C7790
Počítačová chemie a molekulové modelování I

C7800 Počítačová chemie a molekulové modelování I - cvičení

8. Struktura

Petr Kulhánek

kulhanek@chemi.muni.cz

Národní centrum pro výzkum biomolekul, Přírodovědecká fakulta
Masarykova univerzita, Kotlářská 2, CZ-61137 Brno

Konfigurační prostor

$E(\mathbf{R})$

R = bod v 3N rozměrném prostoru (N je počet atomů)

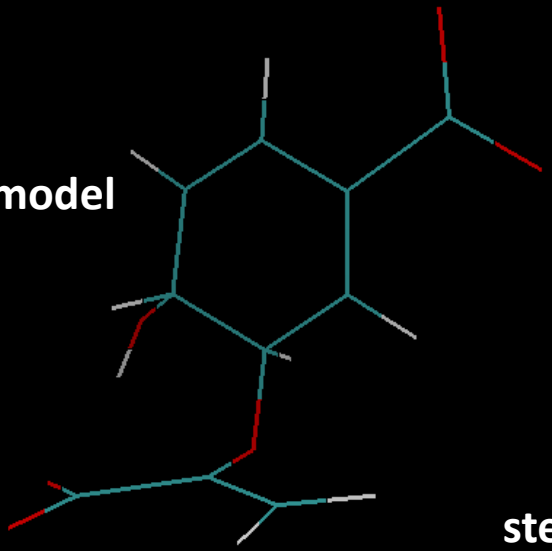
$$\mathbf{R} = \{x_1, y_1, z_1, x_2, y_2, z_2, \ldots, x_N, y_N, z_N\}$$

kartézské souřadnice prvního atomu

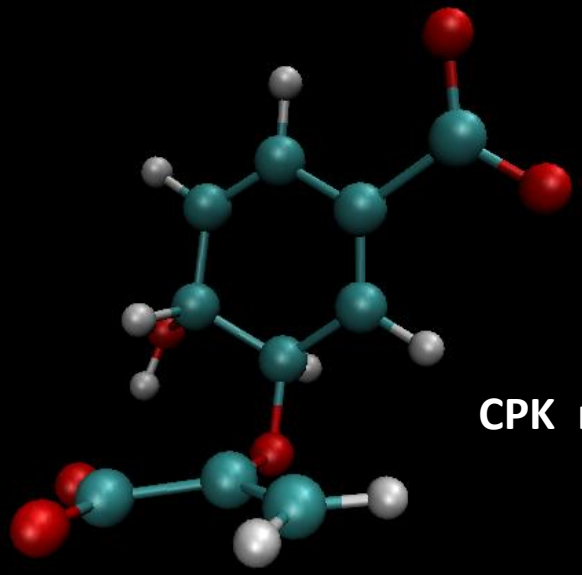
Jednotlivé body tvoří konfigurační prostor. **Každý bod** v konfiguračním prostoru pak představuje **unikátní strukturu** daného systému.

Modely – malé molekuly

čárový model

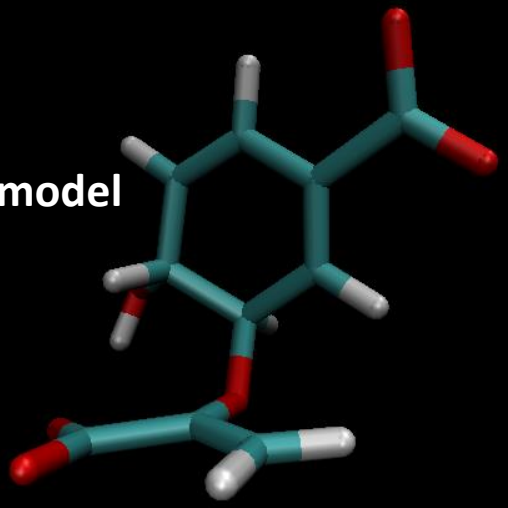


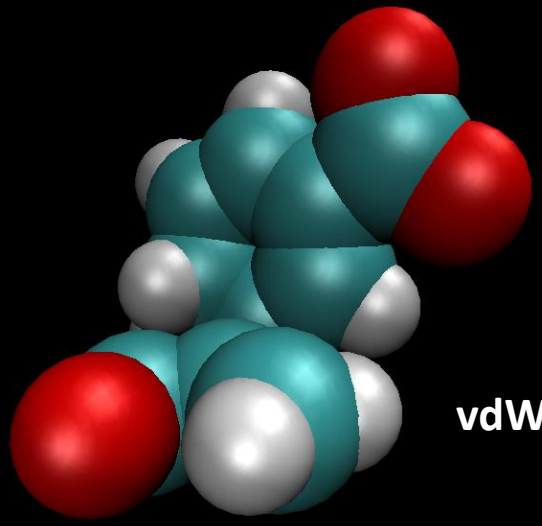
stejná struktura
jiná vizualizace



CPK model

tyčinkový model

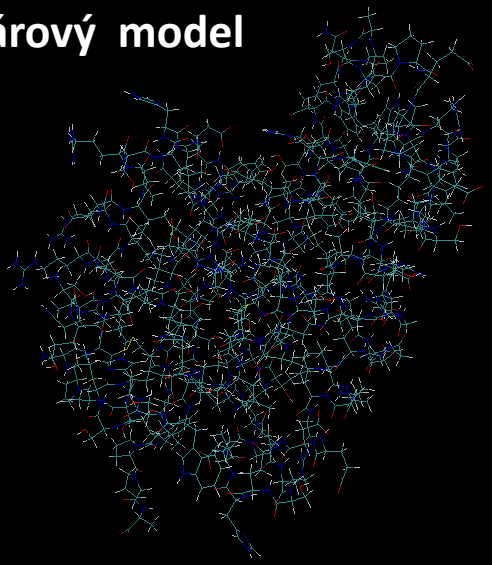




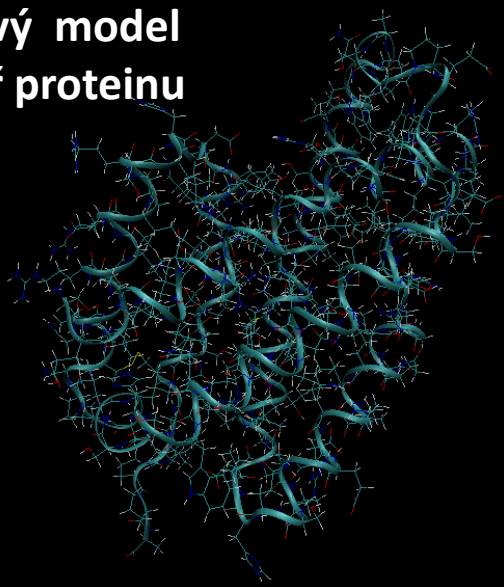
vdW model

Modely – biomolekuly

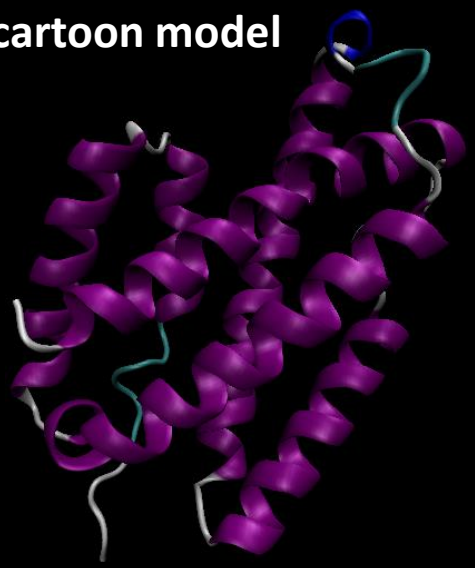
čárový model



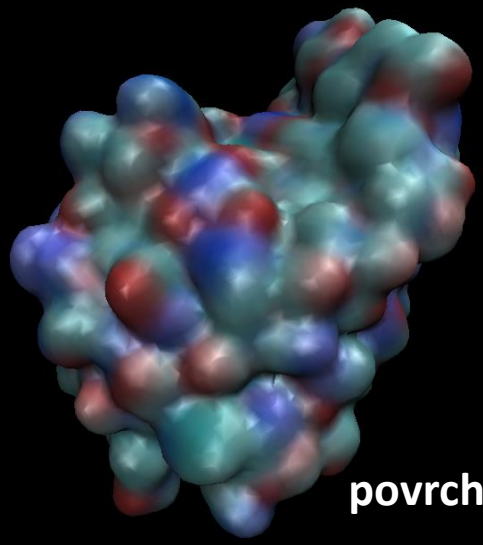
čárový model
páteř proteinu



cartoon model



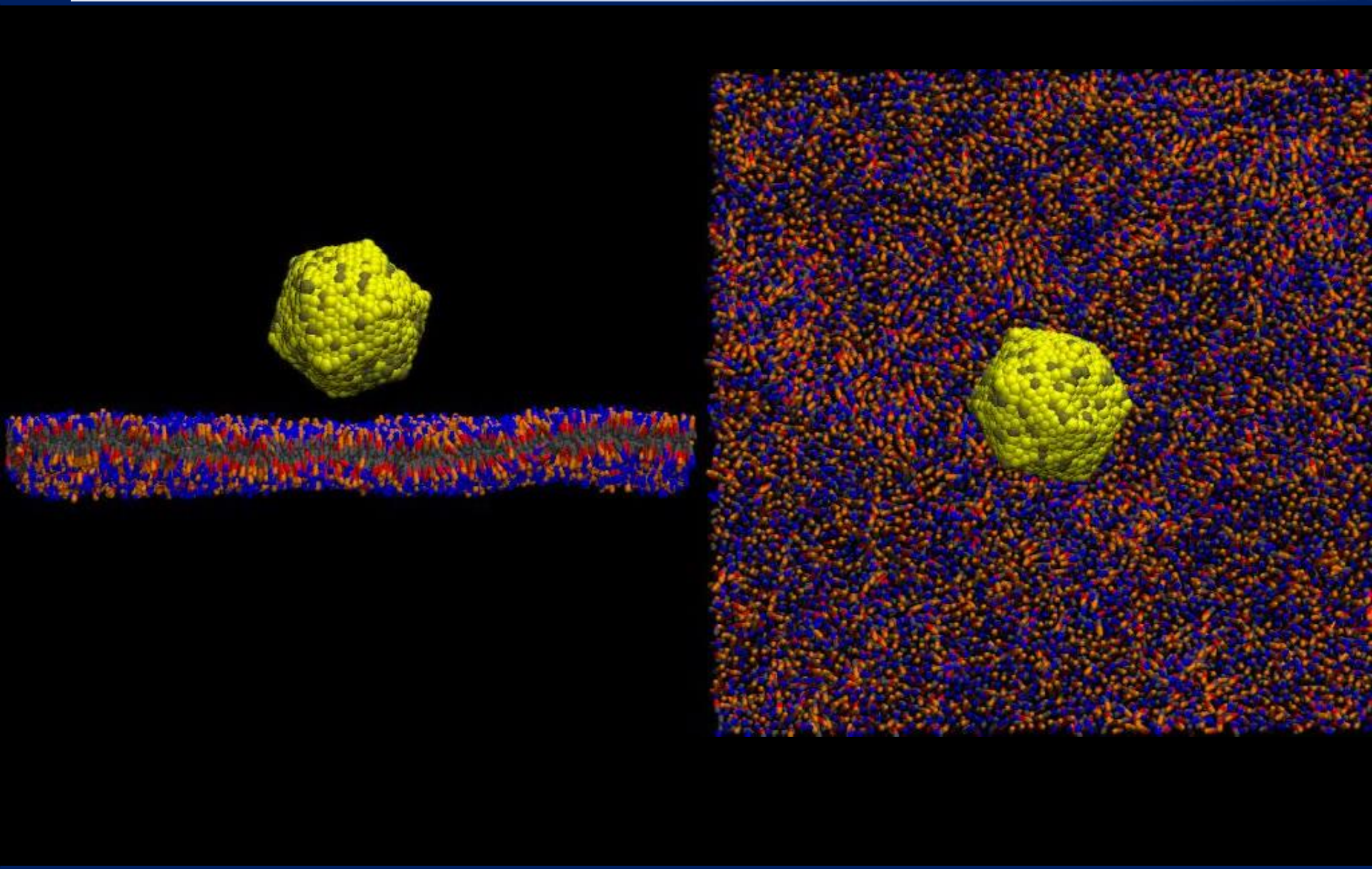
stejná struktura
jiná vizualizace



povrch biomolekuly

Různé modely slouží k zvýraznění určité strukturní informace nebo vnitřní vlastnosti molekuly či uskupení molekul, které pak usnadňuje snadnější pochopení studovaného problému.

Hrubozrné modely



Počítačová reprezentace struktury

Strukturu lze reprezentovat různým způsobem. V chemii se používá více jak 100 formátů, jedná se buď o textové nebo binární soubory. Formát popisuje geometrii systému, jména atomů, skupin atomů, konektivitu mezi atomy (vazby) a další informace.

Geometrie systému může být uvedena v:

- kartézských souřadnicích
- interních souřadnicích
- varianty interních souřadnich

Kartézské vs interní souřadnice

Kartézské souřadnice

```
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

Počet stupňů volnosti:

3N

Interní souřadnice (Z-matrix)

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

vazebná délka vazebný úhel torzní úhel

Počet stupňů volnosti:

3N-6

3N-5 (lineární dvouatomová molekula)

Interní souřadnice

		vazebná délka (a)		vazebný úhel (b)		torzní úhel (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

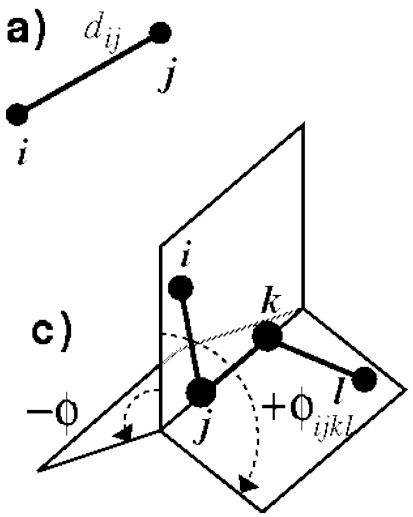
2-1

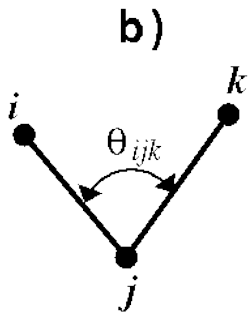
3-1-2

4-3-1-2

4-3

4-3-1



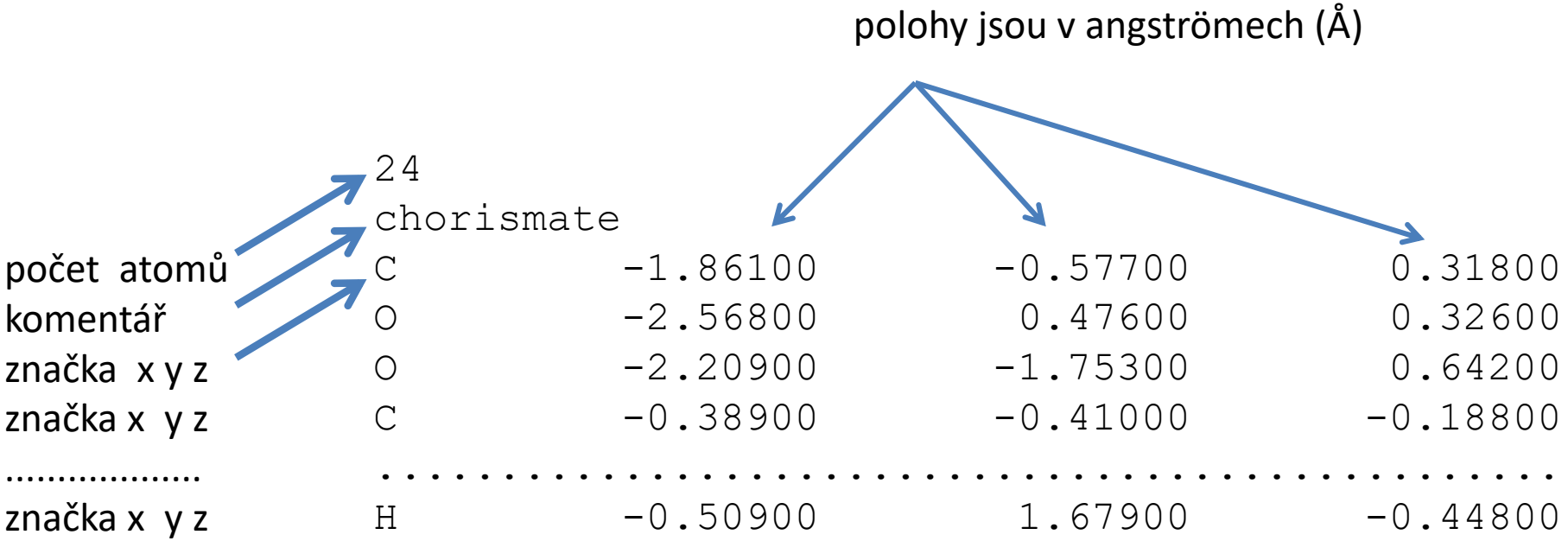


http://www.ccl.net/cca/documents/molecular-modeling/node4.html

Domácí úkol

1. Zamyslete se nad výhodou a nevýhodou reprezentace geometrie systému pomocí kartézských a interních souřadnic.

Formát XYZ

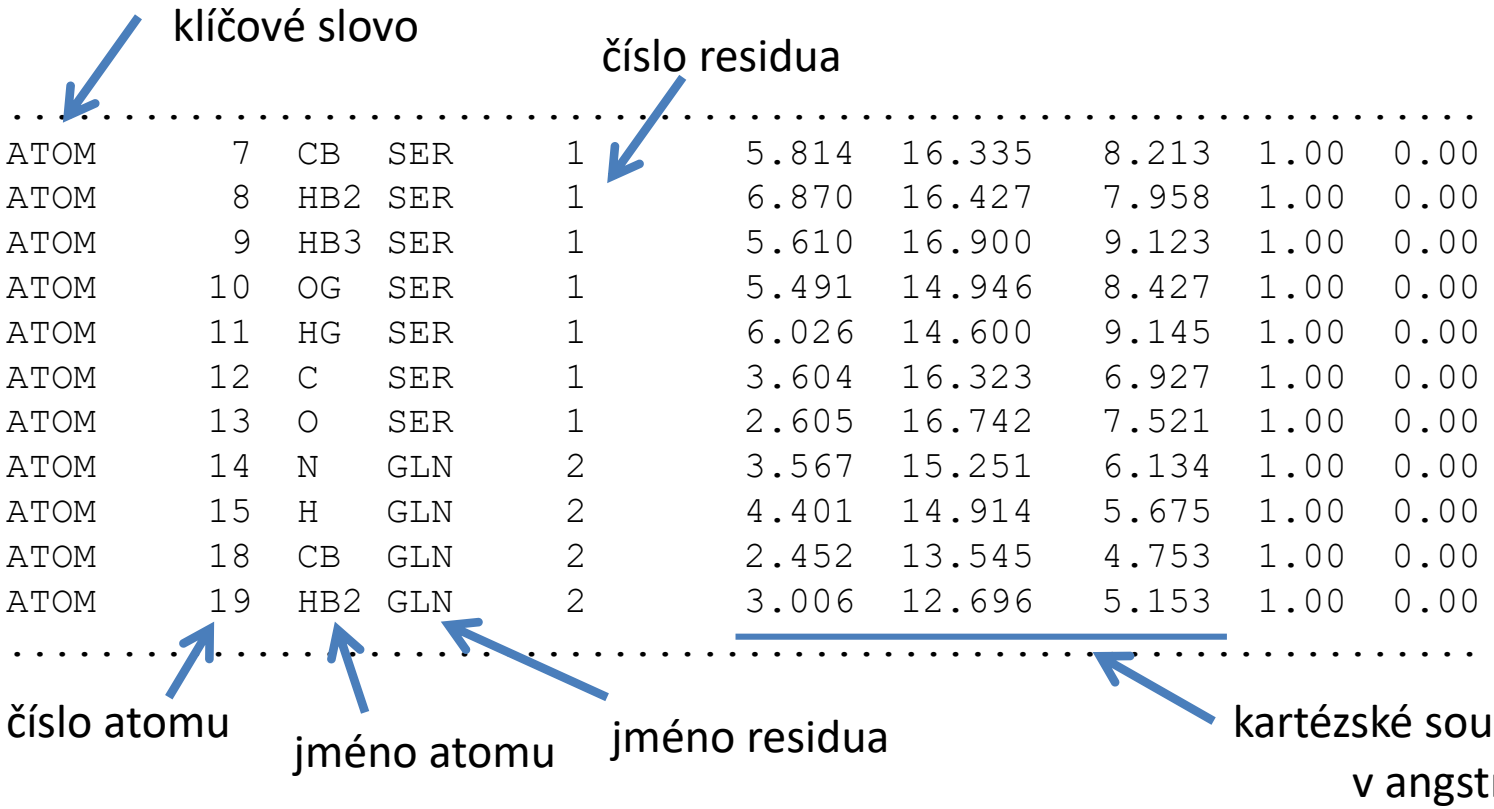


Formát **xyz** je textový soubor s volným formátováním (hodnoty ve sloupcích mohou být odděleny libovolným počtem mezer nebo jiných bílých znaků).

Formát popisuje pouze geometrii systému, neobsahuje informace o vazbách v systému. Program, který s formátem pracuje, musí tyto informace dopočítat (např. pomocí atomových poloměrů).

Formát PDB

Formát pdb se používá k ukládání struktur biomolekul a jejich komplexů.



Formát neobsahuje informace o vazbách v systému. Program, který s formátem pracuje, musí tyto informace dopočítat (na základě templátových struktur). U nestandardních residuích je možné použít klíčové slovo CONECT.

Džungle formátů 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

Džungle formátů 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

Džungle formátů 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

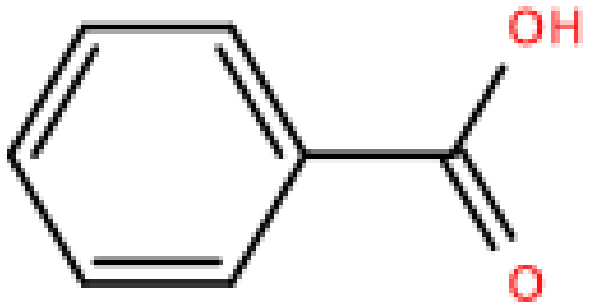
Výše uvedené formáty obsahují většinou kromě 3D/2D struktury také doprovodné informace jako jsou konektivita, parametry silových polí, náboje, různé vlastnosti apod.

OpenBabel prostředí pro konverzi mezi formáty včetně metod pro výpočet celé řady molekulárních vlastností (*chemoinformatika*)

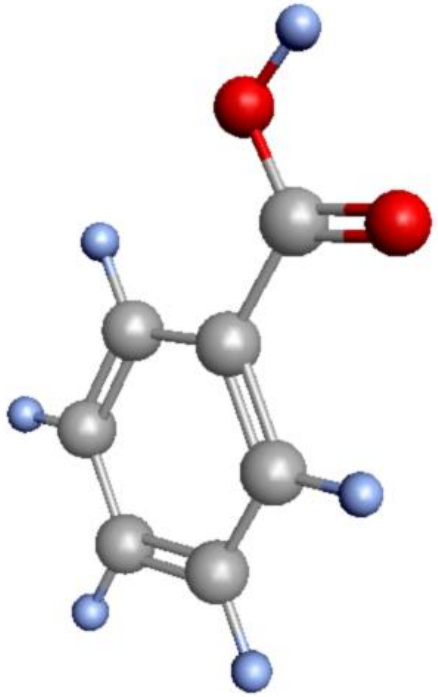
Open Babel is a chemical toolbox designed to speak the many languages of chemical data. It's an open, collaborative project allowing anyone to search, convert, analyze, or store data from molecular modeling, chemistry, solid-state materials, biochemistry, or related areas.

http://openbabel.org/wiki/Main_Page

2D versus 3D struktura



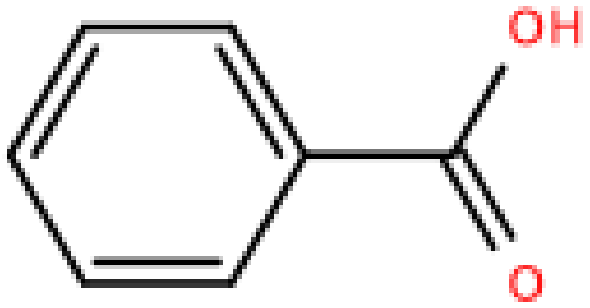
kyselina benzoová



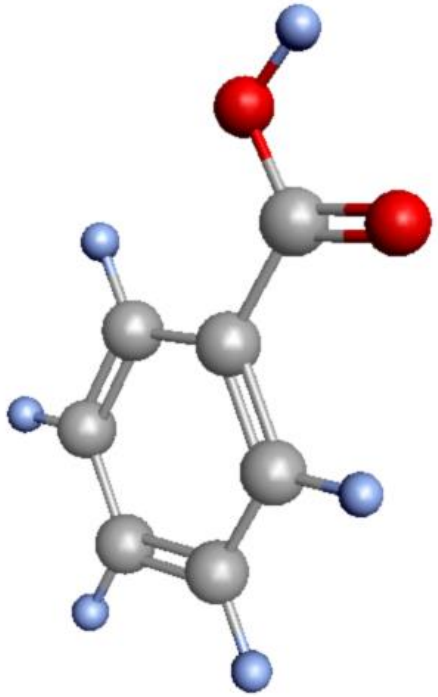
2D struktura obsahuje informaci o atomech a vazbách, kterými jsou spojeny. Tato informace popisuje konstituci (topologii) systému.

3D struktura obsahuje informaci o prostorovém rozmístění atomů. Ostatní informace (např. vazby) jsou dopočitatelné.

3D -> 2D převod



kyselina benzoová

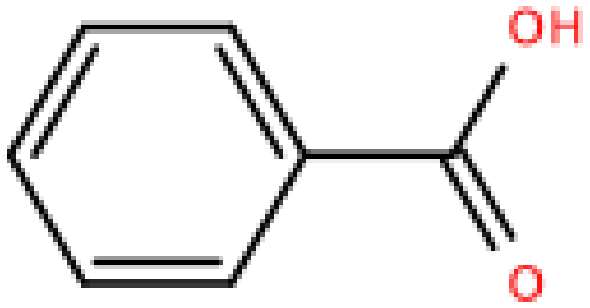


2D struktura obsahuje informaci o atomech a vazbách, kterými jsou spojeny. Tato informace popisuje konstituci (topologii) systému.

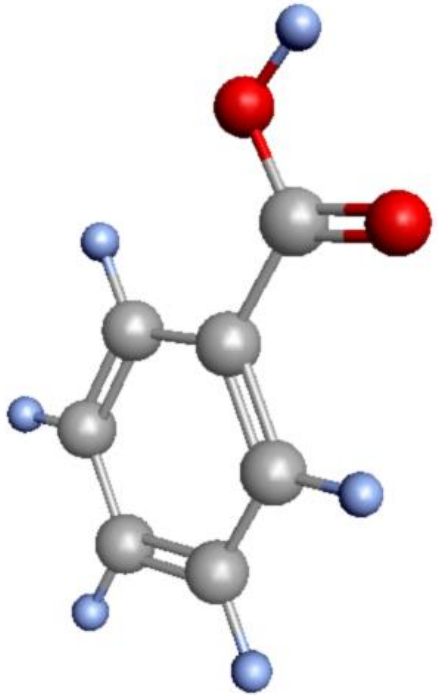
3D struktura obsahuje informaci o prostorovém rozmístění atomů. Ostatní informace (např. vazby) jsou dopočitatelné.

převod je snadný

2D -> 3D převod



kyselina benzoová

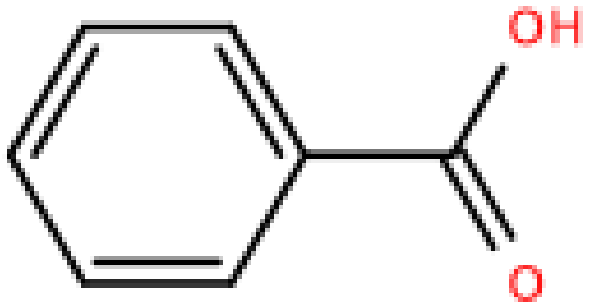


2D struktura obsahuje informaci o atomech a vazbách, kterými jsou spojeny. Tato informace popisuje konstituci (topologii) systému.

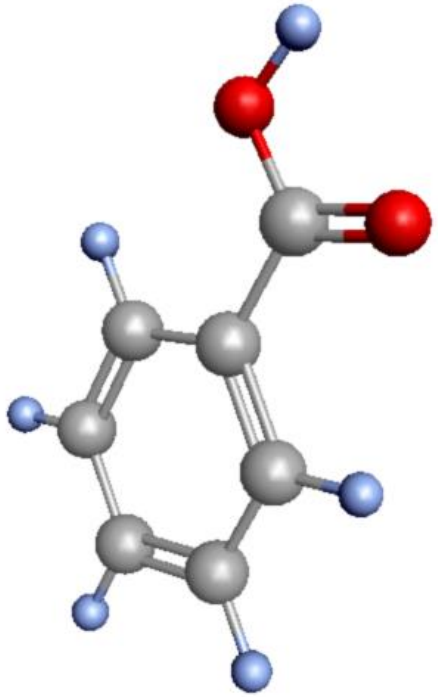
3D struktura obsahuje informaci o prostorovém rozmístění atomů. Ostatní informace (např. vazby) jsou dopočitatelné.

převod je komplikovaný

2D -> 3D převod



kyselina benzoová

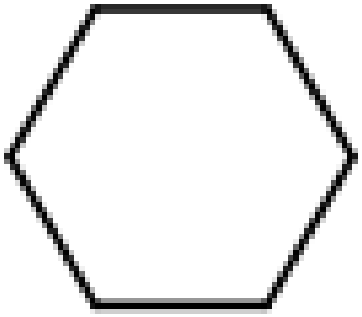


2D struktura obsahuje informaci o atomech a vazbách, kterými jsou spojeny. Tato informace popisuje konstituci (topologii) systému.

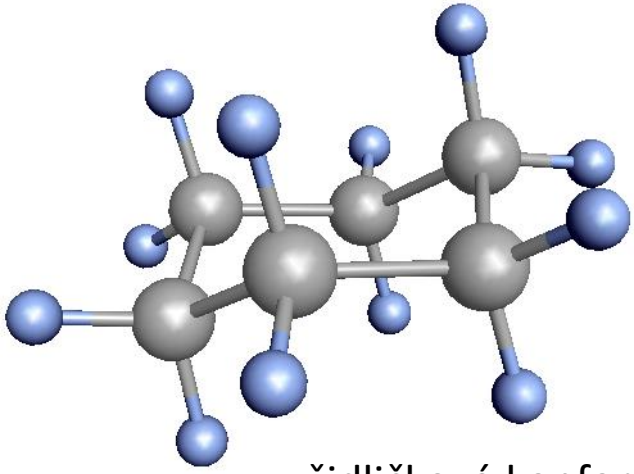
3D struktura obsahuje informaci o prostorovém rozmístění atomů. Ostatní informace (např. vazby) jsou dopočitatelné.

převod je komplikovaný
u velkých systémů nemusí být jednoznačný v důsledku existence více konformerů

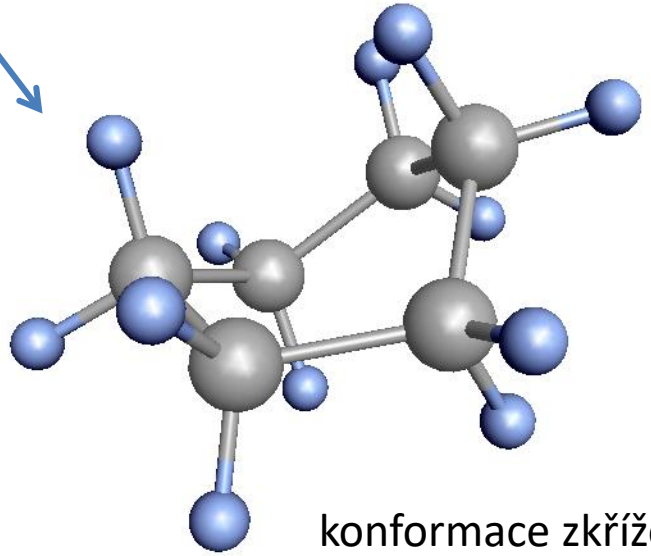
2D -> 3D převod, komplikace



cyklohexan



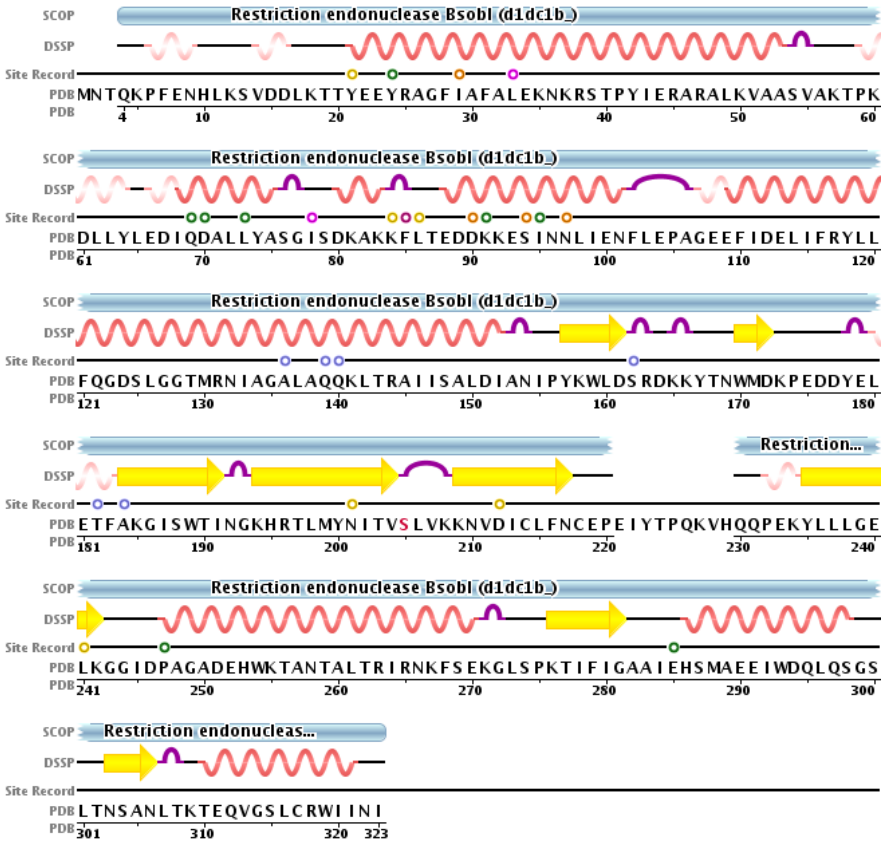
židličková konformace

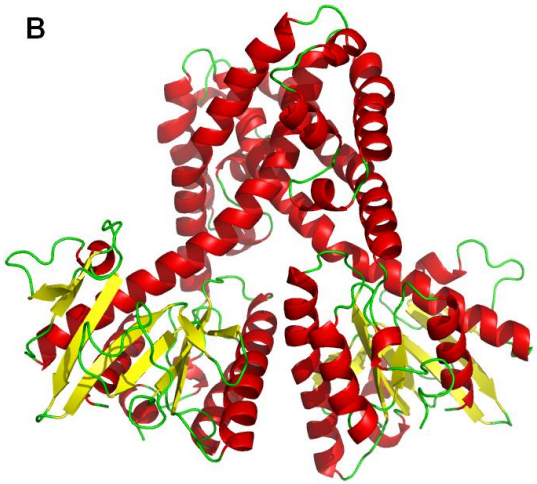


konformace zkřížená vanička

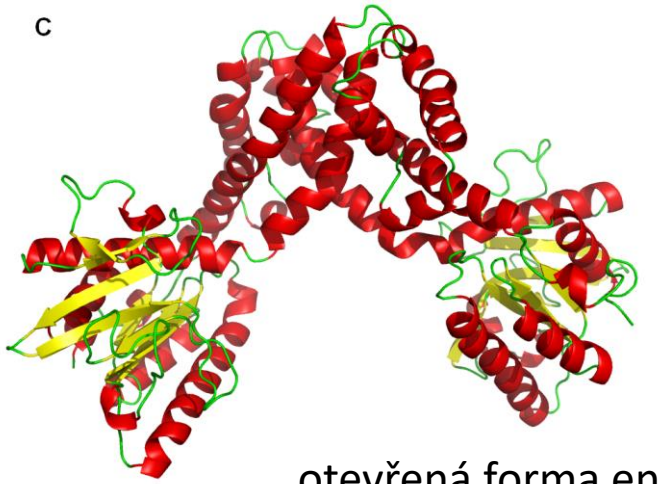
2D -> 3D převod, komplikace

Stejná primární struktura (sekvence aminokyselin).





zavřená forma enzymu



otevřená forma enzymu

Využití 2D struktur

Representace molekul ve 2D formátech se využívá převážně pro ukládání informací do databází a jejich prohledávání, dále k předpovědi chemických vlastností molekul pomocí chemoinformatických přístupů.

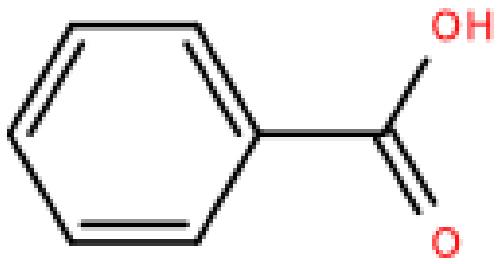
Nejrozšířenější formáty:

- **SMILES** (Simplified Molecular-Input Line-Entry System)

```
C(=O)(O)c1ccccc1
```

- **InChI** (IUPAC International Chemical Identifier)

```
InChI=1S/C7H6O2/c8-7(9)6-4-2-1-3-5-6/h1-5H,(H,8,9)
```



kyselina benzoová

Zdroje 3D struktur - experiment

Cambridge Structural Database (CSD)

http://www.ccdc.cam.ac.uk/Solutions/CSDSystem/Pages/CSD.aspx

Obsahuje zhruba půl miliónu struktur malých molekul určených pomocí rentgenové a neutronové difrakce. Software pro práci s daty: Mercury http://www.ccdc.cam.ac.uk/Solutions/CSDSystem/Pages/Mercury.aspx

Protein Data Bank (PDB)

http://www.pdb.org

Obsahuje zhruba 94 tisíc struktur biomolekulárních systémů určených převážně pomocí rentgenostrukturní analýzy.

Experimentální metoda	Proteiny (P)	Nucleové kyseliny (NA)	P/NA komplexy	Jiné	Celkově
X-ray	77445	1481	4069	3	82998
NMR	8851	1046	193	7	10097
elektronová mikroskopie	469	45	129	0	643

stav v září 2013

Zdroje 3D struktur – *in silico*

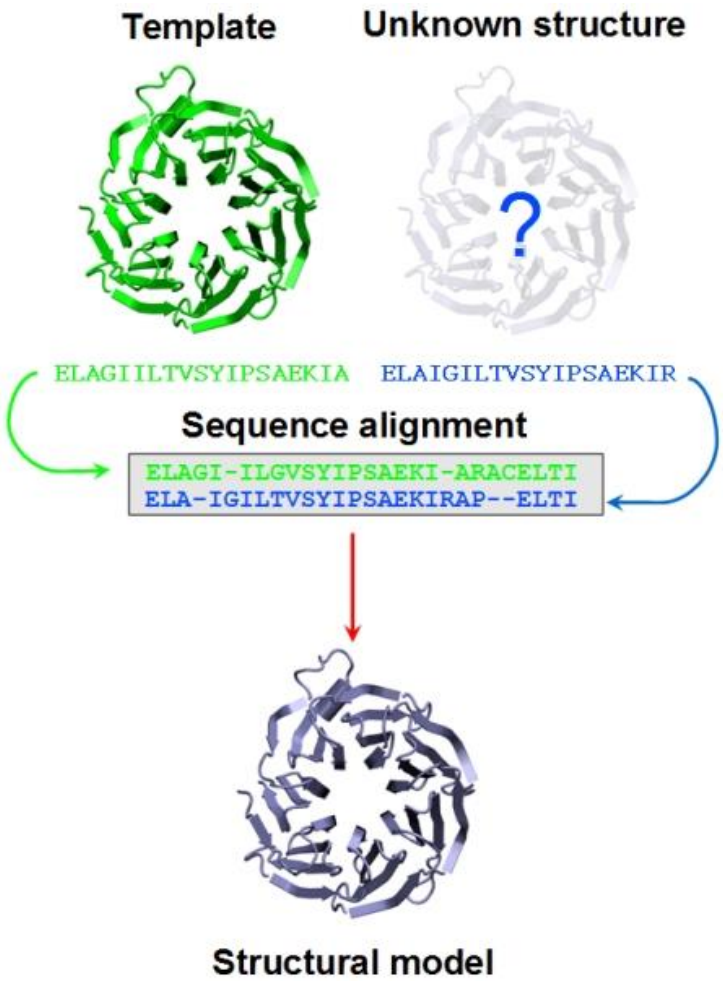
Výpočetní metody

- molekulové modelování
- homologní modelování

In silico modelování se provádí i u experimentálních struktur, které jsou neúplné:

- nedostatečné rozlišení – X-Ray
- jsou určeny jen částečné strukturní informace - NMR

Homologní modelování



http://www.unil.ch/pmf/en/home/menuinst/technologies/homology-modeling.html