

C7790 Introduction to Molecular Modelling

TSM Modelling Molecular Structures

C9087 Computational Chemistry for Structural Biology

Lesson 10 Structure

JS/2022 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

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)

microstate \neq microworld

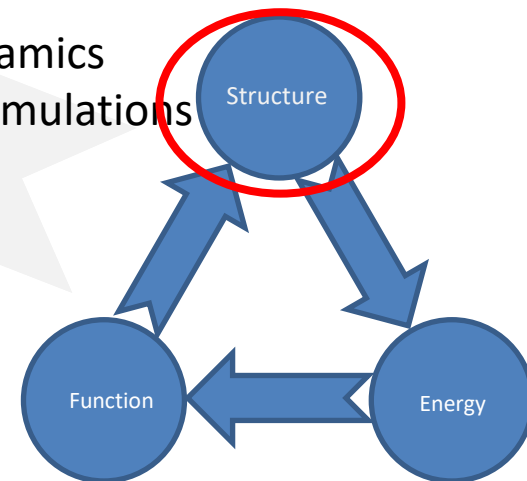
microworld

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
- ...



Structure

Configuration Space

$E(\mathbf{R})$

\mathbf{R} = point in $3N$ dimensional space (N is the number of atoms)

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

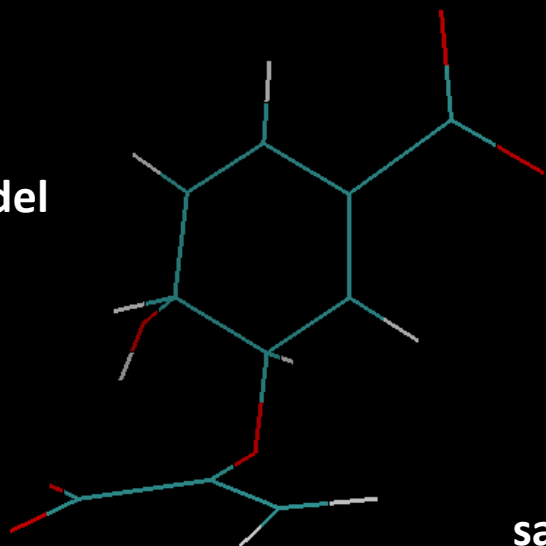
Cartesian coordinates
of the first atom

Cartesian coordinates
of the last atom

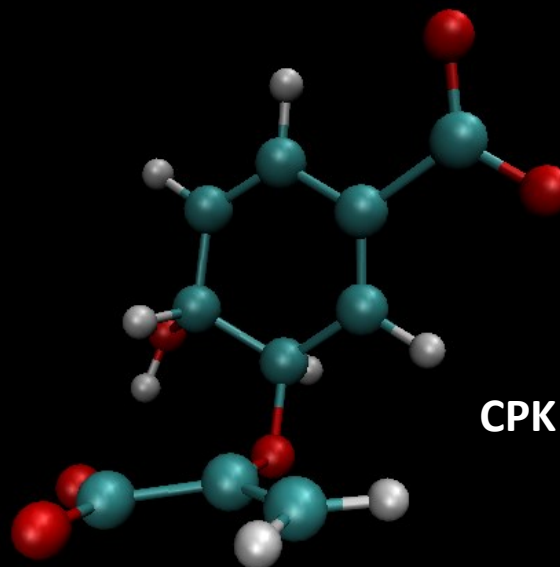
Every point in the configuration space represents
a **unique structure** of the system.

Models - Small Molecules

line model

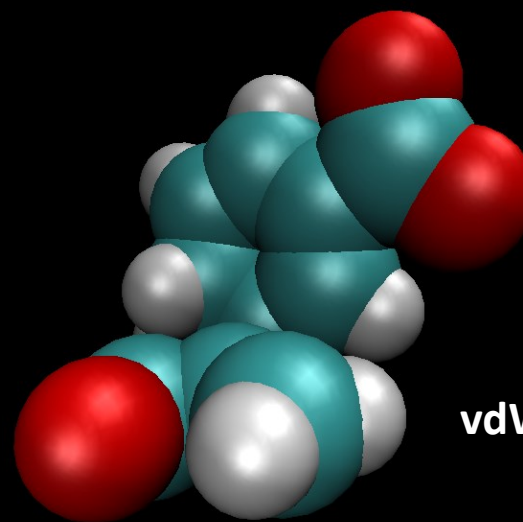
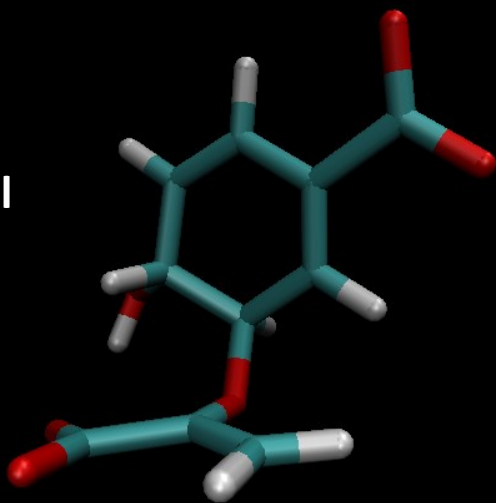


same structure
other visualization



CPK model

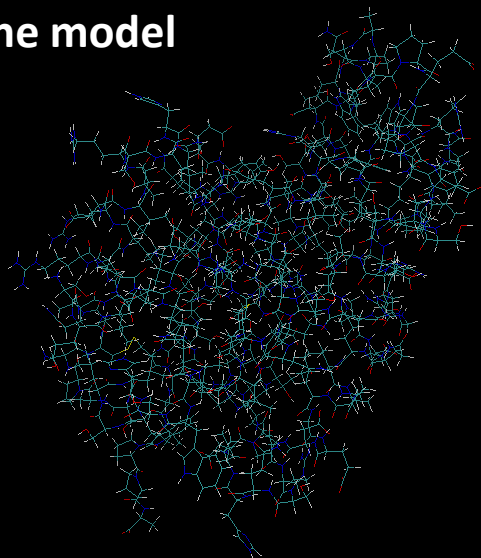
tube model



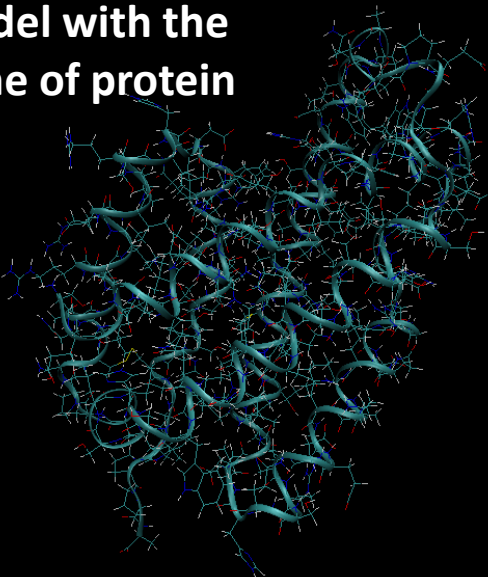
vdW model

Models - Biomolecules

line model



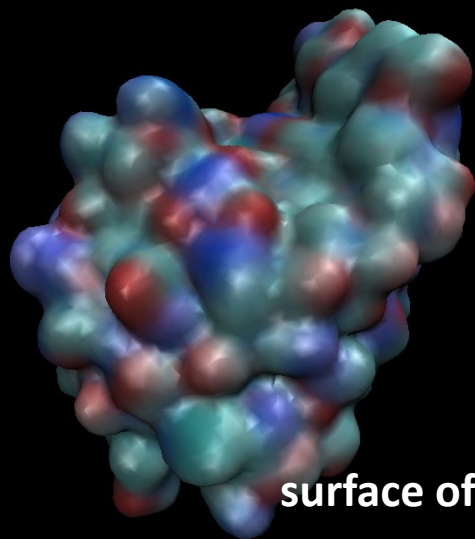
line model with the backbone of protein



Cartoon model



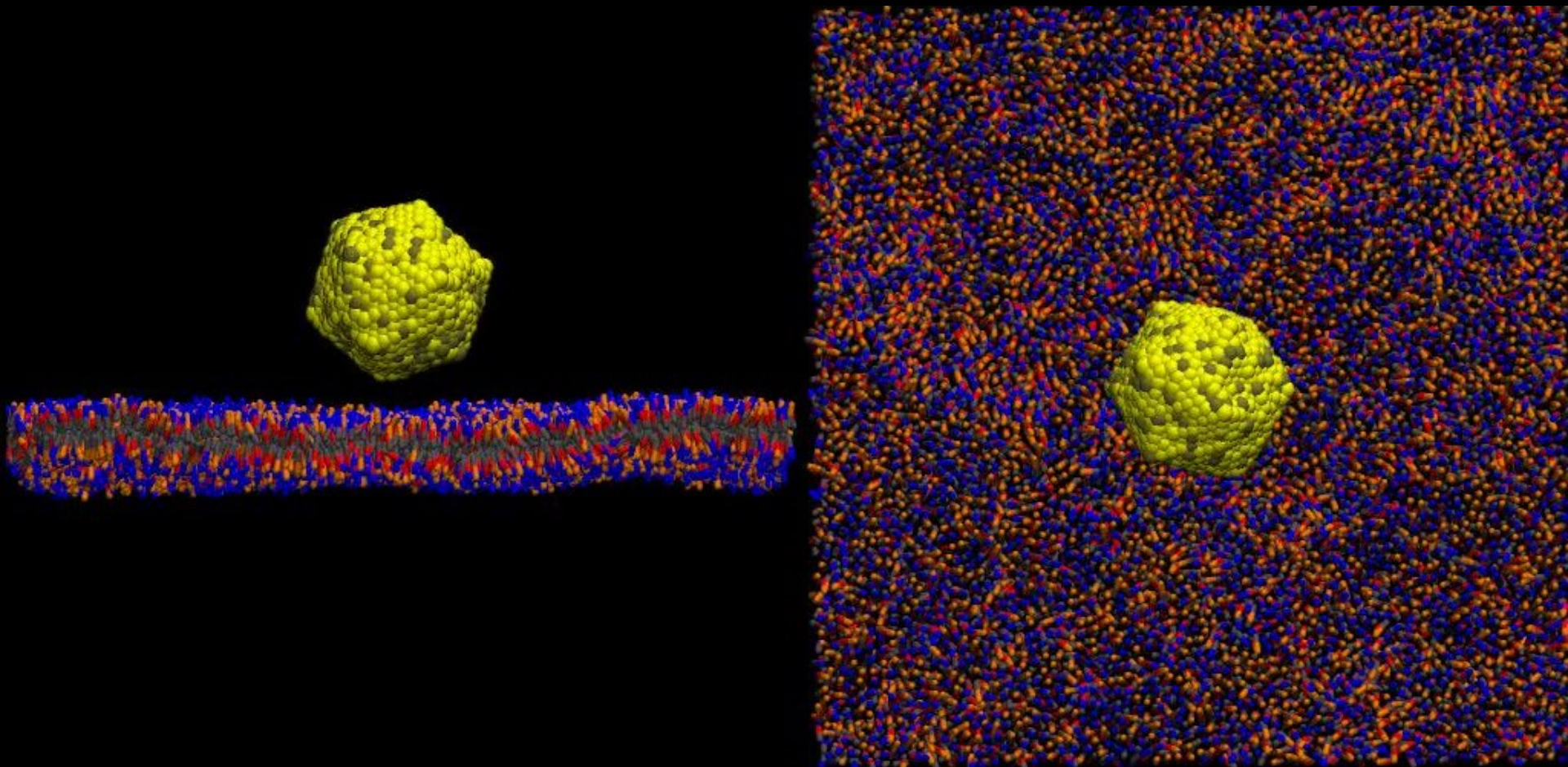
same structure
other visualization



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

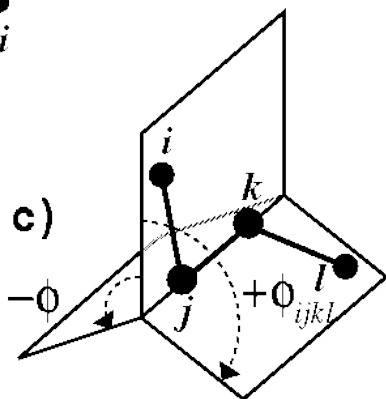
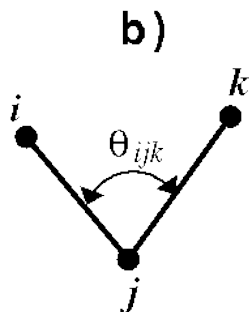
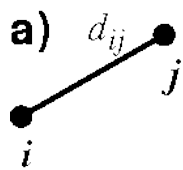
3N-5 (linear diatomic molecule)

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 (blue arrow from 2-1 to bond length 1)
 3-1-2 (blue arrow from 3-1-2 to bond angle 2)
 4-3-1-2 (blue arrow from 4-3-1-2 to torsion angle 2)
 4-3 (blue arrow from 4-3 to bond length 3)
 4-3-1 (blue arrow from 4-3-1 to bond angle 1)



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

XYZ format

positions are in angstroms (Å)

number of atoms	24			
comment	chorismate			
element xyz	C	-1.86100	-0.57700	0.31800
element xyz	O	-2.56800	0.47600	0.32600
element xyz	O	-2.20900	-1.75300	0.64200
element xyz	C	-0.38900	-0.41000	-0.18800
.....
element xyz	H	-0.50900	1.67900	-0.44800

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.

keyword	atom number	atom name	residue name	residue number	Cartesian coordinates of atoms in angstroms (Å)				
ATOM	7	CB	SER	1	5.814	16.335	8.213	1.00	0.00
ATOM	8	HB2	SER	1	6.870	16.427	7.958	1.00	0.00
ATOM	9	HB3	SER	1	5.610	16.900	9.123	1.00	0.00
ATOM	10	OG	SER	1	5.491	14.946	8.427	1.00	0.00
ATOM	11	HG	SER	1	6.026	14.600	9.145	1.00	0.00
ATOM	12	C	SER	1	3.604	16.323	6.927	1.00	0.00
ATOM	13	O	SER	1	2.605	16.742	7.521	1.00	0.00
ATOM	14	N	GLN	2	3.567	15.251	6.134	1.00	0.00
ATOM	15	H	GLN	2	4.401	14.914	5.675	1.00	0.00

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

Djungle of formats I

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

Djungle of formats II

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

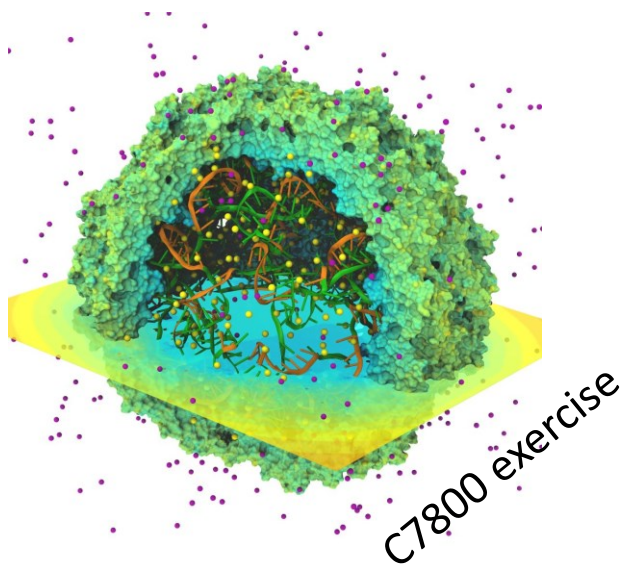
OpenBabel is a chemical toolbox designed to speak the many languages of chemical data. It's an open, collaborative project allowing anyone to search, convert, analyze, or store data from molecular modeling, chemistry, solid-state materials, biochemistry, or related areas.

http://openbabel.org/wiki/Main_Page

Software for visualizations

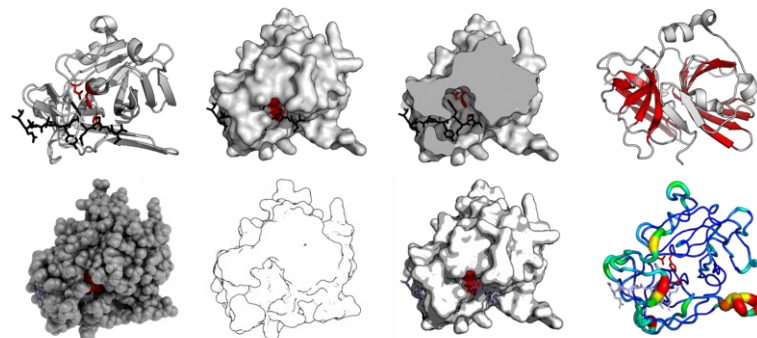
In addition to molecular modelling software (Avogadro, Nemesis, etc.), there are special software serving only for visualizing structures and results.

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?