# C7790 Introduction to Molecular Modelling
# TSM Modelling Molecular Structures

## Lesson 10
## Structure

**PS/2021 Present Form of Teaching: Rev2**

Petr Kulhánek

kulhanek@chemi.muni.cz

National Centre for Biomolecular Research, Faculty of Science
Masaryk University, Kamenice 5, CZ-62500 Brno

# Context

**macroworld** **microworld**

**states**

(thermodynamic properties, G, T,…)

**phenomenological thermodynamics**

equilibrium (equilibrium constant)
kinetics (rate constant)

free energy
(Gibbs/Helmholtz)

partition function

**statistical thermodynamics**

**microstates**

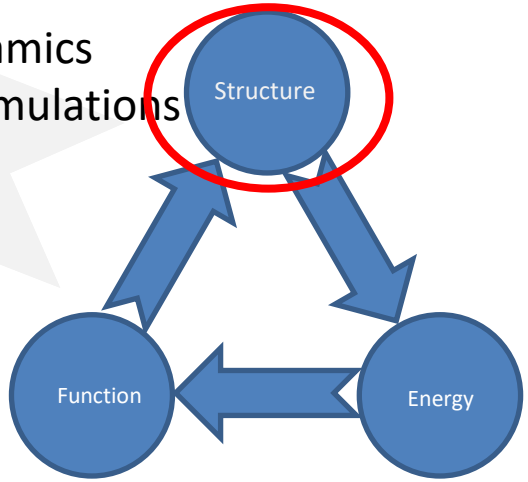
(mechanical properties, E)

**Description levels (model chemistry):**

- quantum mechanics
  - semiempirical methods
  - ab initio methods
  - post-HF methods
  - DFT methods
- molecular mechanics
- coarse-grained mechanics

**Simulations:**

- molecular dynamics
- Monte Carlo simulations
- docking
- …




microstate ≠ microworld

# Structure

# Configuration Space

$E(\mathbf{R})$

R = point in 3N dimensional space (N is the number of atoms)

$$\mathbf{R} = \{x_1, y_1, z_1, x_2, y_2, z_2, \ldots, x_N, y_N, z_N\}$$

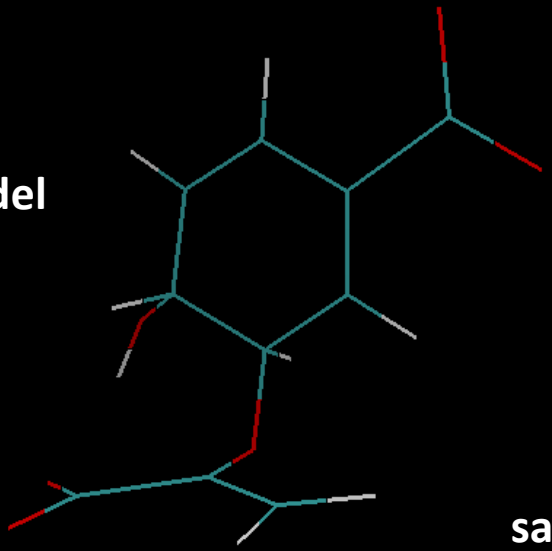
Cartesian coordinates of the first atom

Cartesian coordinates of the last atom

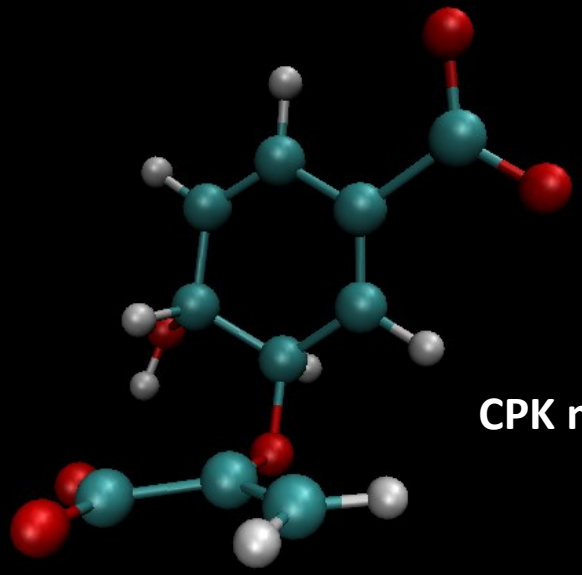
**Every point** in the configuration space represents a **unique structure** of the system.

# Models - Small Molecules

line model

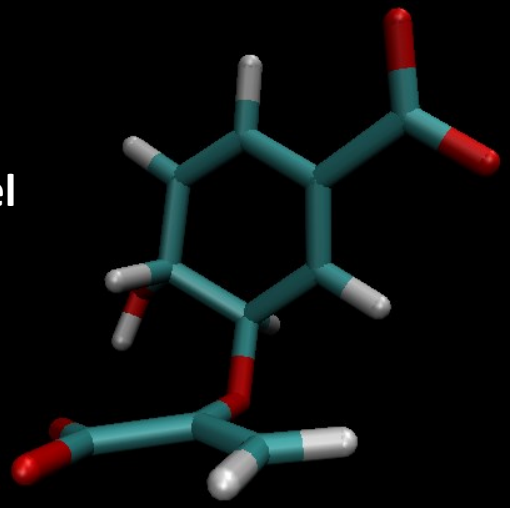


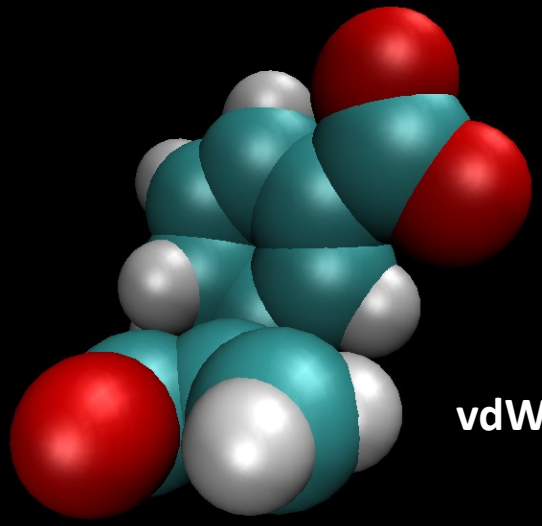
same structure
other visualization



CPK model

tube model

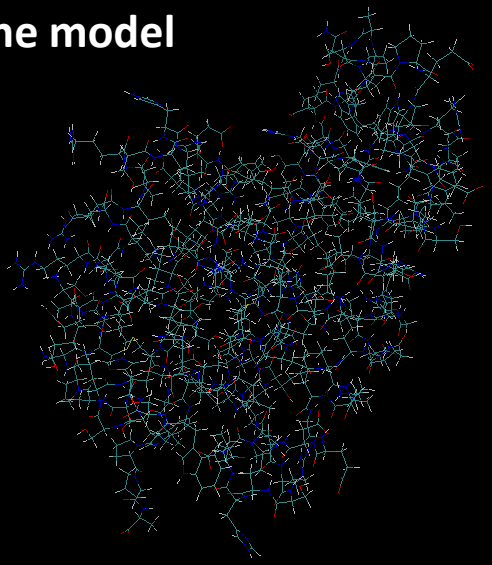




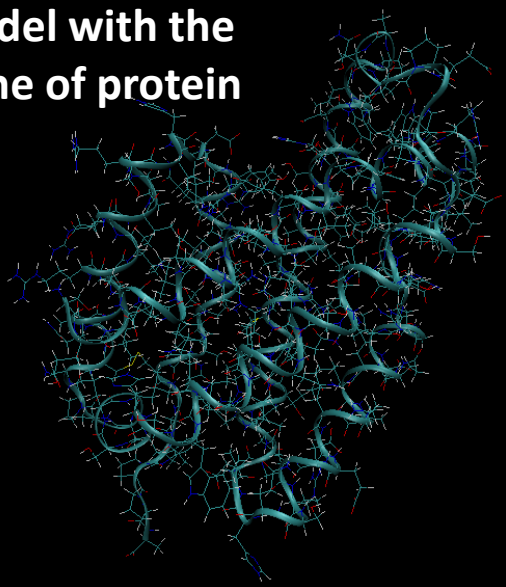
vdW model

# Models - Biomolecules

line model

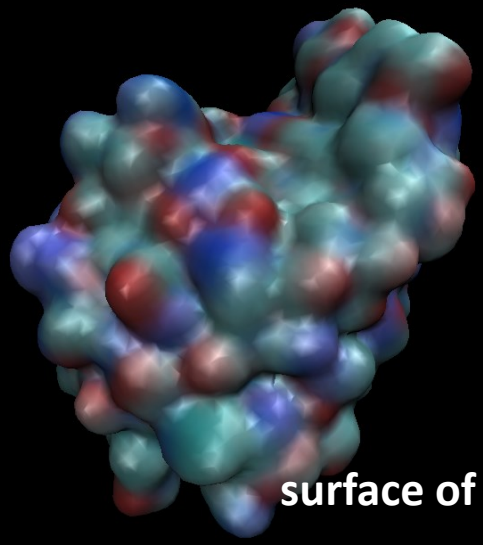


line model with the backbone of protein



Cartoon model

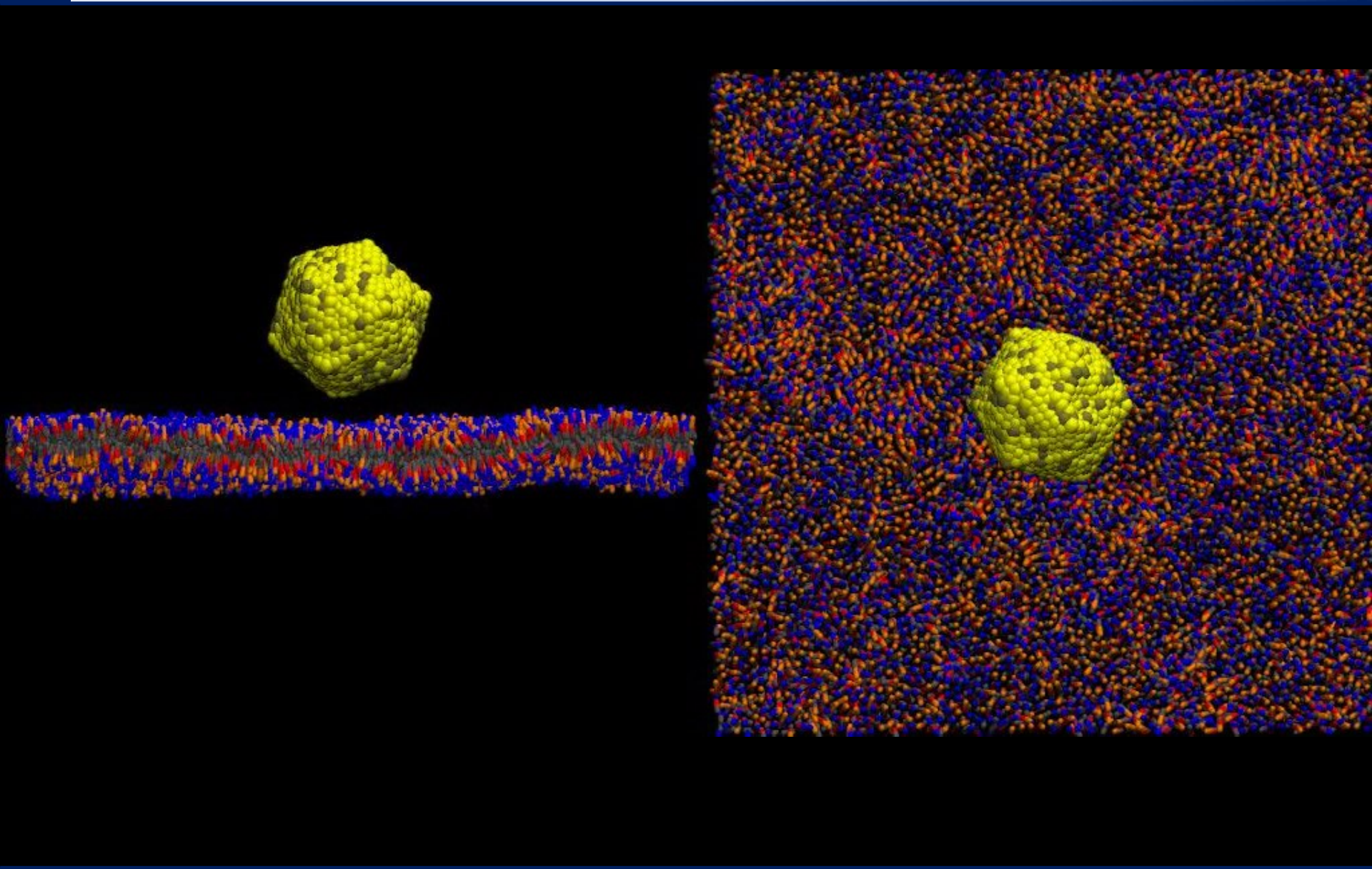
same structure other visualization



surface of the biomolecule

Different models are used to highlight certain structural information or internal properties of a molecule or group of molecules, which facilitates an easier understanding of the studied problem.

# Coarse-grained Models



# Computer Representation of Models

The structure can be represented in various ways. More than 100 formats are used in chemistry. They are either text or binary files.

The format describes:

- the geometry of the system
- the names of atoms
- groups of atoms
- connectivity between atoms (bonds)
- and other information

The system geometry is usually provided as:

- Cartesian coordinates
- internal coordinates
- variants of internal coordinates

# Cartesian vs Internal Coordinates

## Cartesian coordinates

```
O -0.180077 -0.046023 -0.062789
H  0.196208 -0.747659  0.498793
O  0.006537  1.047922  0.877207
H -0.931885  1.299156  0.951390
```

x y z

Number of degrees of freedom:

**3N**

## Internal coordinates (Z-matrix)

```
O
H 1 0.974298
O 1 1.454349 2 96.868054
H 3 0.974298 1 96.868054 2 239.552651
```

bond length bond angle torsion angle

Number of degrees of freedom:

**3N-6**

**3N-5** (linear diatomic molecule)

# Internal Coordinates

bond length (a) bond angle (b) torsion angle (c)

```
1 O
2 H 1 0.974298
3 O 1 1.454349 2 96.868054
4 H 3 0.974298 1 96.868054 2 239.552651
```

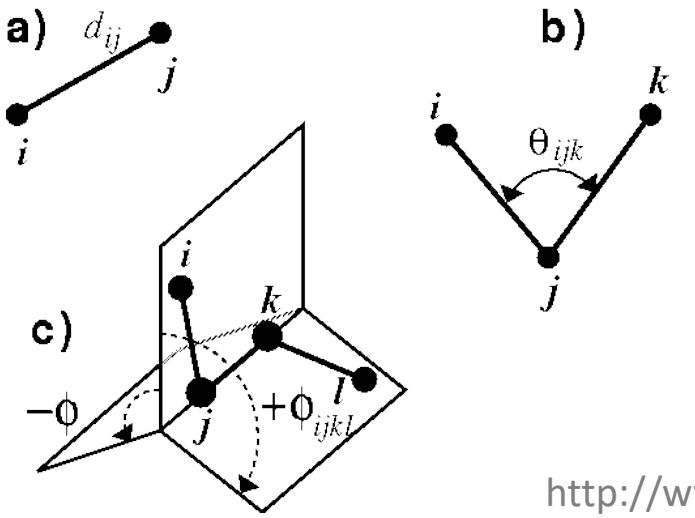
2-1

3-1-2

4-3-1-2

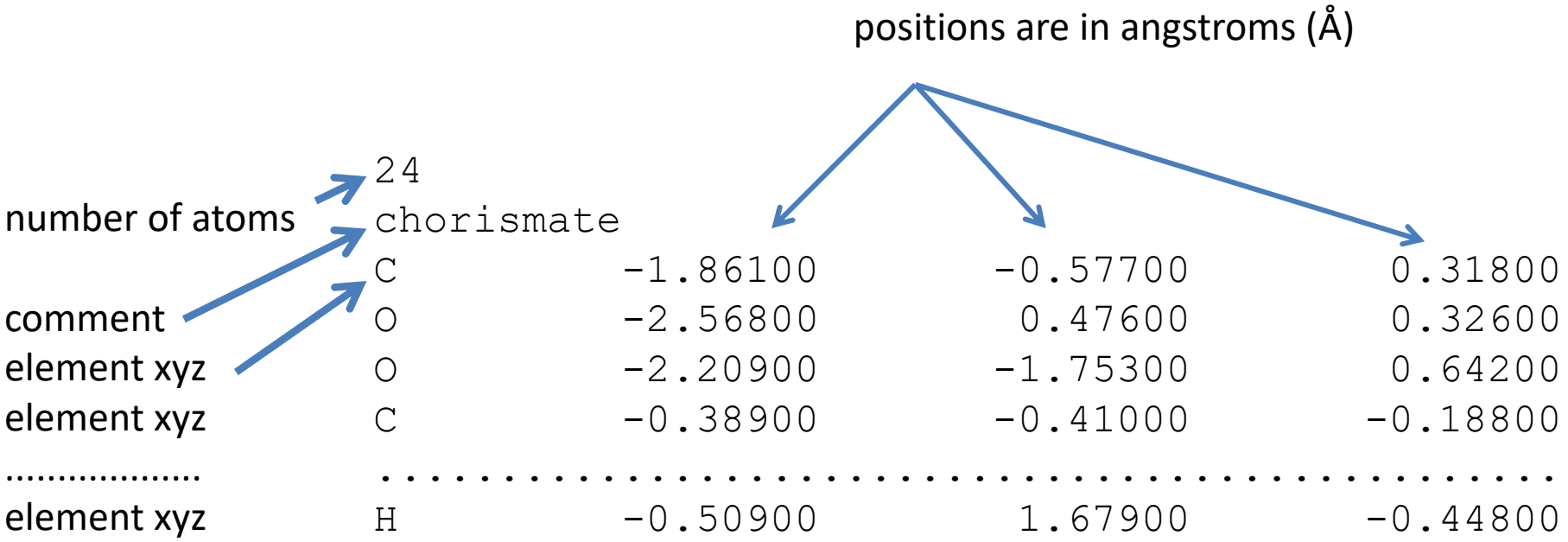
4-3

4-3-1




http://www.ccl.net/cca/documents/molecular-modeling/node4.html

# XYZ format

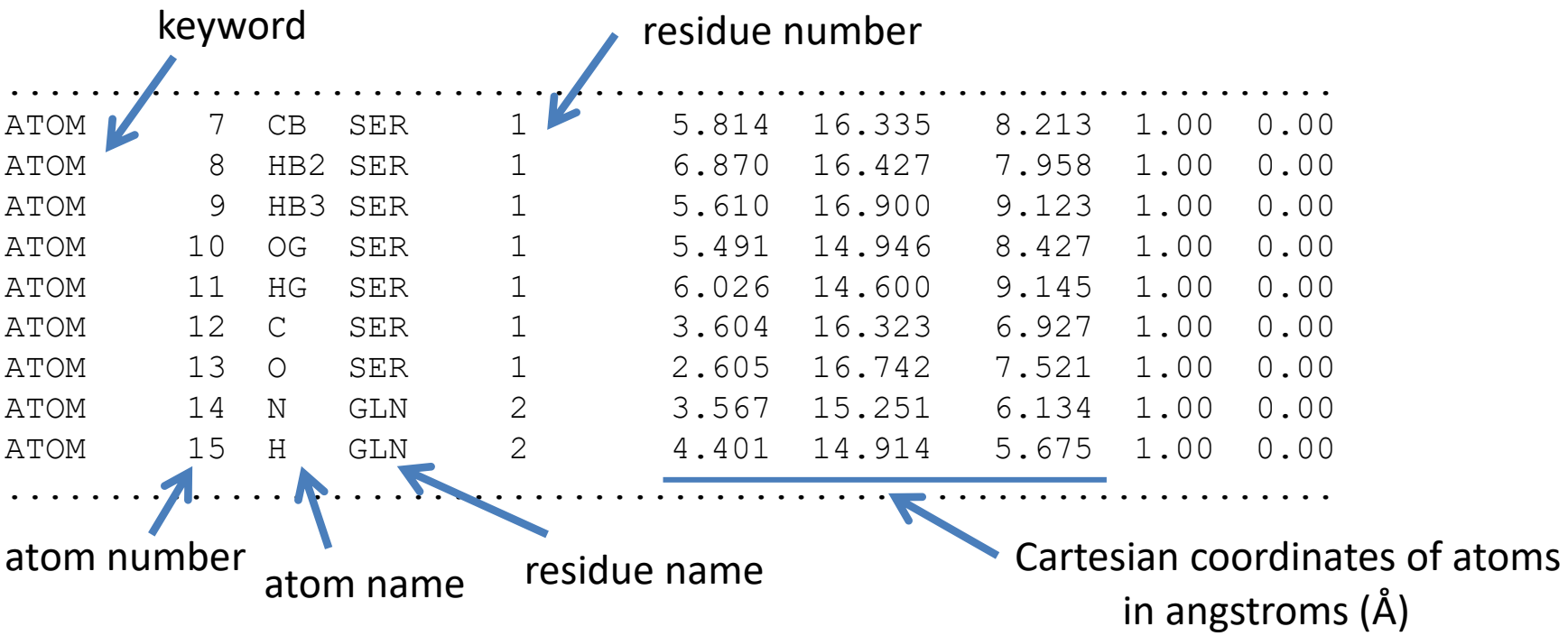



The **xyz** format is a free-formatting text file (values in columns can be separated by any number of spaces or other whitespace).

The format only describes the geometry of the system. It does not contain information about bonds in the system. A program that works with the format must calculate this information (e.g., using atomic radii).

# PDB format

The **pdb** format is employed to store the structures of biomolecules and their complexes. **It is widely used but it has several limitations.** Therefore, it is slowly substituted by more advanced formats such as PDBx/mmCIF and others.




The pdb format does not usually contain information about bods in the system. The program that works with the format must calculate this information (based on template structures). For non-standard residues, the CONECT keyword can be used.

# Djungle of formats I

| | |
|---|---|
| acr | -- ACR format |
| adf | -- ADF cartesian input format |
| adfout | -- ADF output format |
| alc | -- Alchemy format |
| arc | -- Accelrys/MSI Biosym/Insight II CAR format |
| bgf | -- MSI BGF format |
| box | -- Dock 3.5 Box format |
| bs | -- Ball and Stick format |
| c3d1 | -- Chem3D Cartesian 1 format |
| c3d2 | -- Chem3D Cartesian 2 format |
| cac | -- CAChe MolStruct format |
| caccrt | -- Cacao Cartesian format |
| cache | -- CAChe MolStruct format |
| cacint | -- Cacao Internal format |
| can | -- Canonical SMILES format. |
| car | -- Accelrys/MSI Biosym/Insight II CAR format |
| ccc | -- CCC format |
| cdx | -- ChemDraw binary format |
| cdxml | -- ChemDraw CDXML format |
| cht | -- Chemtool format |
| cif | -- Crystallographic Information File |
| ck | -- ChemKin format |
| cml | -- Chemical Markup Language |
| cmlr | -- CML Reaction format |
| com | -- Gaussian 98/03 Input |
| copy | -- Copies raw text |
| crk2d | -- Chemical Resource Kit diagram(2D) |
| crk3d | -- Chemical Resource Kit 3D format |
| csr | -- Accelrys/MSI Quanta CSR format |
| cssr | -- CSD CSSR format |
| ct | -- ChemDraw Connection Table format |
| cub | -- OpenDX cube format for APBS |
| cube | -- OpenDX cube format for APBS |
| dmol | -- DMol3 coordinates format |
| dx | -- OpenDX cube format for APBS |
| ent | -- Protein Data Bank format |
| fa | -- FASTA format |
| fasta | -- FASTA format |
| fch | -- Gaussian formatted checkpoint file format |
| fchk | -- Gaussian formatted checkpoint file format |
| fck | -- Gaussian formatted checkpoint file format |
| feat | -- Feature format |
| fh | -- Fenske-Hall Z-Matrix format |
| fix | -- SMILES FIX format |
| fpt | -- Fingerprint format |
| fract | -- Free Form Fractional format |
| fs | -- FastSearching |
| fsa | -- FASTA format |
| g03 | -- Gaussian98/03 Output |
| g92 | -- Gaussian98/03 Output |
| g94 | -- Gaussian98/03 Output |
| g98 | -- Gaussian98/03 Output |
| gal | -- Gaussian98/03 Output |
| gam | -- GAMESS Output |
| gamin | -- GAMESS Input |
| gamout | -- GAMESS Output |

# Djungle of formats II

| | |
|---|---|
| gau | -- Gaussian 98/03 Input |
| gjc | -- Gaussian 98/03 Input |
| gjf | -- Gaussian 98/03 Input |
| gpr | -- Ghemical format |
| gr96 | -- GROMOS96 format |
| gukin | -- GAMESS-UK Input |
| gukout | -- GAMESS-UK Output |
| gzmat | -- Gaussian Z-Matrix Input |
| hin | -- HyperChem HIN format |
| inchi | -- InChI format |
| inp | -- GAMESS Input |
| ins | -- ShelX format |
| jin | -- Jaguar input format |
| jout | -- Jaguar output format |
| k | -- Compare molecules using InChI |
| mcdl | -- MCDL format |
| mcif | -- Macromolecular Crystallographic Information |
| mdl | -- MDL MOL format |
| ml2 | -- Sybyl Mol2 format |
| mmcif | -- Macromolecular Crystallographic Information |
| mmd | -- MacroModel format |
| mmod | -- MacroModel format |
| mol | -- MDL MOL format |
| mol2 | -- Sybyl Mol2 format |
| molden | -- Molden input format |
| molreport | -- Open Babel molecule report |
| moo | -- MOPAC Output format |
| mop | -- MOPAC Cartesian format |
| mopcrt | -- MOPAC Cartesian format |
| mopin | -- MOPAC Internal |
| mopout | -- MOPAC Output format |
| mpc | -- MOPAC Cartesian format |
| mpd | -- Sybyl descriptor format |
| mpqc | -- MPQC output format |
| mpqcin | -- MPQC simplified input format |
| msi | -- Accelrys/MSI Cerius II MSI format |
| msms | -- M.F. Sanner's MSMS input format |
| nw | -- NWChem input format |
| nwo | -- NWChem output format |
| outmol | -- DMol3 coordinates format |
| pc | -- PubChem format |
| pcm | -- PCModel Format |
| pdb | -- Protein Data Bank format |
| png | -- PNG files with embedded data |
| pov | -- POV-Ray input format |
| pqr | -- PQR format |
| pqs | -- Parallel Quantum Solutions format |
| prep | -- Amber Prep format |
| qcin | -- Q-Chem input format |
| qcout | -- Q-Chem output format |
| report | -- Open Babel report format |
| res | -- ShelX format |
| rsmi | -- Reaction SMILES format |
| rxn | -- MDL RXN format |
| sd | -- MDL MOL format |
| sdf | -- MDL MOL format |

# Djungle of formats III

| | | | |
|---|---|---|---|
| smi | -- SMILES format | txyz | -- Tinker MM2 format |
| smiles | -- SMILES format | unixyz | -- UniChem XYZ format |
| sy2 | -- Sybyl Mol2 format | vmol | -- ViewMol format |
| t41 | -- ADF TAPE41 format | xed | -- XED format |
| tdd | -- Thermo format | xml | -- General XML format |
| test | -- Test format | xtc | -- XTC format |
| therm | -- Thermo format | xyz | -- XYZ cartesian coordinates format |
| tmol | -- TurboMole Coordinate format | yob | -- YASARA.org YOB format |
| txt | -- Title format | zin | -- ZINDO input format |

The formats usually contain, in addition to the 3D/2D structure, also accompanying information such as connectivity, force field parameters, atomic partial charges, various properties, etc.

**How to convert?**

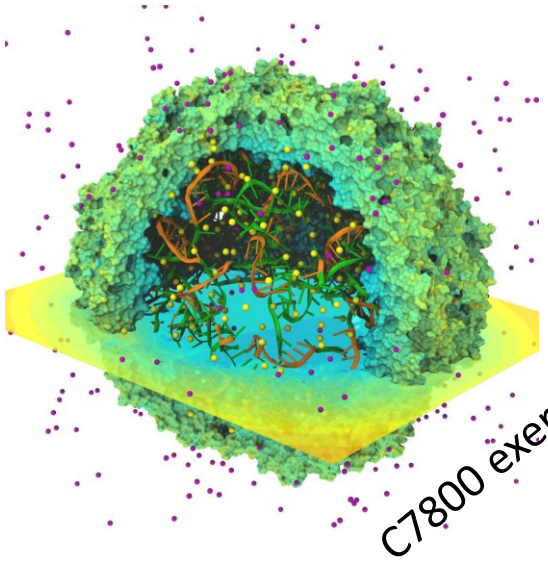
**OpenBabel** is a chemical toolbox designed to speak the many languages of chemical data. It's an open, collaborative project allowing anyone to search, convert, analyze, or store data from molecular modeling, chemistry, solid-state materials, biochemistry, or related areas.

**http://openbabel.org/wiki/Main_Page**

# Software for visualizations

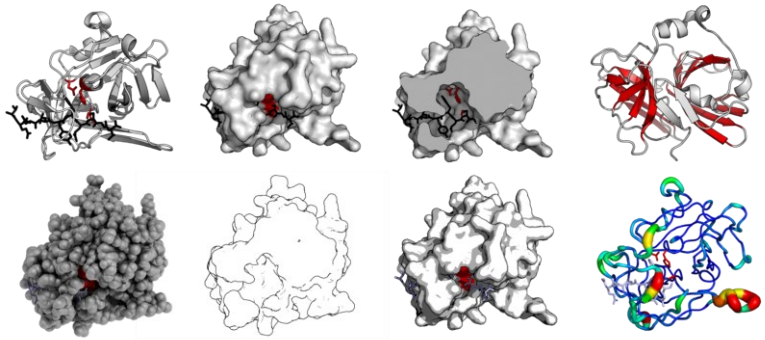
In addition to molecular modelling software (Avogadro, Nemesis, etc.), there are special software serving only for visualizing structures and results.

## Visual Molecular Dynamics (VMD)




https://www.ks.uiuc.edu/Research/vmd

## PyMOL



https://en.wikipedia.org/wiki/PyMOL

- scriptable (TCL, Python)
- advanced rendering
- available for MS Windows, Linux, macOS

## Overview of software:

https://en.wikipedia.org/wiki/List_of_molecular_graphics_systems

# Summary

- Structures (Models) can be visualized in different ways.
- Visualization type is typically based on the intended description of studied phenomenon/property.
- Geometry can be represented in Cartesian and/or internal coordinates.
- Computational chemistry (molecular modelling) employs huge number of formats describing models, which complicates interoperability between different software.

Show some movies ☺

**Homework:**

Is there a principal advantage of using Cartesian or internal coordinates?