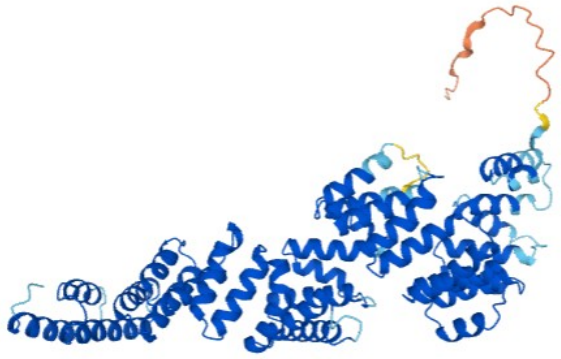
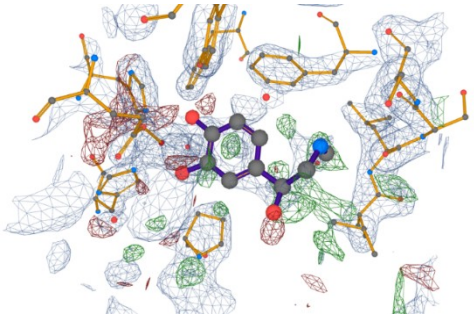
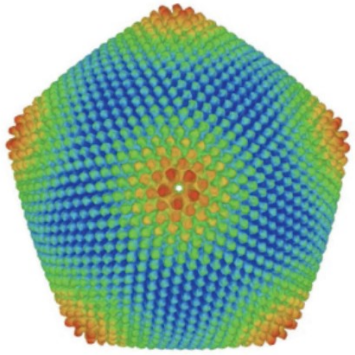
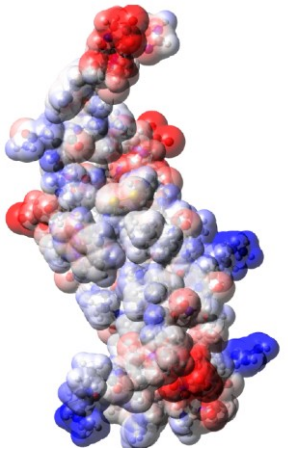
2D structure of a molecule



Radka Svobodová

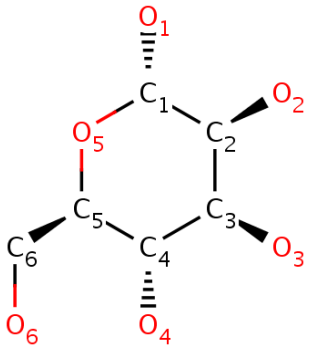


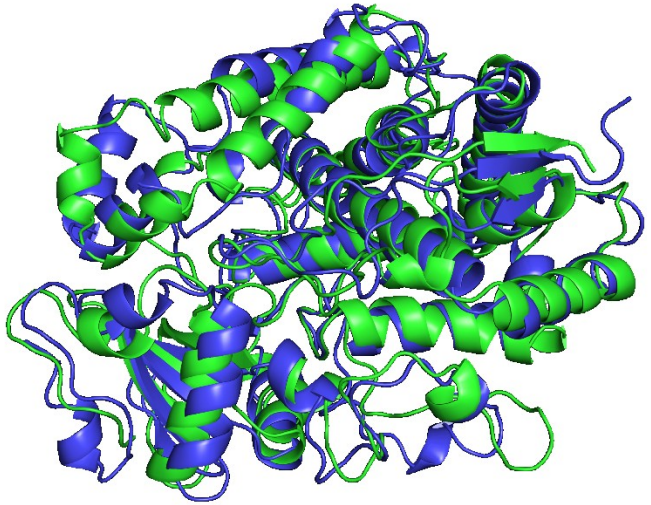




Content

- **Introduction:** concept of chemoinformatics, content of the subject, history of the field
- **Computer model of a molecule:** 1D, 2D and 3D structure, molecule representation using graph and matrix
- **2D structure (topology) of a molecule:**
  - writing a molecule using a string (SMILES, InChi, InChiKey)
  - **Molecular graphs:** Isomorphism and canonical indexing
- **3D structure (geometry) of the molecule:**
  - representation using Cartesian and internal coordinates, data formats, geometry comparison





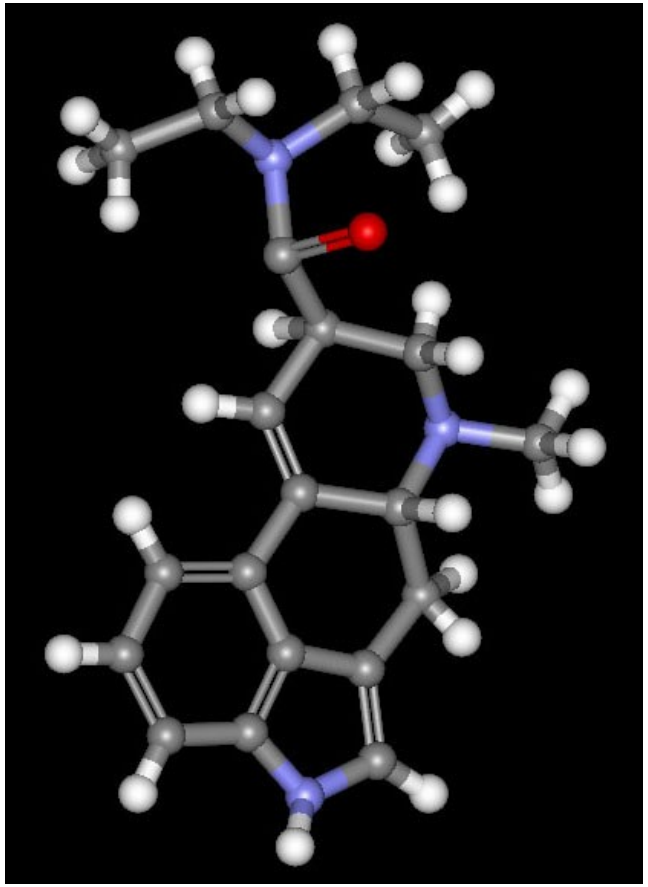
Model of molecule for computer processing

Atoms:

- Points in space
- Chemical symbol of the element listed for each

Bonds:

- Pairs of atoms that are bonded
- Bond order



Databases

Small molecules:

- PubChem
- DrugBank
- LigandExpo

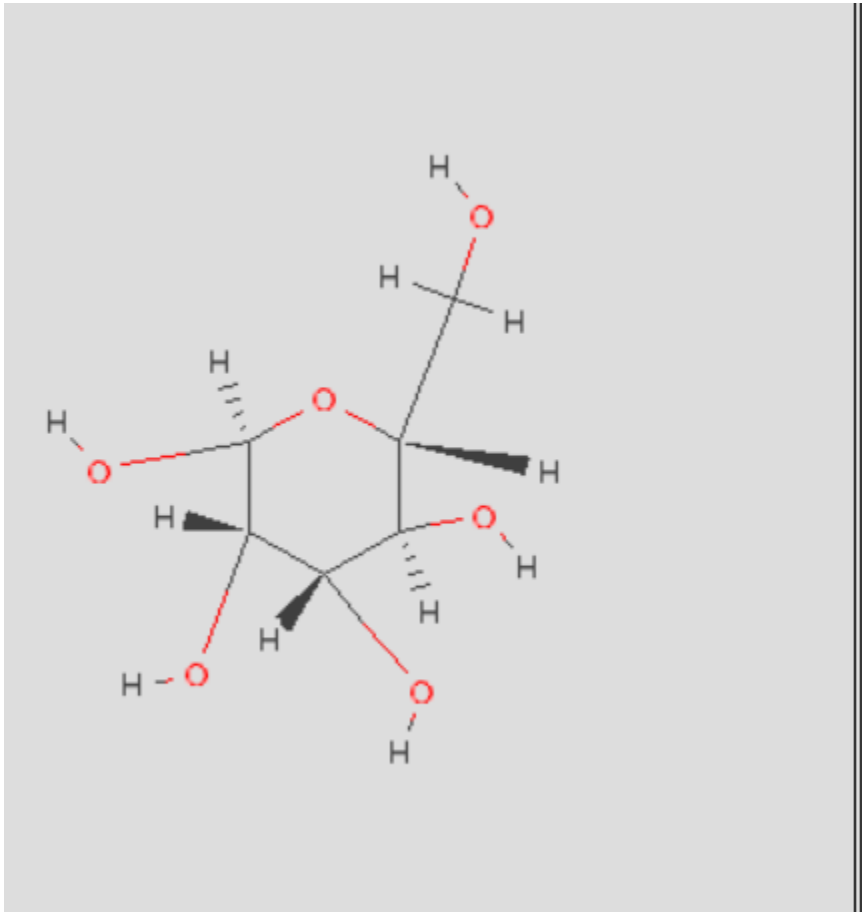
Proteins and nucleic acids:

- Protein Data Bank
- AlphaFold DB

Models of molecule for computations

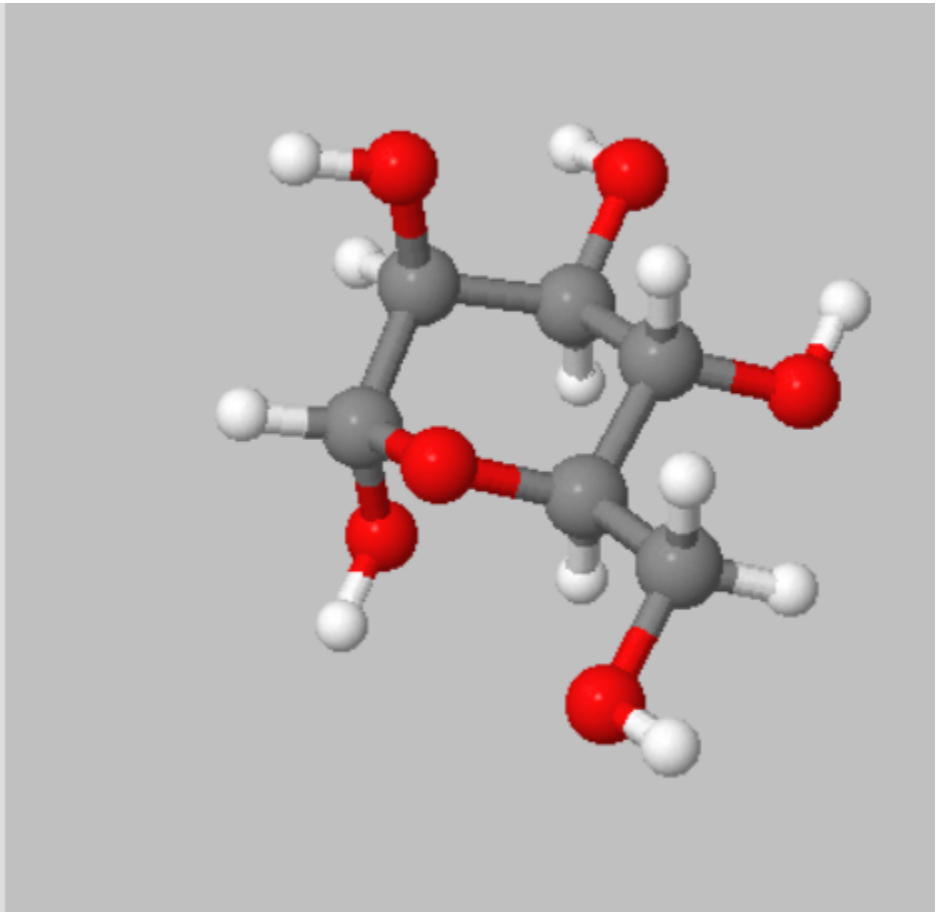
2D structure

topology



3D structure

geometry, conformation

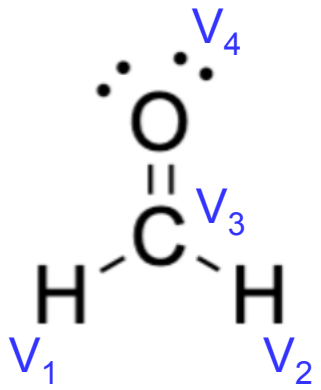


Data structures for storing of 2D structure

- Matrix
- Graph
- List of bonds
- Set of points

2D structure as a matrix

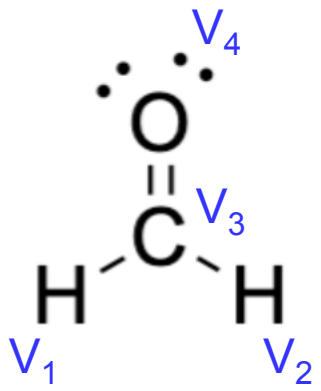
Molecule of formaldehyde:

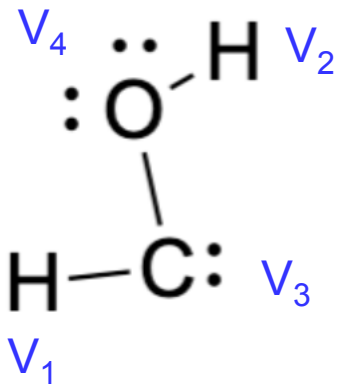


Neighbourhood matrix of formaldehyde:

$$A = \begin{array}{c|cccc|c} & & & & & \\ 0 & 0 & 1 & 0 & v_1 \\ 0 & 0 & 1 & 0 & v_2 \\ 1 & 1 & 0 & 2 & v_3 \\ 0 & 0 & 2 & 4 & v_4 \\ & & & & \\ \hline v_1 & v_2 & v_3 & v_4 & \end{array}$$

2D structure as a matrix
Chemical distance

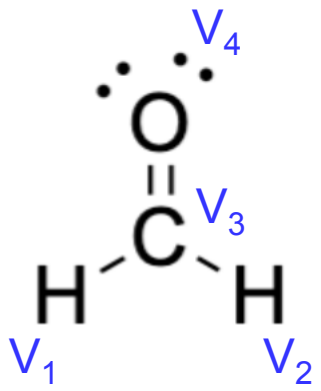


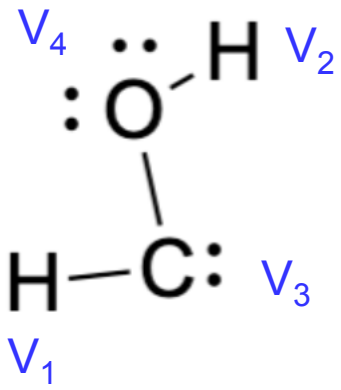


	v1	v2	v3	v4
v1	0	0	1	0
v2	0	0	1	0
v3	1	1	0	2
v4	0	0	2	4

	v1	v2	v3	v4
v1	0	0	1	0
v2	0	0	0	1
v3	1	0	2	1
v4	0	1	1	4

2D structure as a matrix
Chemical distance



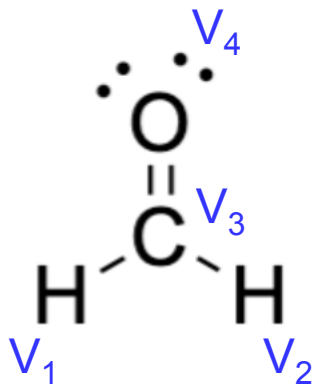


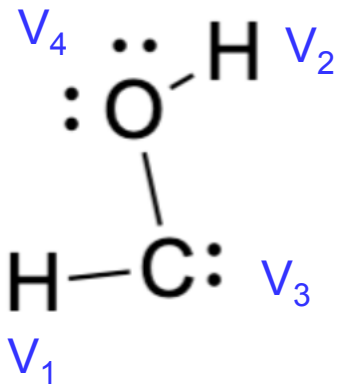
Chemical distance:The smallest number of valence electrons that must be moved to get the second molecule from the first molecule

	v1	v2	v3	v4
v1	0	0	1	0
v2	0	0	1	0
v3	1	1	0	2
v4	0	0	2	4

	v1	v2	v3	v4
v1	0	0	1	0
v2	0	0	0	1
v3	1	0	2	1
v4	0	1	1	4

2D structure as a matrix
Chemical distance





Chemical distance:The smallest number of valence electrons that must be moved to get the second molecule from the first molecule

	v1	v2	v3	v4
v1	0	0	1	0
v2	0	0	1	0
v3	1	1	0	2
v4	0	0	2	4

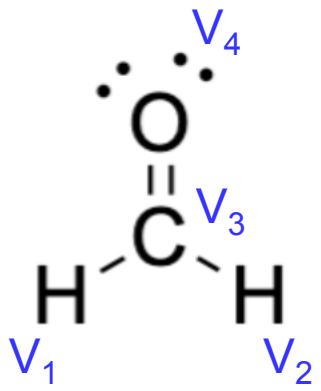
-

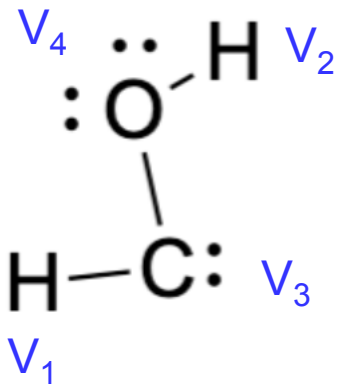
	v1	v2	v3	v4
v1	0	0	1	0
v2	0	0	0	1
v3	1	0	2	1
v4	0	1	1	4

=

$$\begin{vmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & 1 & -2 & 1 \\ 0 & -1 & 1 & 0 \end{vmatrix}$$

2D structure as a matrix
Chemical distance





Chemical distance:The smallest number of valence electrons that must be moved to get the second molecule from the first molecule

	v1	v2	v3	v4
v1	0	0	1	0
v2	0	0	1	0
v3	1	1	0	2
v4	0	0	2	4

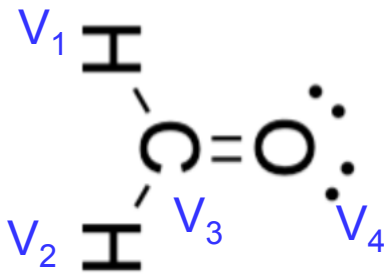
$-$

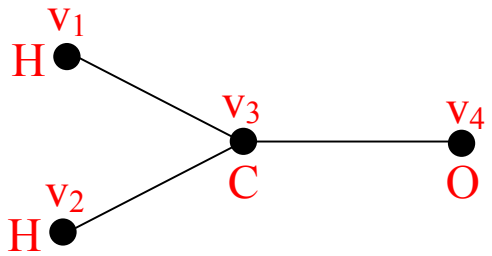
	v1	v2	v3	v4
v1	0	0	1	0
v2	0	0	0	1
v3	1	0	2	1
v4	0	1	1	4

$=$

$$\begin{vmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & 1 & -2 & 1 \\ 0 & -1 & 1 & 0 \end{vmatrix} = 8$$

2D structure as a graph





Atoms = **nodes**
Bonds = **edges**

Is it enough?
NO, we need more:

- **Element symbols**
- **Bond orders**
- **Free electrons**

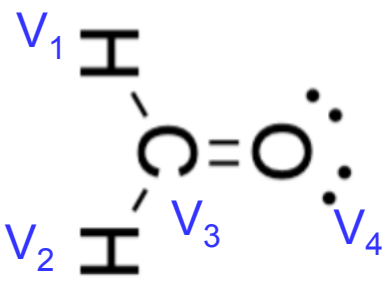
2D structure as a graph – molecular graph :-)

Definition of molecular graph:

$G = (V, E, L, \varphi, \beta)$

- V – set of nodes (atoms)
- E – multiset of edges (bonds)
- L – multiset of loops (free electron pairs)
- φ – function for naming of nodes by element symbols
- β – set of element symbols

Molecule:



Description of its mol. graph:

$V = \{v_1, v_2, v_3, v_4\}$

$E = \{\{v_1, v_3\}, \{v_2, v_3\}, \{v_3, v_4\}, \{v_3, v_4\}\}$

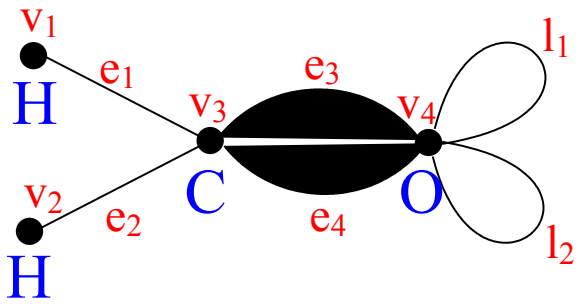
$L = \{\{v_4, v_4\}, \{v_4, v_4\}\}$

$\varphi(v_1) = H, \varphi(v_2) = H,$

$\varphi(v_3) = C, \varphi(v_4) = O$

$\beta = \{C, O, H\}$

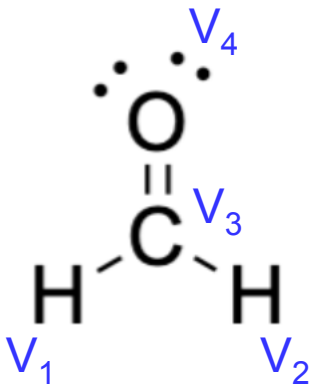
Figure of the molecular graph:



CEITEC

2D structure as a list of bonds

Molecule of formaldehyde:

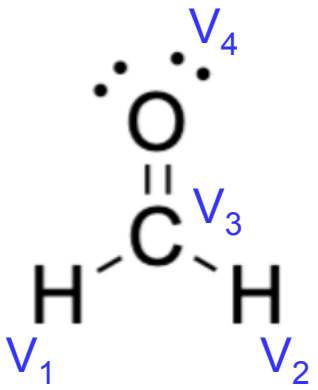


List of bonds:

1 3 1
2 3 1
3 4 2

2D structure as a set of points

Molecule of formaldehyde:



Set of points:

3.4030	-0.3700	0.0000 H
2.8660	0.5600	0.0000 H
2.8660	-0.0600	0.0000 C
2.0000	-0.5600	0.0000 O

2D structure – data formats

2D structure can be described in 3D data formats:

- SDF/MOL data format
- PDB data format
- mmCIF data format

2D structure has also specialised data formats:

- SMILES, SMIRKS, SSMARTS
- InChi, InChiKey
- CHUCKLES, CHORTLES, and CHARTS

MOL and SDF data format - example

number of atoms

number of bonds

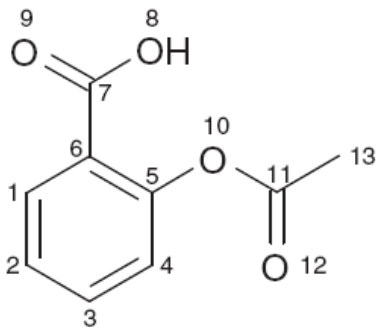
the first atom is a carbon

```
  -ISIS-  09270222202D

 13 13  0  0  0  0  0  0  0  0999 V2000
   -3.4639   -1.5375    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -3.4651   -2.3648    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -2.7503   -2.7777    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -2.0338   -2.3644    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -2.0367   -1.5338    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -2.7521   -1.1247    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -2.7545   -0.2997    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -2.0413    0.1149    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
   -3.4702    0.1107    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
   -1.3238   -1.1186    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
   -0.6125   -1.5292    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.6167   -2.3542    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    0.1000   -1.1125    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  1  2  2  0  0  0  0
  6  7  1  0  0  0  0
  3  4  2  0  0  0  0
  7  8  1  0  0  0  0
  7  9  2  0  0  0  0
  4  5  1  0  0  0  0
  5 10  1  0  0  0  0
  2  3  1  0  0  0  0
 10 11  1  0  0  0  0
  5  6  2  0  0  0  0
 11 12  2  0  0  0  0
  6  1  1  0  0  0  0
 11 13  1  0  0  0  0
M END
```

the first three numbers are the *x*, *y* and *z* coordinates of the atom

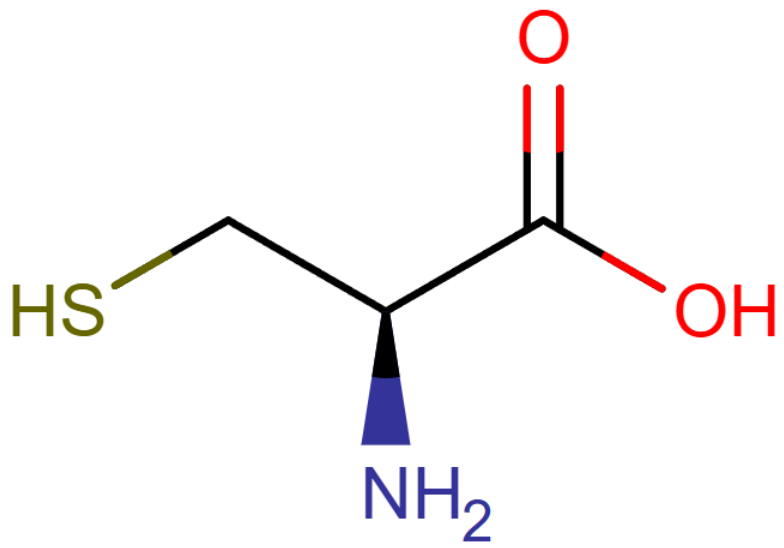
the first bond is between atoms 1 and 2 and has order 2



CEITEC

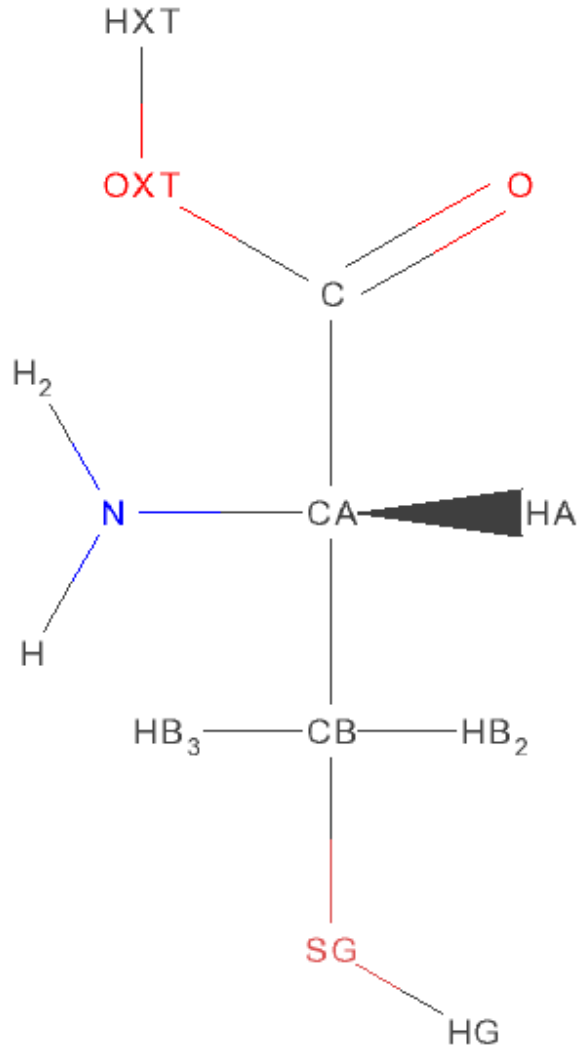
PDB data format - example
(from 2016 replaced by CIF/mmCIF formátem)

```
HETATM    1  C   CYS     0       1.143   0.440   0.000  0.00  0.00           C+0
HETATM    2  C   CYS     0      -0.191  -0.330   0.000  0.00  0.00           C+0
HETATM    3  C   CYS     0      -1.524   0.440   0.000  0.00  0.00           C+0
HETATM    4  O   CYS     0       2.477  -0.330   0.000  0.00  0.00           O+0
HETATM    5  O   CYS     0       1.143   1.980   0.000  0.00  0.00           O+0
HETATM    6  N   CYS     0      -0.191  -1.870   0.000  0.00  0.00           N+0
HETATM    7  S   CYS     0      -2.858  -0.330   0.000  0.00  0.00           S+0
CONECT    1    2    5    4
CONECT    2    1    3    6
CONECT    3    2    7
CONECT    4    1
CONECT    5    1
CONECT    6    2
CONECT    7    3
```



mmCIF data format - example

```
CYS N   N   N 0 1 N N N 22.585 13.716 37.715 1.585  0.483  -0.081 N   CY
CYS CA  CA  C 0 1 N N R 22.372 13.468 39.168 0.141  0.450  0.186  CA  CY
CYS C   C   C 0 1 N N N 21.806 14.686 39.893 -0.095 0.006  1.606  C   CY
CYS O   O   O 0 1 N N N 22.614 15.553 40.277 0.685  -0.742 2.143  O   CY
CYS CB  CB  C 0 1 N N N 23.683 13.019 39.828 -0.533 -0.530 -0.774 CB  CY
CYS SG  SG  S 0 1 N N N 25.202 13.440 38.921 -0.247 0.004  -2.484 SG  CY
CYS OXT OXT O 0 1 N Y N 20.565 14.747 40.076 -1.174 0.443  2.275  OXT CY
CYS H   H   H 0 1 N N N 22.963 12.902 37.230 1.928  -0.454 0.063  H   CY
CYS H2  HN2 H 0 1 N Y N 23.171 14.537 37.565 1.693  0.682  -1.065 H2  CY
CYS HA  HA  H 0 1 N N N 21.614 12.654 39.253 -0.277 1.446  0.042  HA  CY
CYS HB2 1HB H 0 1 N N N 23.739 13.412 40.869 -0.114 -1.526 -0.630 HB2 CY
CYS HB3 2HB H 0 1 N N N 23.651 11.923 40.031 -1.604 -0.554 -0.575 HB3 CY
CYS HG  HG  H 0 1 N N N 26.013 13.162 39.329 -0.904 -0.965 -3.145 HG  CY
CYS HXT HXT H 0 1 N Y N 20.212 15.505 40.527 -1.326 0.158  3.186  HXT CY
#
```



Description of 2D structure using 3D formats

Benefits:

- The most general notation for the structure of a molecule
- Easy to use as input for algorithms working with the structure

Disadvantages:

- Takes up a lot of space
- Not suitable for some special types of tasks.

Data formats specialised on 2D structure

- SMILES, SMIRKS, SSMARTS
- InChi, InChiKey
- CHUCKLES, CHORTLES, and CHARTS

Etc.

Chemical Description

Name	alpha-D-mannopyranose
Synonyms	alpha-D-mannose; D-mannose; mannose
Formula	C6 H12 O6
Formal charge	0
Molecular weight	180.156 g/mol
Component type	D-saccharide, alpha linking

Chemical features

Atom count	24
Chiral atom count	5
Chiral atoms	C1 C2 C3 C4 C5
Observed leaving atoms	O1 HO1 HO2 HO3 HO4 HO6
Bond count	24
Aromatic bond count	0

Chemical Identifiers

Systematic name (ACDLabs)	alpha-D-mannopyranose
Systematic name (OpenEye OEToolkits)	(2S,3S,4S,5S,6R)-6-(hydroxymethyl)oxane-2,3,4,5-tetrol

Chemical Descriptors

Stereo SMILES (CACTVS)	OC[C@H]1O[C@H](O)[C@@H](O)[C@@H](O)[C@@H]1O
SMILES (CACTVS)	OC[CH]1O[CH](O)[CH](O)[CH](O)[CH]1O
Stereo SMILES (OpenEye)	C([C@@H]1[C@H]([C@@H]([C@@H]([C@H](O1)O)O)O)O)O
InChI descriptor	InChI=1S/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5+,6+/m1/s1
InChIKey descriptor	WQZGKKKJIJFFOK-PQMKYFCFSA-N

PubChem Morphine (Compound)

CONTENTS

2.1.2 InChI

InChI=1S/C17H19NO3/c1-18-7-6-17-10-3-5-13(20)16(17)21-15-12(19)4-2-9(14(15)17)8-11(10)18/h2-5,10-11,13,16,19-20H,6-8H2,1H3/t10-,11+,13-,16-,17-/m0/s1

Computed by InChI 1.0.5 (PubChem release 2019.06.18)

PubChem

2.1.3 InChI Key

BQJCRHHNABKAKU-KBQPJGBKSA-N

Computed by InChI 1.0.5 (PubChem release 2019.06.18)

PubChem

2.1.4 Canonical SMILES

CN1CCC23C4C1CC5=C2C(=C(C=C5)O)OC3C(C=C4)O

Computed by OEChem 2.1.5 (PubChem release 2019.06.18)

PubChem

2.1.5 Isomeric SMILES

CN1CC[C@]23[C@@H]4[C@H]1CC5=C2C(=C(C=C5)O)O[C@H]3[C@H](C=C4)O

Computed by OEChem 2.1.5 (PubChem release 2019.06.18)

SMILES

SMILES means the following:

Simplified Molecular Input Line Entry Specification= encoding the structure of a molecule into a string.

Below I will give a brief description of SMILES.

More detailed information can be found e.g. here:http://www.daylight.com/dayhtml/smiles/

SMILES
Syntax of atoms

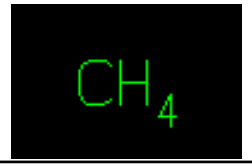
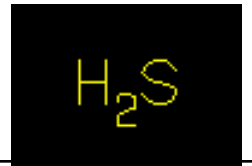
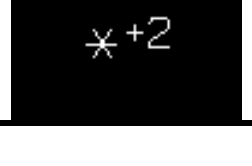
Syntax in SMILES language:

atom : '[' <mass> symbol <chiral> <hcount> <sign<charge>> ']' ;

Description:

symbol	element symbol
	* = any atom
<sign<charge>>	sign and charge
<mass>	atom weight
<chiral>	chirality
<hcount>	number of bound atoms

SMILES – atoms - example

Figure	SMILES string	Description
	[S]	Element sulphur
	C	Methan
	S	Hydrogen sulphide
	[OH-], [OH-1]	Hydroxide anion
	[235U]	Uran isotoph 235
	[*+2]	Any atom with charge 2+

SMILES
Syntax of bonds

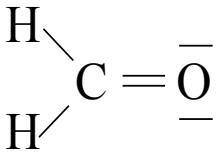
Syntax in SMILES:

bond : *<empty>* | '-' | '=' | '#' | ':' ;

Popis:

<empty>	any bond
-	simple bond
=	double bond
#	triple bond
:	aromatic bond

SMILES – bonds - example

Figure	SMILES string	Description
CH_3-CH_3	CC, C-C, [CH3]-[CH3]	Ethan
	C=O, O=C	Formaldehyde
H-C≡N	C#N, N#C	Hydrogen cyanide
CH_2=CH_2	C=C (lze i cc)	Ethen
CH_2=CH-CH=CH_2	C=C-C=C (lze i cccc)	1,3-butadien
?	ccc	Undefined bond type

SMILES – braneches - syntax

Syntax in smiles:

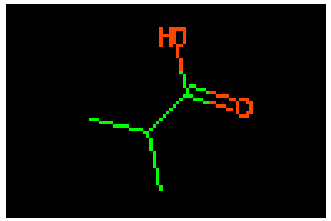
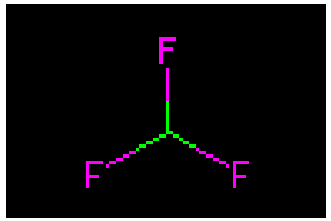
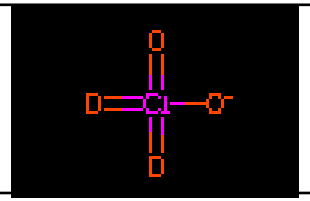
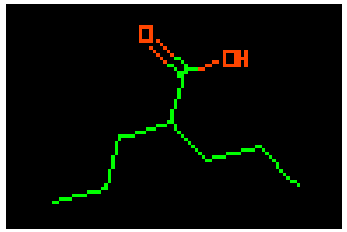
branch : '(' <chain> ')'
| '(' <chain> <branch> ')'
| '(' <branch> <chain> ')'
| '(' <chain> <branch> <chain> ')' ;

Description:

<chain> string

<branch> branch

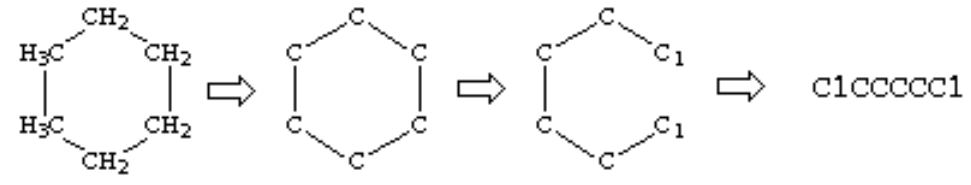
SMILES – branches - examples

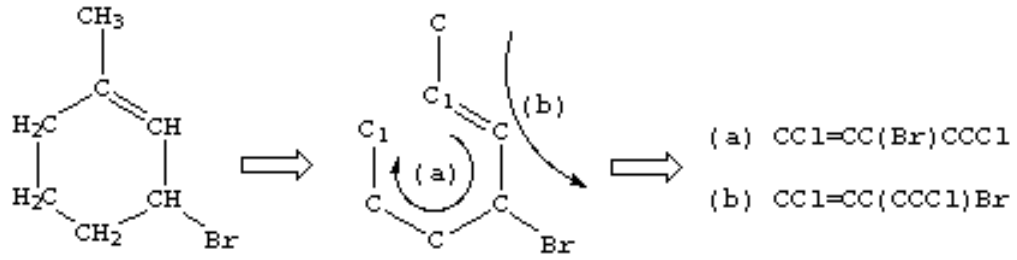
Figure	SMILES string	Description
	CC(C)C(=O)O	Isobutan acid
	FC(F)F, C(F)(F)F	Fluoroform
	O=Cl(=O)(=O)[O-], Cl(=O)(=O)(=O)[O-]	Perchlorat anion
	CCCC(C(=O)O)CCC	4-heptan acid

SMILES – cycles

Choose any bond in the cycle and denote its terminal atoms by a number. We break the cycle at the site of the bond and write it as a linear sequence of atoms.

Examples:





SMILES - summary

Benefits of SMILES:

- Space compression
- Possibility to write a molecule using a regular expression

Disadvantages of SMILES:

- Ambiguity (there is no "correct" order of atoms, 1 fact can be written in multiple ways).
- Necessity to create a complete listing before any algorithm can be applied to the molecule (isomorphism, cycles, etc.)

SMILES – applications

Use of SMILES:

- Nomenclature and automatic name generation.
- Searching parts of molecules using regular expressions.

SMILES extension:

A more advanced version of SMILES strings are SMARTS strings. They are defined in the same way as SMILES + contain additional rules.

More about SMARTS:

http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html

Exercise

Find the SMILES entry for aspirin.

Chemical Description

Name	alpha-D-mannopyranose
Synonyms	alpha-D-mannose; D-mannose; mannose
Formula	C6 H12 O6
Formal charge	0
Molecular weight	180.156 g/mol
Component type	D-saccharide, alpha linking

Chemical features

Atom count	24
Chiral atom count	5
Chiral atoms	C1 C2 C3 C4 C5
Observed leaving atoms	O1 HO1 HO2 HO3 HO4 HO6
Bond count	24
Aromatic bond count	0

Chemical Identifiers

Systematic name (ACDLabs)	alpha-D-mannopyranose
Systematic name (OpenEye OEToolkits)	(2S,3S,4S,5S,6R)-6-(hydroxymethyl)oxane-2,3,4,5-tetrol

Chemical Descriptors

Stereo SMILES (CACTVS)	OC[C@H]1O[C@H](O)[C@@H](O)[C@@H](O)[C@@H]1O
SMILES (CACTVS)	OC[CH]1O[CH](O)[CH](O)[CH](O)[CH]1O
Stereo SMILES (OpenEye)	C([C@@H]1[C@H]([C@@H]([C@@H]([C@H](O1)O)O)O)O)O
InChI descriptor	InChI=1S/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5+,6+/m1/s1
InChIKey descriptor	WQZGKKKJIJFFOK-PQMKYFCFSA-N

InChI

- InChI = IUPAC International Chemical Identifier
- this format is unique for all chemical substances
- created as an IUPAC standard in 2005
- freely usable and redistributable under the LGPL license
- stores more information than SMILES
- is still readable by a person with sufficient practice

InChI – which information stores?

- atoms and bonds
- tautomerism
- isometrism
- stereometrism
- electric charges

How to translate molecule to InChI?

The algorithm for converting the structure to InChI is performed in three stages:

- **Normalization** - in this stage all redundant information is removed.
- **Canonicalization** - in this step, a unique number is assigned to each atom
- The last stage is **serialization**, which generates a string of characters.

InChI layers

- Each InChI is quoted by the string "InChI=„
- This is followed by the version number used (currently "1")
- This is followed by the letter S if this InChI meets the standard.
- The remaining information is divided into six layers and sublayers
- Each of these layers contains different specific information
- The layer separator is "/" and starts with a characteristic prefix, except for the main layer.

Main layer

Must be included in every InChI

Summary formula: carbons are written first, then hydrogens, then other atoms, which are in alphabetical order

Bonds of atoms (prefix: "c"): describes the bonds between atoms in the order they were numbered, these bonds are contained only once

Hydrogen atom bonds (prefix: "h"): describes to which atoms the hydrogen atoms are bonded

Charge layer

Protons (prefix: "p"): used when there are positive charges in the molecule

Electrons (prefix: "q"): used when there are negative charges in the molecule.

Stereochemical layer

- double bonds and cumulenes (prefix: "b")
- tetrahedral stereometry of atoms and allenes (prefix: "t", "m")
- other type of stereometric information (prefix: "s").

Izotop layer

- (prefix: "i", "h")
- It also uses stereometric layer prefixes when isotopic stereochemistry is involved.

Firm layer H

- (prefix: "f")
- This layer is no longer used in standard InChI because it accumulated information from the above layers.

Reconnectable layer

- (prefix: "r")
- This layer is no longer used in standard InChI because it accumulated information from the above layers.

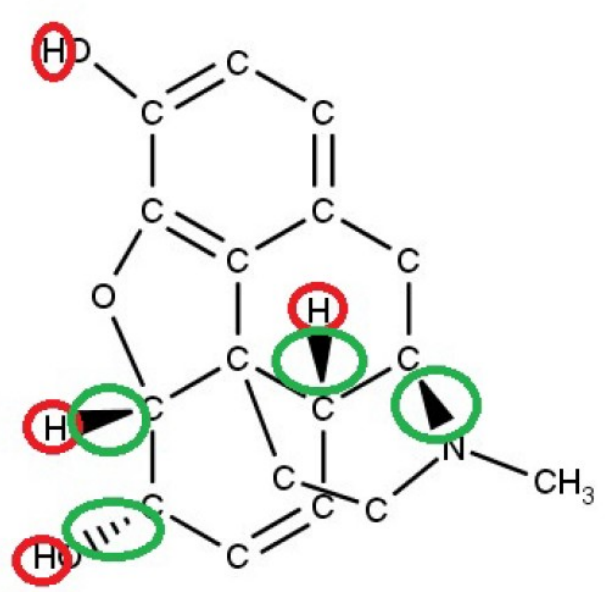
Example

InChI of Morphine :

InChI=1S/C17H19NO3/c1-18-7-6-17-10-3-5-13(20)16(17)21-15-12(19)4-2-9(14(15)17)8-11(10)18/h2-5,10-11,13,16,19-20H,6-8H2,1H3/t10-,11+,13-,16-,17-/m0/s1.

Red text coincides with the hydrogens in the pictureThis part is part of the main layer and must always be present.

Green text are stereometric information - they are marked with green circles



InChIkey

- InChI is a very long identifier of undeclared length
- this complicates its storage and further work with it
- an alternative has been developed based on it
- a condensed 27 character long InChIKey
- HASHInChI (SHA-256 algorithm)
- If an InChIKey is created from an InChI that is standard, the InChIKey is also standard
- Due to the algorithm used, there is very little chance of duplication between structures.

InChIkey

- InChIKey is divided into several parts:
  AAAAAAAAAAAAAAAAAA-BBBBBBBBBBFV-P
- "A" denotes the first 14 characters and is created by hashing the binding information of the molecule. It is terminated by a hyphen.
- "B" describes the next eight characters and is created by hashing the rest of the InChI
- "F" is followed by a character identifying the type of InChIKey
- "V" is the version identifier, currently
- "A" is used for the first version, with future plans to continue the alphabet for other versions
- "P" is the proton identifier

Example:
InChIKey of morphine:
BQJCRHHNABKAKU-KBQPJGBKSA-N

Exercise

Find InChIKey of aspirine

Chemical Description

Name	alpha-D-mannopyranose
Synonyms	alpha-D-mannose; D-mannose; mannose
Formula	C6 H12 O6
Formal charge	0
Molecular weight	180.156 g/mol
Component type	D-saccharide, alpha linking

Chemical features

Atom count	24
Chiral atom count	5
Chiral atoms	C1 C2 C3 C4 C5
Observed leaving atoms	O1 HO1 HO2 HO3 HO4 HO6
Bond count	24
Aromatic bond count	0

Chemical Identifiers

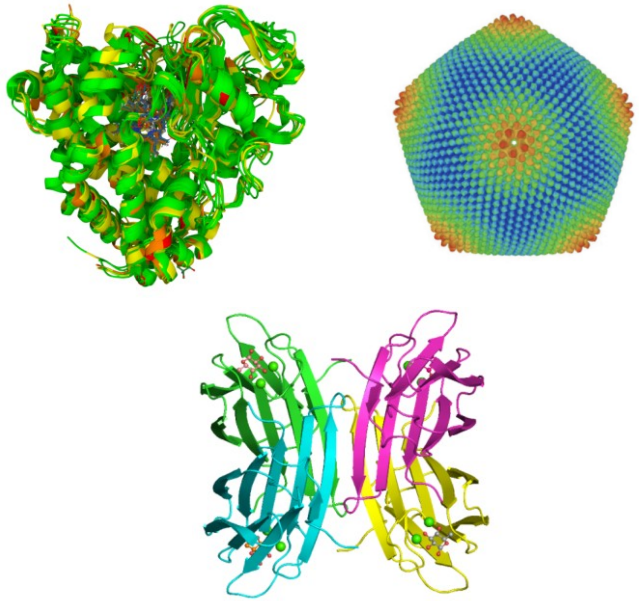
Systematic name (ACDLabs)	alpha-D-mannopyranose
Systematic name (OpenEye OEToolkits)	(2S,3S,4S,5S,6R)-6-(hydroxymethyl)oxane-2,3,4,5-tetrol

Chemical Descriptors

Stereo SMILES (CACTVS)	OC[C@H]1O[C@H](O)[C@@H](O)[C@@H](O)[C@@H]1O
SMILES (CACTVS)	OC[CH]1O[CH](O)[CH](O)[CH](O)[CH]1O
Stereo SMILES (OpenEye)	C([C@@H]1[C@H]([C@@H]([C@@H]([C@H](O1)O)O)O)O)O
InChI descriptor	InChI=1S/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5+,6+/m1/s1
InChIKey descriptor	WQZGKKKJIJFFOK-PQMKYFCFSA-N

Thank you for your attention

Bioinformatics



Tools
Databases
AI

Chemoinformatics