Introduction to Chemoinformatics

www.dq.fct.unl.pt/cadeiras/qc

Prof. João Aires-de-Sousa
Email: jas@fct.unl.pt
Recommended reading

CHEMOINFORMATICS

Definition (wikipedia)

Cheminformatics (also known as chemoinformatics and chemical informatics) is the use of computer and informational techniques, applied to a range of problems in the field of chemistry.

These in silico techniques are used in pharmaceutical companies in the process of drug discovery.

In the U.S., recent NIH emphasis has been placed on developing public domain Cheminformatics research by creating six Exploratory Centers for Cheminformatics Research (ECCRs) as part of the NIH Molecular Libraries Initiative.
# Size of the domain

## The Latest CAS Registry Number® and Substance Count

<table>
<thead>
<tr>
<th>Date</th>
<th>Mon Feb 12 13:40:11 EST 2007</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>30,603,610 organic and inorganic substances</td>
</tr>
<tr>
<td></td>
<td>58,543,576 sequences</td>
</tr>
<tr>
<td>CAS RN</td>
<td>920490-65-9 is the most recent CAS Registry Number</td>
</tr>
</tbody>
</table>

## Specialized Substance Collections Count

<table>
<thead>
<tr>
<th>Collection</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>CASREACT®</td>
<td>12,201,564 Single- and multi-step reactions</td>
</tr>
<tr>
<td>CHEMLIST®</td>
<td>244,763 Inventoried/regulated substances</td>
</tr>
<tr>
<td>CHEMCATS®</td>
<td>12,648,818 Commercially available chemicals</td>
</tr>
<tr>
<td>MARPAT®</td>
<td>710,651 Searchable Markush structures</td>
</tr>
</tbody>
</table>
Types of information

- **Molecular structures** (compounds)
- **Properties** (physical, chemical, biological)
  - m.p., viscosity, solubility, spectra,…
  - electrophilicity, stability, …
  - toxicity, pharmacological activity, …
- **Reactions**
Types of learning

- **Deductive learning** (quantum methods, molecular mechanics)

- **Inductive learning** (model building from experimental data): artificial intelligence, machine learning, statistics, structure-property relationships
Introduction to Chemoinformatics

1. Representation of molecular structures
### A hierarchy of structure representations

<table>
<thead>
<tr>
<th>Name</th>
<th>(S)-Tryptophan</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>2D Structure</strong></td>
<td><img src="image" alt="2D Structure" /></td>
</tr>
<tr>
<td><strong>3D Structure</strong></td>
<td><img src="image" alt="3D Structure" /></td>
</tr>
<tr>
<td><strong>Molecular surface</strong></td>
<td><img src="image" alt="Molecular surface" /></td>
</tr>
</tbody>
</table>
Storing molecular structures in a computer

MarvinSketch

MarvinView

The Source - MDL SDfile

Marvin 02120719232D

15 16 0 0 0 0 999
0.6856 2.9339 0.0000 C
0.6856 3.7589 0.0000 O
-0.0290 1.6963 0.0000 N
-0.0289 2.5213 0.0000 C

The Source - MDL SDfile

Marvin 02120719243D

15 16 0 0 0 0 999
0.5469 -0.7768 2.4931 C
1.7238 -0.9935 2.1852 O
-1.1098 0.7331 1.8186 N
-0.4806 -0.5507 1.4843 C
0.0308 -0.5663 0.0249 C
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-[[2(2S,4R,5S)-3-acetamido-5-[[2(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

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:

\[ \text{H}_3\text{C} \quad \text{OH} \]

SMILES representation: CCCO

Example:

\[ \text{H}_3\text{C} \quad \text{Cl} \quad \text{CH}_2 \quad \text{C} = \text{C} \]

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: $\text{CCC(Cl)C=C}$
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: \textbf{Nc1ccccc1}
The SMILES notation

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

- Is it unique??

  ![Chemical Structure](image)

  SMILES: \textbf{Nc1ccccccc1}
  
  but also \textbf{c1ccccccc1N}
  
  or \textbf{c1cc(N)ccc1}

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

SMILES notation in MarvinSketch

Paste

CCC1=CNC=C1

Ln 1 Col 12  1 # WR Rec Off No Wrap DOS INS
SMILES notation in MarvinSketch
**The InChI notation**
*(IUPAC International Chemical Identifier)*

Example:

```
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
```

-L-ascorbic acid

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.

The InChI notation
(IUPAC International Chemical Identifier)

Example:

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

L-ascorbic acid

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.
Matrix representations

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

Adjacency matrix: indicates which atoms are bonded.
Matrix representations

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

**Adjacency matrix**: indicates which atoms are bonded.
Matrix representations

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

**Adjacency matrix**: indicates which atoms are bonded.
**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.

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.

![Chemical structure diagram]

Matrix:

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</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).

![Chemical structure diagram](image)

<table>
<thead>
<tr>
<th>List of atoms</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 C</td>
</tr>
<tr>
<td>2 C</td>
</tr>
<tr>
<td>3 C</td>
</tr>
<tr>
<td>4 Cl</td>
</tr>
<tr>
<td>5 C</td>
</tr>
<tr>
<td>6 C</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>List of bonds</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>5</td>
</tr>
</tbody>
</table>
The MDL Molfile format


Description of an atom

Nr of bonds

Nr of atoms

Description of a bond

H3C

Cl

CH2
The MDL Molfile format

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

Header Block
Counts Line
Connection Table (Ctab)
Bond Block
Properties Block

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
The atom block

L-Alanine

<table>
<thead>
<tr>
<th>Field</th>
<th>Meaning</th>
<th>Values</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>xyz</td>
<td>atom coordinates</td>
<td>entry in periodic table or L for atom list, A, Q, * for unspecified atom, and LP for lone pair, or R# for Rgroup label</td>
<td>[Generic]</td>
</tr>
<tr>
<td>aaa</td>
<td>atom symbol</td>
<td></td>
<td>[Generic, Query, 3D, Rgroup]</td>
</tr>
<tr>
<td>dd</td>
<td>mass difference</td>
<td>-3, -2, -1, 0, 1, 2, 3, 4 (0 if value beyond these limits)</td>
<td>[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.</td>
</tr>
<tr>
<td>ccc</td>
<td>charge</td>
<td>0 = uncharged or value other than these, 1 = +3, 2 = +2, 3 = +1, 4 = doublet radical, 5 = -1, 6 = -2, 7 = -3</td>
<td>[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.</td>
</tr>
<tr>
<td>sss</td>
<td>atom stereo parity</td>
<td>0 = not stereo, 1 = odd, 2 = even, 3 = either or unmarked stereo center</td>
<td>[Generic] Ignored when read.</td>
</tr>
</tbody>
</table>
The atom block

<table>
<thead>
<tr>
<th>Field</th>
<th>Meaning</th>
<th>Values</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>atom coordinates</td>
<td></td>
<td>C 0 0 2 0 0</td>
<td>[Generic]</td>
</tr>
<tr>
<td>atom symbol</td>
<td>entry in periodic table or L for atom list, A, O, * for unspecified atom, and LP for lone pair, or R# for Rgroup label</td>
<td>C 0 0 0 0 0</td>
<td>[Generic, Query, 3D, Rgroup]</td>
</tr>
<tr>
<td>mass difference</td>
<td>-3, -2, -1, 0, 1, 2, 3, 4 (0 if value beyond these limits)</td>
<td>0.6622 -0.3000 0.0000</td>
<td>[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.</td>
</tr>
<tr>
<td>charge</td>
<td>0 = uncharged or value other than these, 1 = +3, 2 = +2, 3 = +1, 4 = doublet radical, 5 = -1, 6 = -2, 7 = -3</td>
<td>-0.7207 2.0817 0.0000</td>
<td>[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.</td>
</tr>
<tr>
<td>atom stereo parity</td>
<td>0 = not stereo, 1 = odd, 2 = even, 3 = either or unmarked stereo center</td>
<td>-1.8622 -0.3695 0.0000</td>
<td>[Generic] Ignored when read.</td>
</tr>
</tbody>
</table>

Chiral

N^+ 1 2 O^− 6

N 4 O 5

L-Alanine

C 13

www.dq.fct.unl.pt/staff/jas/qc
### The atom block

<table>
<thead>
<tr>
<th>Field</th>
<th>Meaning</th>
<th>Values</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>xyz</td>
<td>atom coordinates</td>
<td>entry in periodic table or L for atom list, A, O, * for unspecified atom, and LP for lone pair, or R# for Rgroup label</td>
<td>[Generic]</td>
</tr>
<tr>
<td>atom symbol</td>
<td></td>
<td></td>
<td>[Generic, Query, 3D, Rgroup]</td>
</tr>
<tr>
<td>dd</td>
<td>mass difference</td>
<td>-3, -2, -1, 0, 1, 2, 3, 4 (0 if value beyond these limits)</td>
<td>[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.</td>
</tr>
<tr>
<td>ccc</td>
<td>charge</td>
<td>0 = uncharged or value other than these, 1 = +3, 2 = +2, 3 = +1, 4 = doublet radical, 5 = -1, 6 = -2, 7 = -3</td>
<td>[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.</td>
</tr>
<tr>
<td>sss</td>
<td>atom stereo parity</td>
<td>0 = not stereo, 1 = odd, 2 = even, 3 = either or unmarked stereo center</td>
<td>[Generic] Ignored when read.</td>
</tr>
</tbody>
</table>
# The atom block

<table>
<thead>
<tr>
<th>Field</th>
<th>Meaning</th>
<th>Values</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>x y z</td>
<td>atom coordinates</td>
<td></td>
<td>[Generic]</td>
</tr>
<tr>
<td>aaa</td>
<td>atom symbol</td>
<td>entry in periodic table or L for atom list, A, Q, * for unspecified atom, and LP for lone pair, or R# for Rgroup label</td>
<td>[Generic, Query, 3D, Rgroup]</td>
</tr>
<tr>
<td></td>
<td>mass difference</td>
<td>-3, -2, -1, 0, 1, 2, 3, 4 (0 if value beyond these limits)</td>
<td>[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.</td>
</tr>
<tr>
<td>ccc</td>
<td>charge</td>
<td>0 = uncharged or value other than these, 1 = +3, 2 = +2, 3 = +1, 4 = doublet radical, 5 = -1, 6 = -2, 7 = -3</td>
<td>[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.</td>
</tr>
<tr>
<td>sss</td>
<td>atom stereo parity</td>
<td>0 = not stereo, 1 = odd, 2 = even, 3 = either or unmarked stereo center</td>
<td>[Generic] Ignored when read.</td>
</tr>
</tbody>
</table>

L-Alanine

Chiral
The atom block

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
<th>z</th>
<th>C</th>
<th>0</th>
<th>0</th>
<th>2</th>
<th>0</th>
<th>0</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.6622</td>
<td>0.5342</td>
<td>0.0000</td>
<td>C</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.6622</td>
<td>-0.3000</td>
<td>0.0000</td>
<td>C</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>-0.7207</td>
<td>2.0817</td>
<td>0.0000</td>
<td>C</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>-1.8622</td>
<td>-0.3695</td>
<td>0.0000</td>
<td>N</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.6220</td>
<td>-1.8037</td>
<td>0.0000</td>
<td>O</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1.9464</td>
<td>0.4244</td>
<td>0.0000</td>
<td>O</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

**Field** | **Meaning** | **Values** | **Notes**
----------|-------------|-------------|-----------------------------------------------
xyz       | atom coordinates | 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] |
aaa       | atom symbol | 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. | [Generic, Query, 3D, Rgroup] |
dd        | mass difference | -3, -2, -1, 0, 1, 2, 3, 4 (0 if value beyond these limits) | [Generic] |
charge    | 0 = uncharged or value other than these, 1 = +3, 2 = +2, 3 = +1, 4 = doublet radical, 5 = -1, 6 = -2, 7 = -3 | [Generic] |
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 (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

M CHG 2 4 1 6 -1
M ISO 1 3 13
M END
The bond block

111222tttssxxxrrrrccc

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

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

© João Aires de Sousa
### The bond block

The bond block is a representation used in molecular modeling to describe the connectivity and properties of atoms in a molecule. It is typically used in the context of quantum chemistry and computational chemistry. The bond block diagram and table provide a detailed description of how atoms are connected and the properties of those connections.

#### Diagram

![L-Alanine Diagram](image)

#### Table

<table>
<thead>
<tr>
<th>Field</th>
<th>Meaning</th>
<th>Values</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>first atom number</td>
<td>1 - number of atoms</td>
<td>[Generic]</td>
</tr>
<tr>
<td>2</td>
<td>second atom number</td>
<td>1 - number of atoms</td>
<td>[Generic]</td>
</tr>
<tr>
<td>ttt</td>
<td>bond type</td>
<td>1 = Single, 2 = Double, 3 = Triple, 4 = Aromatic, 5 = Single or Double, 6 = Single or Aromatic, 7 = Double or Aromatic, 8 = Any</td>
<td>[Query] Values 4 through 8 are for SSS queries only.</td>
</tr>
<tr>
<td>sss</td>
<td>bond stereo</td>
<td>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</td>
<td>[Generic] The wedge (pointed) end of the stereo bond is at the first atom (Field 111 above)</td>
</tr>
<tr>
<td>xxx</td>
<td>not used</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rrr</td>
<td>bond topology</td>
<td>0 = Either, 1 = Ring, 2 = Chain</td>
<td>[Query] SSS queries only.</td>
</tr>
<tr>
<td>ccc</td>
<td>reacting center status</td>
<td>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</td>
<td>[Reaction, Query]</td>
</tr>
</tbody>
</table>
## The bond block

### L-Alanine

<table>
<thead>
<tr>
<th>Field</th>
<th>Meaning</th>
<th>Values</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>111</td>
<td>first atom number</td>
<td>1 = number of atoms</td>
<td>[Generic]</td>
</tr>
<tr>
<td>222</td>
<td>second atom number</td>
<td>1 = number of atoms</td>
<td>[Generic]</td>
</tr>
<tr>
<td></td>
<td>bond type</td>
<td>1 = Single, 2 = Double, 3 = Triple, 4 = Aromatic, 5 = Single or Double, 6 = Single or Aromatic, 7 = Double or Aromatic, 8 = Any</td>
<td>[Query] Values 4 through 8 are for SSS queries only.</td>
</tr>
<tr>
<td>sss</td>
<td>bond stereo</td>
<td>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</td>
<td>[Generic] The wedge (pointed) end of the stereo bond is at the first atom (Field 111 above)</td>
</tr>
<tr>
<td>xxx</td>
<td>not used</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rrr</td>
<td>bond topology</td>
<td>0 = Either, 1 = Ring, 2 = Chain</td>
<td>[Query] SSS queries only.</td>
</tr>
<tr>
<td>ccc</td>
<td>reacting center status</td>
<td>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</td>
<td>[Reaction, Query]</td>
</tr>
</tbody>
</table>
The bond block

111222ttttxxxrrrrccc

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

L-Alanine (13C)
GSMACCS-1110169115362D 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

Header Block
Counts Line
Atom Block
Connection Table (Ctab)
Bond Block
The properties block

2 charged atoms
The properties block

2 charged atoms

atom 4: charge +1
atom 6: charge -1
The properties block

1 entry for an isotope

M  CHG  2  4  1  6  -1
M  END

L-Alanine

Chiral

N

1 2 3 4 5 6

13C

© João Aires de Sousa
www.dq.fct.unl.pt/staff/jas/gc
The properties block

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 SDFFile (.SDF) format

Example

Molfile1
Associated data
$$$$

Molfile2
Associated data
$$$$
...

Associated data (molecular)
The SDFile (.SDF) format

Example

<table>
<thead>
<tr>
<th>61283-01-8</th>
<th>021307180300D</th>
</tr>
</thead>
<tbody>
<tr>
<td>11 10 0 0 0 0</td>
<td>999 12000</td>
</tr>
<tr>
<td>1.6947 -0.2675 -0.0016 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td>0.5343 0.6242 0.0001 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td>3.0073 0.3083 -0.0014 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td>1.5355 -1.4702 0.0019 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td>0.7895 1.9478 0.0018 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td>-1.2132 -0.0983 -0.0002 Br 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td>3.5067 -0.7060 0.0000 H 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td>4.1016 0.7277 0.0000 H 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td>2.6679 1.3226 0.0000 H 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td>1.7228 2.3671 0.0000 H 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td>-0.1688 2.6166 0.0000 H 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0</td>
<td></td>
</tr>
</tbody>
</table>

H END

> <Anes test categorisation>
nutagen

> <EXACTMASS>
147.952377

> $$$$ 598-55-0

Associated data (atomic)
The SDFFile (.SDF) format

Example

Molfile1
Associated data
$$$$
Molfile2
Associated data
$$$$
...

Associated data (molecular)
# The SDFFile (.SDF) format

**Example**

```
61203-01-8
Harvin 021307103030

11 10 0 0 0 0 0 0 0 0 0 999 2000
1.6947 -0.2675 -0.0016 C 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
3.0873 -0.3083 -0.0014 C 0 0 0 0 0 0 0 0 0
1.5355 -1.4702 0.0019 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
-1.2132 -0.0983 -0.0002 Br 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
4.1016 0.7277 0.0000 H 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
1.7228 2.3671 0.0000 H 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

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

H END

> <anes test categorisation>
nutagen

> <CHARGE>
0,11;0,95;0,95;0,95;0,29;0,02;0,05;0,03;0,03;0,03;0,06;0,06

> <EXACTMASS>
147.952377

$$$$
598-55-0
Harvin 021307103030

10 9 0 0 0 0 0 0 0 0 0 999 2000
```

**Delimiter**

```
Beginning of Molfile2

```
### The SDFFile (.SDF) format

**Example**

```
> <NAME>
147.952377

> <CHARGECAT>
3.40

> <CLASS>
nonmutagen

> <CHARGE>
0.04; -0.20; -0.05; -0.26; -0.05; -0.13; -0.13; -0.05; +0.05; +0.05

> <EXACTMASS>
75.02028

> $$$$$
```

Molfile1
Associated data $$$$$
Molfile2
Associated data $$$$$
...

```
```

```
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

Aromatize
Replaces the alternating single/double bonds of aromatic rings with aromatic bond types. The aromatic rings of the current compound are determined by the Hückel's rule. Two methods are provided, the GENERAL approach operates on the SSSR (smallest set of smallest rings) and considers some mesomers as well (for example, pyridine is aromatic). The BASIC method detects all aromatic rings (not just the SSSR) and operates on the current structure (pyridone is non-aromatic).

Click here for online help...

Press 'Next' to continue...
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.

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.

‘Hashed Fingerprints’
or Daylight fingerprints

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

- Length (in bytes)
- Nr of bits activated by a pattern
- Maximum size of patterns
- Output file
- Input file
‘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.

- \( a \) = number of bits ‘on’ in X but not in Y.
- \( b \) = number of bits ‘on’ in Y but not in X.
- \( c \) = number of bits ‘on’ both in X and in Y.
- \( d \) = number of bits ‘off’ both in X and in Y.

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

**Euclidean coefficient** :

\[
( c + d ) / n \quad \text{(common bits in X and Y)}
\]

**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

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 > CO₂H > CH₃ > H

(R) - lactic acid

(S) - lactic acid
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

```
xxxxx.xxxxxyyyyy.yyyyyyyyyyy.zzzz aaaddccccssshhhbvbvbvVHHhrrriiiimmnnnneee
-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
```

<table>
<thead>
<tr>
<th>Field</th>
<th>Meaning</th>
<th>Values</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>sss</td>
<td>atom stereo parity</td>
<td>0 = not stereo, 1 = odd, 2 = even, 3 = either or unmarked stereo center</td>
<td>[Generic] Ignored when read.</td>
</tr>
</tbody>
</table>

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

![L-Alanine](image)

```
111222tttssssxxxrrrcce
```

### Bond Block Field Descriptions

<table>
<thead>
<tr>
<th>Field</th>
<th>Meaning</th>
<th>Values</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>111</td>
<td>first atom number</td>
<td>1 - number of atoms</td>
<td>[Generic]</td>
</tr>
<tr>
<td>222</td>
<td>second atom number</td>
<td>1 - number of atoms</td>
<td>[Generic]</td>
</tr>
<tr>
<td>ttt</td>
<td>bond type</td>
<td>1 = Single, 2 = Double, 3 = Triple, 4 = Aromatic, 5 = Single or Double, 6 = Single or Aromatic, 7 = Double or Aromatic, 8 = Any</td>
<td>[Query] Values 4 through 8 are for SSS queries only.</td>
</tr>
<tr>
<td>sss</td>
<td>bond stereo</td>
<td>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</td>
<td>[Generic] The wedge (pointed) end of the stereo bond is at the first atom (Field 111 above)</td>
</tr>
<tr>
<td>xxx</td>
<td>not used</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rrr</td>
<td>bond topology</td>
<td>0 = Either, 1 = Ring, 2 = Chain</td>
<td>[Query] SSS queries only.</td>
</tr>
<tr>
<td>ccc</td>
<td>reacting center status</td>
<td>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</td>
<td>[Reaction, Query]</td>
</tr>
</tbody>
</table>
**Representation of stereochemistry in SMILES notation**

**Chirality in a tetrahedral center** is specified by ‘@’ (anticlockwise direction) or ‘@@’ (clockwise 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-dichloroethene - 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**

S **Stereochemistry around a double bond** (cis/trans) is specified with characters ‘\’ and ‘/’.

\[ \text{Two cis substituents} \quad C\backslash C(F) = C(/C)\text{Cl} \]

- Bond goes down
- Bond goes up

\[ \text{Bond goes down} \quad \text{Bond goes up} \]
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
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.

```
The Source - XYZ

<table>
<thead>
<tr>
<th>File</th>
<th>Edit</th>
<th>Format</th>
</tr>
</thead>
<tbody>
<tr>
<td>L2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

C  -0.85493  -1.08954  -0.31429
C  0.04142   -0.25622  0.63089
C  1.48250   -0.07374  0.08904
Cl  2.47103   0.03584  1.22712
Cl  1.52914   0.79699  -1.44386
O  -2.15904   -1.23574  0.23262
H  -0.45118   -2.10113  -0.35634
H  -0.99610   -0.52044  -1.23316
H  0.13470   -0.83867  1.54749
H  -0.39862   0.74055  0.66096
H  1.74981   -1.12601  -0.00780
H  -2.10656   -1.22985  1.19117
```
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:

```
C  0.00  0.00  0.00    0  0  0
C  1.35  0.00  0.00    1  0  0
C  1.35  60.00 0.00    2  1  0
H  1.10 110.00 120.00  3  2  1
H  1.10 110.00 240.00  3  2  1
H  1.10 110.00 120.00  2  1  3
H  1.10 110.00 240.00  2  1  3
H  1.10 110.00 120.00  1  2  3
H  1.10 110.00 240.00  1  2  3
```
Generation of a 3D structure

Theoretical methods:

*ab initio* (e.g. Gaussian)

semi-empirical (e.g. Mopac, ArgusLab)

molecular mechanics (e.g. Chem3D, ArgusLab)

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

Online Demonstrations

Demo - CORINA Interactively
Please enter a structure as SMILES string and an identifier in the form below and press the Submit button (or just use "alanin" for demonstration). CORINA will generate 3D coordinates for the given structure. A new page will be generated showing the 3D molecular model if you have RASMOL, CHIME, or some similar program installed on your computer).

N[C@H](Cc1c[nH]c2cccc12)c

Triptofano

Reset to Demo

Submit

Download 3D structure as PDB file

Rotation:  
Start  
Stop  

Display:  
Wireframe  
Stick model  
Ball & Stick Model  
Space filling  

Background Color:  
Black  
White  
Grey  

Done
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