C7790 Introduction to Molecular Modelling

TSM Modelling Molecular Structures

C9087 Computational Chemistry for Structural Biology

Lesson 9
Model

JS/2022 Present Form of Teaching: Rev3

Petr Kulhánek

kulhanek@chemi.muni.cz

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

Reality *vs* Simulation

Is it possible to accurately simulate the reality around us?

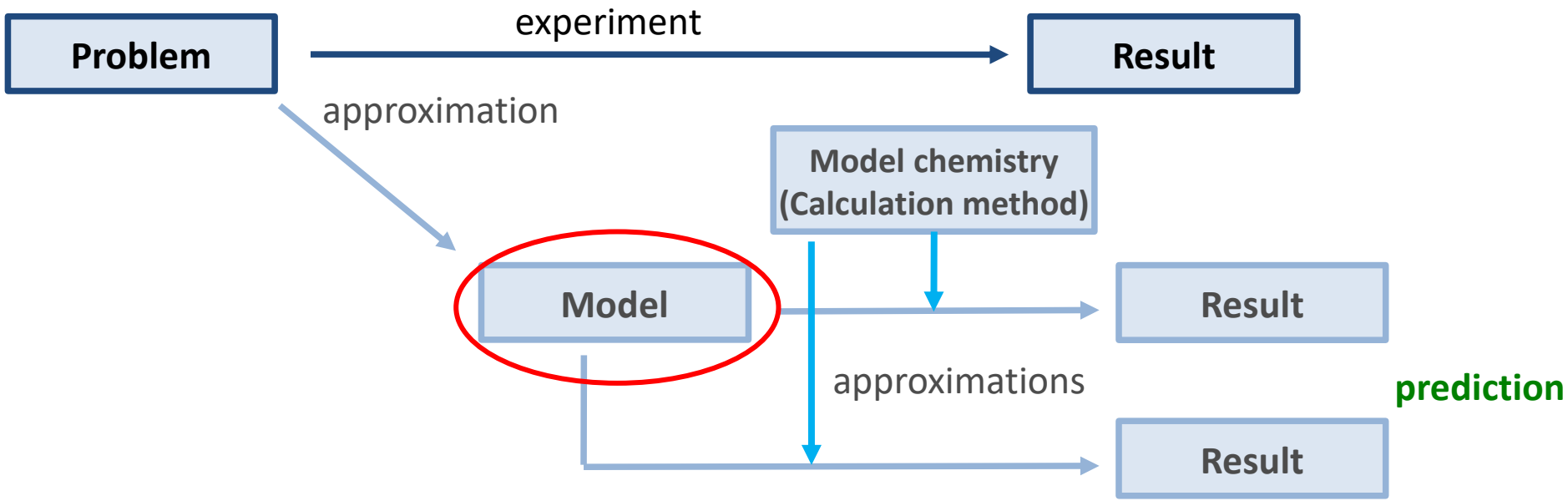
Unfortunately, no :-(

Why?

- incomplete theory
- insufficient performance of current and future (?) computers

Solution ...

- use **approximation** for solution of problems using the available computing capacity



Do we need a model?

The only input: nuclei and electrons and their description of interactions and motions

$$\hat{H}\psi_k(\mathbf{r}) = E_k\psi_k(\mathbf{r})$$

Solutions: energy and wavefunction describing QM states (microstates)

In theory, no model is needed because it is outcome of SR equation solution.

In practice, we need to employ the BO approximation, which then requires a model (**R**).

Born-Oppenheimer Approximation

$$\hat{H}\psi(\mathbf{r},\mathbf{R}) = E\psi(\mathbf{r},\mathbf{R}) \qquad \psi(\mathbf{r},\mathbf{R}) = \Psi(\mathbf{r},\mathbf{R})\chi(\mathbf{R})$$

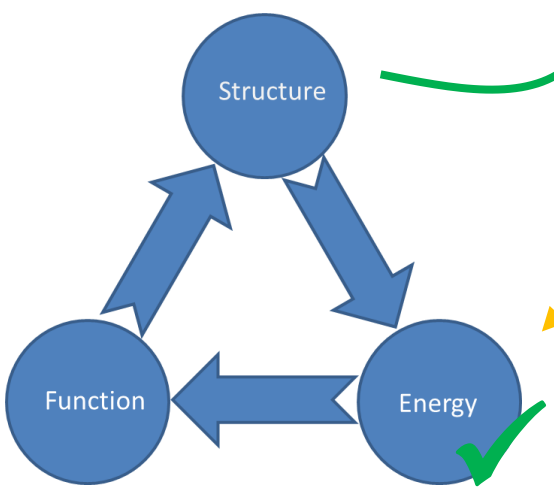
Born-Oppenheimer approximation

electronic properties of the molecule

vibrational, rotational, translational motions of a molecule

$$\hat{H}_e\Psi(\mathbf{r},\mathbf{R}) = E_e(\mathbf{R})\Psi(\mathbf{r},\mathbf{R})$$

$$\hat{H}_R\chi(\mathbf{R}) = E_{VRT}\chi(\mathbf{R})$$



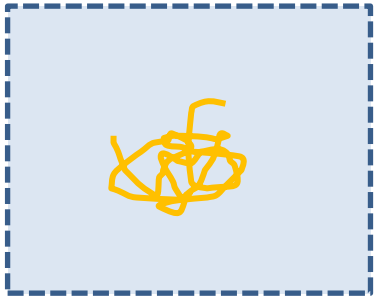
We need a structure (model) to calculate energies.

What is a model?

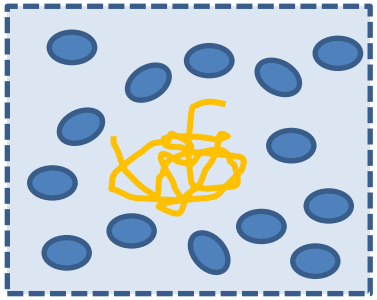
A model is smallest representation of studied system, which can describe studied phenomena by chosen computational method (model chemistry).

Main model types:

molecule (molecules) in **vacuum**



molecule (molecules) in **implicit environment**



molecule (molecules) in **explicit environment**

environment (typically solvent, membrane, etc.)

is implicitly modelled as a mean field environment representation (e.g., polarizable dielectric)

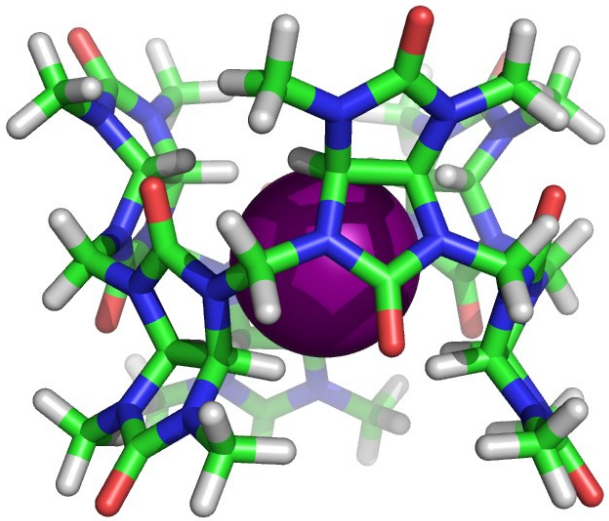
is explicitly modelled by atoms/molecules

Is environment important?

In molecular modelling, it is not good idea to neglect environment even when qualitative outcome is required.

! Neglecting environment can lead to wrong conclusions !

Modelling of molecules in vacuum (only) must be carefully justified.



bambus[6]uril/anion interaction

vacuum binding affinities:

$F^- > Cl^- > Br^- > I^-$

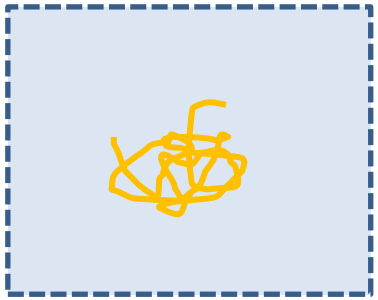
solvent ($MeOH/CH_3Cl$) binding affinities:

$I^- > Br^- > Cl^- > F^-$

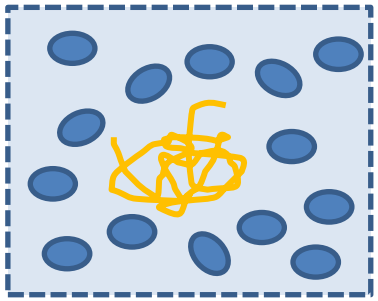
the order is changed due to anion desolvation energies

SLAVÍK, Jan. Počítačové modelování glykolurilových struktur, Bachelor's thesis. Masaryk University, Brno, 2010

Is environment important? cont.



molecule (molecules) in
implicit environment



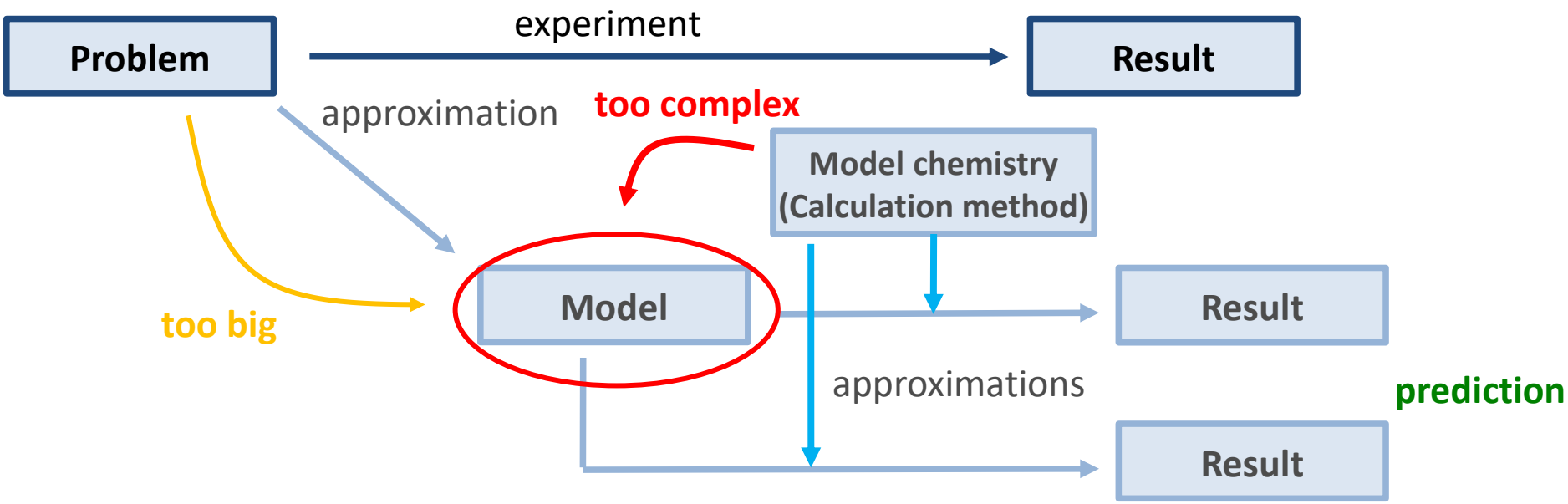
molecule (molecules) in
explicit environment

- **Mainly used in QM calculations**
  - **PCM** (Polarizable continuum model)
  - **COSMO** (Conductor like screening model)
  - …
- but also in MM
  - **PB** (Poisson–Boltzmann solvent model)
  - **GB** (Generalized Born solvent model)
  - **3D-RISM** (3D reference interaction site model)
  - …

- Too complex for QM calculations (rarely used).
- **Typically used in MM and MD**
  - **TIP3P** (water model)
  - **SPC/E** (water model)
  - …

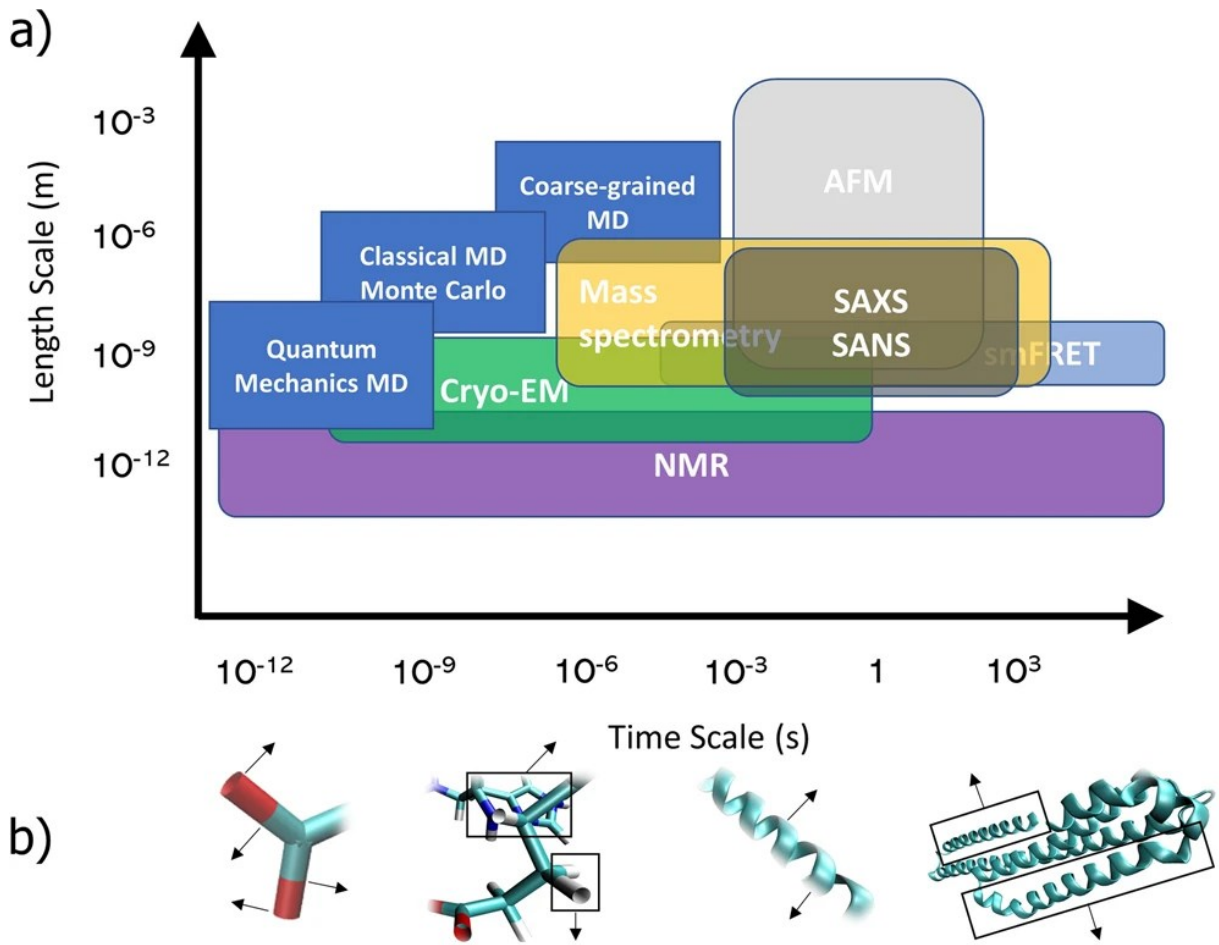
Homework:
How accurately different solvent types (implicit/explicit) can describe interactions at solute/solvent interface?

Model is a compromise



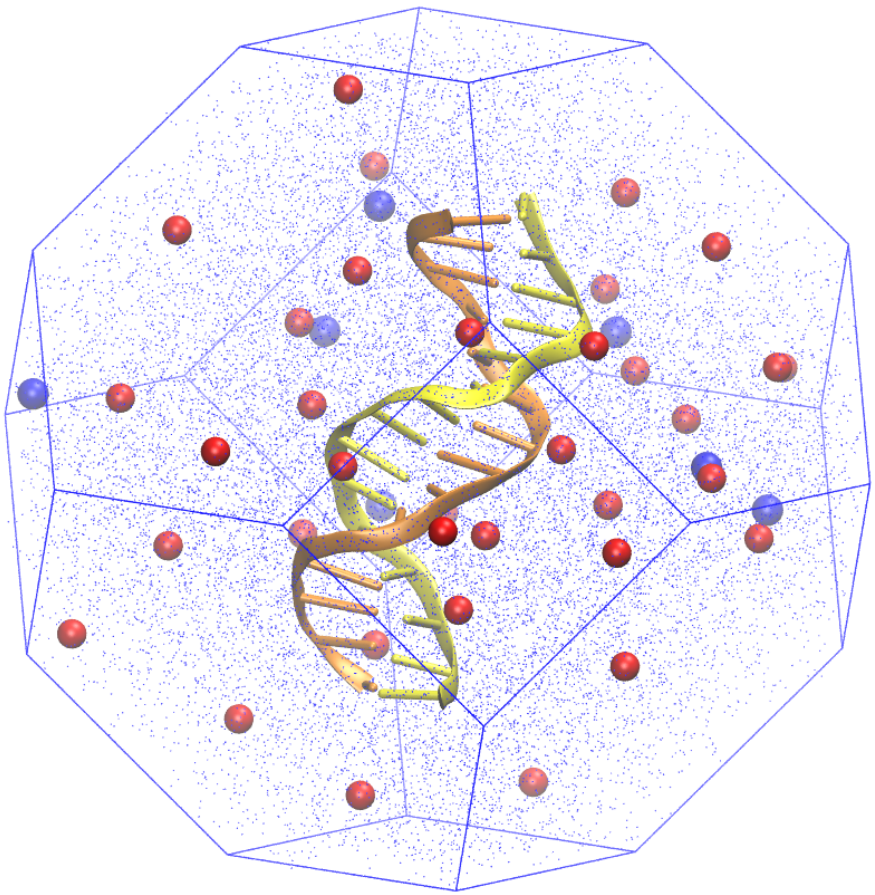
We need to choose between **accuracy** (computational feasibility) and **model reliability** (reasonable representation of studied system).

Model sizes and time scales



Hsu, C.C., Buehler, M.J. & Tarakanova, A. The Order-Disorder Continuum: Linking Predictions of Protein Structure and Disorder through Molecular Simulation. Sci Rep 10, 2068 (2020). https://doi.org/10.1038/s41598-020-58868-w

Example



- DNA (15-nt long)
  - **948** atoms
  - c(DNA)=7 mM
- explicit ions
  - n(Na^+)=35, c(Na^+)=244 mM
  - n(Cl^-)=7, c(Cl^-)=49 mM
  - effective c(NaCl)=154 mM*
- explicit water (TIP3P model)
  - n(H_20)=7592
  - **22776** atoms
- PBC with truncated octahedral box
  - largest subscribed sphere R_{in}=29 Å

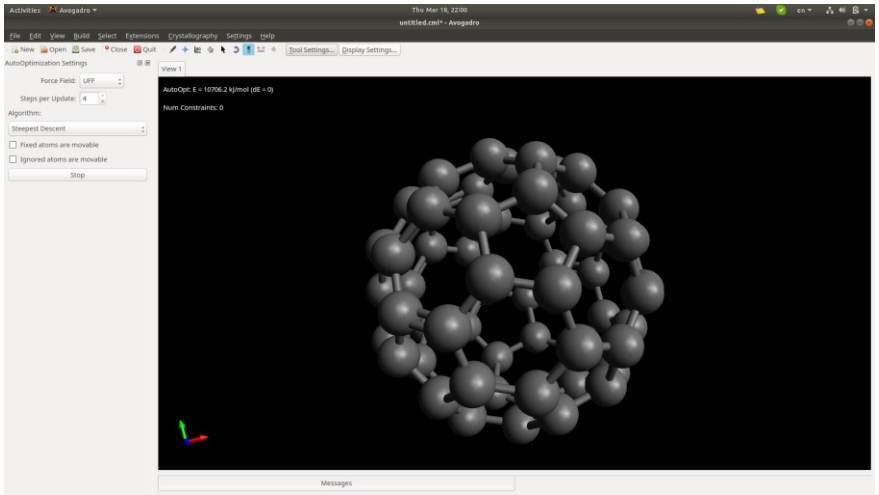
*Machado,M.R. and Pantano,S. (2020) Split the Charge Difference in Two! A Rule of Thumb for Adding Proper Amounts of Ions in MD Simulations. *J. Chem. Theory Comput.*, **16**, 1367–1372

Where to get a model?

- **In silico modelling**
  - small molecules
  - 2D -> 3D conversions (high-throughput modelling, virtual screening)
  - *ab initio* prediction of biomolecular structures
- **Modelling based on experimental structures**
  - small molecules
  - large molecules (proteins, DNA, biomolecular complexes, …)
- **Experimentally guided modelling**
  - NMR (NOE contacts, …)
  - cryoEM, SAXS (electron density, shape, …)
- **Similarity modelling**
  - in silico modification of experimental structures
  - homology modelling

In silico modelling

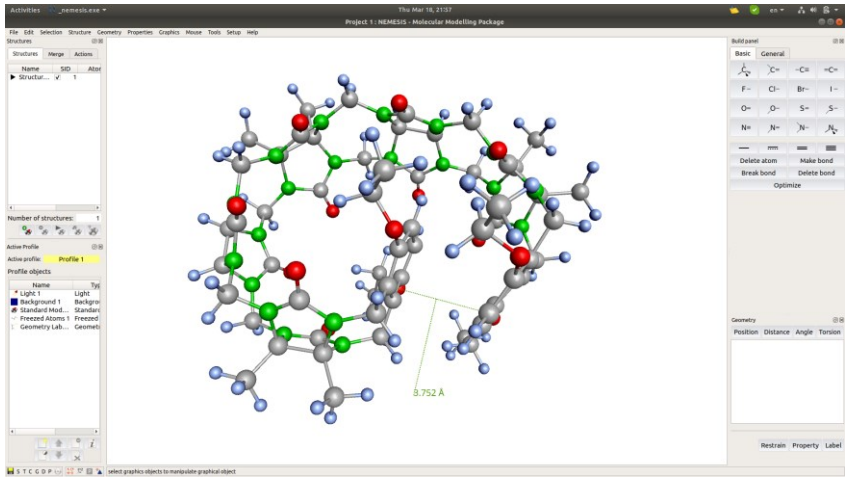
Avogadro



free drawing of molecular structures

C7800 exercise

Nemesis



piecewise assembly of molecular structures

C7800 exercise

For modelling, we need 3D structures.

$$E(\mathbf{R})$$

Other software (commercial):

- Spartan, Hyperchem
- SCM (ADF)
- …

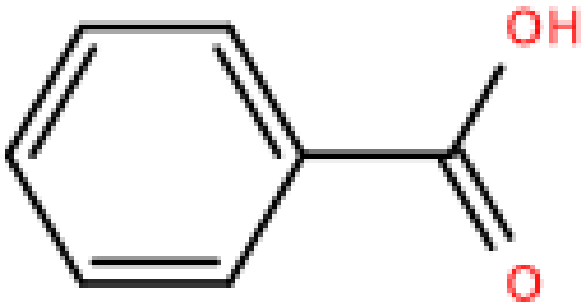
Overview of software:

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

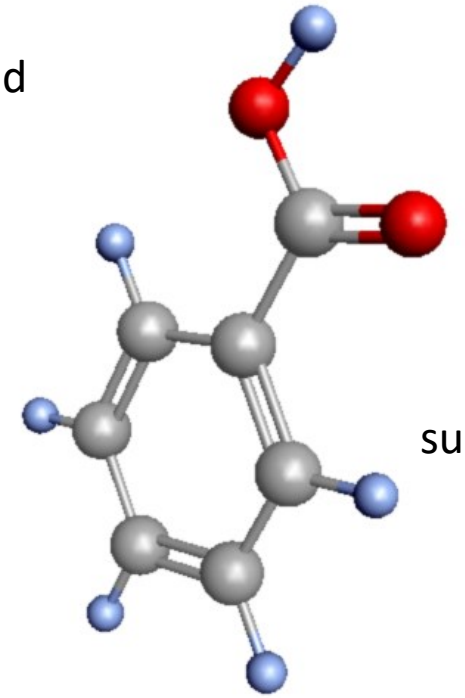
2D vs 3D structure

2D

benzoic acid



3D



✓ suitable for modelling

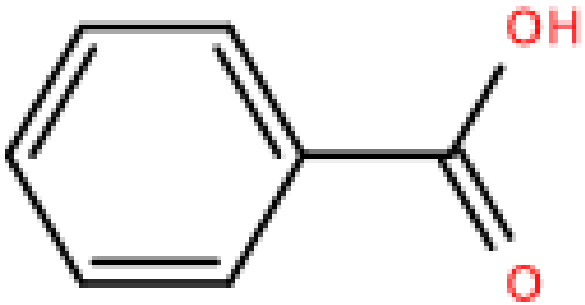
2D structure contains information about the atoms and bonds. This information describes the constitution (topology) of the system.

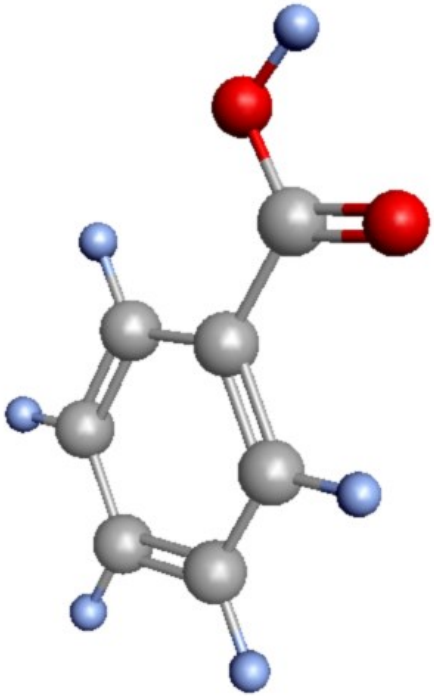
3D structure contains information on the spatial distribution of atoms in space. Other information (e.g., bonds) is computable.

3D <-> 2D conversions

2D **3D**

benzoic acid

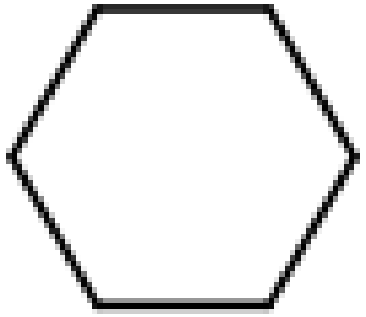




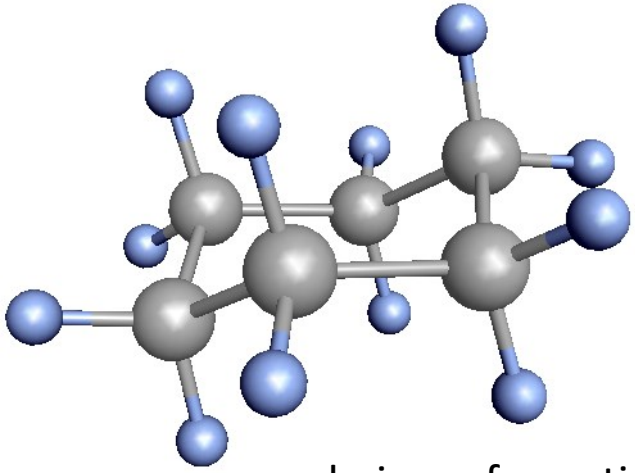
conversion is easy

conversion is difficult or impossible

3D/2D conversions, complications

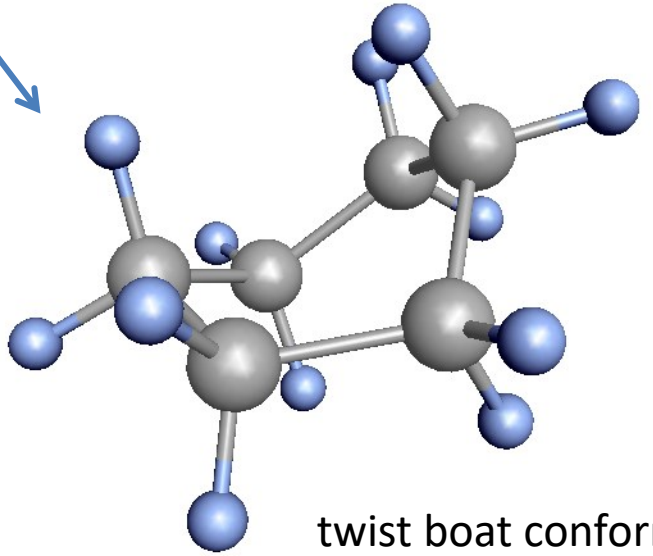


cyclohexane



chair conformation ✓

For small molecules, 2D->3D conversion is possible. Usually, the most stable conformer is modelled.



twist boat conformation

2D structure usage

Representation of molecules in 2D formats is employed mainly for:

- storing information in databases
- searching in such databases (InChiKey and other variants)
- predicting the chemical properties of molecules using chemoinformatic approaches (machine learning)
- automatic structure generation, generating libraries of molecules (computer aided combinatorial chemistry) - virtual screening

Most common formats:

- **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)`

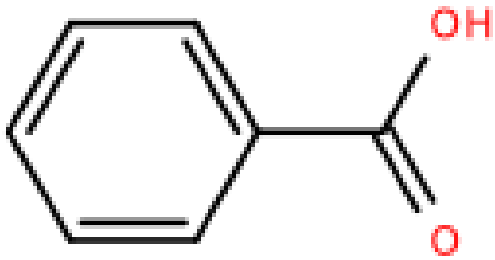
hash

- **InChIKey** (IUPAC International Chemical Identifier Key)

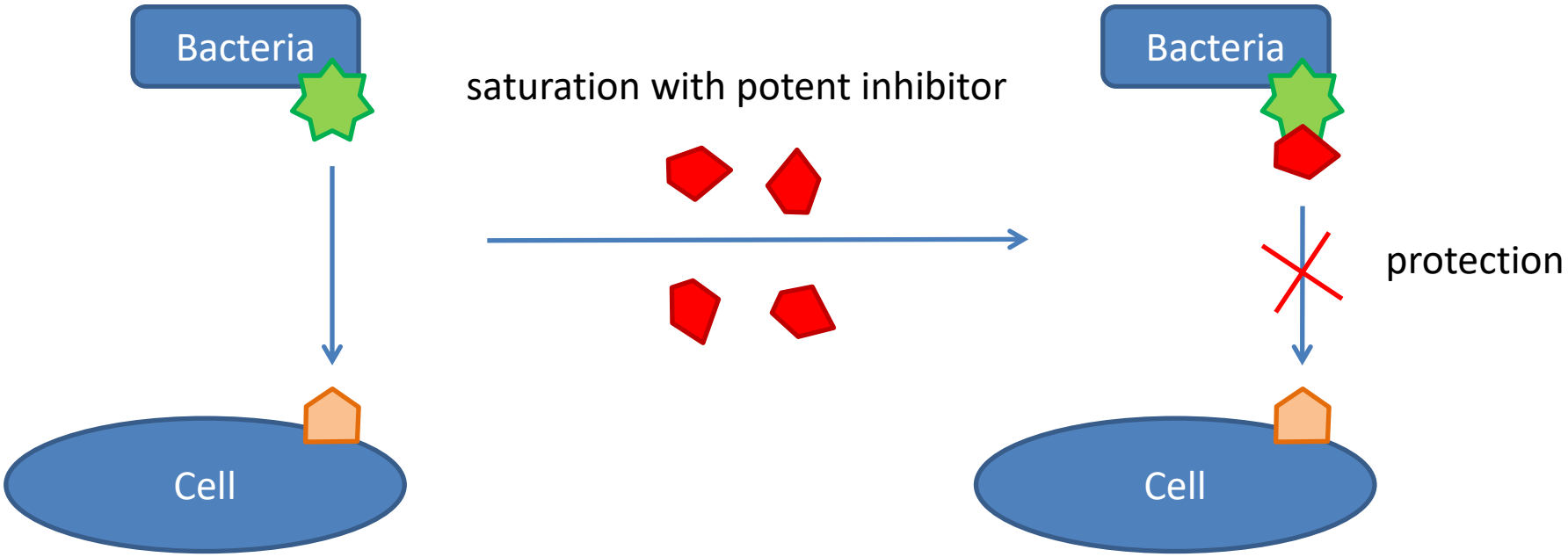
  `WPYMKLBDIGXBTP-UHFFFAOYSA-N`

  constant length, possible collisions

benzoic acid



Virtual screening (motivation)



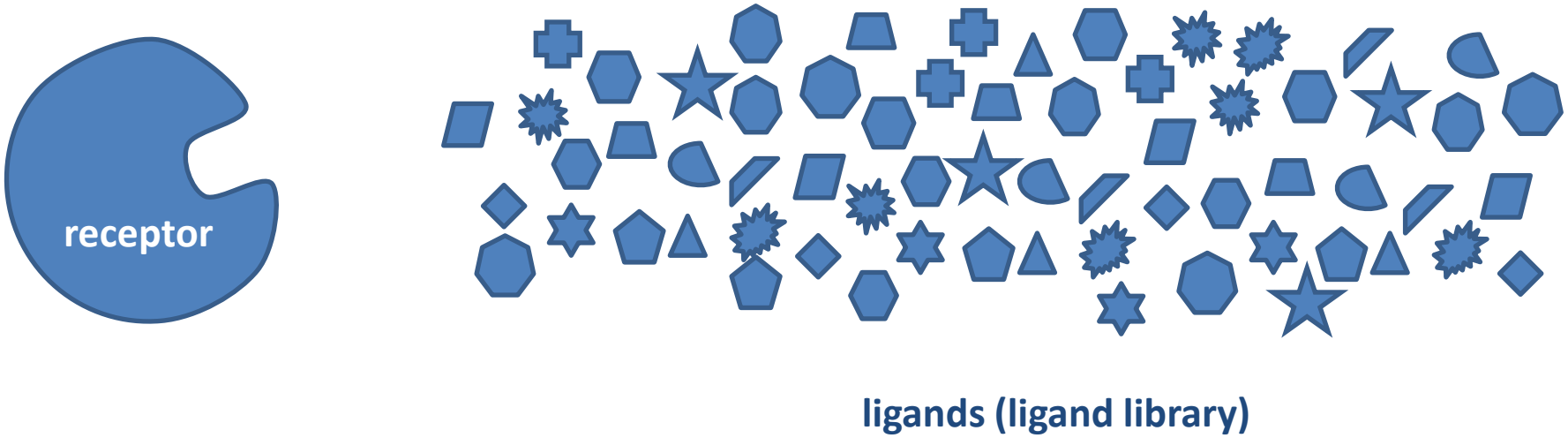
Early inhibition of bacterial lectin surface hinder bacterial adhesion to host cells.

Potent inhibitor (glycomimetics) can be used in treatment of bacterial infections.

(development of new antibiotics)

Virtual screening

Which of them is the best?



Docking

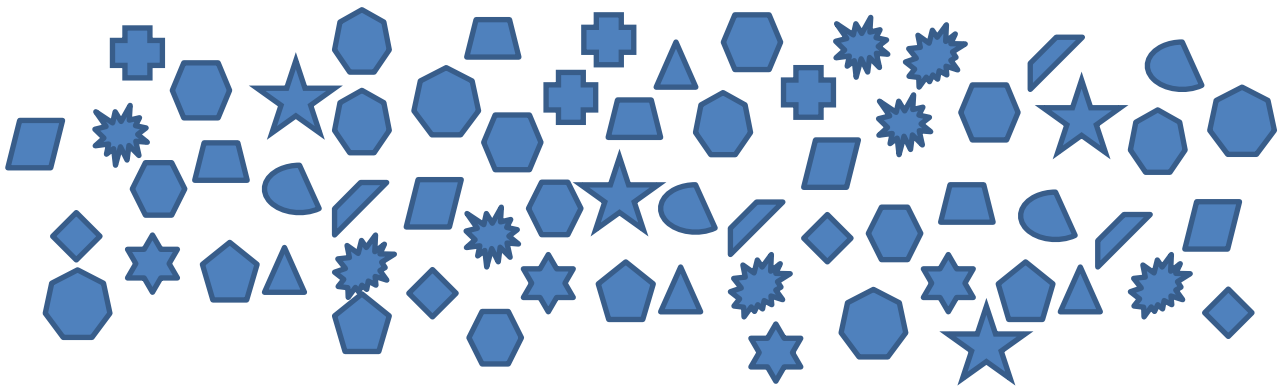
- method that tries to find geometry of ligand/receptor complex

Virtual screening

- identification of compounds with highest affinity towards receptor
- plus special properties …

Screening library

How to obtain the library?

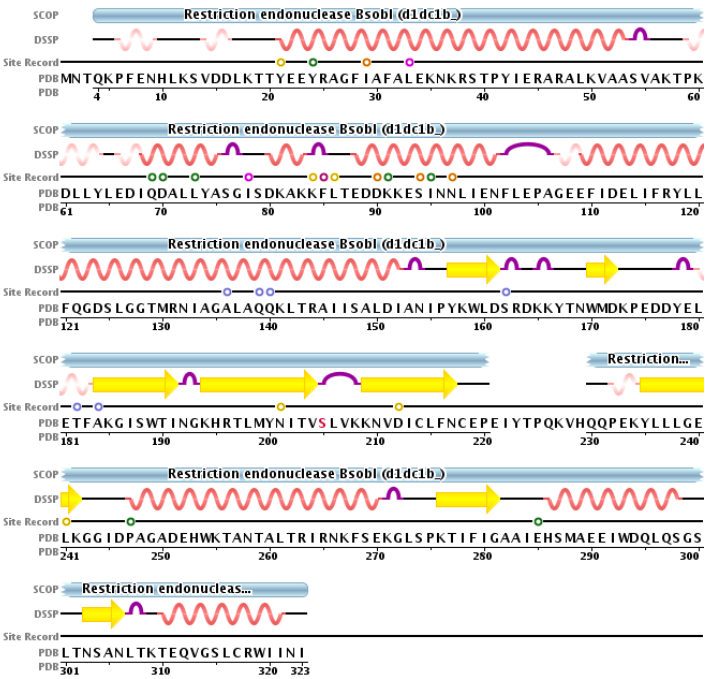


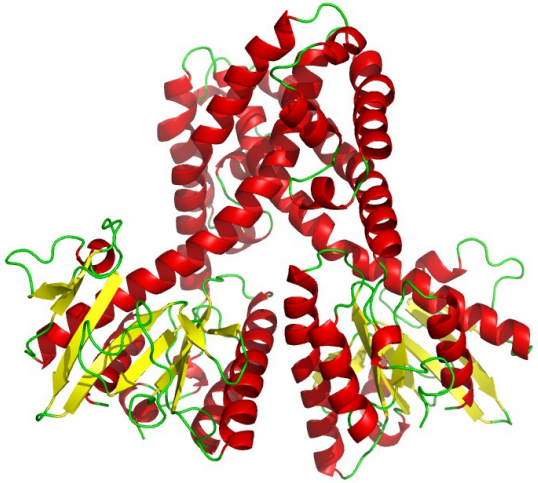
Potential ligand sources:

- **in silico modelling** (2D -> 3D conversion)
- precalculated/experimental structure libraries
  - PubChem (https://pubchem.ncbi.nlm.nih.gov/)
  - ZINC (https://zinc.docking.org/)

3D/2D conversions, biomolecules

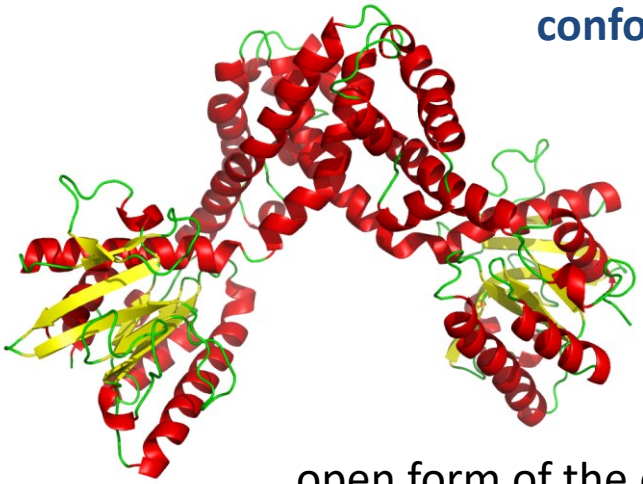
The same primary structure (amino acid sequence).





closed form of the enzyme

different conformations



open form of the enzyme

In molecular modelling, ***ab initio* modelling (prediction)** of biomolecular structures is challenging task.

Experimental structures

Cambridge Structural Database (CSD)

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

It contains about half a million structures of small molecules determined by X-ray and neutron diffraction. Software for working with data:Mercury
http: // www.ccdc.cam.ac.uk/Solutions/CSDSystem/Pages/Mercury.aspx

Protein Data Bank (PDB)

http: //www.pdb.org

It contains about 94 thousand structures biomolecular systems determined mainly by X-ray structural analysis.

Experimental method	Proteins (P)	Nucleic acids (NA)	P / NA complexes	Other	Overall
X ray	77445	1481	4069	3	82998
NMR	8851	1046	193	7	10097
electron microscopy	469	45	129	0	643

status in September 2013

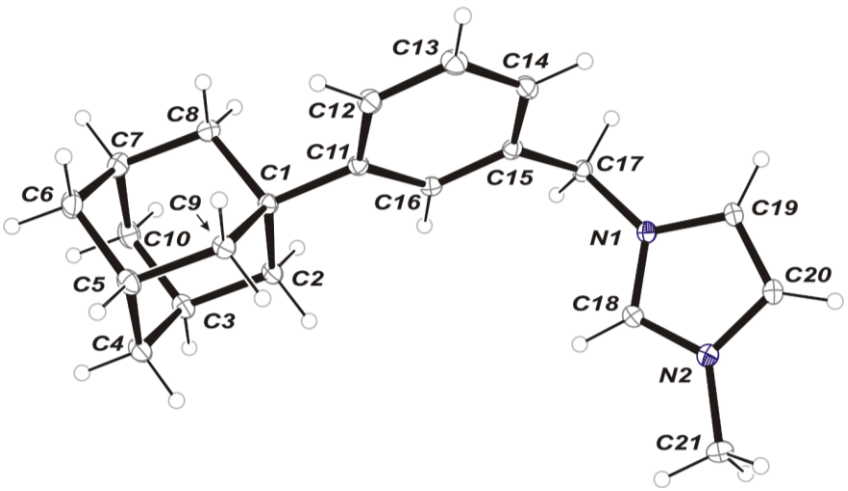
Experimental structures, cont.

- Experimental structures are usually sources for models of biomolecular structures or complicated small molecules.
- Due to low resolution and molecular flexibility, some parts might be unresolved.
- Missing parts need to be modelled in silico
  - hydrogen atoms (assignment can be sensitive to pH, **PROPKA**, https://github.com/jensengroup/propka)
  - flexible protein loops (**Modeller**, https://salilab.org/modeller/)
- **Structures can be influenced by the crystal packing**.
- It is advisable to check source electron density, especially for low-resolution structures.
- Check B-factors to evaluate structure quality.

B-factors

The Debye–Waller factor (DWF, B-factor, temperature factor) is used in condensed matter physics to describe the attenuation of X-ray scattering or coherent neutron scattering **caused by thermal motion**.

ORTEP diagram drawn with 40% ellipsoid probability for non-H atoms



For protein structures:

The B-factors (B) can be taken as indicating the relative vibrational motion of different parts of the structure.

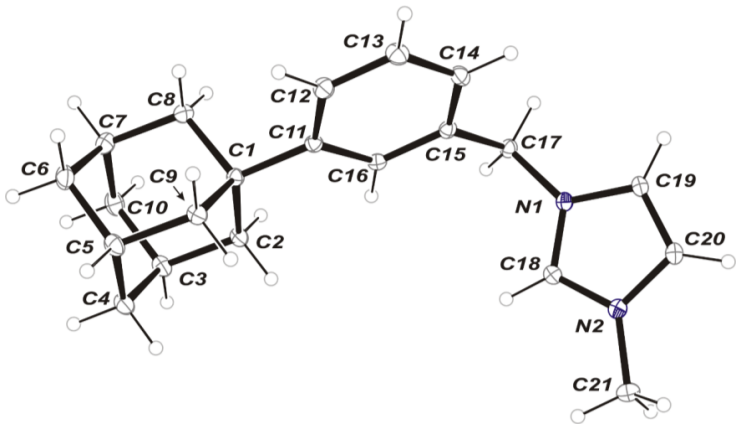
$$B = \frac{8\pi^2}{3} < u^2 >$$

u - displacement of scattering center (atoms)
<> - time or thermal average

- Atoms with **low B-factors** belong to a part of the structure that is **well ordered**.
- Atoms with **large B-factors** generally belong to part of the structure that is **very flexible**.

Similarity modelling

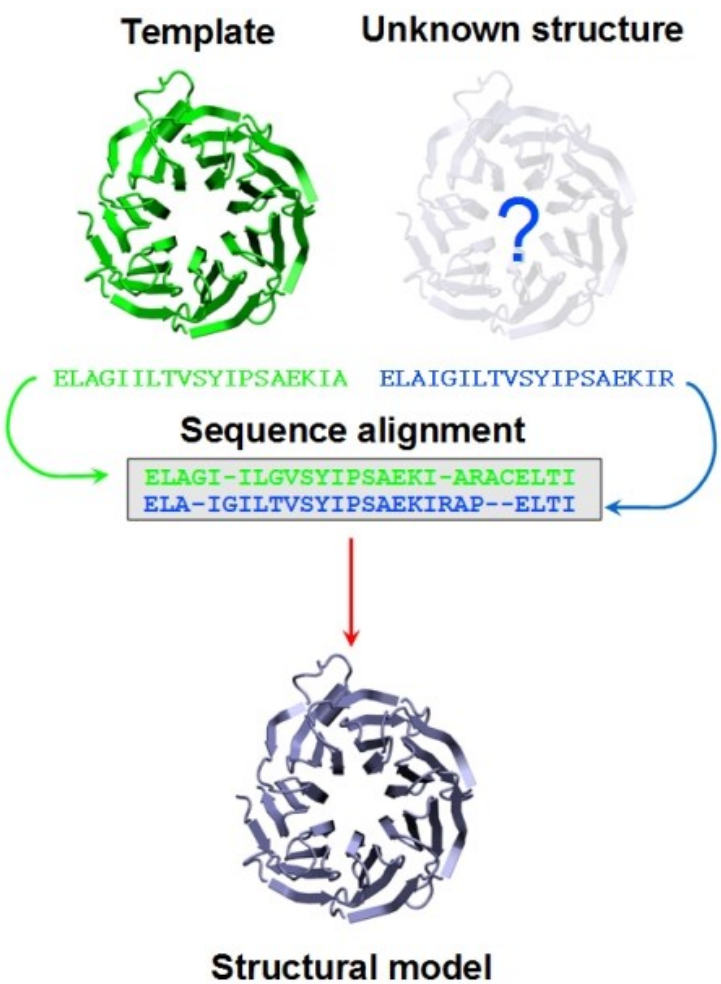
Modifying existing experimental structures



Available experimental structure(s) is (are) modified

- structure substitution
- assembly of complexes
- …

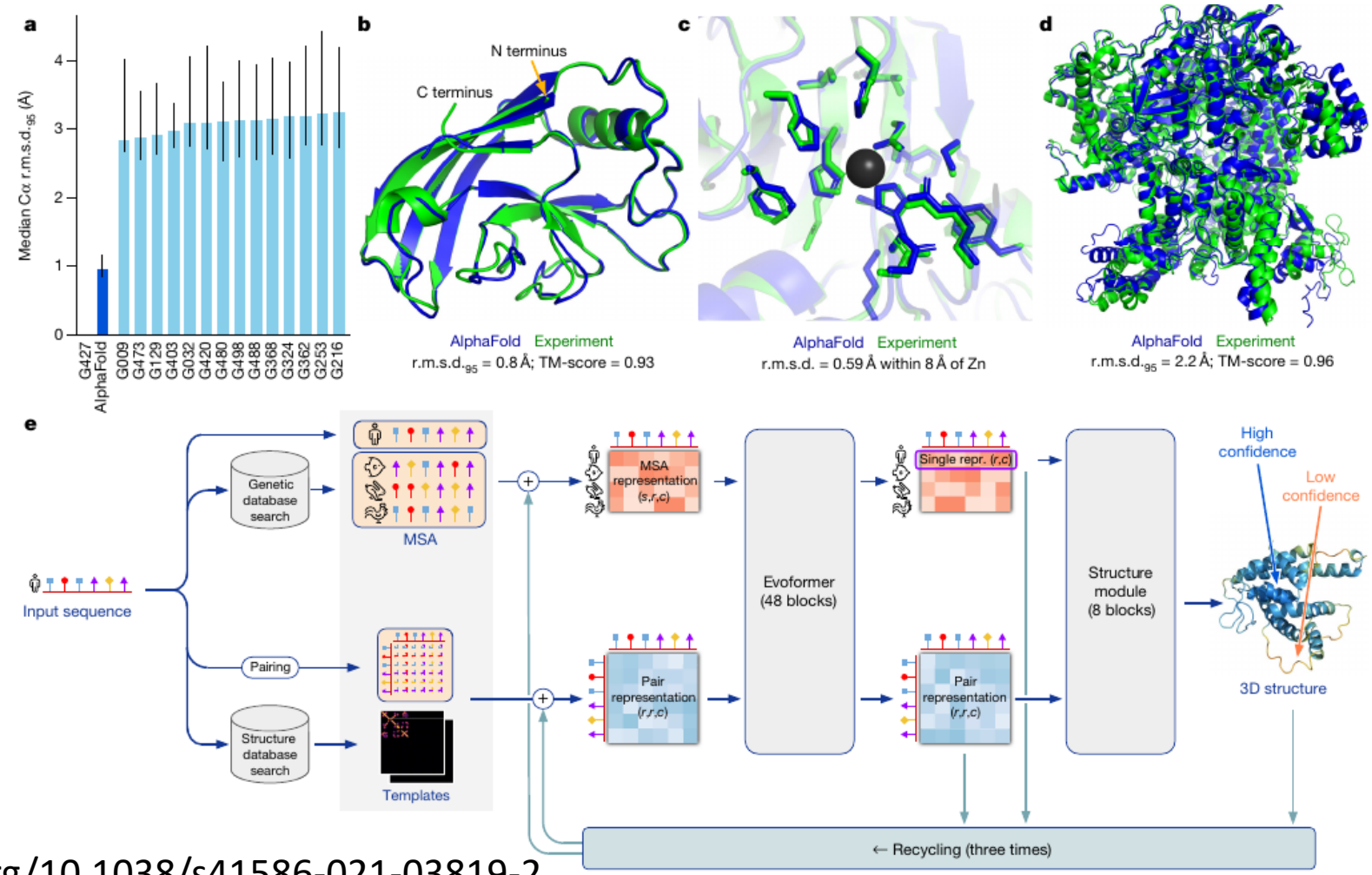
Homology modeling



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

Alphafold

Prediction of the 3D protein structure from amino acid sequence employing machine learning.



https://doi.org/10.1038/s41586-021-03819-2

Summary

- Proper model is a key element for molecular modelling.
  - **Any error in the model propagates to calculated results.**
- Therefore, it is worth to spent some time to check validity of the model (especially for in silico modelled parts).
- It is also advisable to put some effort in cleaning/improving the model (atom names, etc.) as it can save a lot of time in later analyses.