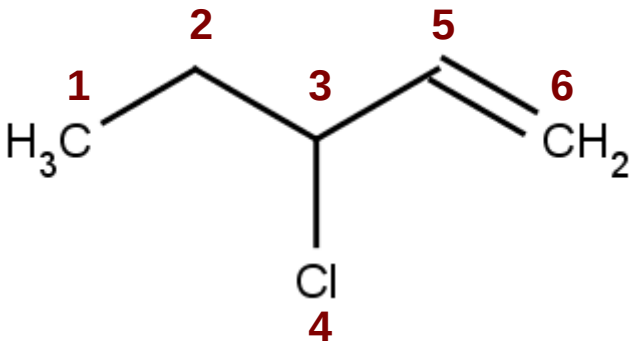


	1	2	3	4	5	6
1	0	1	0	0	0	0
2	1	0	1	0	0	0
3	0	1	0	1	1	0
4	0	0	1	0	0	0
5	0	0	1	0	0	1
6	0	0	0	0	1	0

Matrix representations

A molecular structure with *n* atoms may be represented by an n × n matrix (H-atoms are often omitted).

Adjacency matrix : indicates which atoms are bonded.

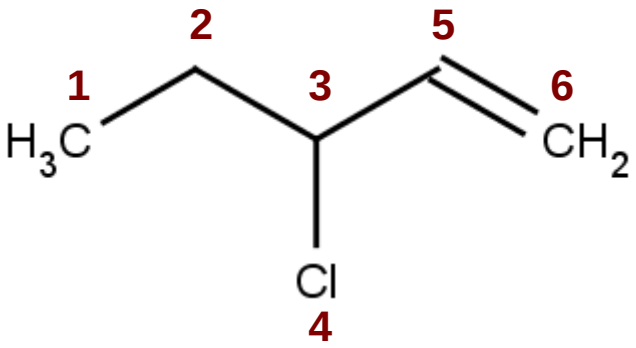


	1	2	3	4	5	6
1		1				
2	1		1			
3		1		1	1	
4			1			
5			1			1
6					1	

Matrix representations

A molecular structure with *n* atoms may be represented by an n × n matrix (H-atoms are often omitted).

Adjacency matrix : indicates which atoms are bonded.

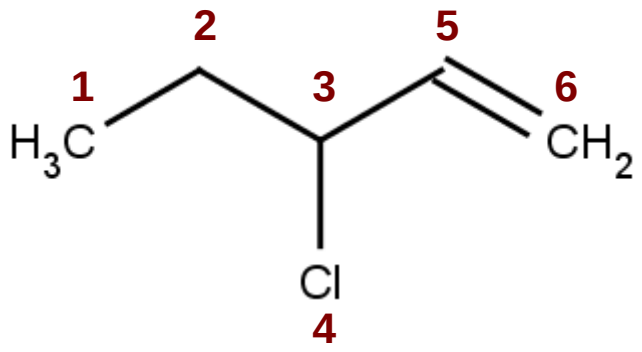


	1	2	3	4	5	6
1		1				
2			1			
3				1	1	
4						
5						1
6						

Matrix representations

Distance matrix : encodes the distances between atoms.

The distance is defined as the number of bonds between atoms on the shortest possible path.

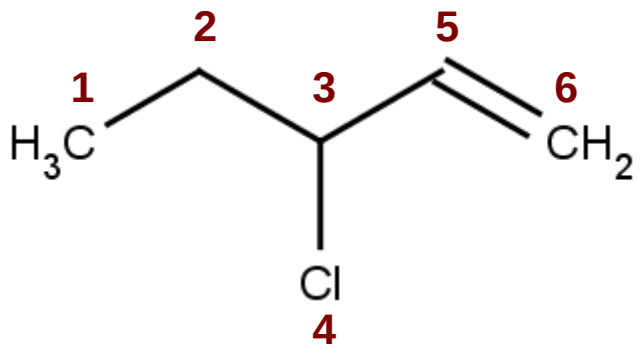


	1	2	3	4	5	6
1	0	1	2	3	3	4
2	1	0	1	2	2	3
3	2	1	0	1	1	2
4	3	2	1	0	2	3
5	3	2	1	2	0	1
6	4	3	2	3	1	0

Distance may also be defined as the 3D distance between atoms.

Matrix representations

Bond matrix : indicates which atoms are bonded, and the corresponding bond orders.

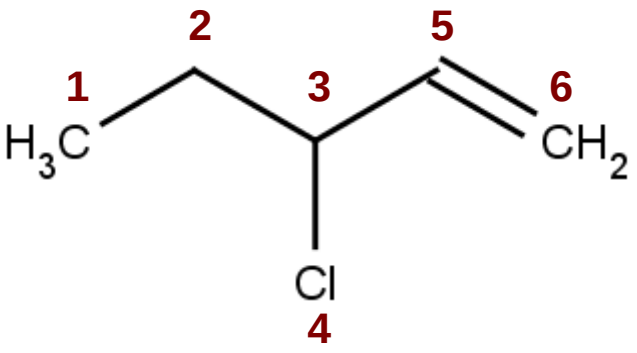


	1	2	3	4	5	6
1	0	1	0	0	0	0
2	1	0	1	0	0	0
3	0	1	0	1	1	0
4	0	0	1	0	0	0
5	0	0	1	0	0	**2**
6	0	0	0	0	**2**	0

Connection table

A disadvantage of matrix representations is that the matrix size increases with the square of the number of atoms.

A **connection table** lists the atoms of a molecule, and the bonds between them (may include or not H-atoms).



List of atoms

1	C
2	C
3	C
4	Cl
5	C
6	C

List of bonds

1st	2nd	order
1	2	1
2	3	1
3	4	1
3	5	1
5	6	2

The MDL Molfile format

(http://www.mdli.com/downloads/public/ctfile/ctfile.jsp)

Nr of bonds

Description of an atom

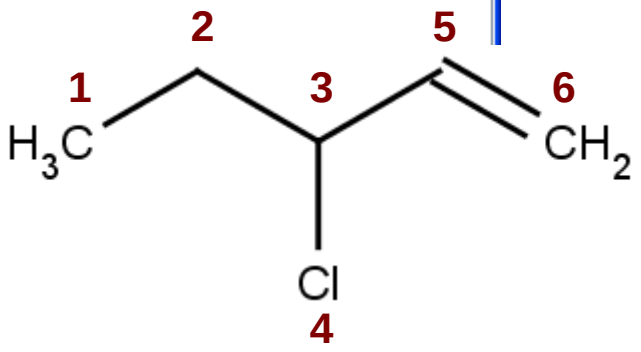
Nr of atoms

Description of a bond

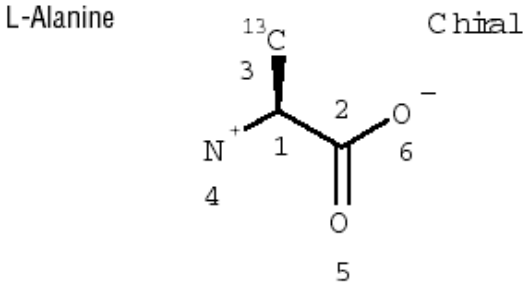
The Source - MDL Molfile or Rxnfile

File Edit Format

```
  Marvin  02130709022D

  6  5  0  0  0  0            999 V2000
    1.4289    0.0000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.7145   -0.4125    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.0000    0.0000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.0000    0.8250    0.0000 Cl  0  0  0  0  0  0  0  0  0  0  0  0
   -0.7145   -0.4125    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -1.4289   -0.0000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  1  2  1  0  0  0  0
  2  3  1  0  0  0  0
  3  4  1  0  0  0  0
  3  5  1  0  0  0  0
  5  6  2  0  0  0  0
M  END
```



The MDL Molfile format




```
L-Alanine (13C)
GSMACCS-II10169115362D 1   0.00366     0.00000     0
  6  5  0  0  1  0               3 V2000
   -0.6622    0.5342    0.0000 C   0  0  2  0  0  0
    0.6622   -0.3000    0.0000 C   0  0  0  0  0  0
   -0.7207    2.0817    0.0000 C   1  0  0  0  0  0
   -1.8622   -0.3695    0.0000 N   0  3  0  0  0  0
    0.6220   -1.8037    0.0000 O   0  0  0  0  0  0
    1.9464    0.4244    0.0000 O   0  5  0  0  0  0
  1  2  1  0  0  0
  1  3  1  1  0  0
  1  4  1  0  0  0
  2  5  2  0  0  0
  2  6  1  0  0  0
M  CHG  2   4   1   6  -1
M  ISO  1   3  13
M  END
```

Header Block

Counts Line

Atom Block

Bond Block

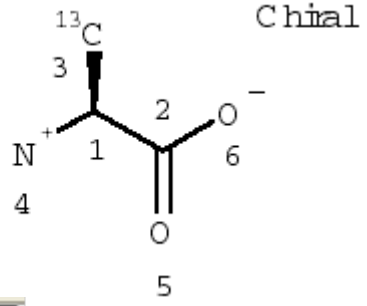
Properties Block

Connection Table (Ctab)

The atom block

```
xxxxx.xxxxyyyyy.yyyyzzzzz.zzzz aaaddcccssshhhbbbvvvHHHrrriiimmmnnneee
-0.6622    0.5342    0.0000 C   0  0  2  0  0  0
 0.6622   -0.3000    0.0000 C   0  0  0  0  0  0
-0.7207    2.0817    0.0000 C   1  0  0  0  0  0
-1.8622   -0.3695    0.0000 N   0  3  0  0  0  0
 0.6220   -1.8037    0.0000 O   0  0  0  0  0  0
 1.9464    0.4244    0.0000 O   0  5  0  0  0  0
```

L-Alanine



Chiral

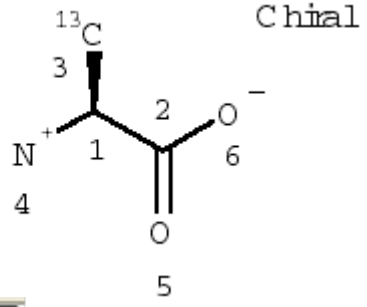
Field	Meaning	Values	Notes
x y z	atom coordinates		**[Generic]**
aaa	atom symbol	entry in periodic table or L for atom list, A, Q, * for unspecified atom, and LP for lone pair, or R# for Rgroup label	**[Generic, Query, 3D, Rgroup]**
dd	mass difference	-3, -2, -1, 0, 1, 2, 3, 4 (0 if value beyond these limits)	**[Generic]** Difference from mass in periodic table. Wider range of values allowed by `M ISO` line, below. Retained for compatibility with older Ctabs, `M ISO` takes precedence.
ccc	charge	0 = uncharged or value other than these, 1 = +3, 2 = +2, 3 = +1, 4 = doublet radical, 5 = -1, 6 = -2, 7 = -3	**[Generic]** Wider range of values in `M CHG` and `M RAD` lines below. Retained for compatibility with older Ctabs, `M CHG` and `M RAD` lines take precedence.
sss	atom stereo parity	0 = not stereo, 1 = odd, 2 = even, 3 = either or unmarked stereo center	**[Generic]** Ignored when read.

The atom block

```
xxxxx.xxxxyyyyy.yyyyzzzzz.zzzz aaaddcccssshhhbbbvvvHHHrrriiimmmnnneee
```

```
-0.6622    0.5342    0.0000 C   0  0  2  0  0  0
 0.6622   -0.3000    0.0000 C   0  0  0  0  0  0
-0.7207    2.0817    0.0000 C   1  0  0  0  0  0
-1.8622   -0.3695    0.0000 N   0  3  0  0  0  0
 0.6220   -1.8037    0.0000 O   0  0  0  0  0  0
 1.9464    0.4244    0.0000 O   0  5  0  0  0  0
```

L-Alanine



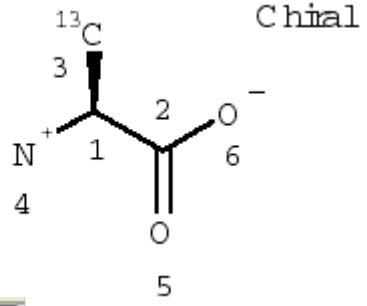
Chiral

Field	Meaning	Values	Notes
x y z	atom coordinates		[Generic]
aaa	atom symbol	entry in periodic table or L for atom list, A, Q, * for unspecified atom, and LP for lone pair, or R# for Rgroup label	[Generic, Query, 3D, Rgroup]
dd	mass difference	-3, -2, -1, 0, 1, 2, 3, 4 (0 if value beyond these limits)	[Generic] Difference from mass in periodic table. Wider range of values allowed by M ISO line, below. Retained for compatibility with older Ctabs, M ISO takes precedence.
ccc	charge	0 = uncharged or value other than these, 1 = +3, 2 = +2, 3 = +1, 4 = doublet radical, 5 = -1, 6 = -2, 7 = -3	[Generic] Wider range of values in M CHG and M RAD lines below. Retained for compatibility with older Ctabs, M CHG and M RAD lines take precedence.
sss	atom stereo parity	0 = not stereo, 1 = odd, 2 = even, 3 = either or unmarked stereo center	[Generic] Ignored when read.

The atom block

```
xxxxx.xxxxyyyyy.yyyyzzzzz.zzzz aaaddcccssshhhbbbvvvHHHrrriiimmmnnneee
-0.6622    0.5342    0.0000 C   0  0  2  0  0  0
 0.6622   -0.3000    0.0000 C   0  0  0  0  0  0
-0.7207    2.0817    0.0000 C   1  0  0  0  0  0
-1.8622   -0.3695    0.0000 N   0  3  0  0  0  0
 0.6220   -1.8037    0.0000 O   0  0  0  0  0  0
 1.9464    0.4244    0.0000 O   0  5  0  0  0  0
```

L-Alanine

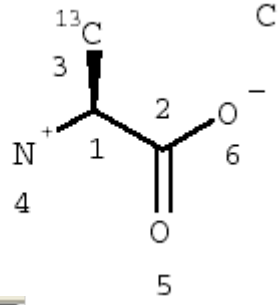


Field	Meaning	Values	Notes
x y z	atom coordinates		**[Generic]**
aaa	atom symbol	entry in periodic table or L for atom list, A, Q, * for unspecified atom, and LP for lone pair, or R# for Rgroup label	**[Generic, Query, 3D, Rgroup]**
dd	mass difference	-3, -2, -1, 0, 1, 2, 3, 4 (0 if value beyond these limits)	**[Generic]** Difference from mass in periodic table. Wider range of values allowed by `M  ISO` line, below. Retained for compatibility with older Ctabs, `M  ISO` takes precedence.
ccc	charge	0 = uncharged or value other than these, 1 = +3, 2 = +2, 3 = +1, 4 = doublet radical, 5 = -1, 6 = -2, 7 = -3	**[Generic]** Wider range of values in `M  CHG` and `M  RAD` lines below. Retained for compatibility with older Ctabs, `M  CHG` and `M  RAD` lines take precedence.
sss	atom stereo parity	0 = not stereo, 1 = odd, 2 = even, 3 = either or unmarked stereo center	**[Generic]** Ignored when read.

The atom block

```
xxxxx.xxxxyyyyy.yyyyzzzzz.zzzz aaaddcccssshhhbbbvvvHHHrrriiimmmnnneee
-0.6622     0.5342     0.0000 C   0  0  2  0  0  0
 0.6622    -0.3000     0.0000 C   0  0  0  0  0  0
-0.7207     2.0817     0.0000 C   1  0  0  0  0  0
-1.8622    -0.3695     0.0000 N   0  3  0  0  0  0
 0.6220    -1.8037     0.0000 O   0  0  0  0  0  0
 1.9464     0.4244     0.0000 O   0  5  0  0  0  0
```

L-Alanine



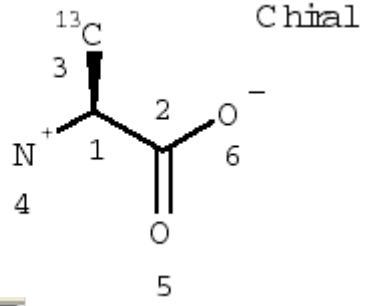
Chiral

Field	Meaning	Values	Notes
x y z	atom coordinates		**[Generic]**
aaa	atom symbol	entry in periodic table or L for atom list, A, Q, * for unspecified atom, and LP for lone pair, or R# for Rgroup label	**[Generic, Query, 3D, Rgroup]**
dd	mass difference	-3, -2, -1, 0, 1, 2, 3, 4 (0 if value beyond these limits)	**[Generic]** Difference from mass in periodic table. Wider range of values allowed by `M  ISO` line, below. Retained for compatibility with older Ctabs, `M  ISO` takes precedence.
ccc	charge	0 = uncharged or value other than these, 1 = +3, 2 = +2, 3 = +1, 4 = doublet radical, 5 = -1, 6 = -2, 7 = -3	**[Generic]** Wider range of values in `M  CHG` and `M  RAD` lines below. Retained for compatibility with older Ctabs, `M  CHG` and `M  RAD` lines take precedence.
sss	atom stereo parity	0 = not stereo, 1 = odd, 2 = even, 3 = either or unmarked stereo center	**[Generic]** Ignored when read.

The atom block

```
xxxxx.xxxxyyyyy.yyyyzzzzz.zzzz aaaddcccssshhhbbbvvvHHHrrriiimmmnnneee
-0.6622    0.5342    0.0000 C   0  0  2  0  0  0
 0.6622   -0.3000    0.0000 C   0  0  0  0  0  0
-0.7207    2.0817    0.0000 C   1  0  0  0  0  0
-1.8622   -0.3695    0.0000 N   0  3  0  0  0  0
 0.6220   -1.8037    0.0000 O   0  0  0  0  0  0
 1.9464    0.4244    0.0000 O   0  5  0  0  0  0
```

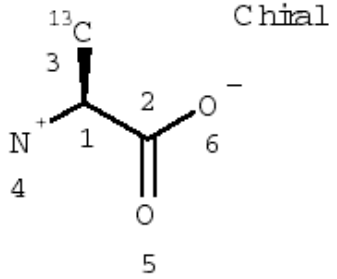
L-Alanine



Field	Meaning	Values	Notes
x y z	atom coordinates		**[Generic]**
aaa	atom symbol	entry in periodic table or L for atom list, A, Q, * for unspecified atom, and LP for lone pair, or R# for Rgroup label	**[Generic, Query, 3D, Rgroup]**
dd	mass difference	-3, -2, -1, 0, 1, 2, 3, 4 (0 if value beyond these limits)	**[Generic]** Difference from mass in periodic table. Wider range of values allowed by `M ISO` line, below. Retained for compatibility with older Ctabs, `M ISO` takes precedence.
ccc	charge	0 = uncharged or value other than these, 1 = +3, 2 = +2, 3 = +1, 4 = doublet radical, 5 = -1, 6 = -2, 7 = -3	**[Generic]** Wider range of values in `M CHG` and `M RAD` lines below. Retained for compatibility with older Ctabs, `M CHG` and `M RAD` lines take precedence.
sss	atom stereo parity	0 = not stereo, 1 = odd, 2 = even, 3 = either or unmarked stereo center	**[Generic]** Ignored when read.

The MDL Molfile format

L-Alanine



Chiral

```
L-Alanine (13C)
GSMACCS-II10169115362D 1   0.00366     0.00000     0
  6  5  0  0  1  0              3 V2000
   -0.6622    0.5342    0.0000 C   0  0  2  0  0  0
    0.6622   -0.3000    0.0000 C   0  0  0  0  0  0
   -0.7207    2.0817    0.0000 C   1  0  0  0  0  0
   -1.8622   -0.3695    0.0000 N   0  3  0  0  0  0
    0.6220   -1.8037    0.0000 O   0  0  0  0  0  0
    1.9464    0.4244    0.0000 O   0  5  0  0  0  0
  1  2  1  0  0  0
  1  3  1  1  0  0
  1  4  1  0  0  0
  2  5  2  0  0  0
  2  6  1  0  0  0
M  CHG  2   4   1   6  -1
M  ISO  1   3  13
M  END
```

Header Block

Counts Line

Atom Block

Bond Block

Properties Block

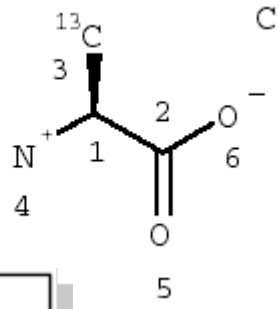
Connection Table (Ctab)

The bond block

```
111222tttsssxxxrrrccc
```

```
  1   2   1   0   0   0
  1   3   1   1   0   0
  1   4   1   0   0   0
  2   5   2   0   0   0
  2   6   1   0   0   0
```

L-Alanine



Chiral

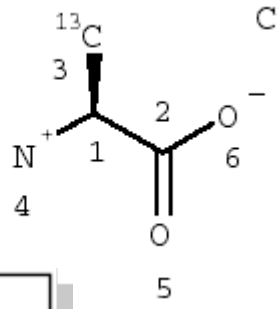
Field	Meaning	Values	Notes
111	first atom number	1 - number of atoms	**[Generic]**
222	second atom number	1 - number of atoms	**[Generic]**
ttt	bond type	1 = Single, 2 = Double, 3 = Triple, 4 = Aromatic, 5 = Single or Double, 6 = Single or Aromatic, 7 = Double or Aromatic, 8 = Any	**[Query]** Values 4 through 8 are for SSS queries only.
sss	bond stereo	Single bonds: 0 = not stereo, 1 = Up, 4 = Either, 6 = Down, Double bonds: 0 = Use x-, y-, z-coords from atom block to determine cis or trans, 3 = Cis or trans (either) double bond	**[Generic]** The wedge (pointed) end of the stereo bond is at the first atom (Field 111 above)
xxx	not used		
rrr	bond topology	0 = Either, 1 = Ring, 2 = Chain	**[Query]** SSS queries only.
ccc	reacting center status	0 = unmarked, 1 = a center, -1 = not a center, Additional: 2 = no change, 4 = bond made/broken, 8 = bond order changes 12 = 4+8 (both made/broken and changes); 5 = (4 + 1), 9 = (8 + 1), and 13 = (12 + 1) are also possible	**[Reaction, Query]**

The bond block

111222tttsssxxxrrrccc

```
  1  2  1  0  0  0
  1  3  1  1  0  0
  1  4  1  0  0  0
  2  5  2  0  0  0
  2  6  1  0  0  0
```

L-Alanine



Chiral

Field	Meaning	Values	Notes
111	first atom number	1 - number of atoms	**[Generic]**
222	second atom number	1 - number of atoms	**[Generic]**
ttt	bond type	1 = Single, 2 = Double, 3 = Triple, 4 = Aromatic, 5 = Single or Double, 6 = Single or Aromatic, 7 = Double or Aromatic, 8 = Any	**[Query]** Values 4 through 8 are for SSS queries only.
sss	bond stereo	Single bonds: 0 = not stereo, 1 = Up, 4 = Either, 6 = Down, Double bonds: 0 = Use x-, y-, z-coords from atom block to determine cis or trans, 3 = Cis or trans (either) double bond	**[Generic]** The wedge (pointed) end of the stereo bond is at the first atom (Field 111 above)
xxx	not used		
rrr	bond topology	0 = Either, 1 = Ring, 2 = Chain	**[Query]** SSS queries only.
ccc	reacting center status	0 = unmarked, 1 = a center, -1 = not a center, Additional: 2 = no change, 4 = bond made/broken, 8 = bond order changes 12 = 4+8 (both made/broken and changes); 5 = (4 + 1), 9 = (8 + 1), and 13 = (12 + 1) are also possible	**[Reaction, Query]**

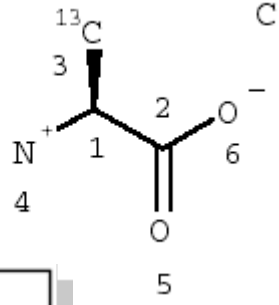
The bond block

111222tttsssxxxrrrccc

```
1  2  1  0  0  0
1  3  1  1  0  0
1  4  1  0  0  0
2  5  2  0  0  0
2  6  1  0  0  0
```

L-Alanine

Chiral



Field	Meaning	Values	Notes
111	first atom number	1 - number of atoms	**[Generic]**
222	second atom number	1 - number of atoms	**[Generic]**
ttt	bond type	1 = Single, 2 = Double, 3 = Triple, 4 = Aromatic, 5 = Single or Double, 6 = Single or Aromatic, 7 = Double or Aromatic, 8 = Any	**[Query]** Values 4 through 8 are for SSS queries only.
sss	bond stereo	Single bonds: 0 = not stereo, 1 = Up, 4 = Either, 6 = Down, Double bonds: 0 = Use x-, y-, z-coords from atom block to determine cis or trans, 3 = Cis or trans (either) double bond	**[Generic]** The wedge (pointed) end of the stereo bond is at the first atom (Field 111 above)
xxx	not used		
rrr	bond topology	0 = Either, 1 = Ring, 2 = Chain	**[Query]** SSS queries only.
ccc	reacting center status	0 = unmarked, 1 = a center, -1 = not a center, Additional: 2 = no change, 4 = bond made/broken, 8 = bond order changes 12 = 4+8 (both made/broken and changes); 5 = (4 + 1), 9 = (8 + 1), and 13 = (12 + 1) are also possible	**[Reaction, Query]**

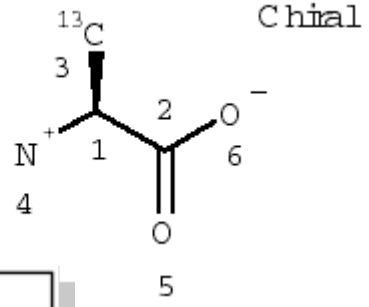
The bond block

111222tttsssxxxrrrccc

```
  1   2   1   0   0   0
  1   3   1   1   0   0
  1   4   1   0   0   0
  2   5   2   0   0   0
  2   6   1   0   0   0
```

L-Alanine

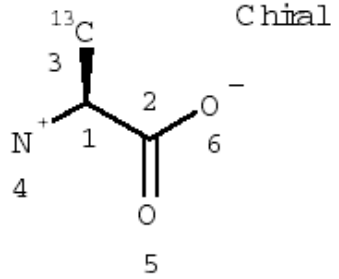
Chiral



Field	Meaning	Values	Notes
111	first atom number	1 - number of atoms	**[Generic]**
222	second atom number	1 - number of atoms	**[Generic]**
ttt	bond type	1 = Single, 2 = Double, 3 = Triple, 4 = Aromatic, 5 = Single or Double, 6 = Single or Aromatic, 7 = Double or Aromatic, 8 = Any	**[Query]** Values 4 through 8 are for SSS queries only.
sss	bond stereo	Single bonds: 0 = not stereo, 1 = Up, 4 = Either, 6 = Down, Double bonds: 0 = Use x-, y-, z-coords from atom block to determine cis or trans, 3 = Cis or trans (either) double bond	**[Generic]** The wedge (pointed) end of the stereo bond is at the first atom (Field 111 above)
xxx	not used		
rrr	bond topology	0 = Either, 1 = Ring, 2 = Chain	**[Query]** SSS queries only.
ccc	reacting center status	0 = unmarked, 1 = a center, -1 = not a center, Additional: 2 = no change, 4 = bond made/broken, 8 = bond order changes 12 = 4+8 (both made/broken and changes); 5 = (4 + 1), 9 = (8 + 1), and 13 = (12 + 1) are also possible	**[Reaction, Query]**

The MDL Molfile format

L-Alanine



Chiral

```
L-Alanine (13C)
GSMACCS-II10169115362D 1   0.00366     0.00000     0

  6  5  0  0  1  0              3 V2000
   -0.6622    0.5342    0.0000 C   0  0  2  0  0  0
    0.6622   -0.3000    0.0000 C   0  0  0  0  0  0
   -0.7207    2.0817    0.0000 C   1  0  0  0  0  0
   -1.8622   -0.3695    0.0000 N   0  3  0  0  0  0
    0.6220   -1.8037    0.0000 O   0  0  0  0  0  0
    1.9464    0.4244    0.0000 O   0  5  0  0  0  0
  1  2  1  0  0  0
  1  3  1  1  0  0
  1  4  1  0  0  0
  2  5  2  0  0  0
  2  6  1  0  0  0
M  CHG  2   4   1   6  -1
M  ISO  1   3  13
M  END
```

Header Block

Counts Line

Atom Block

Bond Block

Properties Block

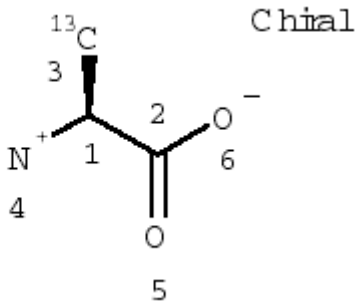
Connection Table (Ctab)

The properties block

```
M  CHG  2   4   1   6  -1
M  ISO  1   3  13
M  END
```

L-Alanine

Chiral



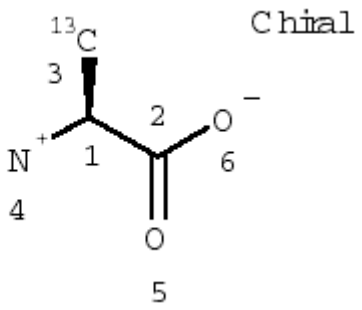
2 charged atoms

The properties block

```
M  CHG  2   4   1   6  -1
M  ISO  1   3  13
M  END
```

L-Alanine

Chiral



2 charged atoms

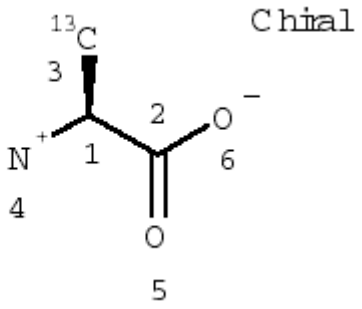
atom 4: charge +1
atom 6: charge -1

The properties block

```
M  CHG  2   4   1   6  -1
M  ISO  1   3  13
M  END
```

L-Alanine

Chiral

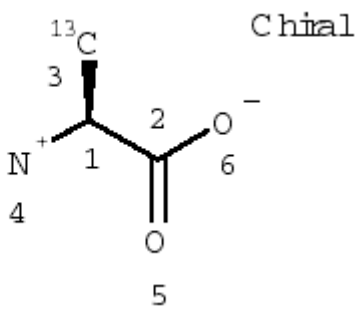


1 entry for an isotope

The properties block

```
M  CHG  2   4   1   6   -1
M  ISO  1   3   13
M  END
```

L-Alanine



Chiral

1 entry for an isotope

atom 3: mass=13

The SDFile (.SDF) format

Includes structural information in the Molfile format
and associated data items for one **or more** compounds.

Molfile1
Associated data
$$$$
Molfile2
Associated data
$$$$
…

The SDFile (.SDF) format

Molfile1
Associated data
$$$$
Molfile2
Associated data
$$$$
…

Example

```
61203-01-8
  Marvin  02130710303D

 11 10  0  0  0  0            999 V2000
    1.6947   -0.2675   -0.0016 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.5343    0.6242    0.0001 C   0  0  0  0  0  0  0  0  0  0  0  0
    3.0873    0.3083   -0.0014 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.5355   -1.4702    0.0019 O   0  0  0  0  0  0  0  0  0  0  0  0
    0.7095    1.9478    0.0018 C   0  0  0  0  0  0  0  0  0  0  0  0
   -1.2132   -0.0983   -0.0002 Br  0  0  0  0  0  0  0  0  0  0  0  0
    3.5067   -0.7060    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
    4.1016    0.7277    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
    2.6679    1.3226    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
    1.7238    2.3671    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
   -0.1608    2.6166    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
  1  2  1  0  0  0  0
  1  3  1  0  0  0  0
  1  4  2  0  0  0  0
  2  5  2  0  0  0  0
  2  6  1  0  0  0  0
  3  7  1  0  0  0  0
  3  8  1  0  0  0  0
  3  9  1  0  0  0  0
  5 10  1  0  0  0  0
  5 11  1  0  0  0  0
M  END
>  <Ames test categorisation>
mutagen

>  <CHARGE>
0,11;0,05;-0,00;-0,29;-0,02;-0,05;0,03;0,03;0,03;0,06;0,06

>  <EXACTMASS>
147,952377

$$$$
598-55-0
  Marvin  02130710303D

 10  9  0  0  0  0            999 V2000
    0.4485   -0.0024   -0.0013 C   0  0  0  0  0  0  0  0  0  0  0  0
```

Associated data (molecular)

The SDFile (.SDF) format

Molfile1
Associated data
$$$$
Molfile2
Associated data
$$$$
…

Example

```
61203-01-8
  Marvin  02130710303D

 11 10  0  0  0  0            999 V2000
    1.6947   -0.2675   -0.0016 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.5343    0.6242    0.0001 C   0  0  0  0  0  0  0  0  0  0  0  0
    3.0873    0.3083   -0.0014 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.5355   -1.4702    0.0019 O   0  0  0  0  0  0  0  0  0  0  0  0
    0.7095    1.9478    0.0018 C   0  0  0  0  0  0  0  0  0  0  0  0
   -1.2132   -0.0983   -0.0002 Br  0  0  0  0  0  0  0  0  0  0  0  0
    3.5067   -0.7060    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
    4.1016    0.7277    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
    2.6679    1.3226    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
    1.7238    2.3671    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
   -0.1608    2.6166    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
  1  2  1  0  0  0  0
  1  3  1  0  0  0  0
  1  4  2  0  0  0  0
  2  5  2  0  0  0  0
  2  6  1  0  0  0  0
  3  7  1  0  0  0  0
  3  8  1  0  0  0  0
  3  9  1  0  0  0  0
  5 10  1  0  0  0  0
  5 11  1  0  0  0  0
M  END
>  <Ames test categorisation>
mutagen
```

Associated data (atomic) ←

```
>  <CHARGE>
0,11;0,05;-0,00;-0,29;-0,02;-0,05;0,03;0,03;0,03;0,06;0,06

>  <EXACTMASS>
147,952377

$$$$
598-55-0
  Marvin  02130710303D

 10  9  0  0  0  0            999 V2000
    0.4485   -0.0024   -0.0013 C   0  0  0  0  0  0  0  0  0  0  0  0
```

The SDFile (.SDF) format

Molfile1
Associated data
$$$$
Molfile2
Associated data
$$$$
…

Example

```
61203-01-8
  Marvin  02130710303D

 11 10  0  0  0  0            999 V2000
    1.6947   -0.2675   -0.0016 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.5343    0.6242    0.0001 C   0  0  0  0  0  0  0  0  0  0  0  0
    3.0873    0.3083   -0.0014 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.5355   -1.4702    0.0019 O   0  0  0  0  0  0  0  0  0  0  0  0
    0.7095    1.9478    0.0018 C   0  0  0  0  0  0  0  0  0  0  0  0
   -1.2132   -0.0983   -0.0002 Br  0  0  0  0  0  0  0  0  0  0  0  0
    3.5067   -0.7060    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
    4.1016    0.7277    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
    2.6679    1.3226    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
    1.7238    2.3671    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
   -0.1608    2.6166    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
  1  2  1  0  0  0  0
  1  3  1  0  0  0  0
  1  4  2  0  0  0  0
  2  5  2  0  0  0  0
  2  6  1  0  0  0  0
  3  7  1  0  0  0  0
  3  8  1  0  0  0  0
  3  9  1  0  0  0  0
  5 10  1  0  0  0  0
  5 11  1  0  0  0  0
M  END
>  <Ames test categorisation>
mutagen

>  <CHARGE>
0,11;0,05;-0,00;-0,29;-0,02;-0,05;0,03;0,03;0,03;0,06;0,06

>  <EXACTMASS>
147,952377

$$$$
598-55-0
  Marvin  02130710303D

 10  9  0  0  0  0            999 V2000
    0.4485   -0.0024   -0.0013 C   0  0  0  0  0  0  0  0  0  0  0  0
```

Associated data (molecular)

The SDFile (.SDF) format

Molfile1
Associated data
$$$$
Molfile2
Associated data
$$$$
…

Example

```
61203-01-8
  Marvin  02130710303D

 11 10  0  0  0  0            999 V2000
    1.6947   -0.2675   -0.0016 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.5343    0.6242    0.0001 C   0  0  0  0  0  0  0  0  0  0  0  0
    3.0873    0.3083   -0.0014 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.5355   -1.4702    0.0019 O   0  0  0  0  0  0  0  0  0  0  0  0
    0.7095    1.9478    0.0018 C   0  0  0  0  0  0  0  0  0  0  0  0
   -1.2132   -0.0983   -0.0002 Br  0  0  0  0  0  0  0  0  0  0  0  0
    3.5067   -0.7060    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
    4.1016    0.7277    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
    2.6679    1.3226    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
    1.7238    2.3671    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
   -0.1608    2.6166    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
  1  2  1  0  0  0  0
  1  3  1  0  0  0  0
  1  4  2  0  0  0  0
  2  5  2  0  0  0  0
  2  6  1  0  0  0  0
  3  7  1  0  0  0  0
  3  8  1  0  0  0  0
  3  9  1  0  0  0  0
  5 10  1  0  0  0  0
  5 11  1  0  0  0  0
M  END
>  <Ames test categorisation>
mutagen

>  <CHARGE>
0,11;0,05;-0,00;-0,29;-0,02;-0,05;0,03;0,03;0,03;0,06;0,06

>  <EXACTMASS>
147,952377

$$$$
598-55-0
  Marvin  02130710303D

 10  9  0  0  0  0            999 V2000
    0.4485   -0.0024   -0.0013 C   0  0  0  0  0  0  0  0  0  0  0  0
```

Delimiter

Beginning of Molfile2

The SDFile (.SDF) format

Molfile1
Associated data
$$$$
Molfile2
Associated data
$$$$
…

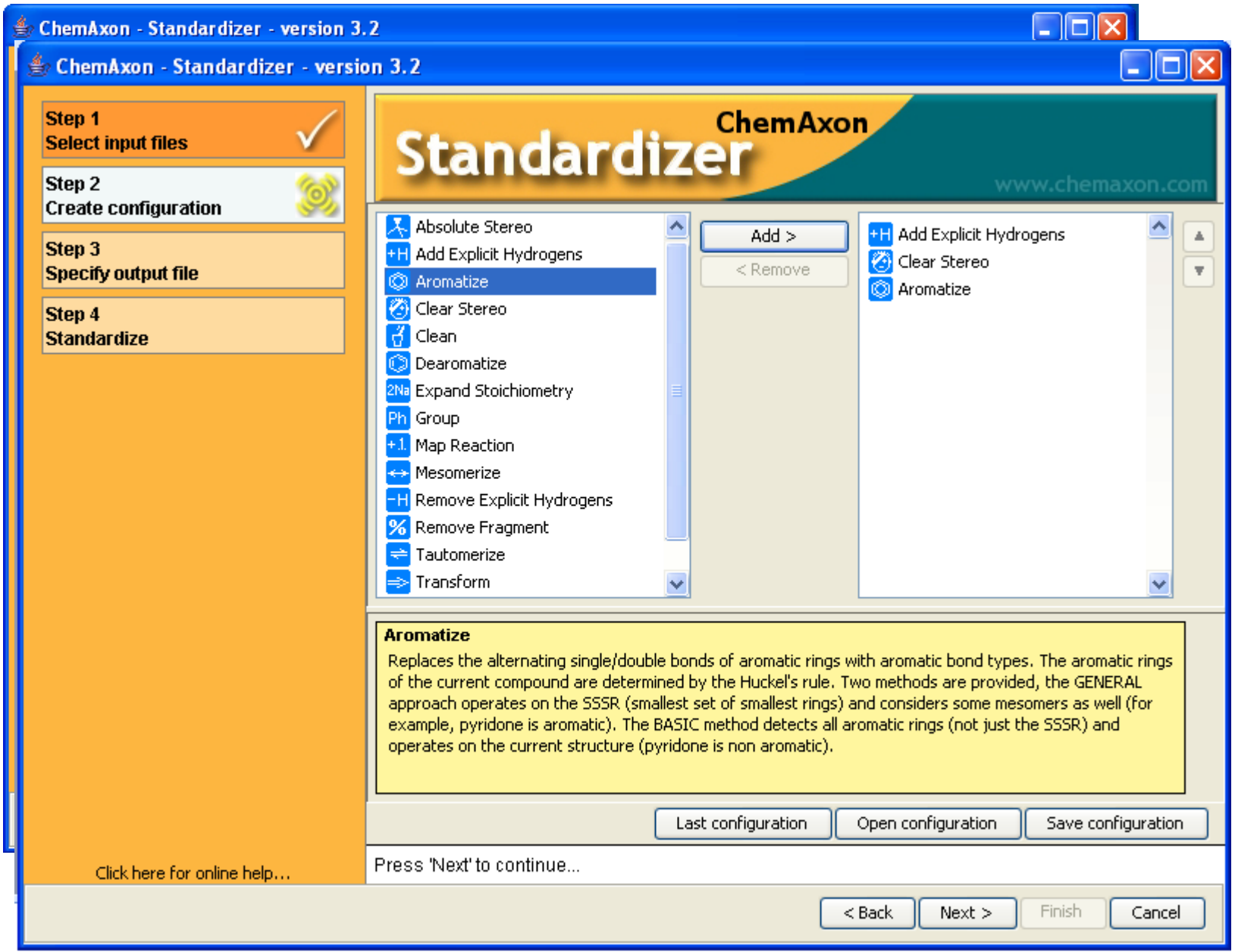
Example

```
,11;0,05;-0,00;-0,29;-0,02;-0,05;0,03;0,03;0,03;0,06;0,06

>  <EXACTMASS>
147,952377

$$$$
598-55-0
  Marvin  02130710303D

 10  9  0  0  0  0            999 V2000
    0.4485   -0.0024   -0.0013 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.7168    0.6719    0.0009 O   0  0  0  0  0  0  0  0  0  0  0  0
    1.6164    0.6704    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
    0.4473   -1.2175    0.0005 O   0  0  0  0  0  0  0  0  0  0  0  0
   -1.9750   -0.0528   -0.0006 C   0  0  0  0  0  0  0  0  0  0  0  0
    2.4909    0.1644    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
    1.6174    1.6807    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
   -2.4793    0.8227    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
   -2.8505   -0.5571    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
   -1.4707   -0.9283    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
  1  2  1  0  0  0  0
  1  3  1  0  0  0  0
  1  4  2  0  0  0  0
  2  5  1  0  0  0  0
  3  6  1  0  0  0  0
  3  7  1  0  0  0  0
  5  8  1  0  0  0  0
  5  9  1  0  0  0  0
  5 10  1  0  0  0  0
M  END
>  <Ames test categorisation>
nonmutagen

>  <CHARGE>
0,04;-0,20;-0,05;-0,26;0,05;0,13;0,13;0,05;0,05;0,05

>  <EXACTMASS>
75,032028

$$$$
```

The ChemAxon Standardize program

- Conversion of file formats
- Generation of unique SMILES strings
- Standardization of structures
- Addition of H-atoms, removal of H-atoms, assignment of aromatic systems, cleaning of stereochemistry, …

The ChemAxon Standardize program

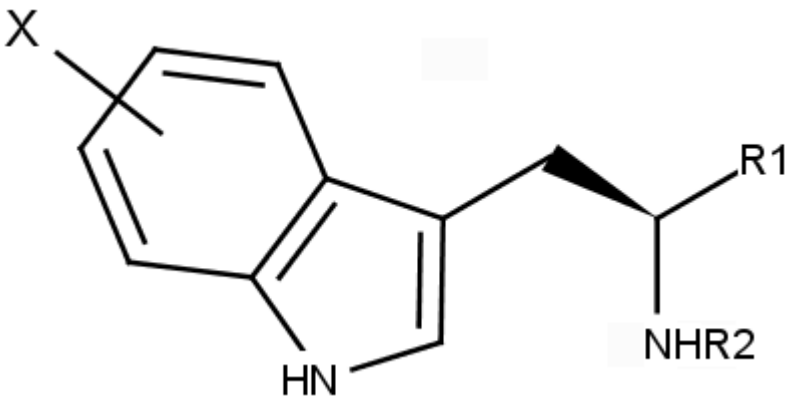


Markush structures

A Markush structures diagram is a type of representation specific for a SERIES of chemical compounds.

The diagram can describe not only a specific molecule, but several families of compounds.

It includes a core and substituents, which are listed as text separately from the diagram.



R1= H, halogen, OH, COOH
R2= H, CH_3
X= Cl, Br, CH3

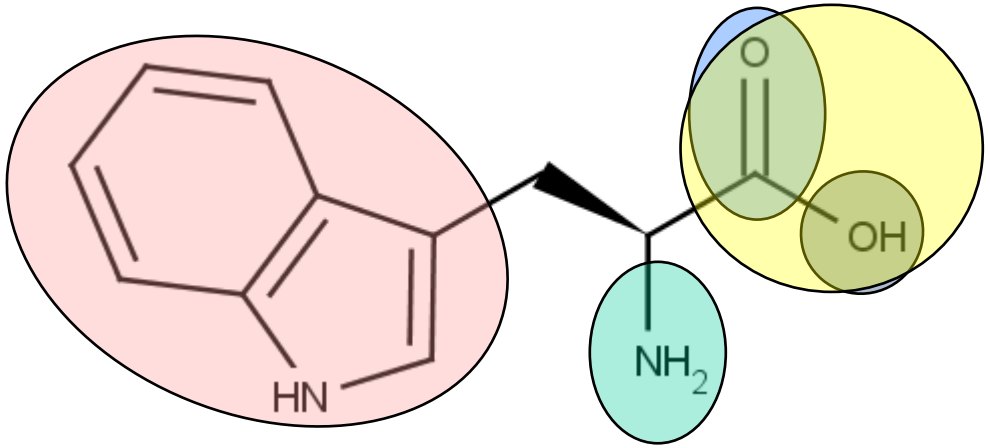
These are mostly used in databases of patents.

Representation of molecular fragments

Just like a text document may be indexed on the basis of specified keywords, a chemical structure may be indexed on the basis of specific chemical characteristics, usually fragments.

Fragments may be, e.g., small groups of atoms, functional groups, rings. These are defined beforehand.

It is an ambiguous representation: different structures may have common fragments.

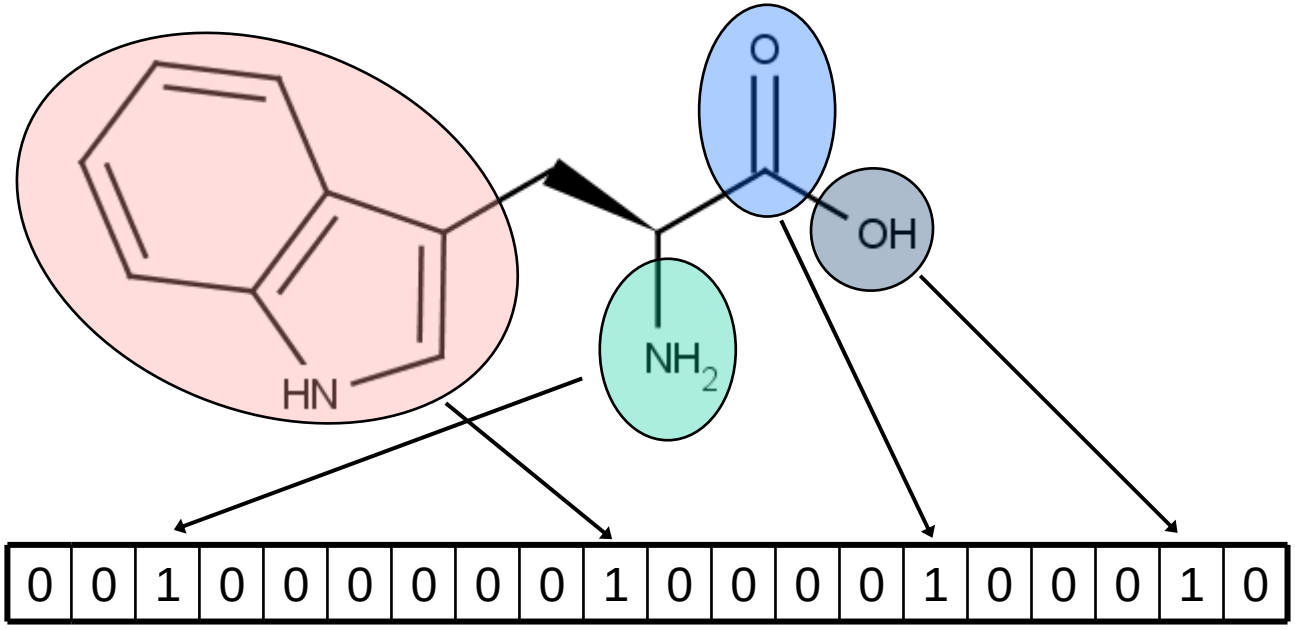


Fragments:

- -OH
- -COOH
- >C=O
- -NH2
- -3-indole

Fingerprints

Fingerprints encode the presence or absence of certain features in a compound, e.g., fragments.



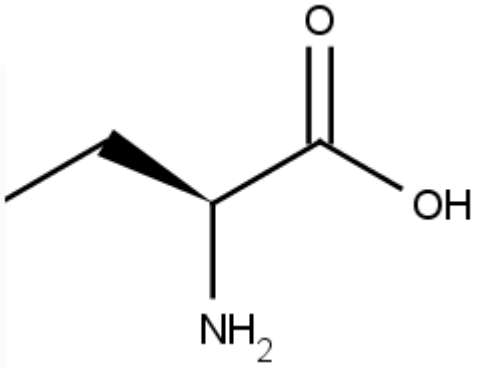
If 20 fragments are defined, the fingerprint has a length of 20.

It is an ambiguous representation.

Allows for similarity searches.

'Hashed Fingerprints'

Encode the presence of sub-structures. **These are not previously defined**.



All patterns are listed consisting of

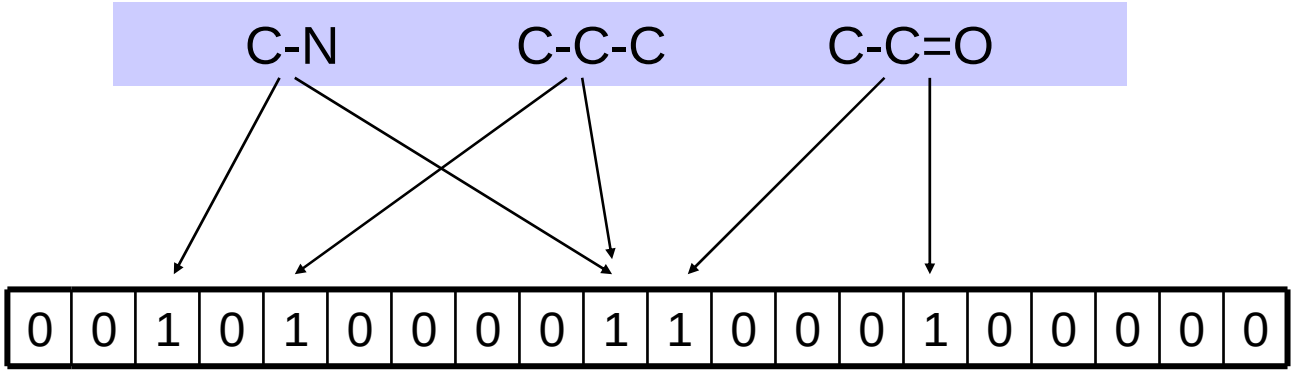
- 1 atom
- 2 bonded atoms and their bond
- Sequences of 3 atoms and their bonds
- Sequences of 4 atoms and their bonds
- …

Patterns up to 3 atoms

- C, N, O
- C-C, C-N, C=O, C-O
- C-C-C, C-C-N, C-C=O, C-C-O, O=C-O

'Hashed Fingerprints'

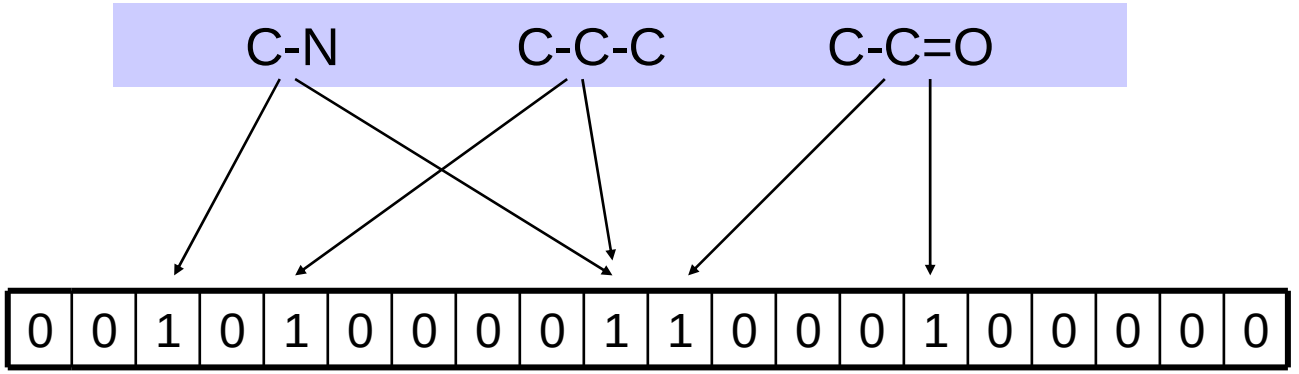
Each pattern activates a certain number of positions (bits) in the fingerprint, in the following example two bits / pattern:



An algorithm determines which bits are activated by a pattern. The same pattern always activates the same bits. The algorithm is designed in such a way that it is always possible to assign bits to a pattern.

There may be collisions. Pre-definition of fragments is not required. But it is not possible to interpret fingerprints.

'Hashed Fingerprints'



H-atoms are omitted. Stereochemistry is not considered.

Parameters to define: fingerprint length, size of patterns, and number of bits activated by each pattern.

Main application: similarity search in large databases.

'Hashed Fingerprints'
Influence of parameters

Length of fingerprint:

- too short ⇒ almost all bits=1, poor discrimination of molecules.
- too large ⇒ too many bits=0, too much disk space required.

Maximum size of patterns:

- too short ⇒ poor discrimination of molecules.
- too large ⇒ ability to discriminate molecules, but many bits=1.

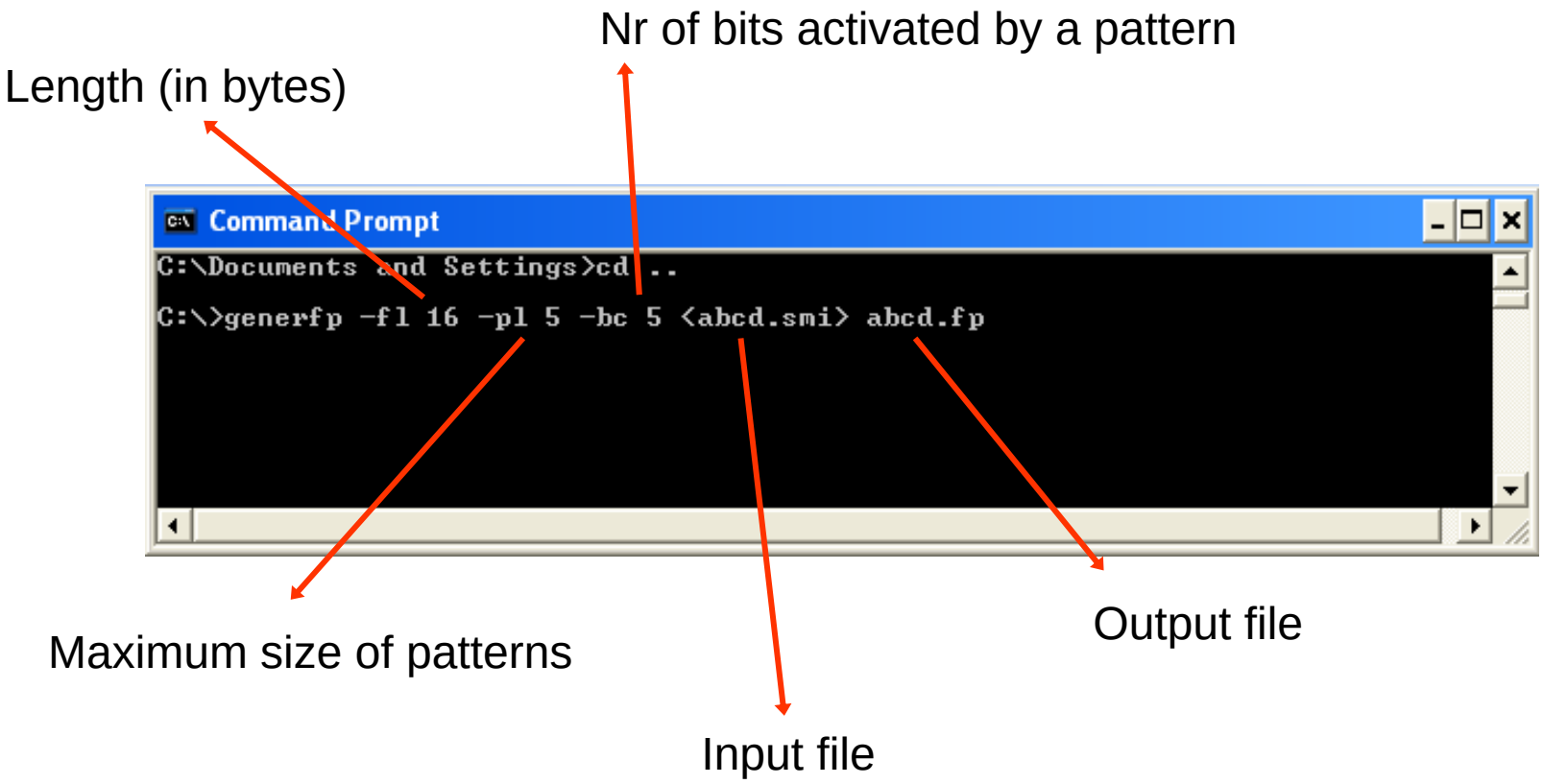
Nr of bits a pattern activates:

- too few ⇒ poor ability to discriminate between patterns.
- too many ⇒ ability to discriminate between patterns, but many bits=1.

More at: **http://www.daylight.com/dayhtml/doc/theory/theory.finger.html**

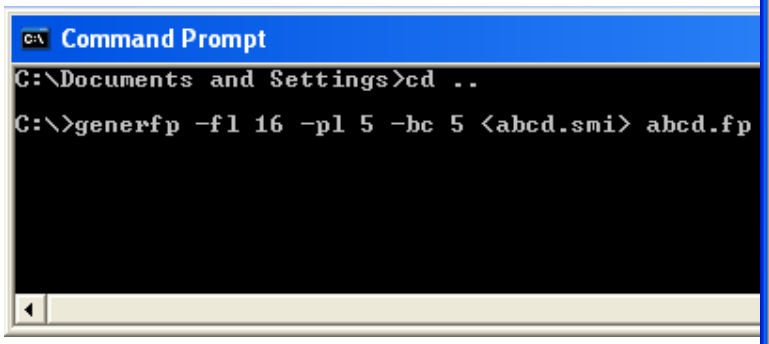
'Hashed Fingerprints'
or Daylight fingerprints

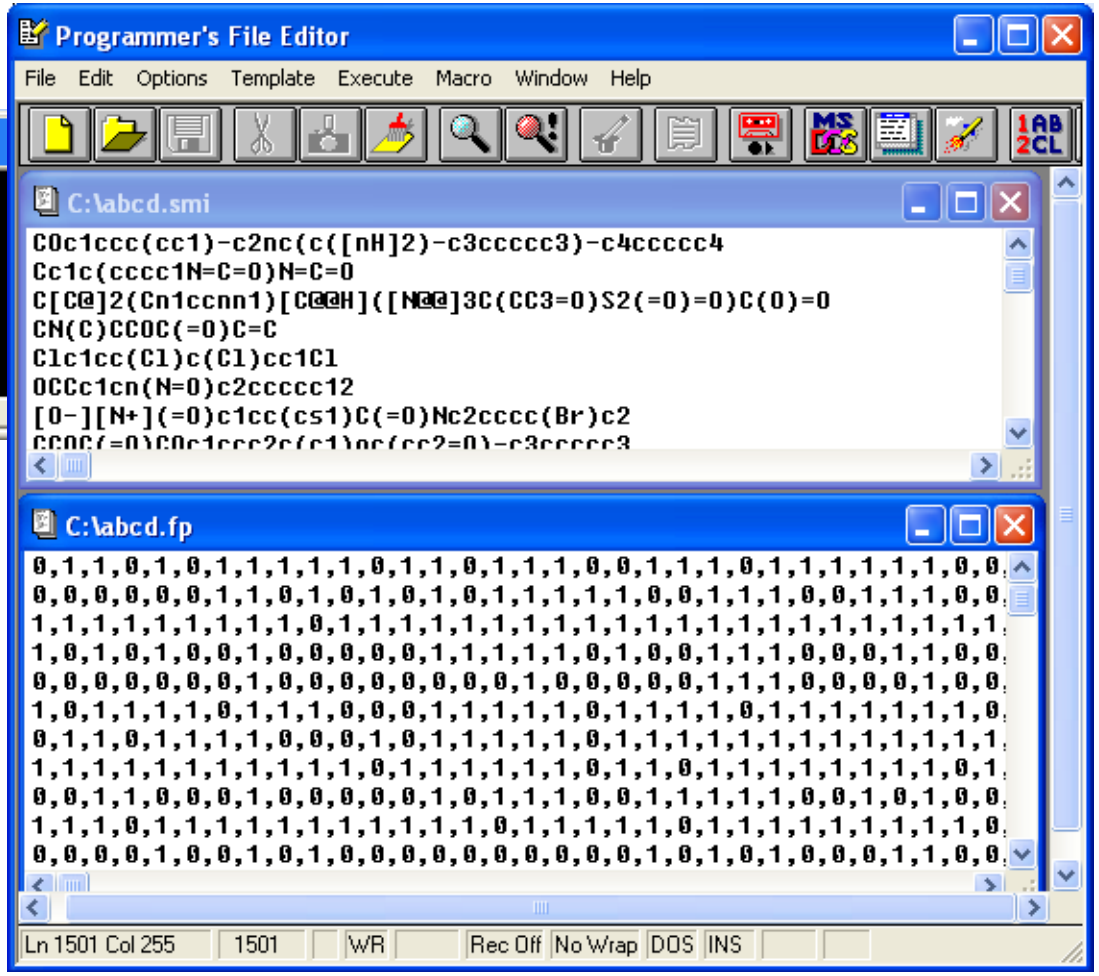
Can be calculated with several software packages, e.g. the *generfp* command of the program JCHEM (Chemaxon).



'Hashed Fingerprints'
or Daylight fingerprints

Can be calculated with the generfp command of the program JCHEM (Chemaxon).





Similarity measures based on fingerprints

Similarity between compounds X and Y can be calculated from the similarity between their fingerprints.

a = nr of bits ‘on’ in X but not in Y.
b = nr of bits ‘on’ in Y but not in X.
c = nr of bits ‘on’ both in X and in Y.
d = nr of bits ‘off’ both in X and in Y.

n = (a + b + c + d) is the total number of bits

Euclidean coefficient :

(c + d) / n (common bits in X and Y)

Tanimoto coefficient :

c / (a + b + c)

‘Hash codes’

Hash codes result from an algorithm that transforms a molecular structure into a sequence of characters or numbers encoding the presence of fragments in the molecule.

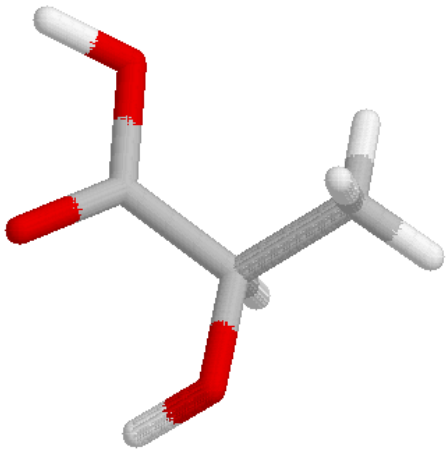
They have a fixed length.

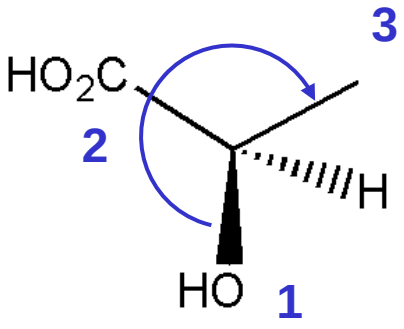
Hash codes are not interpretable. They’re used as unique identifiers of structures, e.g. in large databases of compounds hash codes allow for the fast perception of an exact match between two molecules.

Hash codes can also be defined for atoms, or bonds.

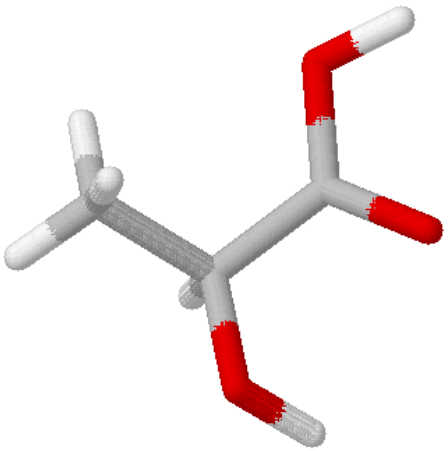
Representation of stereochemistry

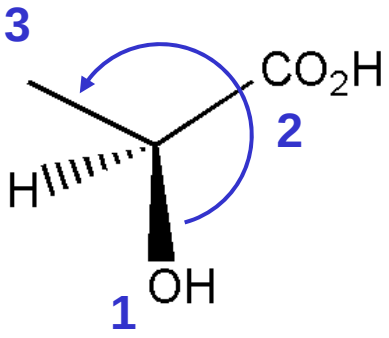
The Cahn-Ingold-Prelog (CIP) rules





(R) - lactic acid





(S) - lactic acid

Useful for nomenclature but difficult to implement: assignment of priorities.

But in a Molfile…
Atoms are ranked.
Priorities can easily be assigned corresponding to the atoms' ranks in the Molfile.

CIP priorities : OH > CO_2H > CH_3 > H

Representation of stereochemistry

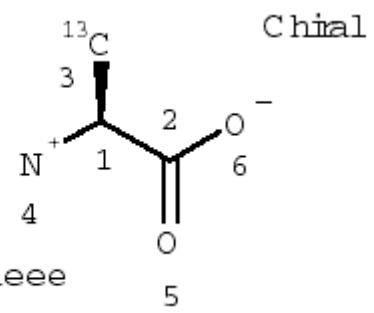
Parity in Molfiles

1. Number the atoms surrounding the stereo center with 1, 2, 3, and 4 in order of increasing atom number (position in the atom block) (a hydrogen atom should be considered atom 4).
2. View the center from a position such that the bond connecting the highest-numbered atom (4) projects behind the plane formed by atoms 1, 2, and 3.
3. Parity ‘1’ if atoms 1-3 are arranged in clockwise direction in ascending numerical order, or parity ‘2’ if counterclockwise.

Representation of stereochemistry

Molfile

L-Alanine




```
xxxxx.xxxxyyyyy.yyyyzzzzz.zzzz aaaddcccssshhhbbbvvvHHHrrriiimmmnnneee
    -0.6622    0.5342    0.0000 C   0  0  2  0  0  0
     0.6622   -0.3000    0.0000 C   0  0  0  0  0  0
    -0.7207    2.0817    0.0000 C   1  0  0  0  0  0
    -1.8622   -0.3695    0.0000 N   0  3  0  0  0  0
     0.6220   -1.8037    0.0000 O   0  0  0  0  0  0
     1.9464    0.4244    0.0000 O   0  5  0  0  0  0
```

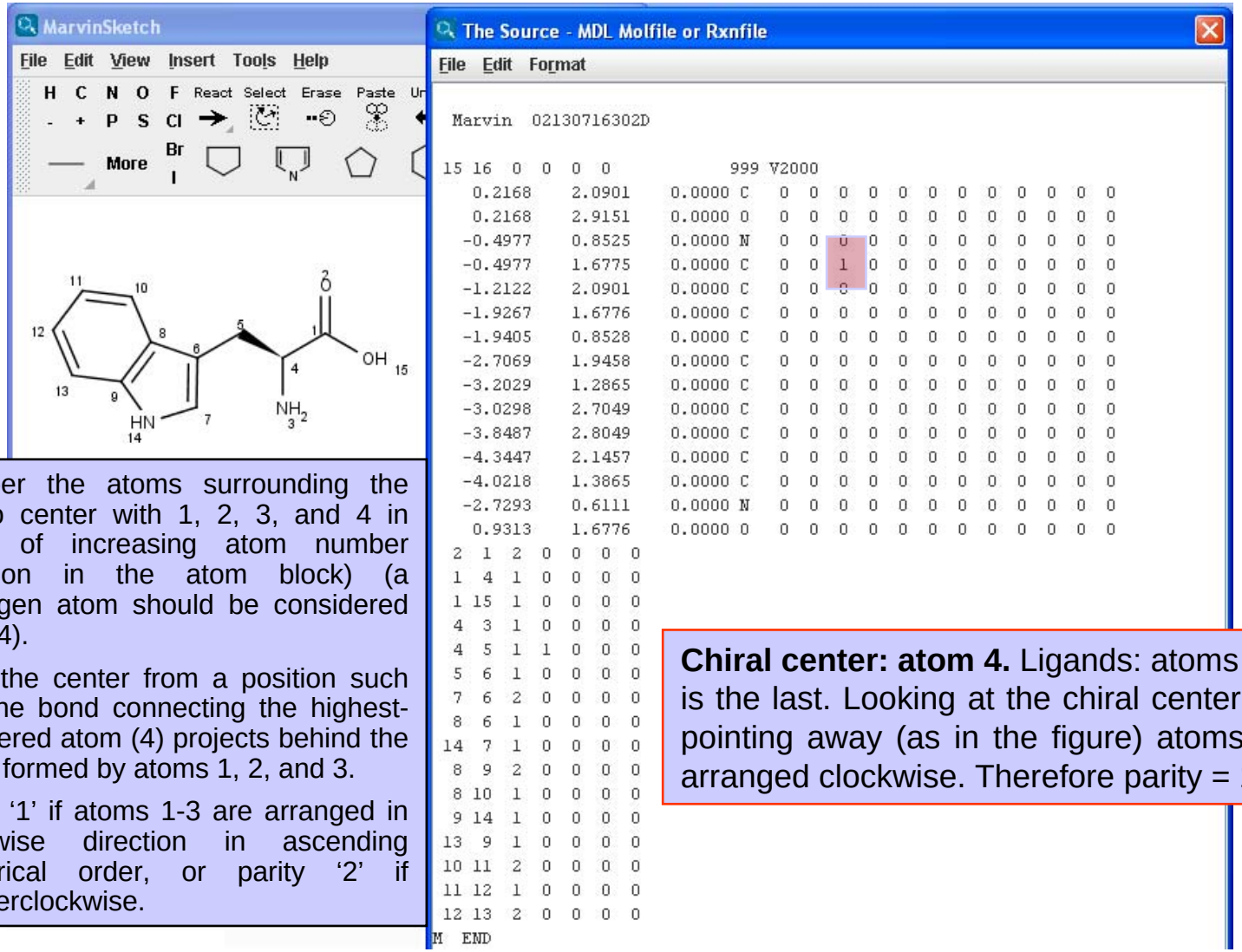
Atom Block

Field	Meaning	Values	Notes
sss	atom stereo parity	0 = not stereo, 1 = odd, 2 = even, 3 = either or unmarked stereo center	**[Generic]** Ignored when read.

Chiral center: atom 1. Ligands: atoms 2, 3, 4 and H. H is the last. Looking at the chiral center with the H-atom pointing away (as in the figure) atoms 2, 3, and 4 are arranged counterclockwise. Therefore parity = 2.

Representation of stereochemistry

Molfile



1. Number the atoms surrounding the stereo center with 1, 2, 3, and 4 in order of increasing atom number (position in the atom block) (a hydrogen atom should be considered atom 4).
2. View the center from a position such that the bond connecting the highest-numbered atom (4) projects behind the plane formed by atoms 1, 2, and 3.
3. Parity ‘1’ if atoms 1-3 are arranged in clockwise direction in ascending numerical order, or parity ‘2’ if counterclockwise.

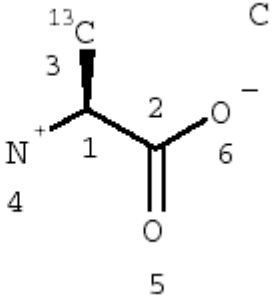
Chiral center: atom 4. Ligands: atoms 1, 3, 5, and H. H is the last. Looking at the chiral center with the H-atom pointing away (as in the figure) atoms 1, 3, and 5 are arranged clockwise. Therefore parity = 1.

Representation of stereochemistry

Molfile - *bond block*

```
111222tttsssxxxrrrccc
```

L-Alanine



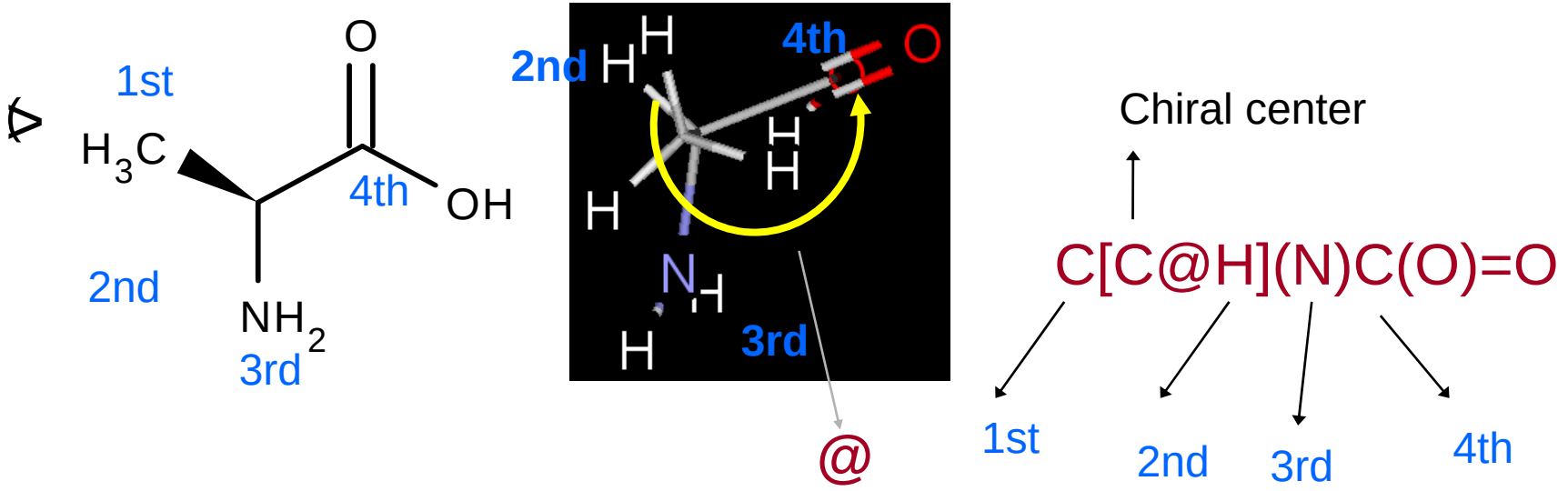
Chiral

1	2	1	0	0	0
1	3	1	1	0	0
1	4	1	0	0	0
2	5	2	0	0	0
2	6	1	0	0	0

Field	Meaning	Values	Notes
111	first atom number	1 - number of atoms	**[Generic]**
222	second atom number	1 - number of atoms	**[Generic]**
ttt	bond type	1 = Single, 2 = Double, 3 = Triple, 4 = Aromatic, 5 = Single or Double, 6 = Single or Aromatic, 7 = Double or Aromatic, 8 = Any	**[Query]** Values 4 through 8 are for SSS queries only.
sss	bond stereo	Single bonds: 0 = not stereo, 1 = Up, 4 = Either, 6 = Down, Double bonds: 0 = Use x-, y-, z-coords from atom block to determine cis or trans, 3 = Cis or trans (either) double bond	**[Generic]** The wedge (pointed) end of the stereo bond is at the first atom (Field 111 above)
xxx	not used		
rrr	bond topology	0 = Either, 1 = Ring, 2 = Chain	**[Query]** SSS queries only.
ccc	reacting center status	0 = unmarked, 1 = a center, -1 = not a center, Additional: 2 = no change, 4 = bond made/broken, 8 = bond order changes 12 = 4+8 (both made/broken and changes); 5 = (4 + 1), 9 = (8 + 1), and 13 = (12 + 1) are also possible	**[Reaction, Query]**

Representation of stereochemistry in SMILES notation

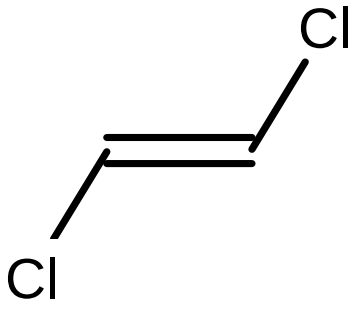
Chirality in a tetrahedral center is specified by '@' (clockwise direction) or '@@' (counterclockwise direction). Looking to the chiral center from the ligand appearing first in the SMILES string, the other three ligands are arranged clockwise or counterclockwise in the order of appearance in the SMILES string.



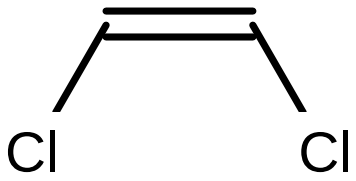
Representation of cis-trans stereochemistry in double bonds

Stereochemistry around a double bond (cis/trans) is specified with characters '\' and '/'.

Example: *trans*-1,2-dichloroethene - **Cl/C=C/Cl** (starting at the 1st Cl, a bond goes up (/) to C=C, and from here goes up (/) to the 2nd Cl).

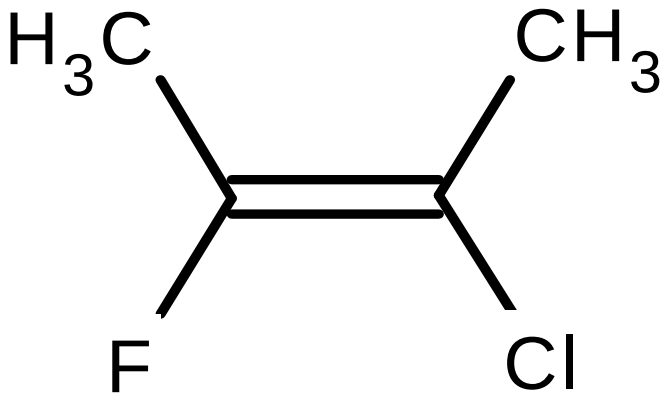


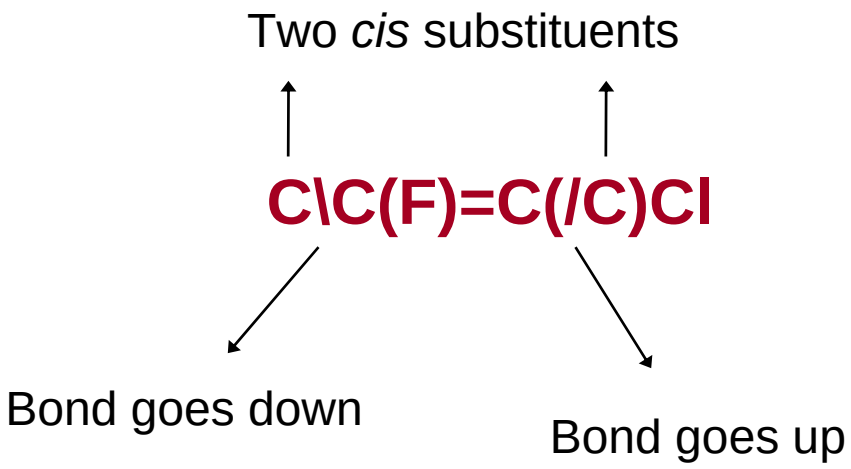
cis-1,2-dicloroeteno - **Cl/C=C\Cl** (starting at the 1st Cl, a bond goes up (/) to C=C, and from here goes down (\) to the 2nd Cl).



Representation of cis-trans stereochemistry in double bonds

Stereochemistry around a double bond (cis/trans) is specified with characters '\' and '/'.





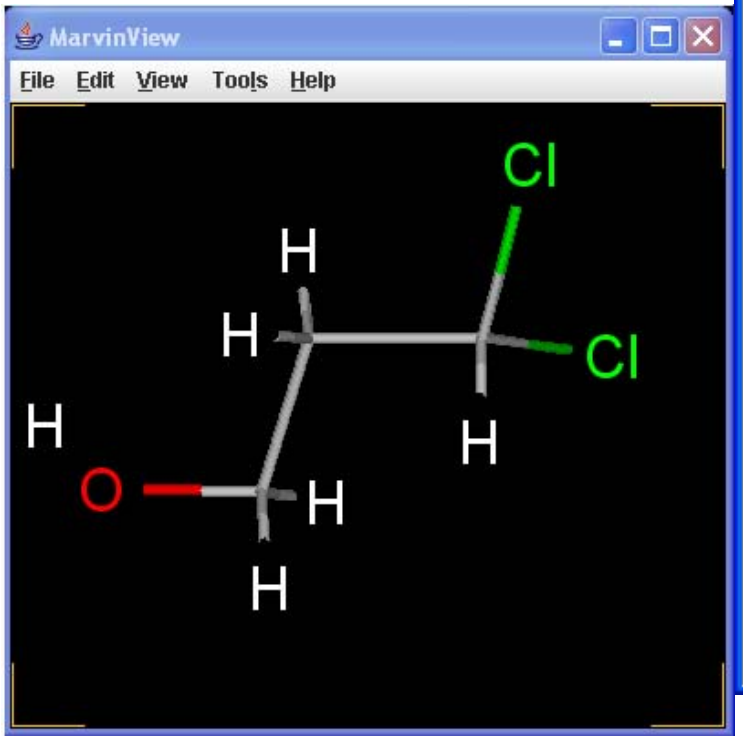
Representation of the 3D structure

The most obvious (and common) representation consists of a Cartesian system, i.e. the x, y, and z coordinates of each atom.

For a given conformation the coordinates depend on the orientation of the structure relative to the reference axes.

In a Molfile, 3D coordinates can be listed.

Representation of the 3D structure in a Molfile



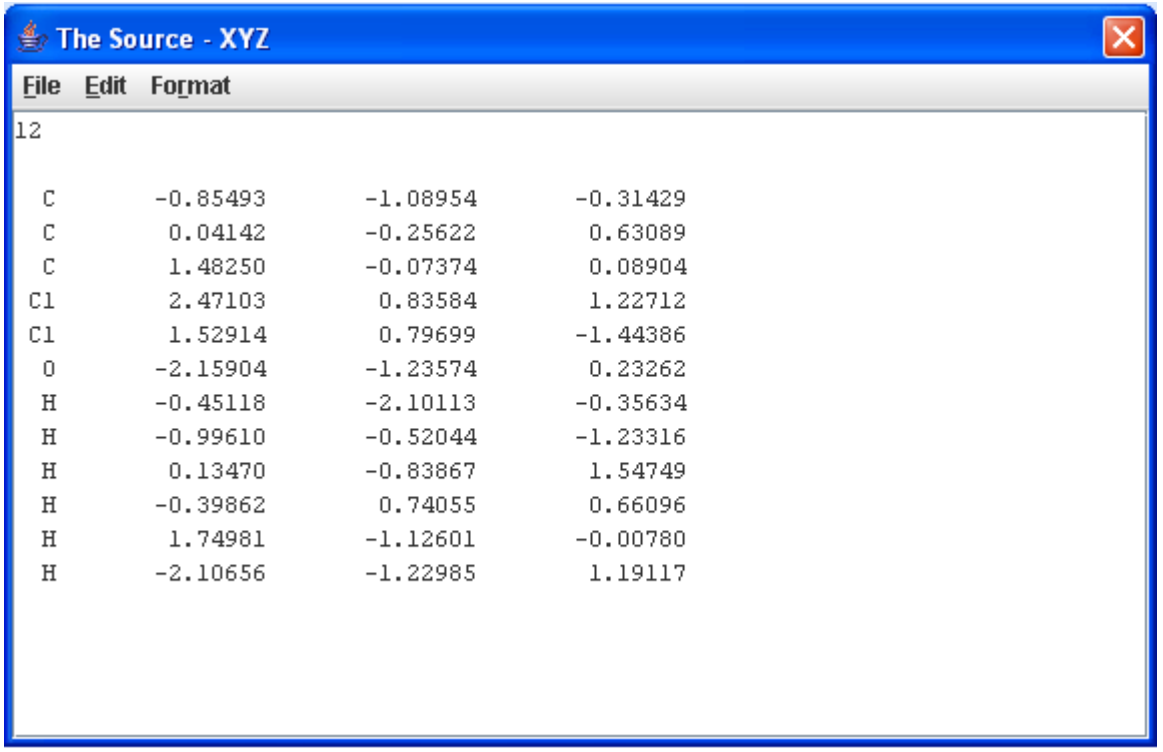
The Source - MDL Molfile or Rxnfile

File Edit Format

```
 Marvin  02150700213D

12 11  0  0  0  0            999 V2000
  -0.8549   -1.0895   -0.3143 C   0  0  0  0  0  0  0  0  0  0  0  0
   0.0414   -0.2562    0.6309 C   0  0  0  0  0  0  0  0  0  0  0  0
   1.4825   -0.0737    0.0890 C   0  0  2  0  0  0  0  0  0  0  0  0
   2.4710    0.8358    1.2271 Cl  0  0  0  0  0  0  0  0  0  0  0  0
   1.5291    0.7970   -1.4439 Cl  0  0  0  0  0  0  0  0  0  0  0  0
  -2.1590   -1.2357    0.2326 O   0  0  0  0  0  0  0  0  0  0  0  0
  -0.4512   -2.1011   -0.3563 H   0  0  0  0  0  0  0  0  0  0  0  0
  -0.9961   -0.5204   -1.2332 H   0  0  0  0  0  0  0  0  0  0  0  0
   0.1347   -0.8387    1.5475 H   0  0  0  0  0  0  0  0  0  0  0  0
  -0.3986    0.7405    0.6610 H   0  0  0  0  0  0  0  0  0  0  0  0
   1.7498   -1.1260   -0.0078 H   0  0  0  0  0  0  0  0  0  0  0  0
  -2.1066   -1.2299    1.1912 H   0  0  0  0  0  0  0  0  0  0  0  0
  1  2  1  0  0  0  0
  2  3  1  0  0  0  0
  3  4  1  0  0  0  0
  3  5  1  0  0  0  0
  1  6  1  0  0  0  0
  1  7  1  0  0  0  0
  1  8  1  0  0  0  0
  2  9  1  0  0  0  0
  2 10  1  0  0  0  0
  3 11  1  0  0  0  0
  6 12  1  0  0  0  0
M  END
```

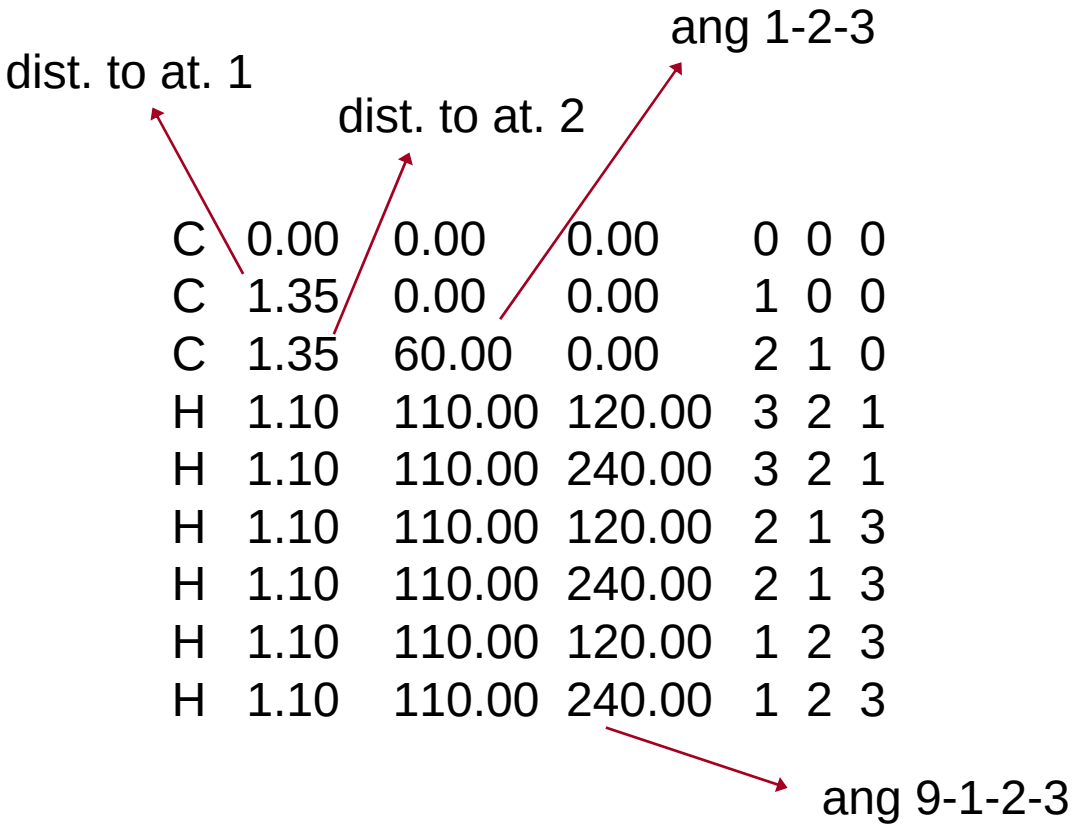
Representation of the 3D structure

It is also possible to represent only coordinates, with no specification of bonds. Bonds may be inferred with reasonable confidence from the 3D interatomic distances. But demands some kind of computer processing.



Representation of the 3D structure

Another representation of the 3D structure is the Z matrix, in which internal coordinates are specified (bond lengths, bond angles and dihedral angles). It is mostly used for the input to quantum chemistry software. Example for cyclopropane:



Generation of a 3D structure

Theoretical methods :

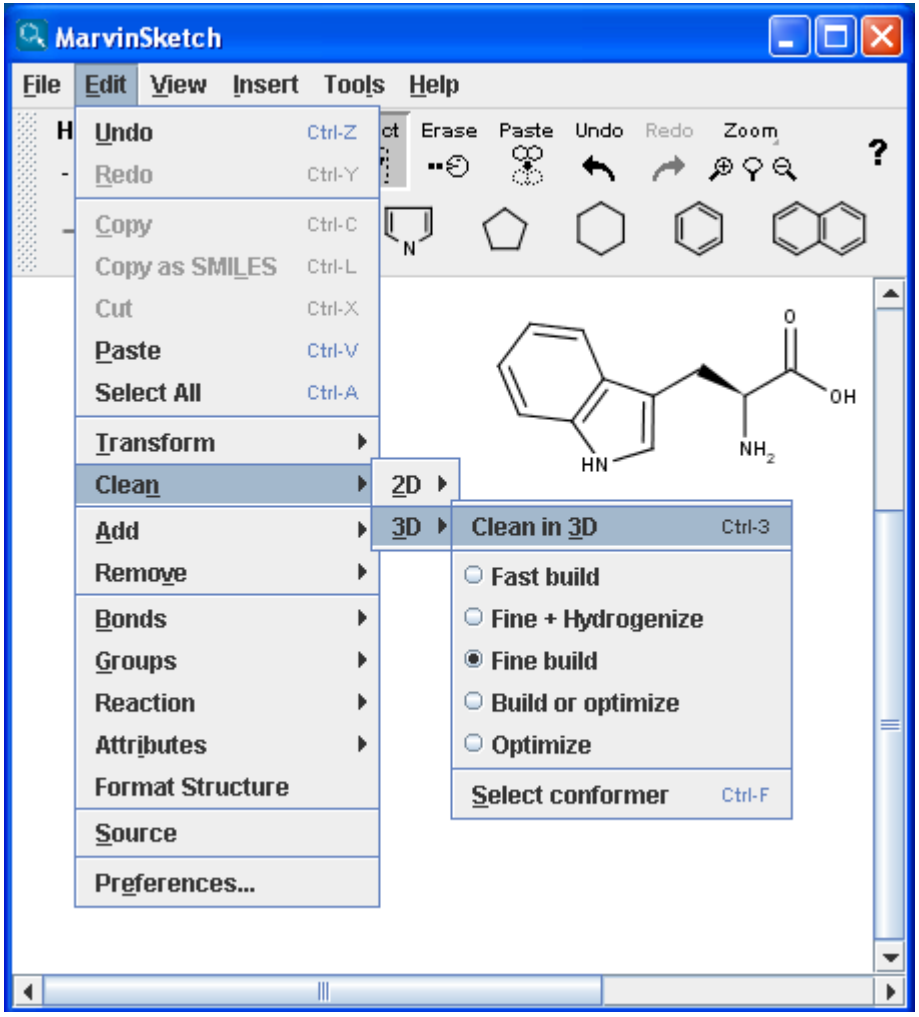
- *ab initio* (e.g. Gaussian)
- semi-empirical (e.g. Mopac)
- molecular mechanics (e.g. Mopac, Chem3D)

Empirical methods (e.g. CONCORD, CORINA) :

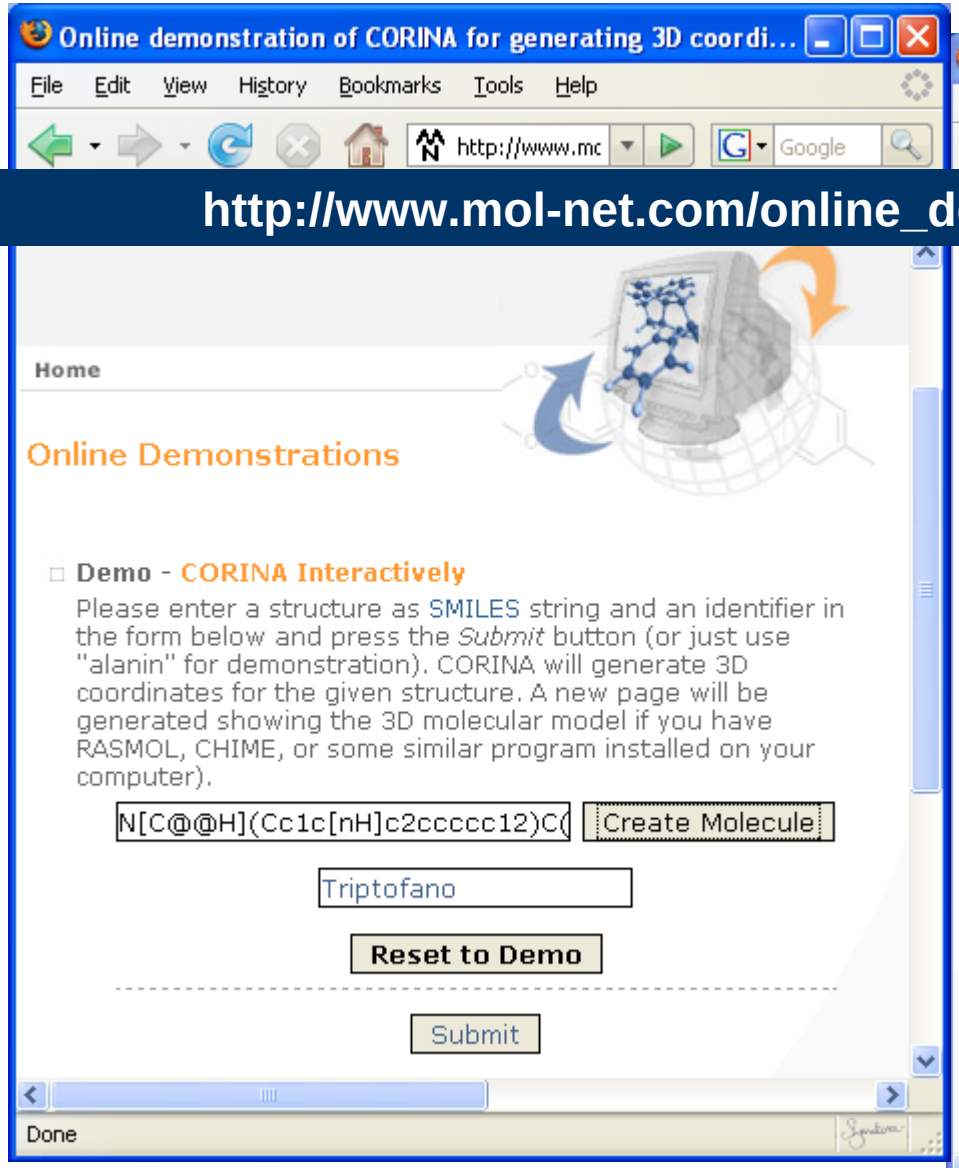
- use fragments with predefined geometries
- use rules
- use databases of geometries
- use simple optimizations

Generation of the 3D structure

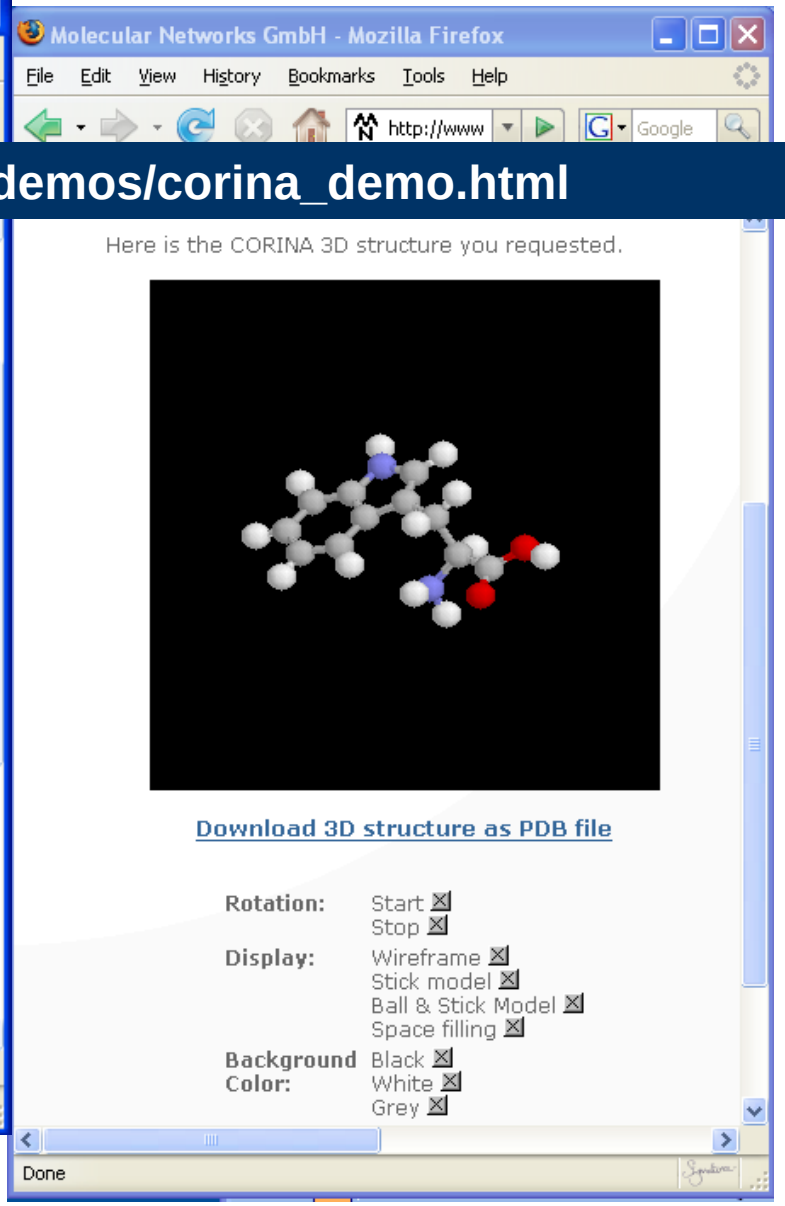
Chemaxon's Marvin



Generation of the 3D structure - CORINA



http://www.mol-net.com/online_demos/corina_demo.html



Representation of molecular surfaces

The 3D structure presented up to here is just the skeleton of the molecule, but a molecule also has a ‘skin’… the molecular surface.

The molecular surface divides the 3D space in an internal volume and an external volume. This is just an analogy with macroscopic objects, since molecules cannot rigorously be approached with classical mechanics. The electronic density is continuous, and there are probabilities of finding electrons at certain locations (it tends to zero at infinite distance from nuclei).

The electronic distribution “at the surface” determines the interactions a molecule can establish with others (e.g. docking to a protein).

Representation of molecular surfaces

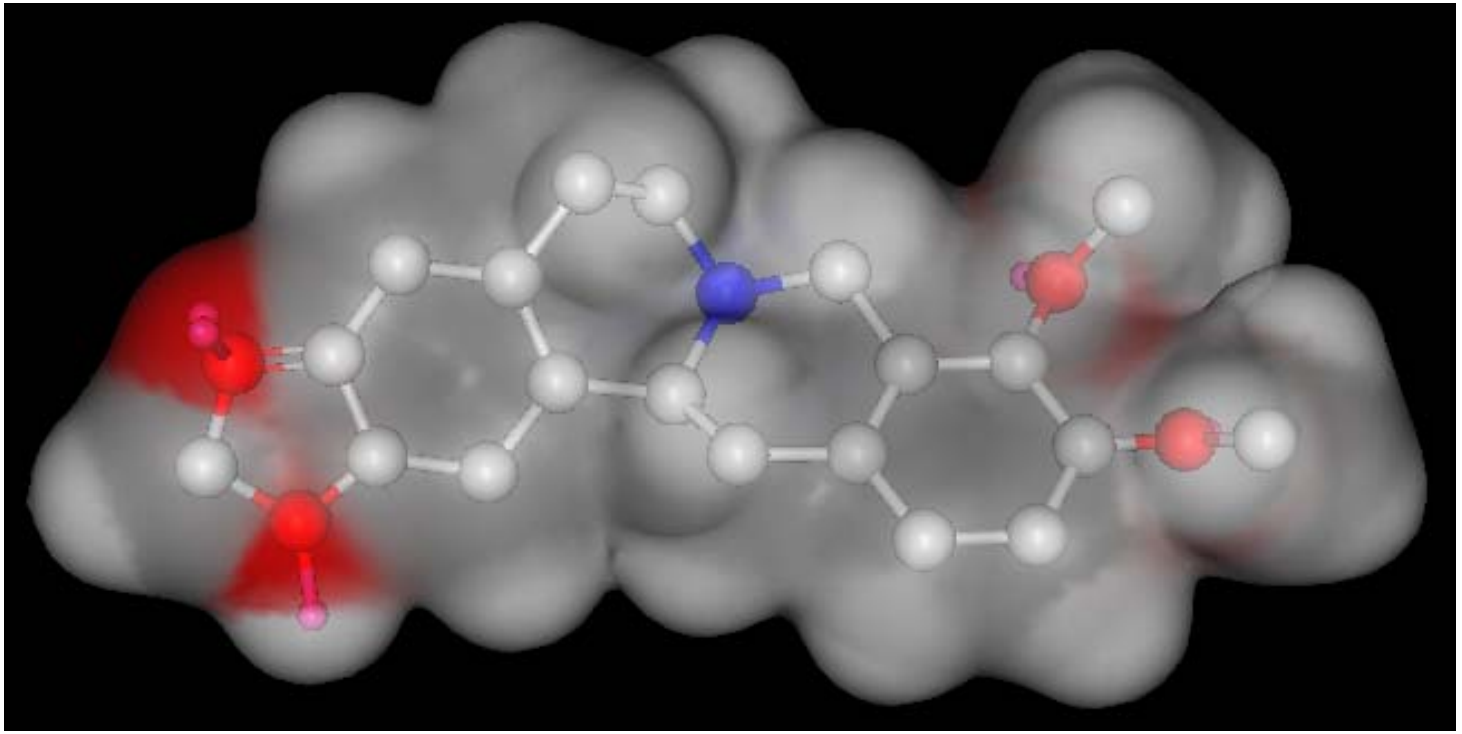
A molecular surface can express different properties, such as charge, electrostatic potential, or hydrophobicity, by means of colors.

Such properties may be experimentally determined (2D NMR, x-ray crystallography and electronic cryomicroscopy give indications about 3D molecular properties), or theoretically calculated.

There are several ways of defining a surface. The most used are: van der Waals surface, surface accessible to a solvent, and Connolly surface.

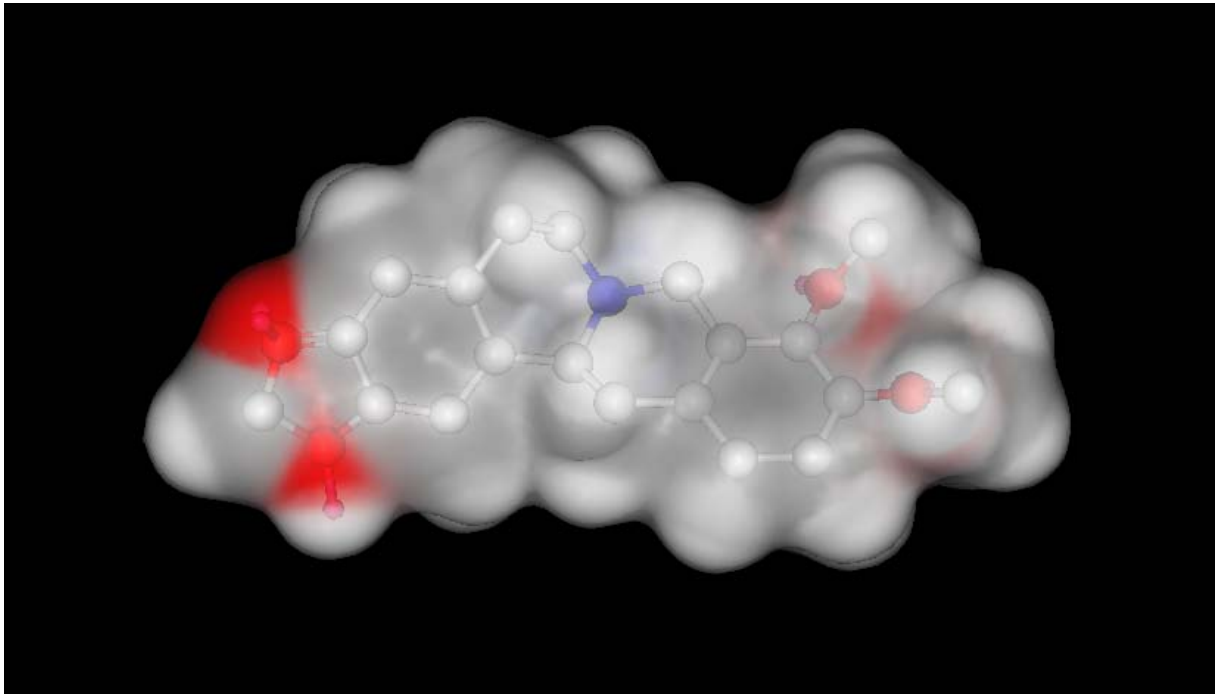
van der Waals surface

It is the simplest surface. It can be determined from the van der Waals radius of all atoms. Each atom is represented by a sphere. The spheres of all atoms are fused – the total volume is the van der Waals volume, and the envelop defines the van der Waals surface. It is fast to be calculated.



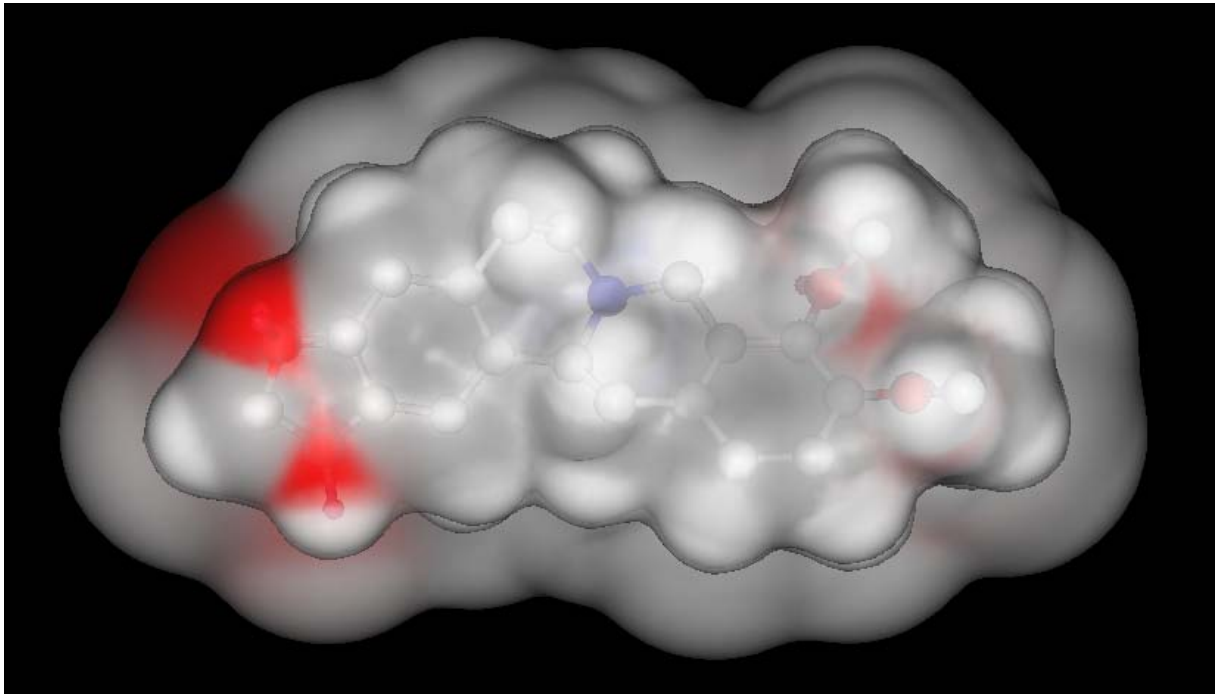
Connolly surface

It is generated by simulating a sphere rolling over the van der Waals surface. The sphere represents the solvent. The radius of the sphere may be chosen (typically it is set at 1.4 Å, the effective radius of water). The Connolly surface has two regions: the convex contact surface (it is a segment of the van der Waals surface) and the concave surface (where the sphere touches two or more atoms).

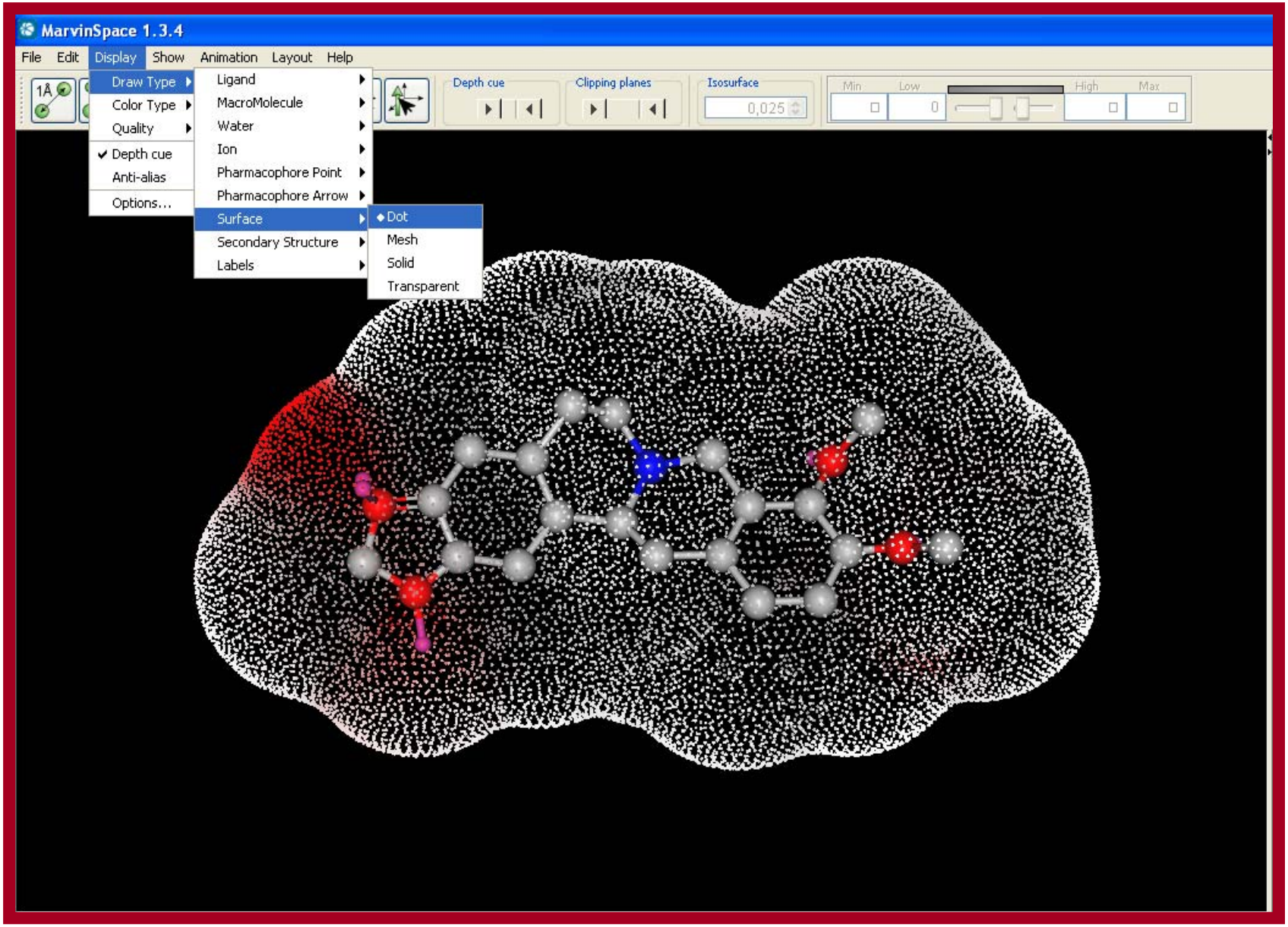


Surface accessible to the solvent

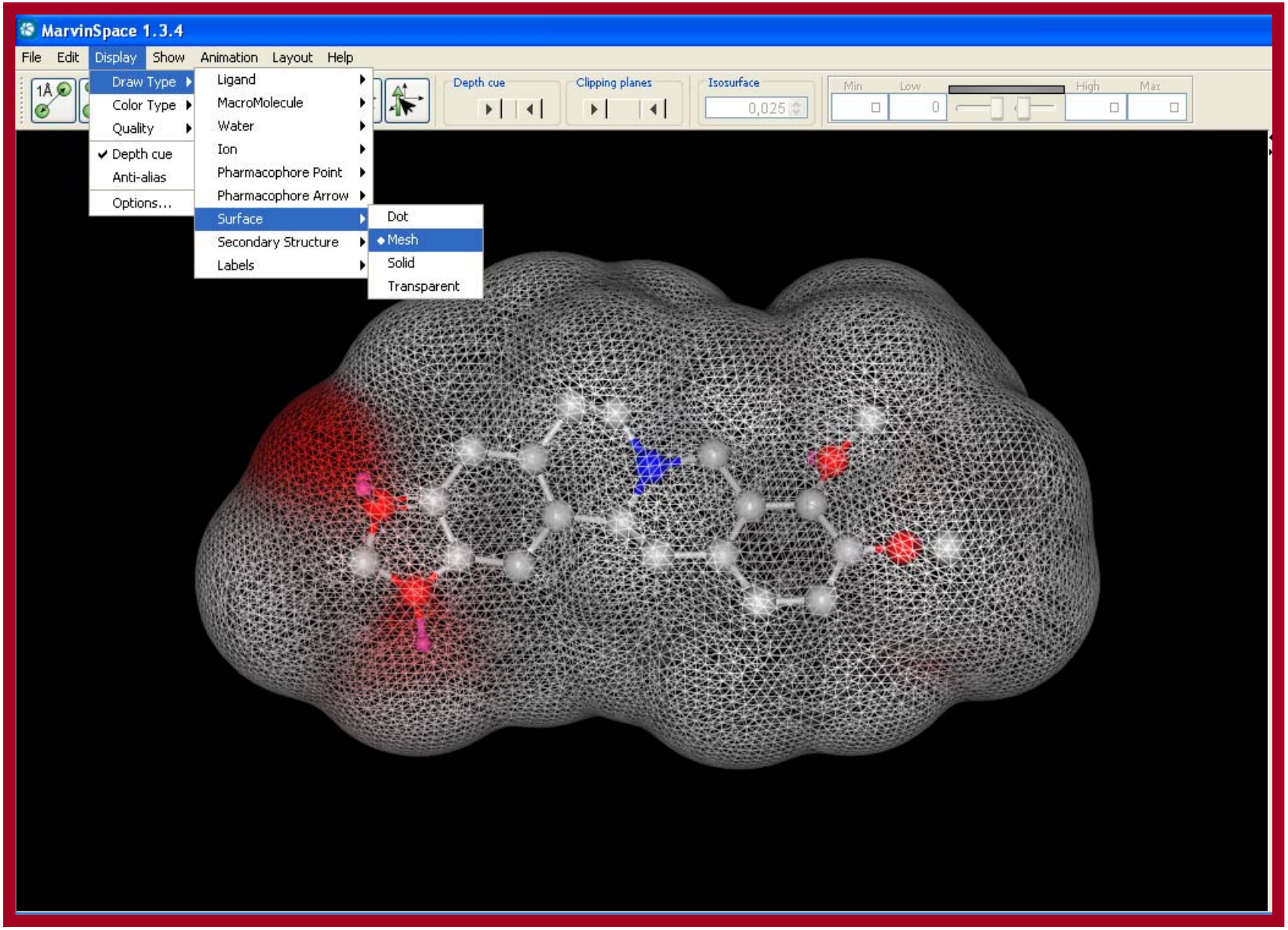
The path of the center of the sphere that generates the Connolly surface defines the surface accessible to the solvent.



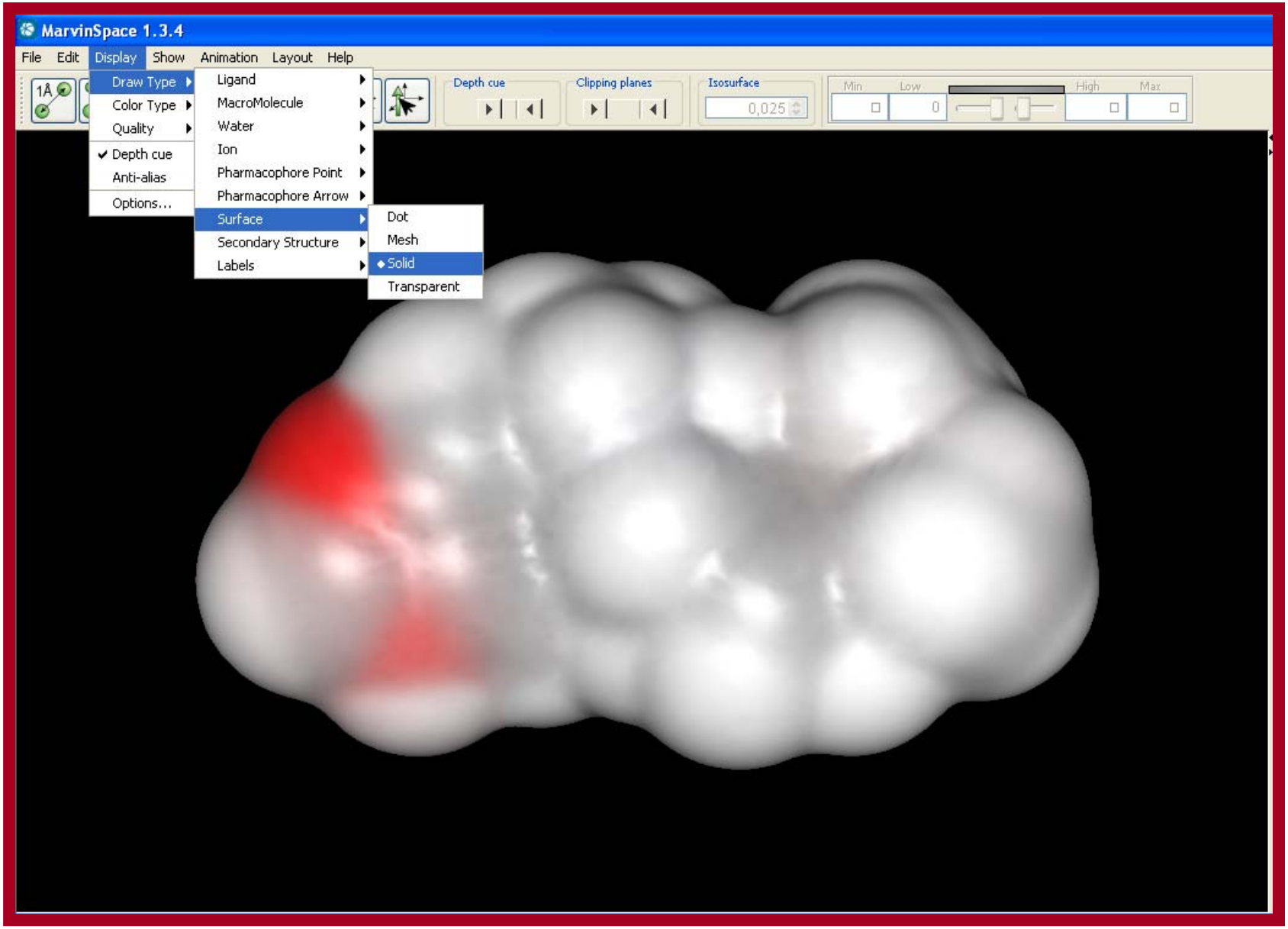
Molecular surfaces with ChemAxon MarvinSpace



Molecular surfaces with ChemAxon MarvinSpace



Molecular surfaces with ChemAxon MarvinSpace



Molecular surfaces with ChemAxon MarvinSpace