

# Computational Techniques in Chemistry

## Session 3: Molecular simulation of biomolecular systems

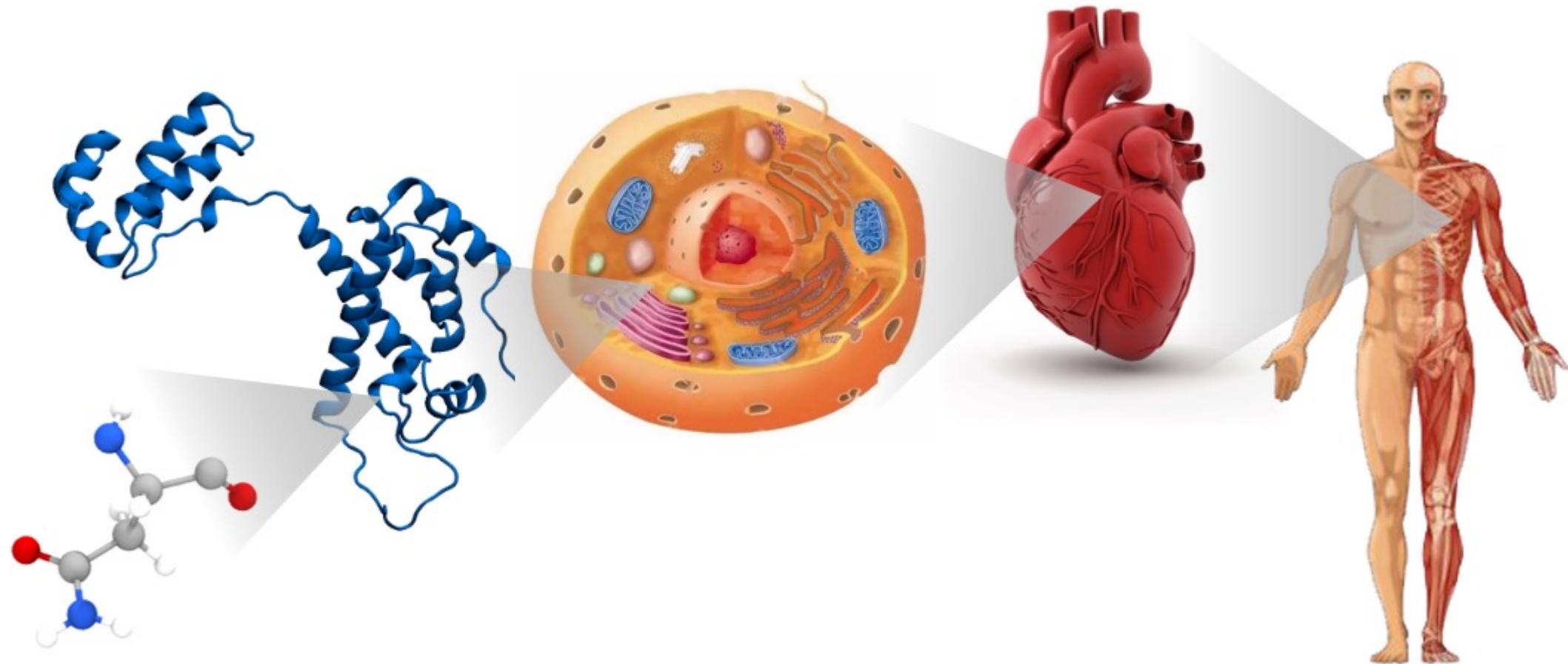
Delivered by **Tim Spankie**  
University of Edinburgh  
January 2025



THE UNIVERSITY of EDINBURGH  
School of Chemistry

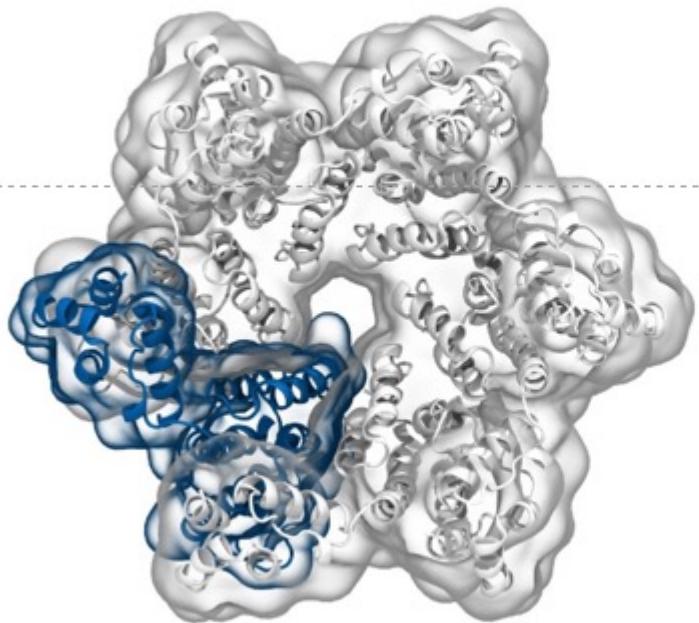
Based on content from a ten-hour series <https://github.com/CCPBioSim/BioSim-analysis-workshop>  
by Matteo Degiacomi ([matteo.degiacomi@ed.ac.uk](mailto:matteo.degiacomi@ed.ac.uk)) and Antonia Mey ([antonia.mey@ed.ac.uk](mailto:antonia.mey@ed.ac.uk))

# The structure determines the (mal)function



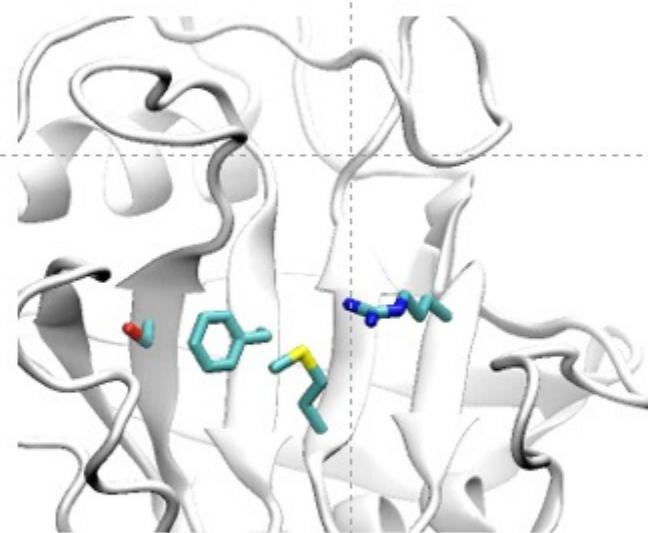
# Structure and dynamics determine protein (mal)function

HIV Capsomer



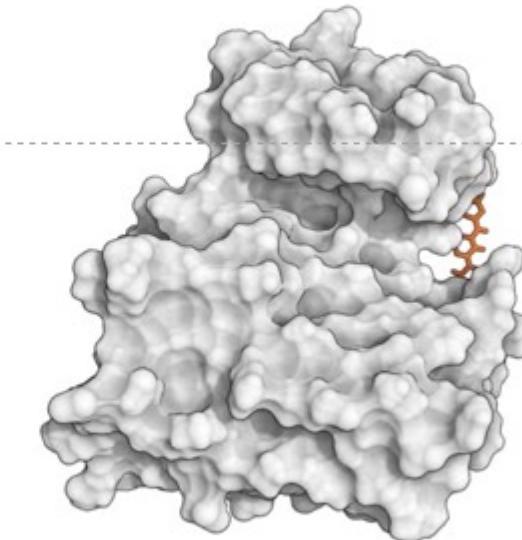
M.T. Degiacomi, *Structure*, 2019

Cyclophilin



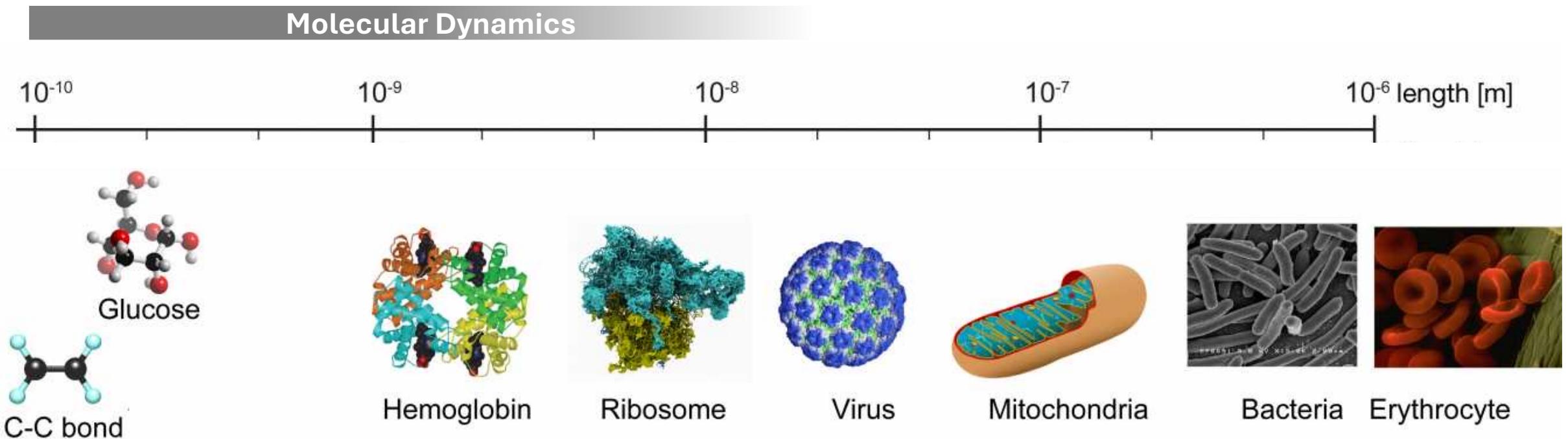
Wapeesittipan, Mey, et al., *Comms. Chem.*, 2019

Tyrosine kinase —dasatanib

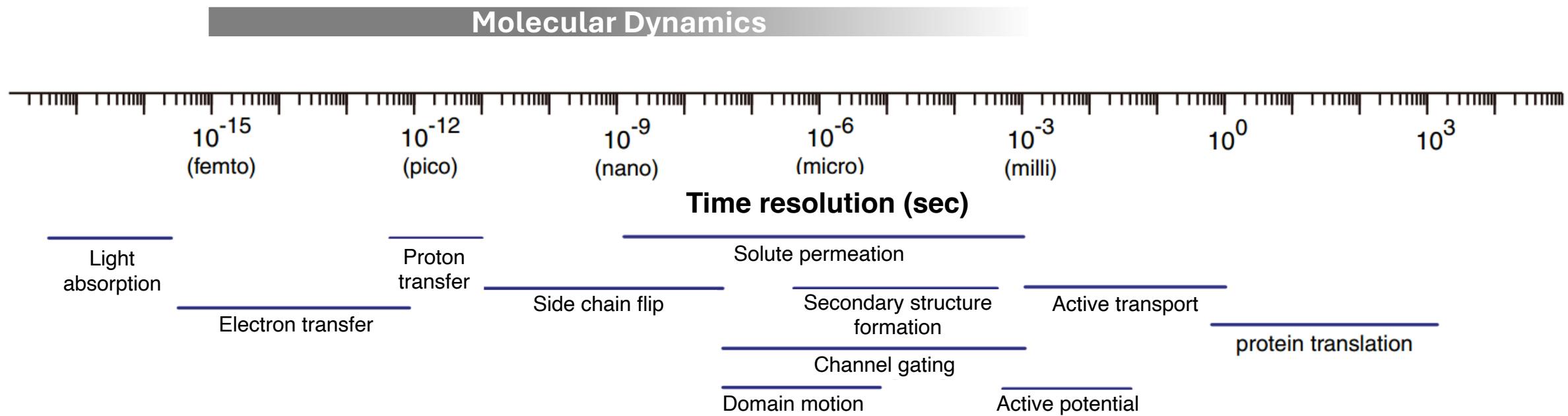


Y Shan et al. *JACS*, 2011

# Sizes in Biochemistry



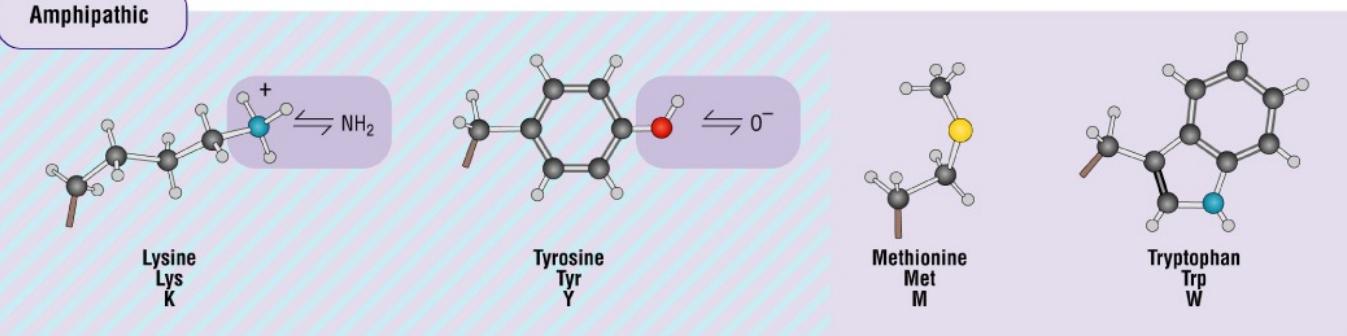
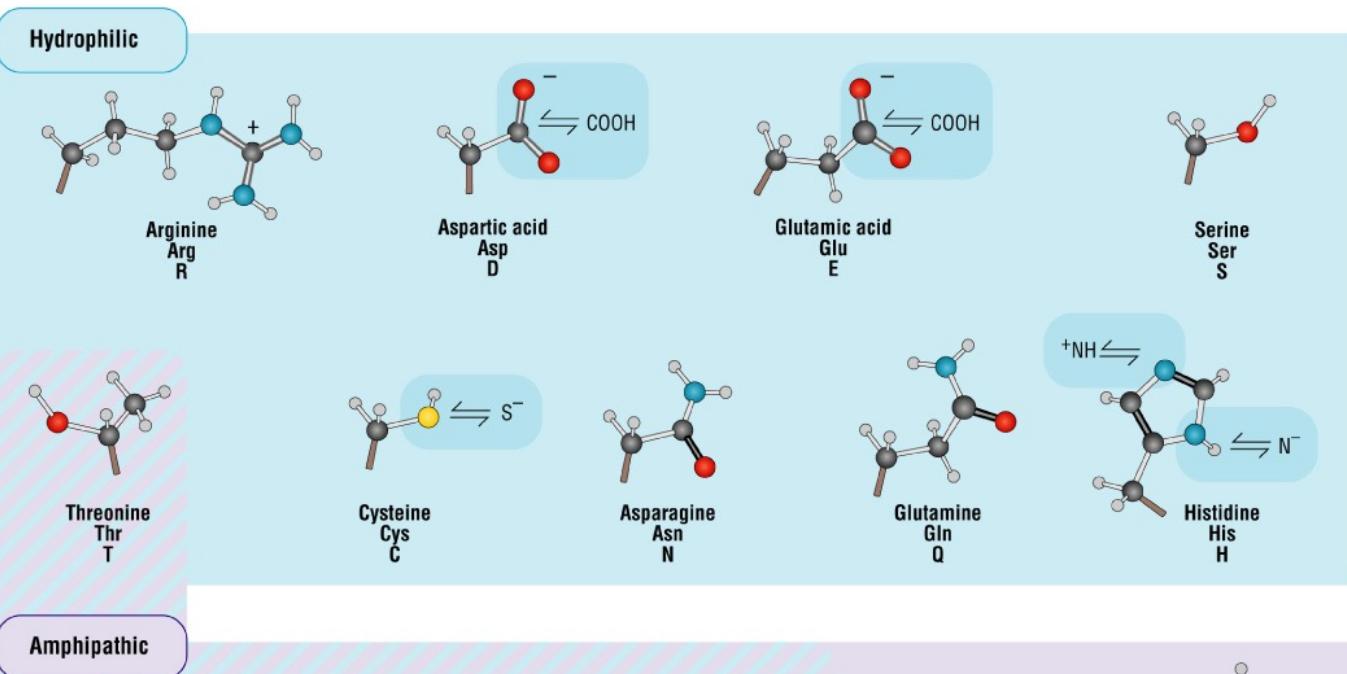
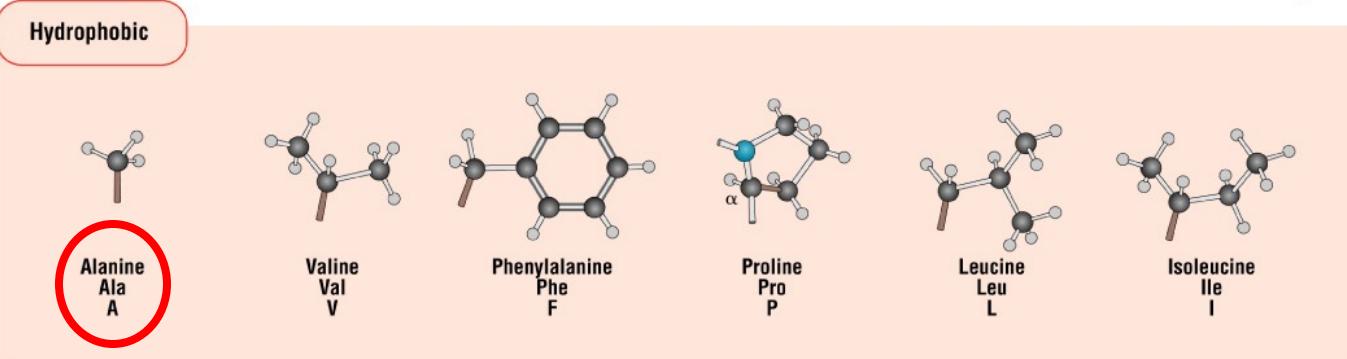
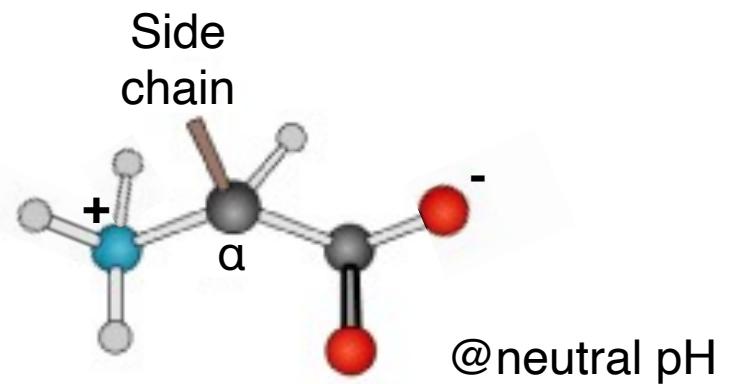
# Timescales in biochemistry



# Proteins are amino acids polymers

Amino acids are composed of:

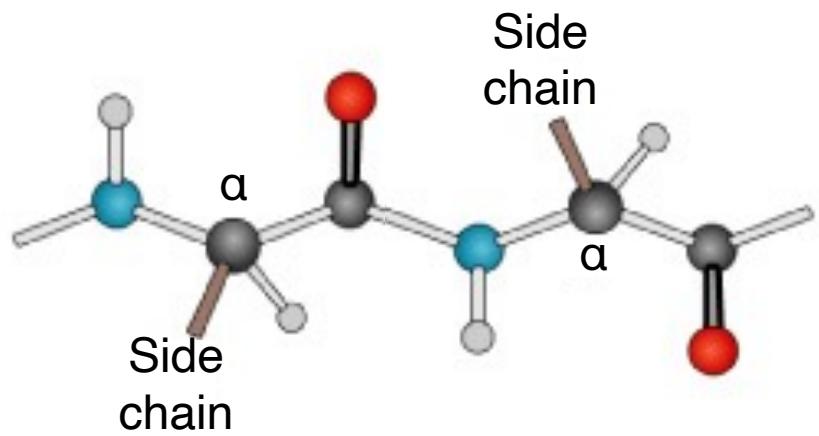
- **Backbone** (conserved)
- **Side chain** (variable)



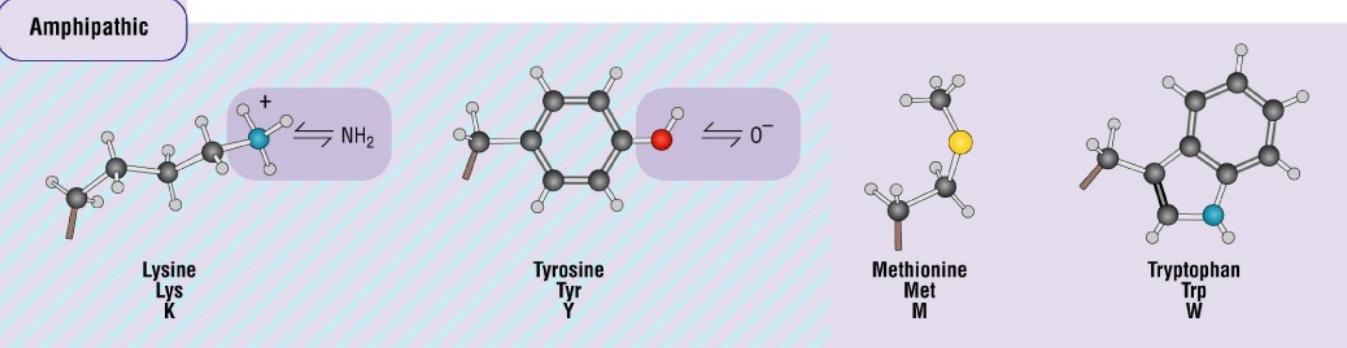
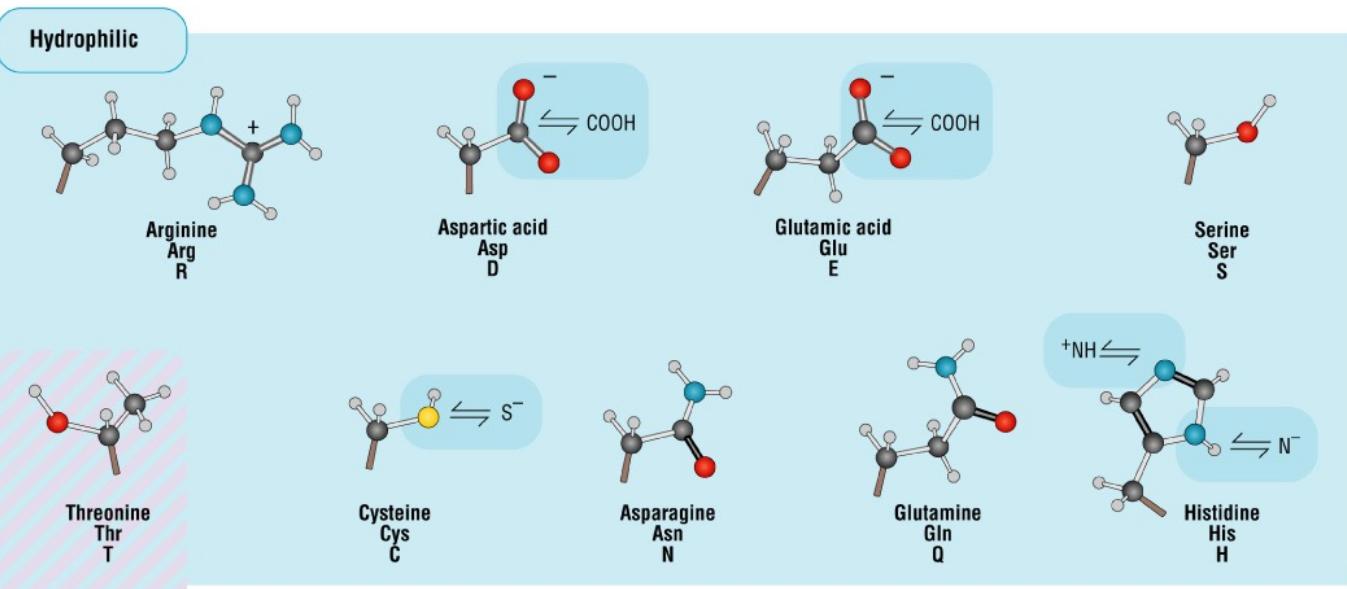
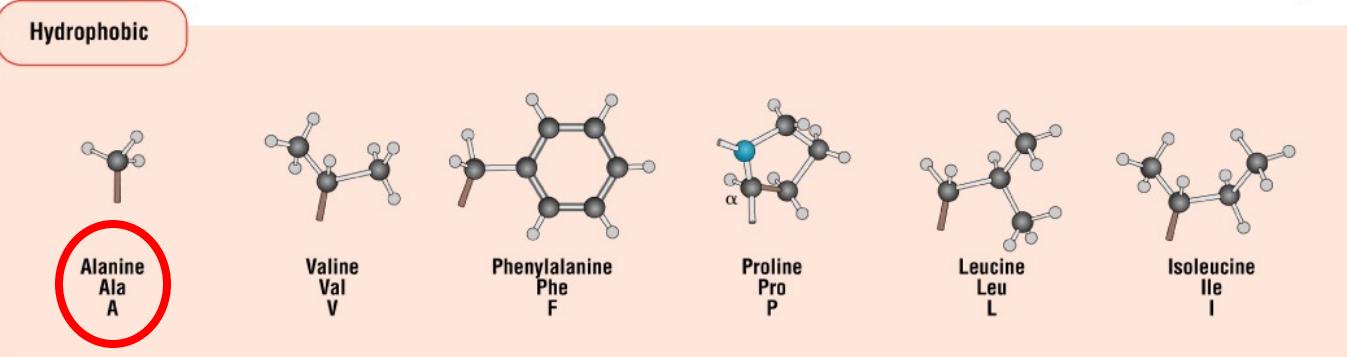
# Proteins are amino acids polymers

Amino acids are composed of:

- **Backbone** (conserved)
- **Side chain** (variable)



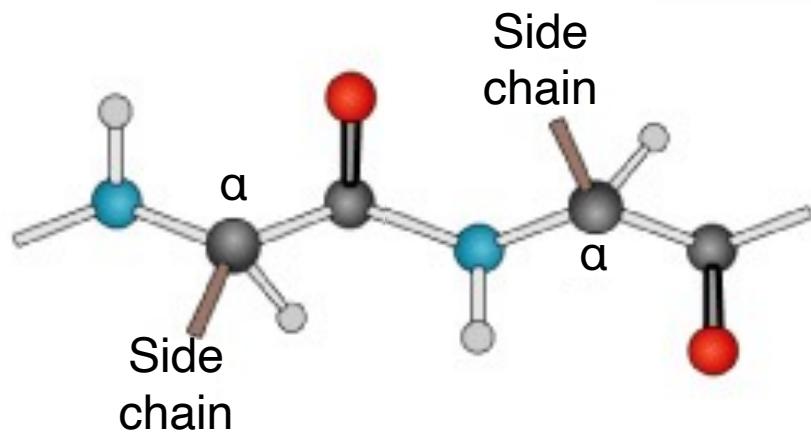
Amino acids polymerize forming a *peptidic bond* (condensation)



# Proteins are amino acids polymers

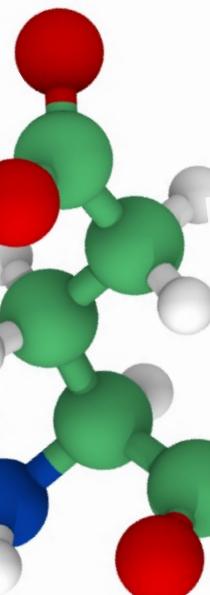
Amino acids are composed of:

- **Backbone** (conserved)
- **Side chain** (variable)

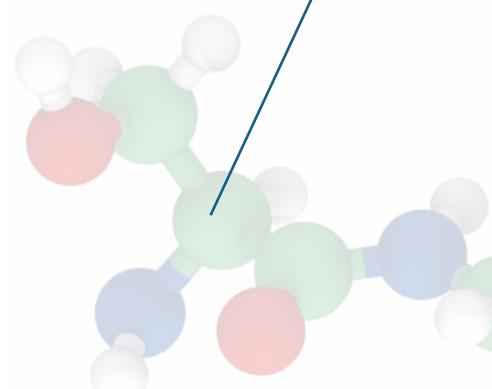


Amino acids polymerize forming a *peptidic bond* (condensation)

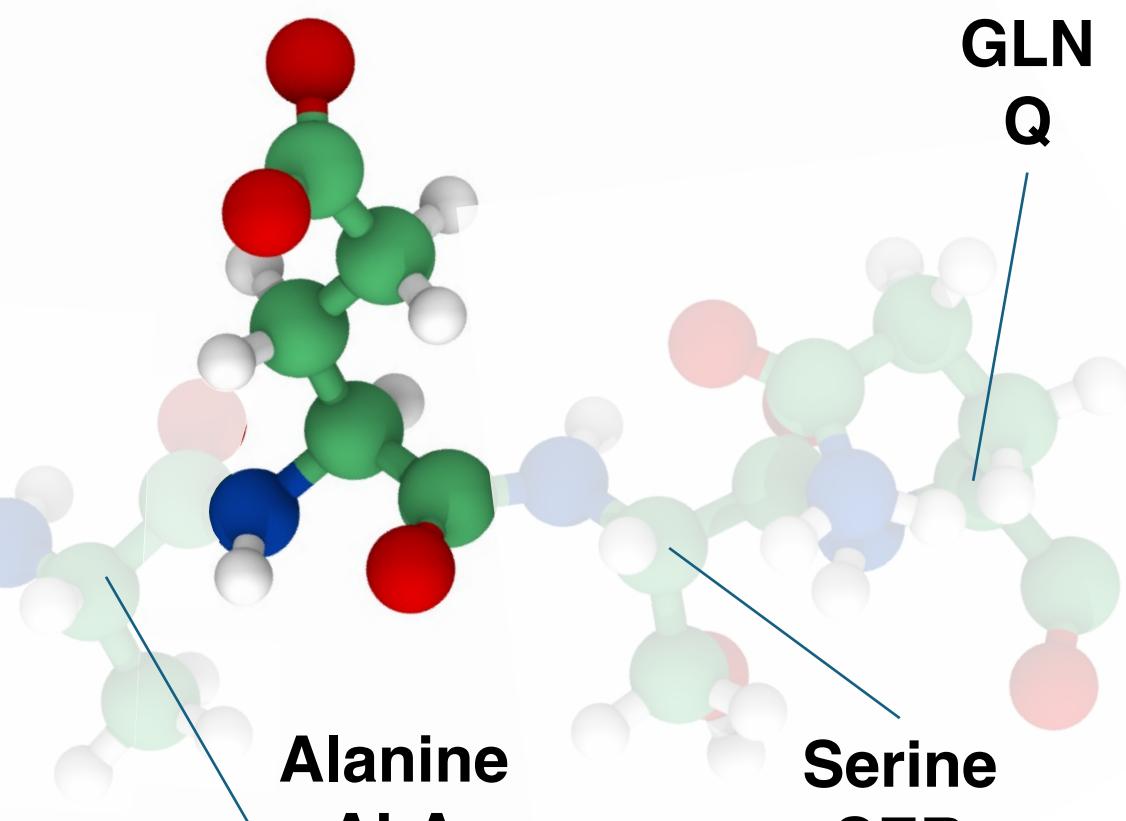
Glutamic acid  
GLU  
E



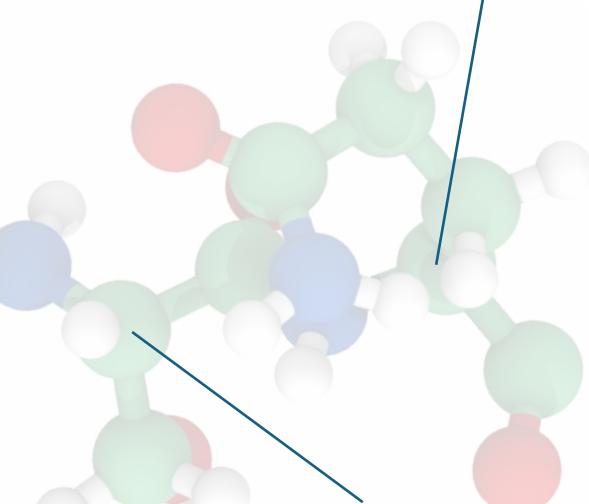
Serine  
SER  
S



Alanine  
ALA  
A



Glutamine  
GLN  
Q



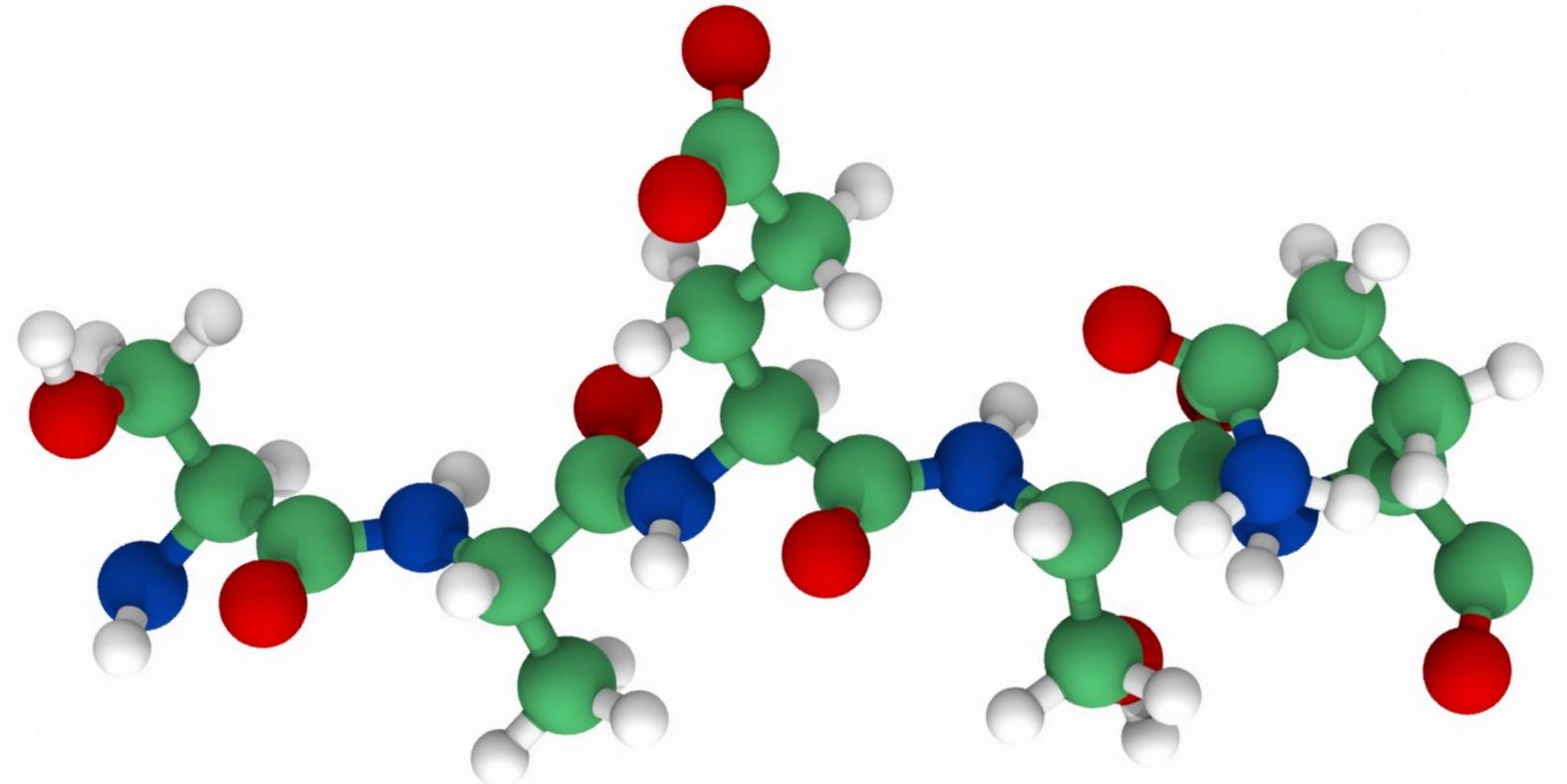
Serine  
SER  
S

# Protein Primary Structure: Sequence

SAESQ

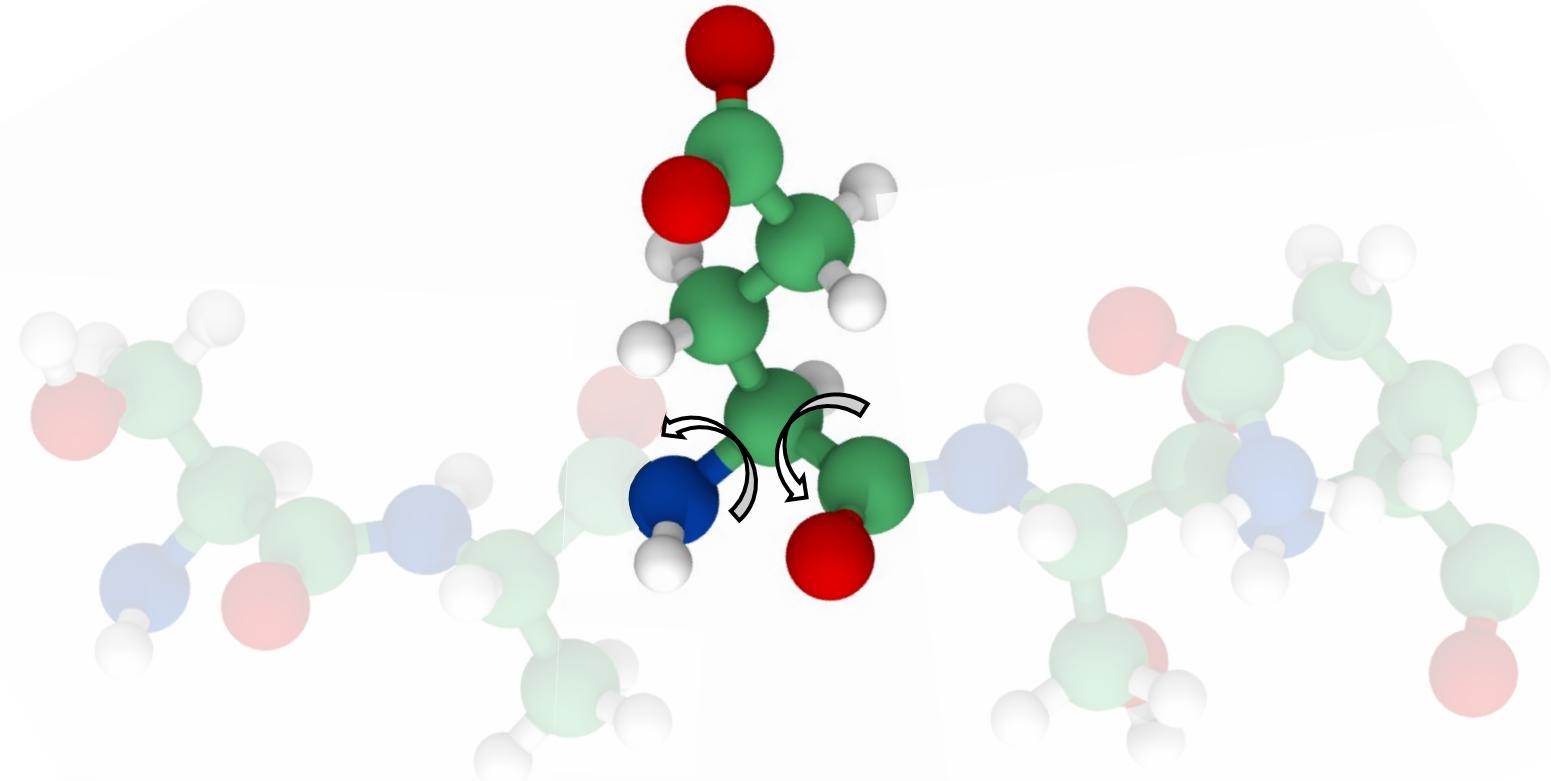
Proteins with similar sequence, likely:

- have similar functions in an organism
- are evolutionarily related



# Protein Secondary Structure

The amino acid chain path is determined by  
backbone torsional angles



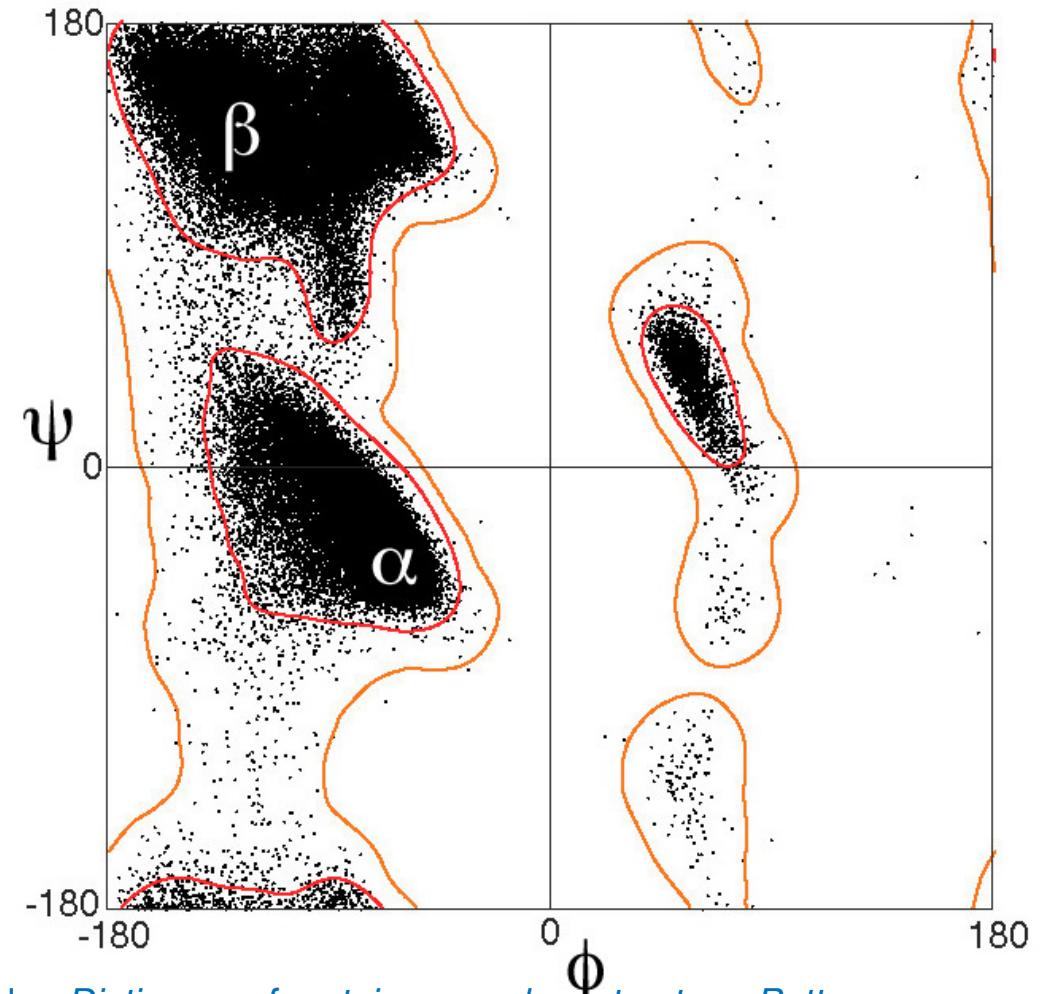
# Protein Secondary Structure

The amino acid chain path is determined by backbone torsional angles

**Ramachandran plot:** scatter plot of amino acids backbone torsional angles

Dictionary of Secondary Structure in Proteins (DSSP) classification defines 7 secondary structure elements (regions in the plot):

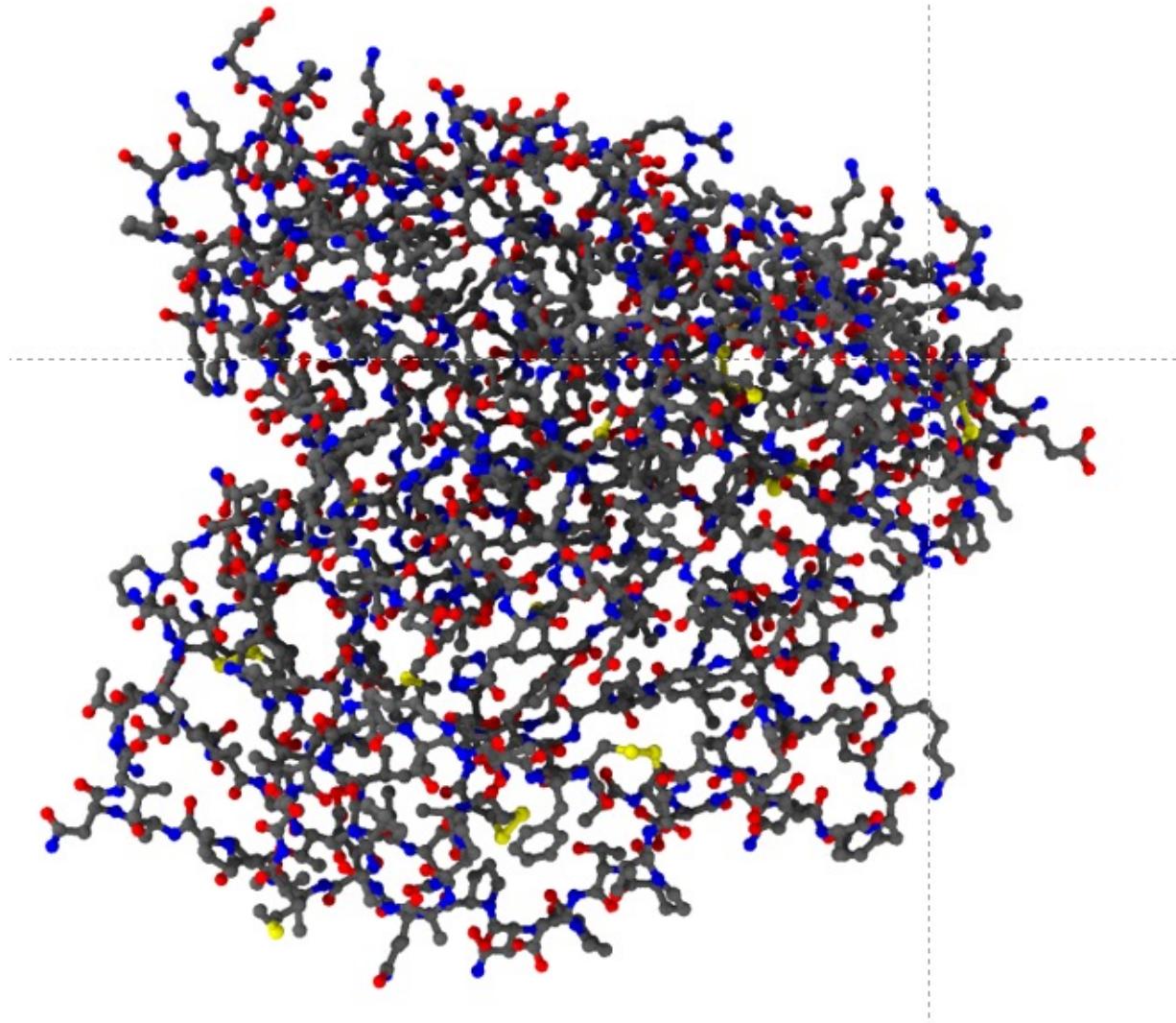
- H =  $\alpha$ -helix
- G =  $3_{10}$  helix
- I =  $\pi$ -helix
- B = residue in isolated  $\beta$ -bridge
- E = extended strand
- T = hydrogen bonded turn
- S = bend



W. Kabsch and C. Sander, *Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features*, *Biopolymers*, 1983

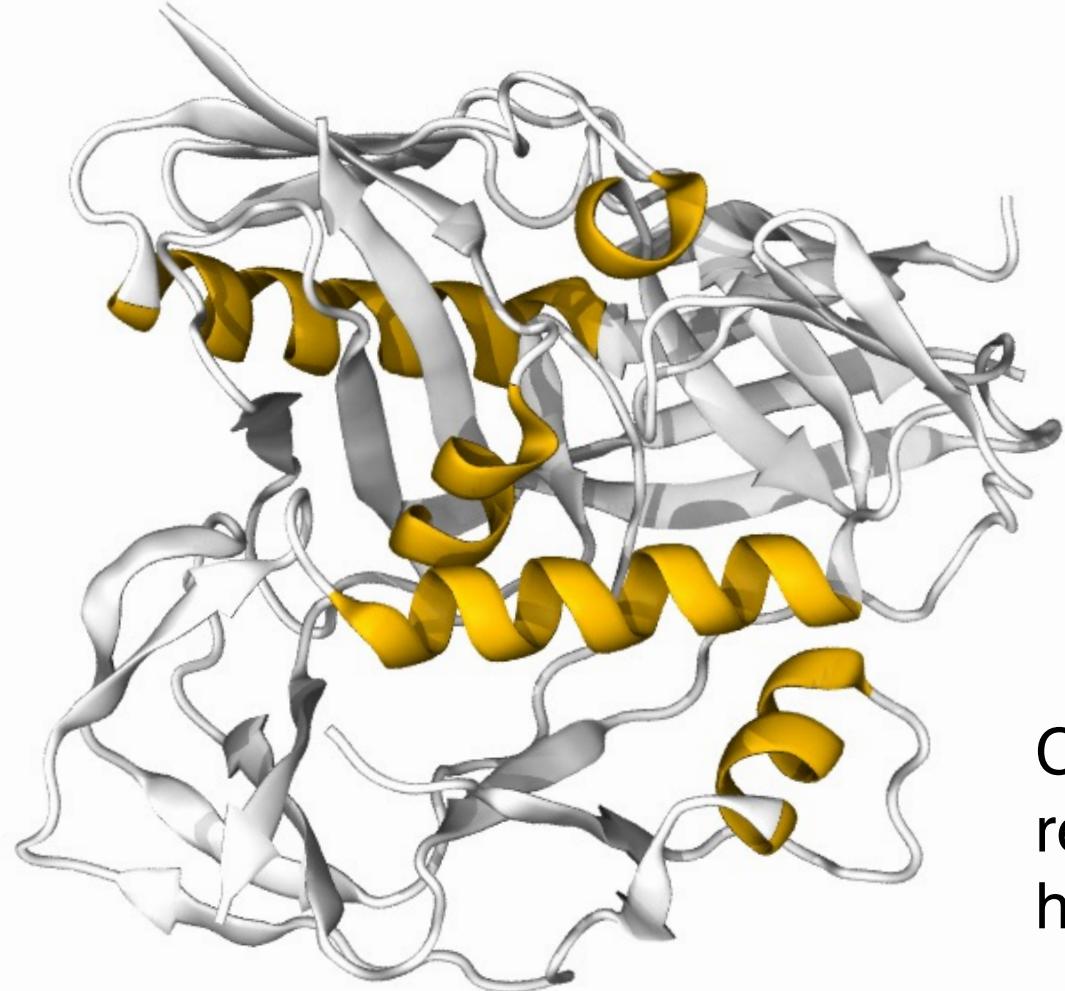
# Protein tertiary structure: folding

Protein («polypeptide»): 10 to >1000 amino acids



# **Protein tertiary structure: folding**

Protein («polypeptide»): 10 to >1000 amino acids



Cartoon  
representation:  
helices

# **Protein tertiary structure: folding**

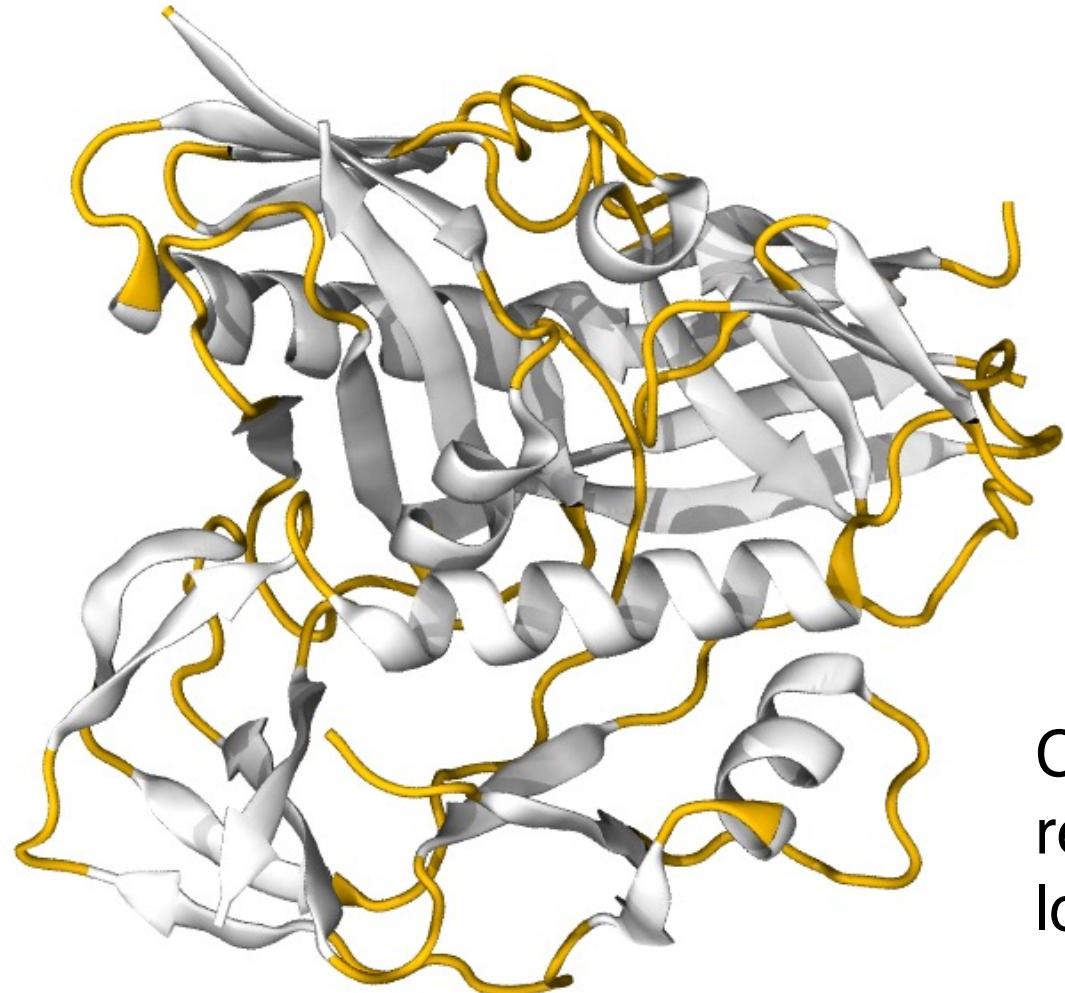
Protein («polypeptide»): 10 to >1000 amino acids



Cartoon  
representation:  
sheets

# **Protein tertiary structure: folding**

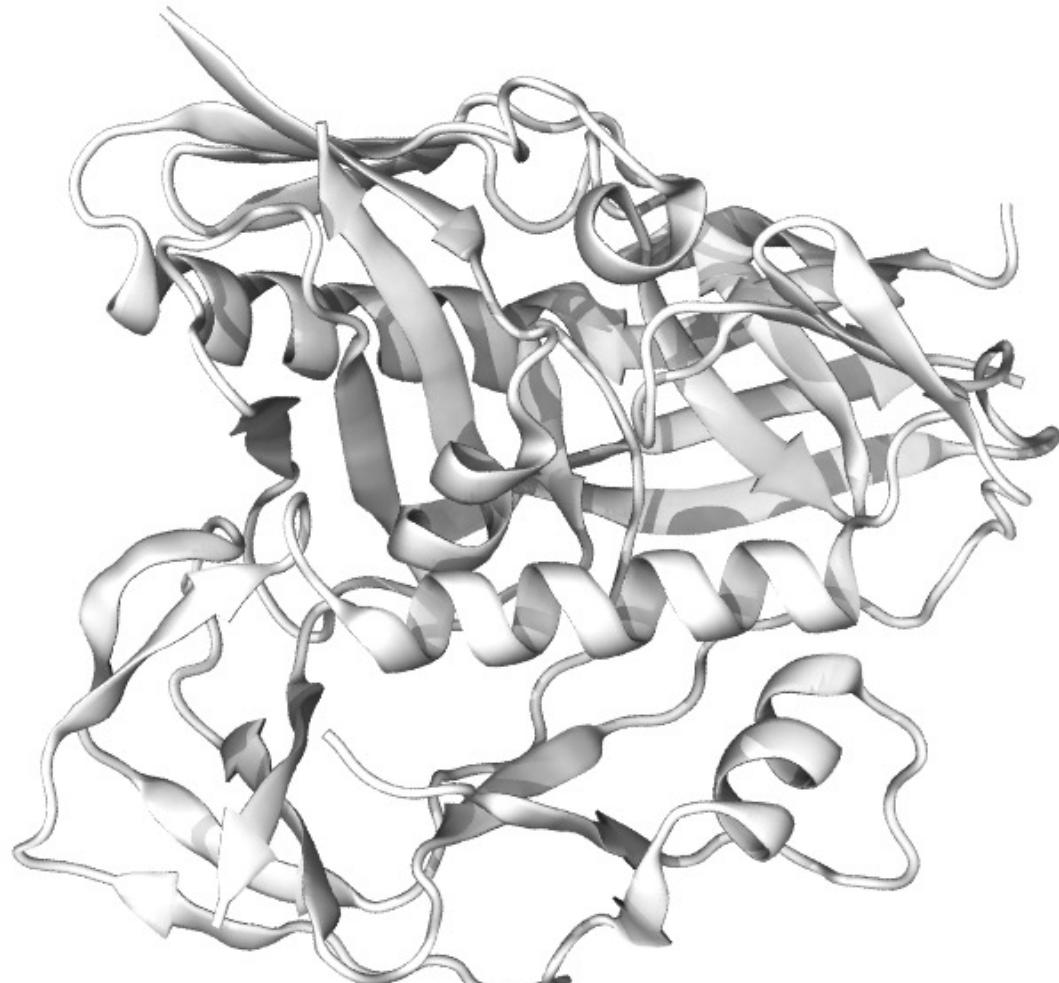
Protein («polypeptide»): 10 to >1000 amino acids



Cartoon  
representation:  
loops

# Protein tertiary structure: folding

Protein («polypeptide»): 10 to >1000 amino acids



## Anfinsen's dogma

The three-dimensional structure of a protein in its native environment is solely determined by its amino acid sequence.



Christian Anfinsen. *Principles that govern the folding of protein chains*, *Science*, 1973

# Hydrogen bonds

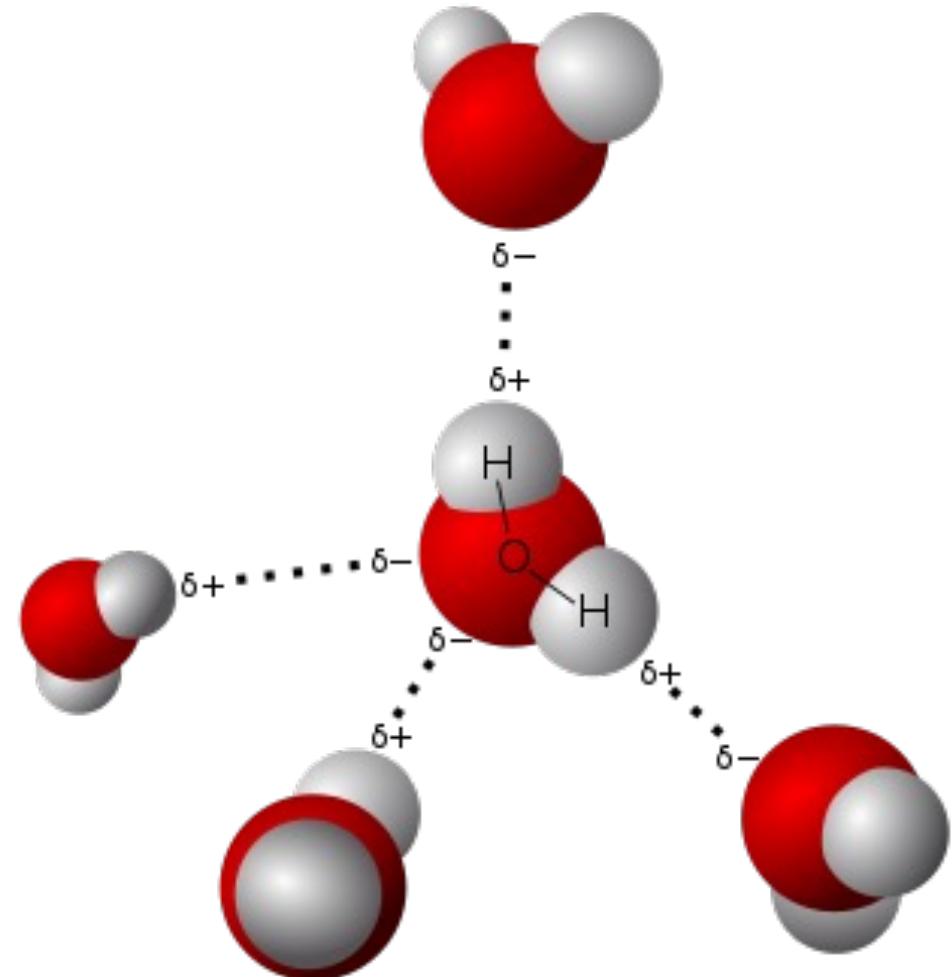
Electrostatic interaction. Structure:

- donor-acceptor distance typically 1.6-2 Å
- donor-acceptor-hydrogen angle must be small

Hydrogen bond energy in biomolecular systems typically 5-25 kJ/mol, e.g.:

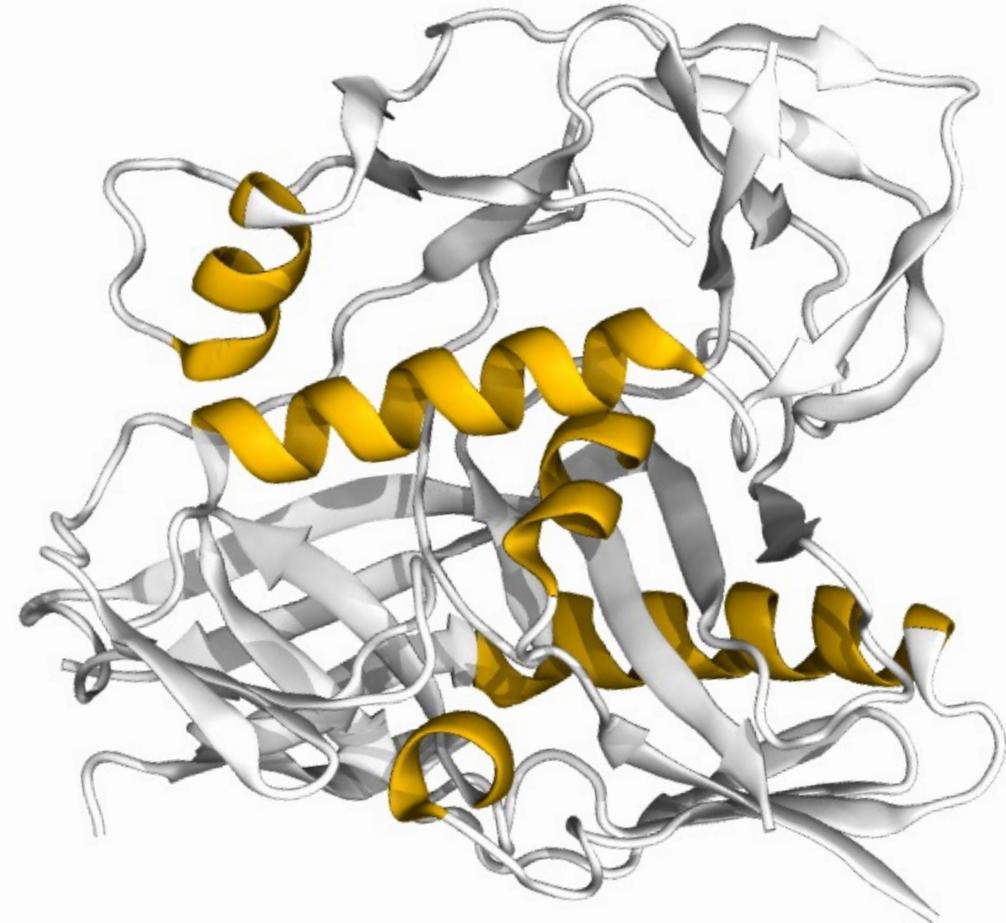
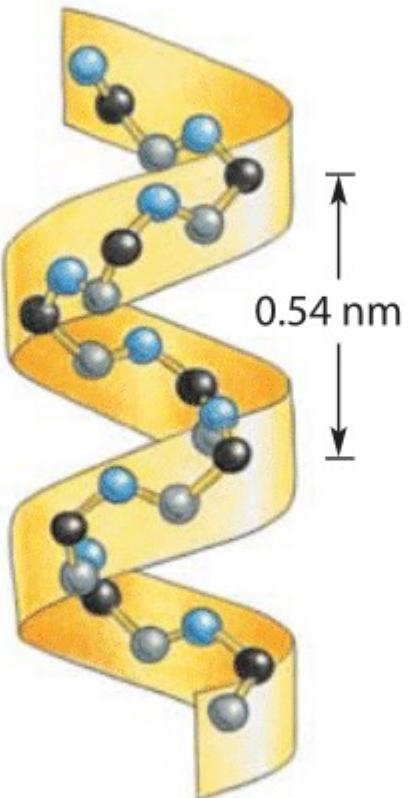
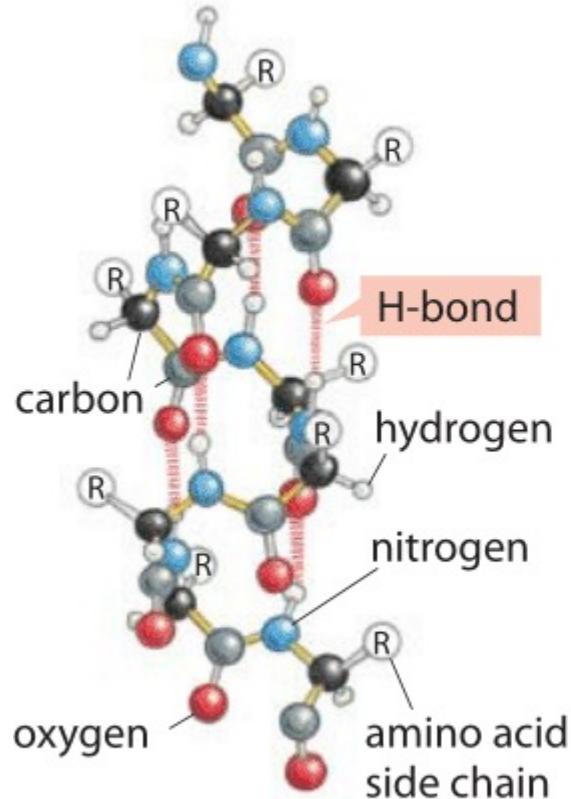
- O–H $\cdots$ :O , 21 kJ/mol (5.0 kcal/mol)
- N–H $\cdots$ :O , 8 kJ/mol (1.9 kcal/mol)

amino acids' backbone and polar side chains can be donor/acceptor



# Hydrogen bonding on protein backbone

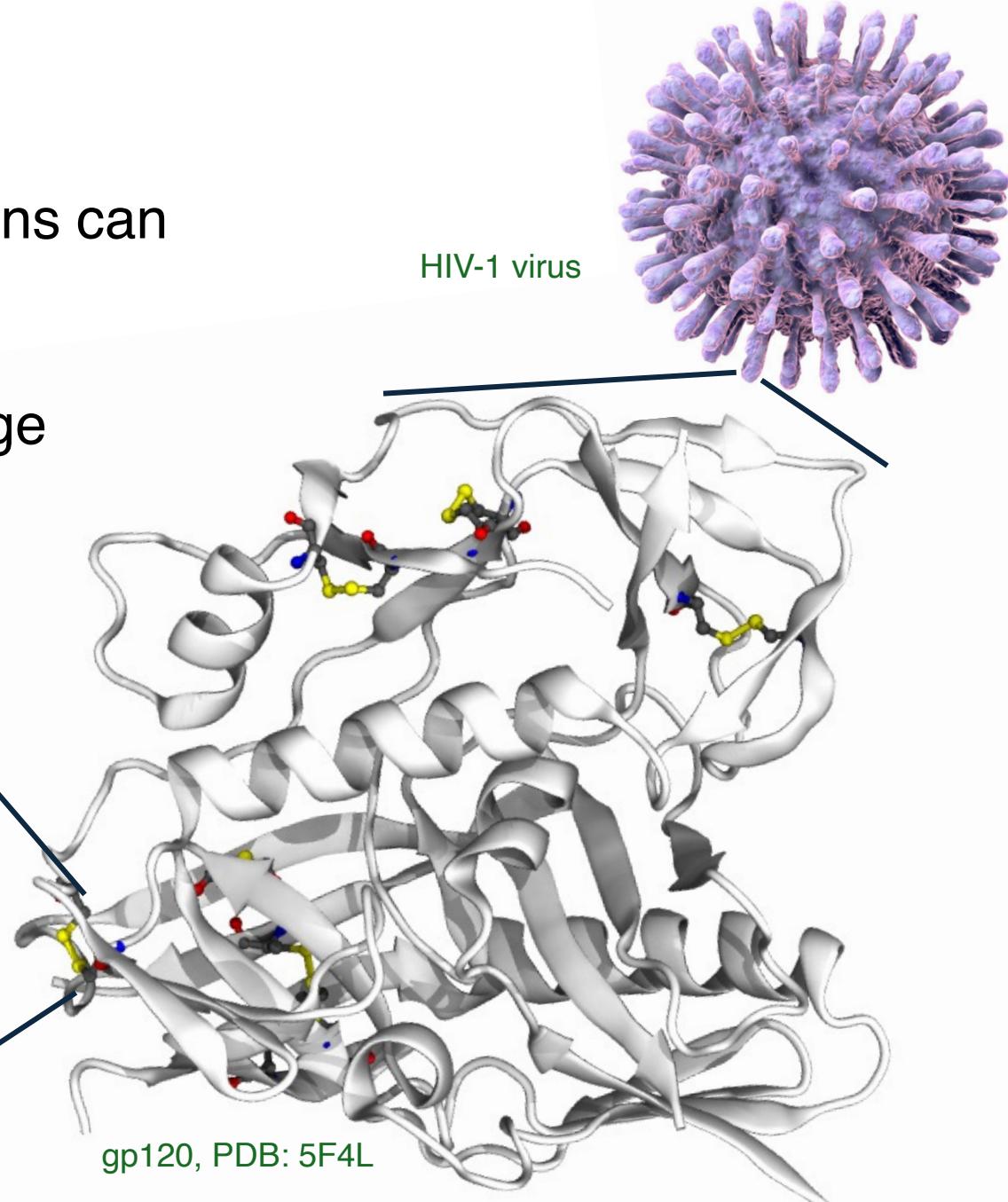
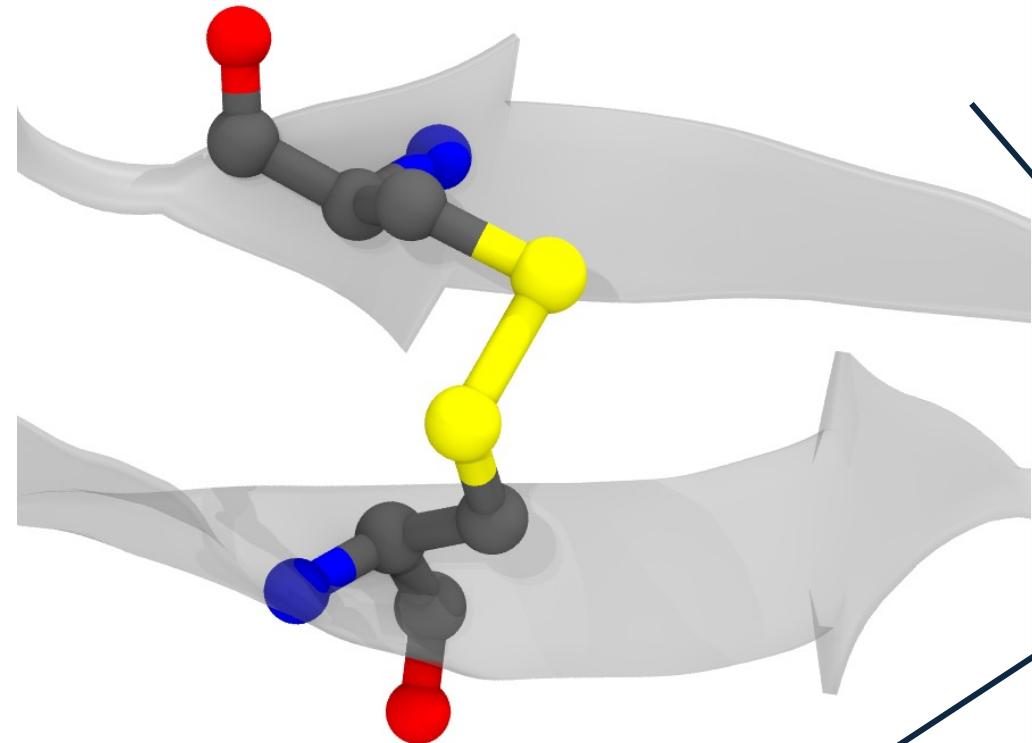
alpha helix



# Disulfide bridges

Under *oxidising* conditions, cysteine side chains can form a *covalent bond*.

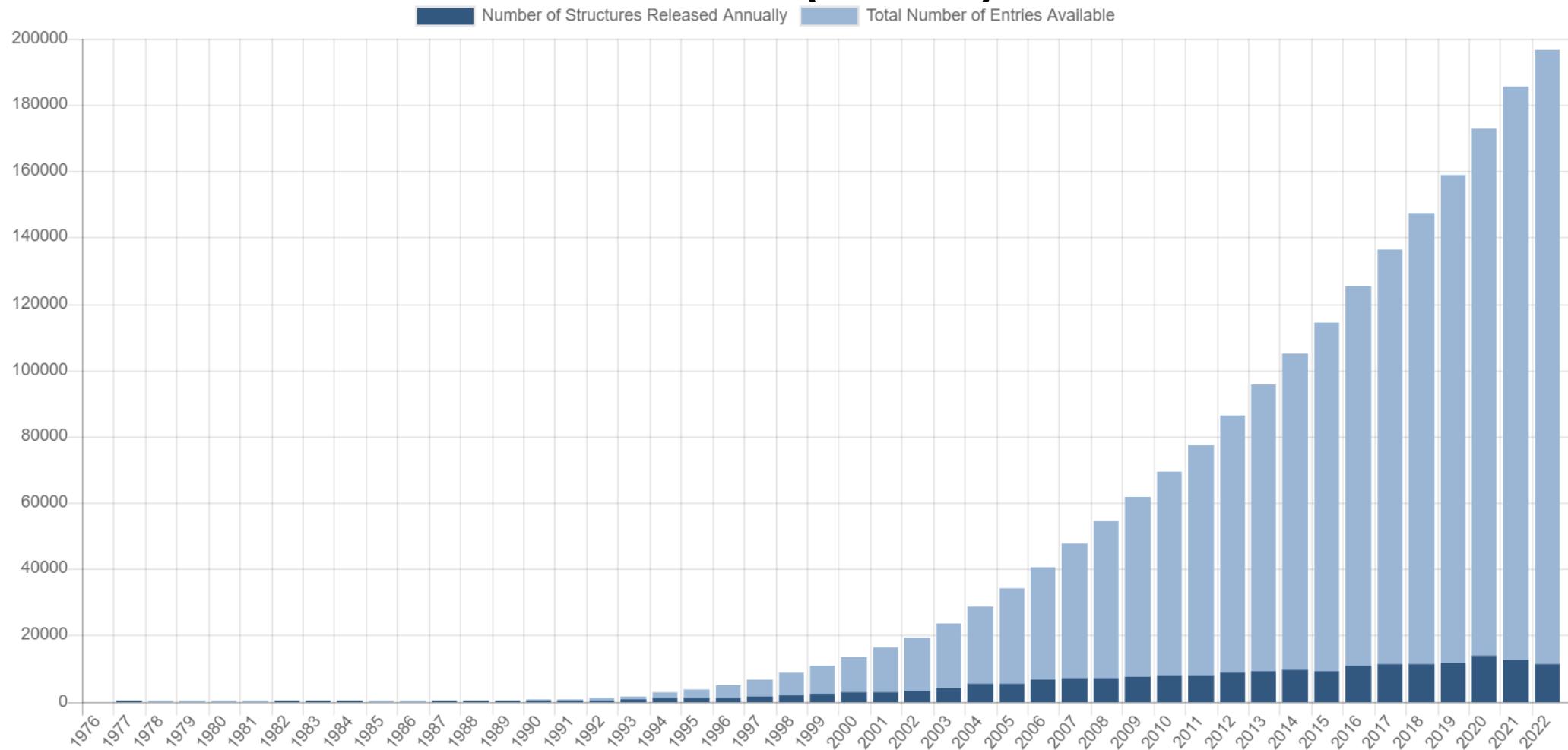
- Protein more resistant to denaturation
- Changes in conditions → structural change



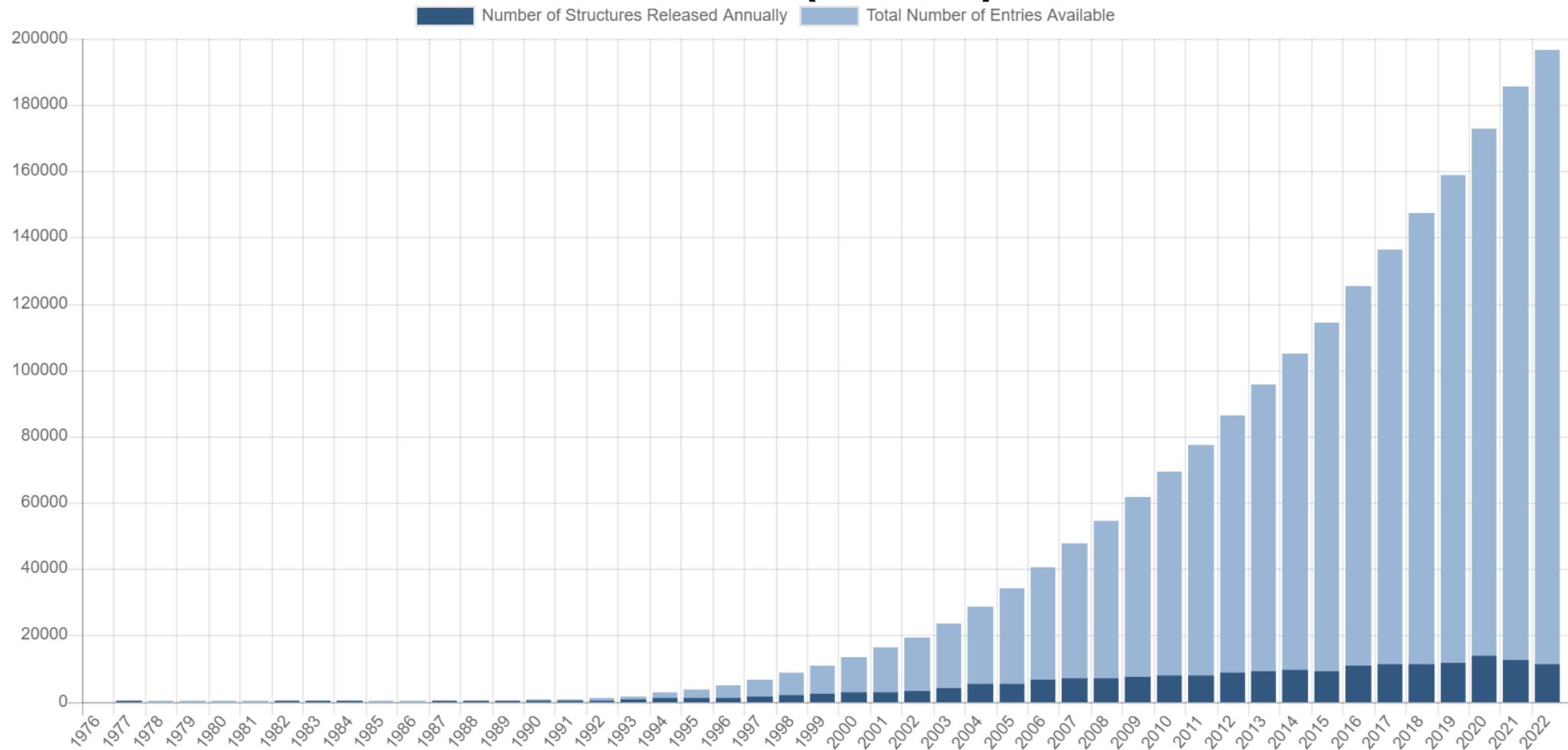
# Proteins, illness, and drug design

- **Proteins in diseases** (e.g., COVID-19, salmonella, flu, ...)
  - pathogen's own metabolism/structure
  - pathogen's weapon
- **Proteins in disorders** (e.g., Cancer, Alzheimer's, ...)
  - own protein misfolds
  - own protein folds, but has different dynamics
- **Drug**
  - Small molecule designed to specifically bind to a protein, so as to affect its function

# The Protein Data Bank (PDB)



# The Protein Data Bank (PDB)



known protein **structures**: ~80'000 (*PDB, 90% identity*)  
known protein **sequences**: ~190'000'000 (*UNIPROT*)

# Protein fold prediction

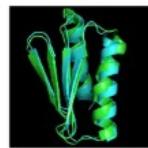
Protein Sequence

SQETRKKCTEMKKFKNCEVRCDESNHCVRCSDTKYTLC



prediction

Structure



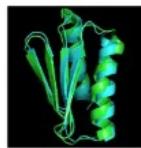
**CASP**, since 1994 biennial competition on protein fold prediction: [predictioncenter.org](http://predictioncenter.org)

# Protein fold prediction

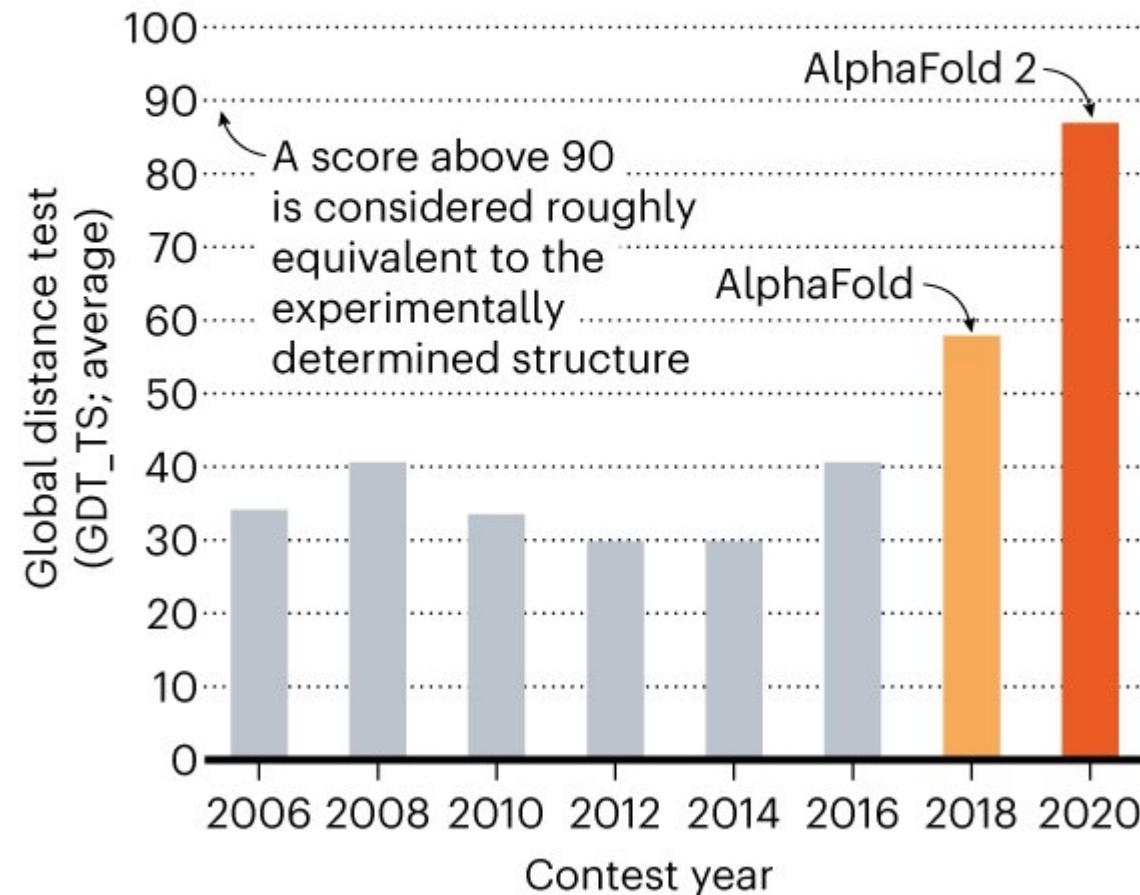
Protein Sequence

SQETRKKCTEMKKFKNCEVRCDESNHCVRCSDTKYTLC

prediction



Structure



A.W. Senior et al., *Improved protein structure prediction using potentials from deep learning*, **Nature**, 2020

J. Jumper et al., *Highly accurate structure prediction with AlphaFold*, **Nature**, 2021

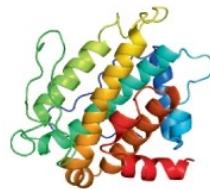
E. Callaway, 'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures, **Nature**, 2021

# Protein fold prediction: AlphaFold

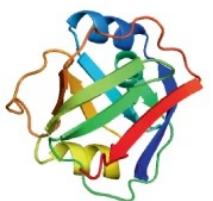
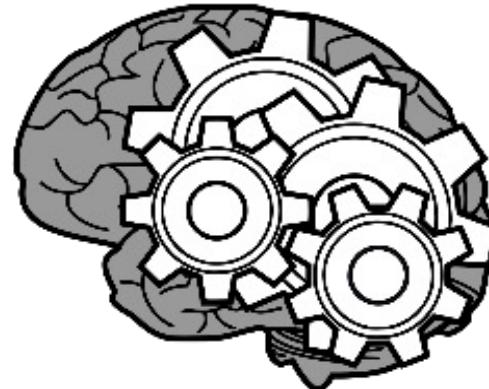
training



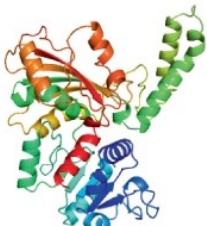
NITPLAKRTYNYRAVL...



HTIWKL...SRLWSLQ...



LIYRICMPGILCYEND...



ALRIKAIVVPRILGPQ...

# Protein fold prediction: AlphaFold

MAPVYEGMAS...

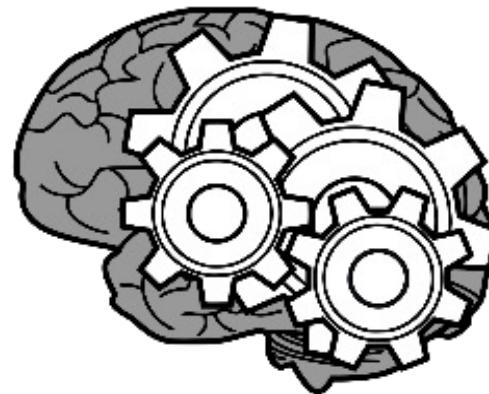
HQQVFSPHTL...

QSSAFCSVKK...

LKIEPSSNWD...

MTGYGSHSKV...

...

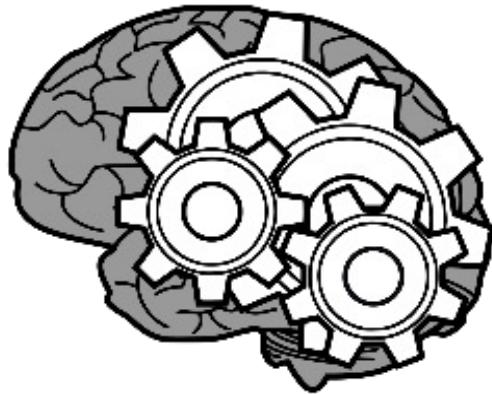


prediction

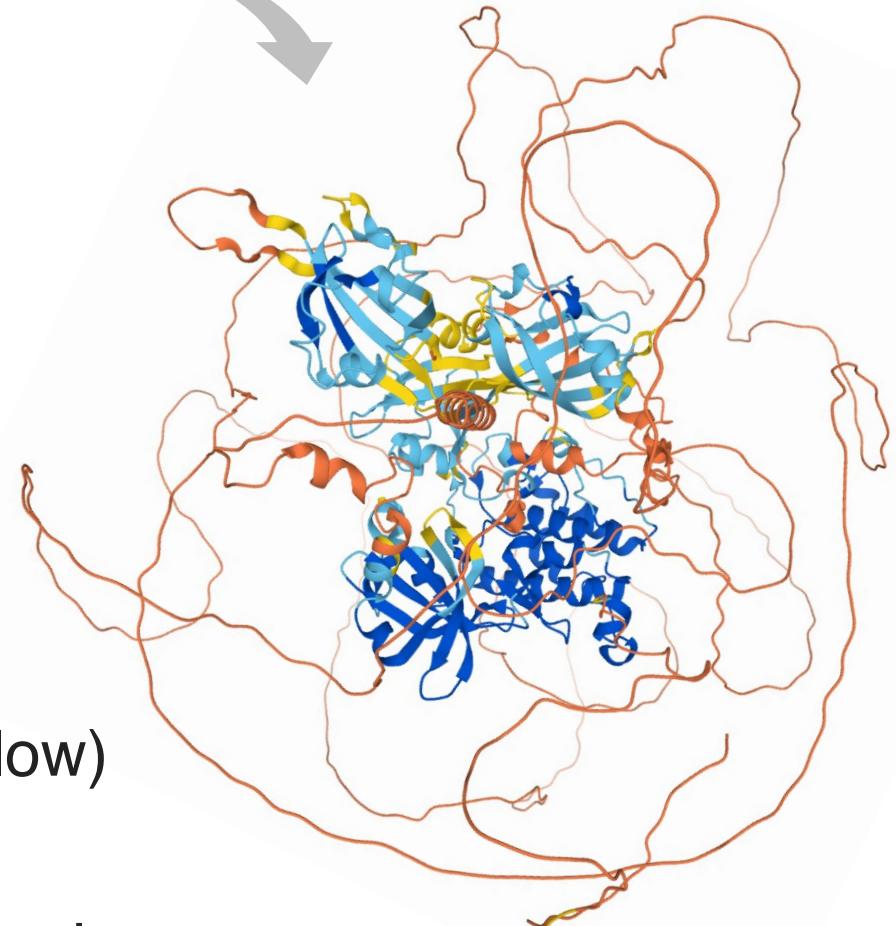


# Protein fold prediction: AlphaFold

YSQSKNIPLS...



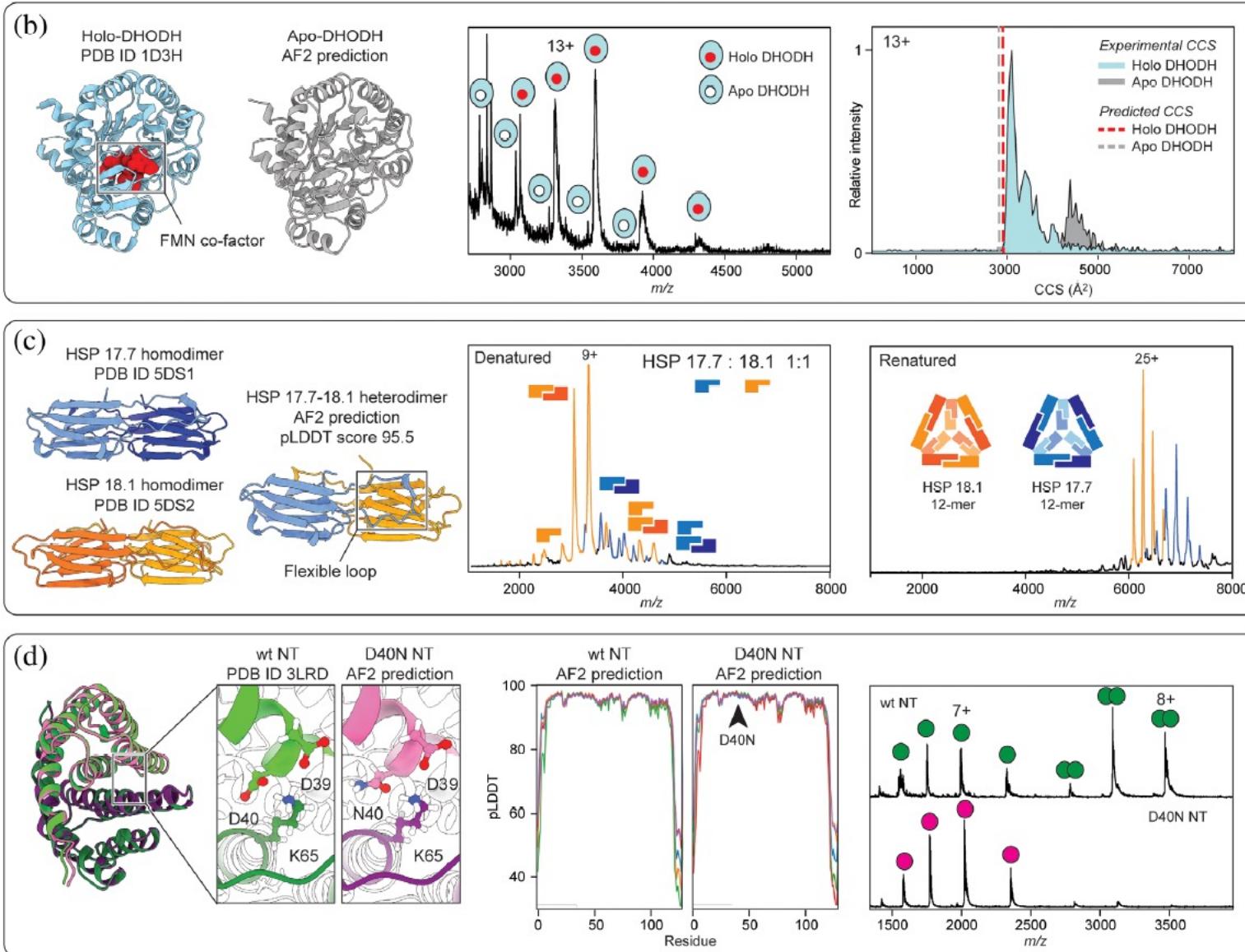
prediction



**PLDDT score:** confidence score (**100** = high, **<50** = low)

Warning: even high-confidence models can be incorrect.

# Protein fold prediction: warning!



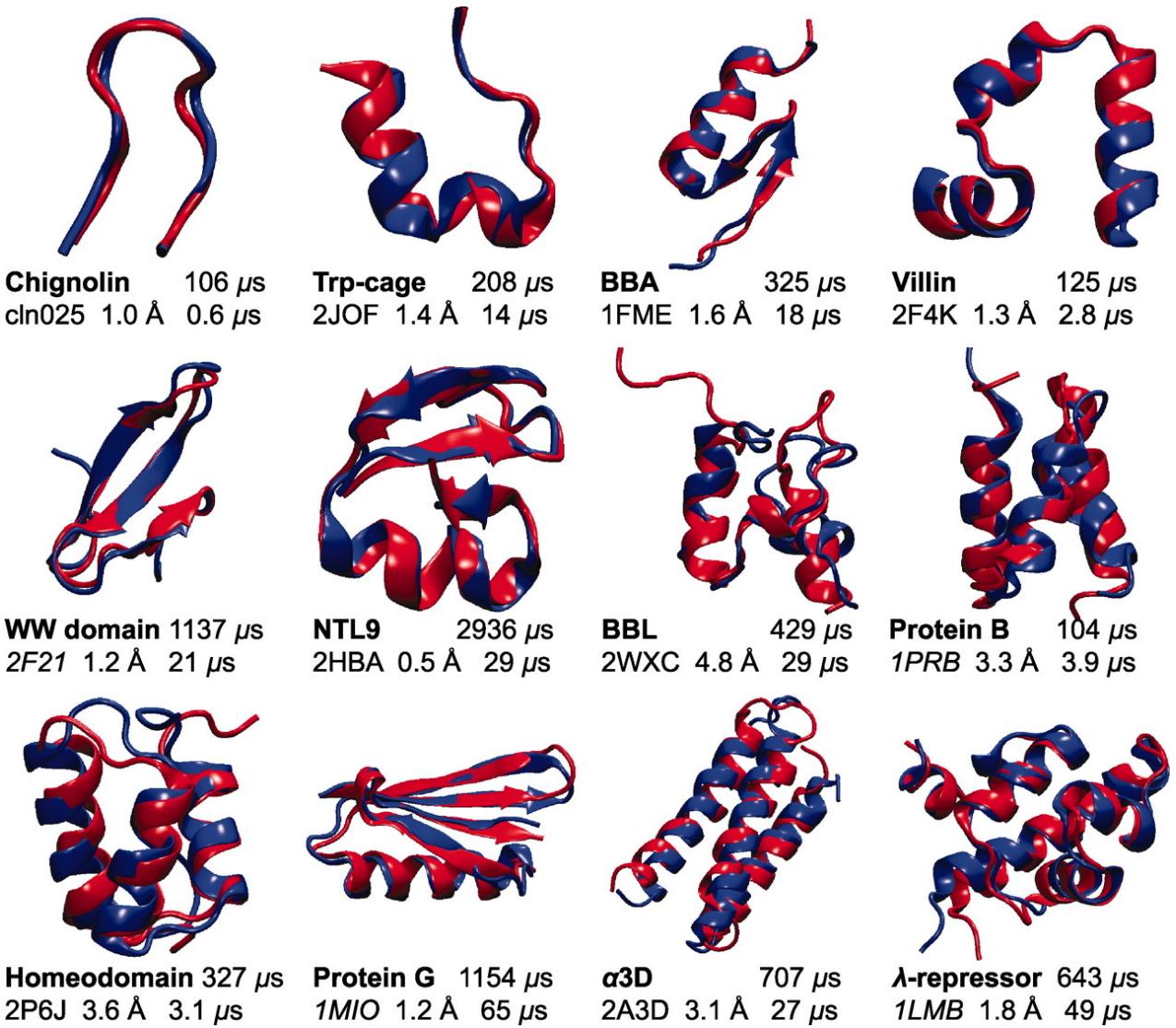
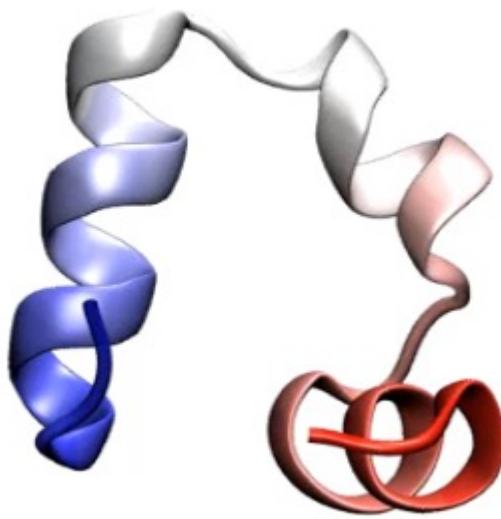
Apo protein predicted folded like holo state, but it should be unfolded

High-confidence hetero multimer predicted, but proteins do not co-assemble

High-confidence homodimer of mutated protein predicted, but mutation abolishes complex formation

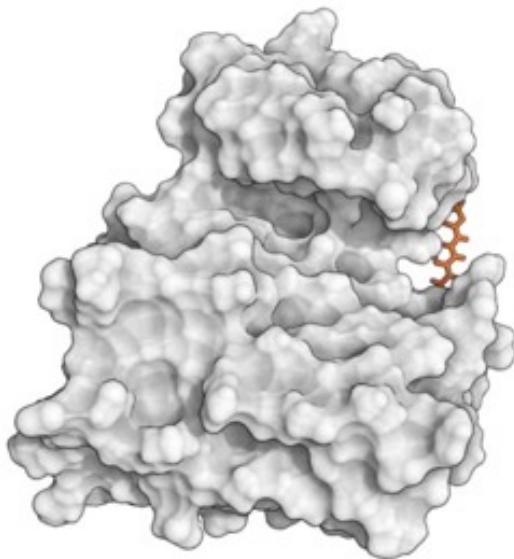
# Watching proteins fold: simulation

- Following experimentally the *folding pathway* of a protein is difficult
- Folding of small fast-folding (<100  $\mu$ s) proteins can be studied via simulation



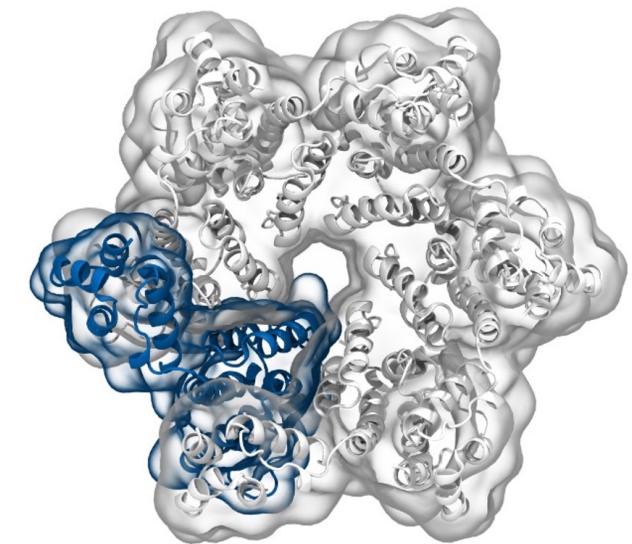
**Next: how to prepare a protein  
structure so that it is ready for  
molecular modelling?**

# Simulation of Biomolecules



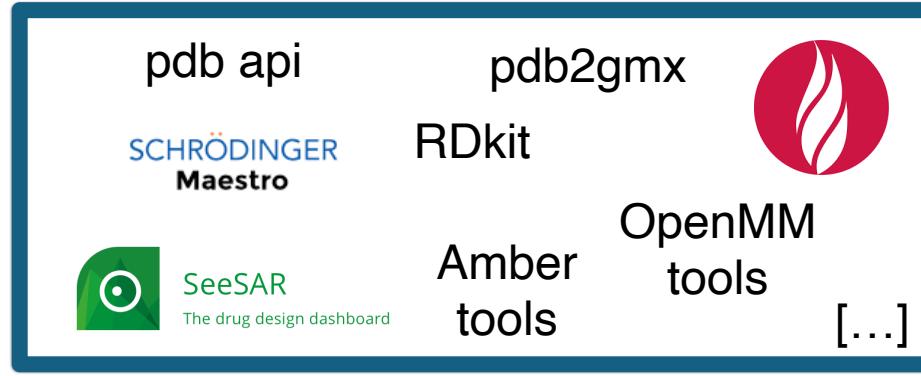
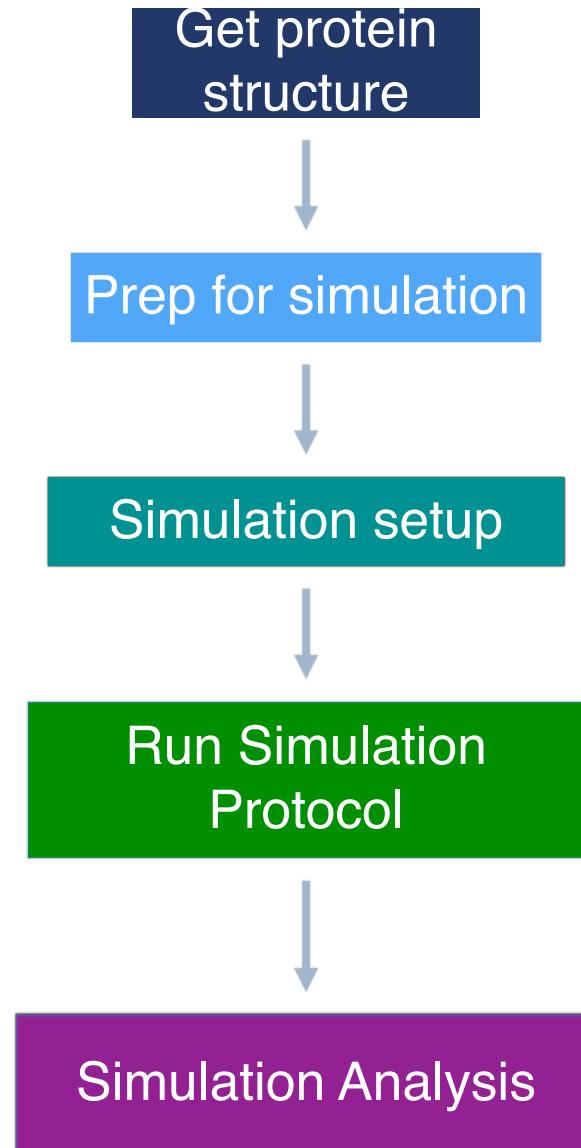
## Protein preparation

Delivered by **Tim Spankie**  
University of Edinburgh  
January 2025

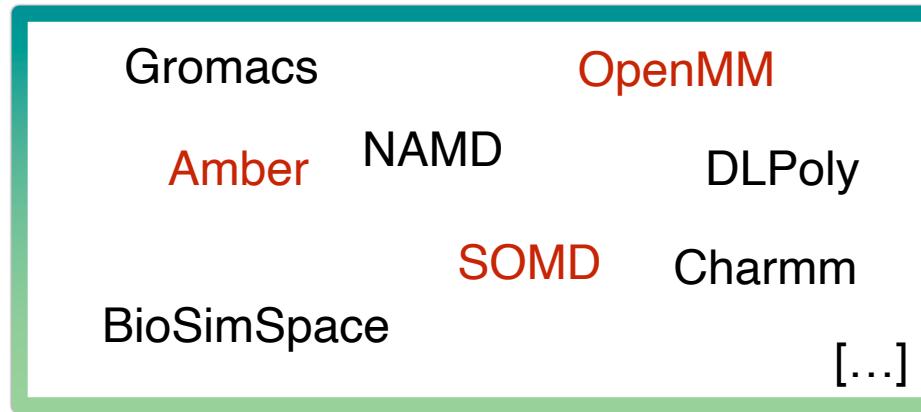


Based on content from a ten-hour series <https://github.com/CCPBioSim/BioSim-analysis-workshop>  
by Matteo Degiacomi ([matteo.degiacomi@ed.ac.uk](mailto:matteo.degiacomi@ed.ac.uk)) and Antonia Mey ([antonia.mey@ed.ac.uk](mailto:antonia.mey@ed.ac.uk))

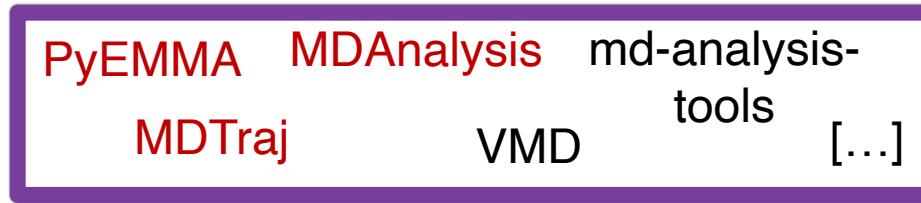
# A typical workflow for molecular dynamics



*GUI, TCL, bash,  
Python, Perl ...  
get creative*

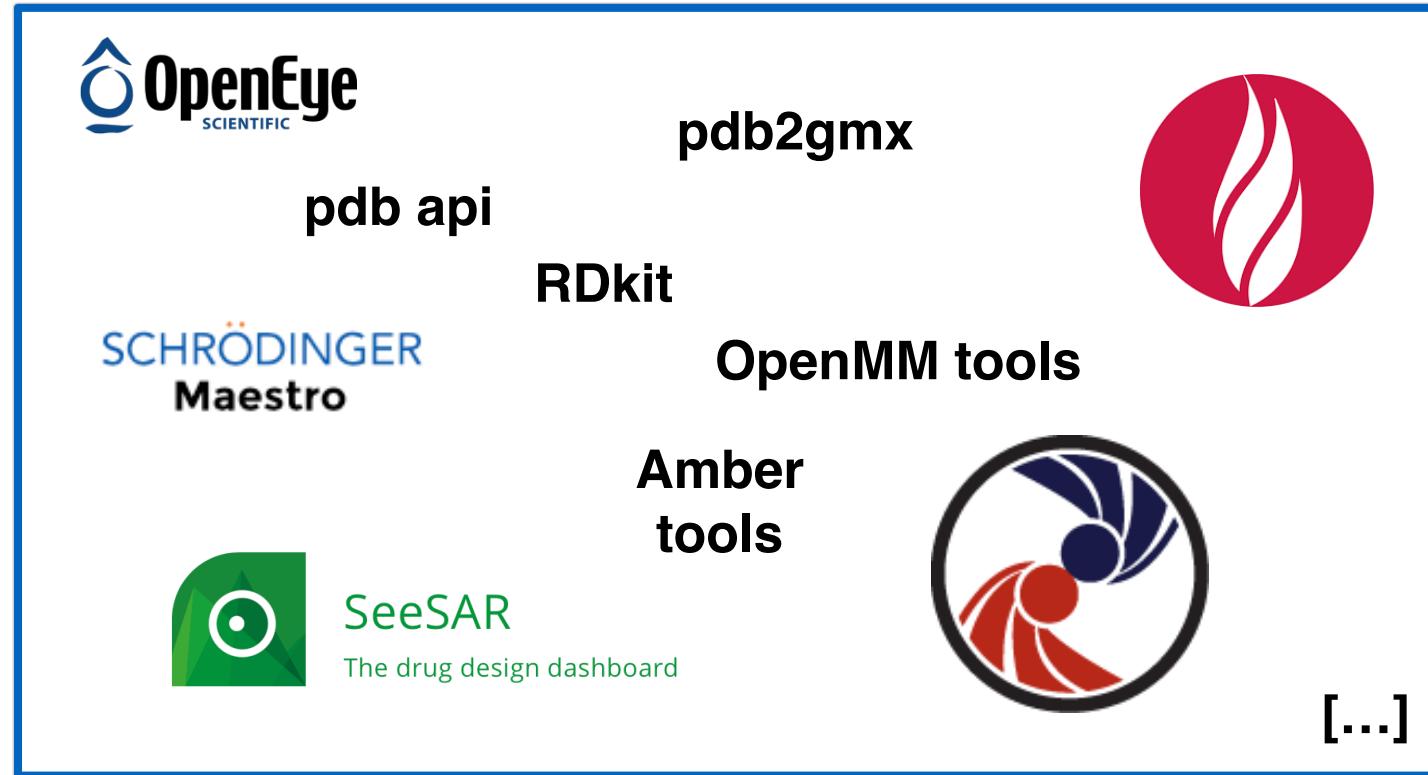
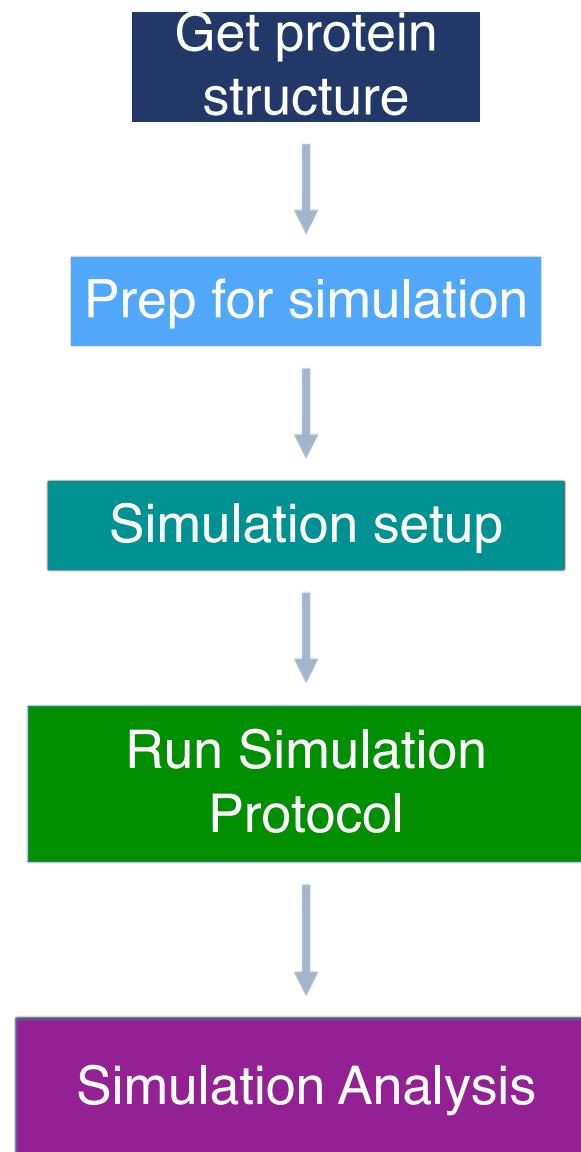


*mostly C++  
command line*



*GUI, TCL, bash,  
Python, Perl, ...  
get creative*

# Let's get started with understanding protein structures



*GUI, TCL, bash, python, Perl, ... get creative*

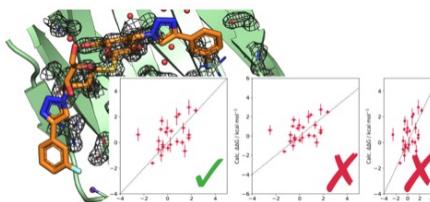
# Crystal structures are models with potential errors

HOME / ARCHIVES / VOL. 4 NO. 1 (2022) / Articles

## Best Practices for Constructing, Preparing, and Evaluating Protein-Ligand Binding Affinity Benchmarks [Article v1.0]

David F. Hahn

Computational Chemistry, Janssen Research & Development, Turnhoutseweg 30,  
Beerse B-2340, Belgium  
<https://orcid.org/0000-0003-2830-6880>



### Examples of errors/oddities in PDB structures:

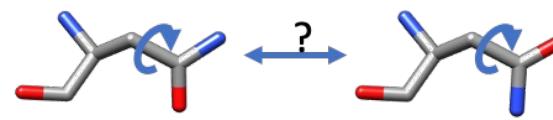
<https://swift.cmbi.umcn.nl/teach/pdbad/>

Gert Vriend (author of WHAT\_CHECK)

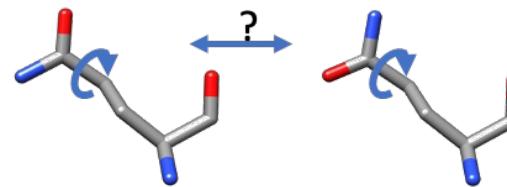
# Alternative conformations of side chains - NGH flips

Typically, the crystallographers will have assigned the orientation based on potential H-bonding interactions, etc.

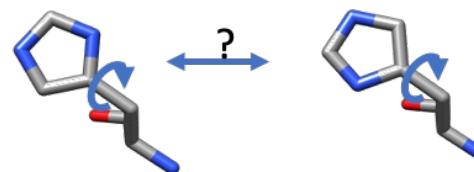
No standard protocol for this: always worth double-checking.



**Asparagine (N)**



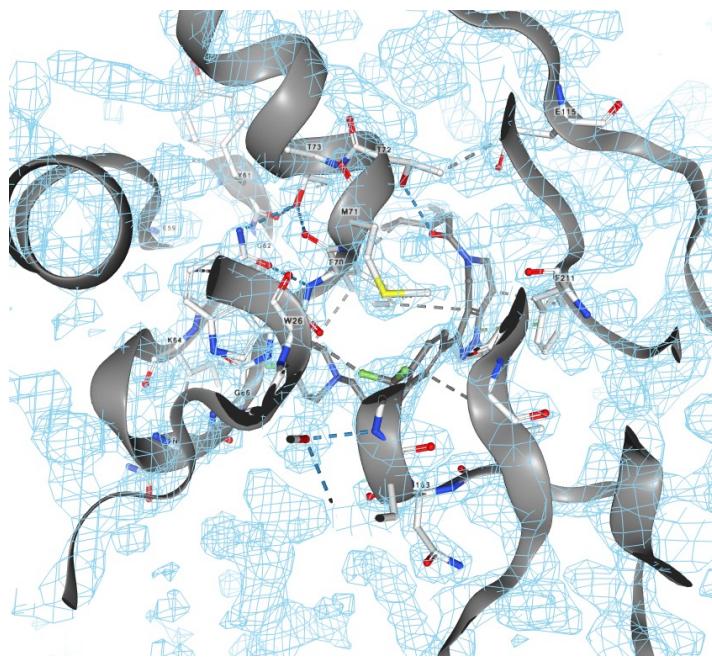
**Glutamine (Q)**



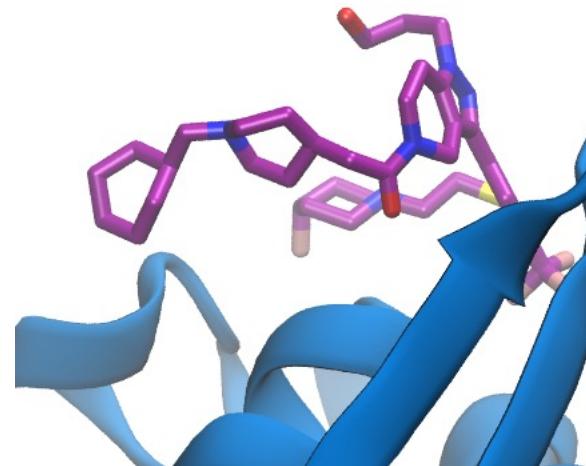
**Histidine (H)**

# Ligand densities can also have creative input from the crystallographer

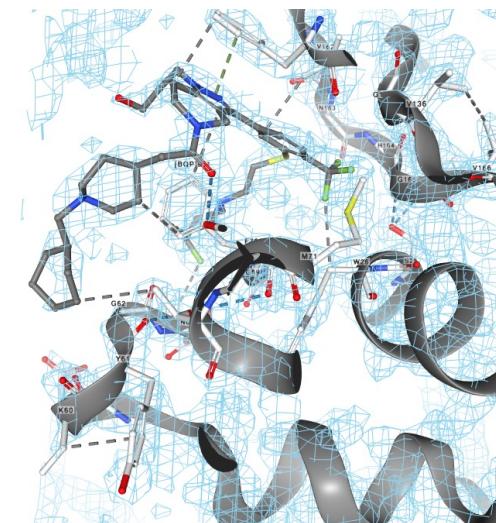
Cathepsin S



Cathepsin S  
with ligand

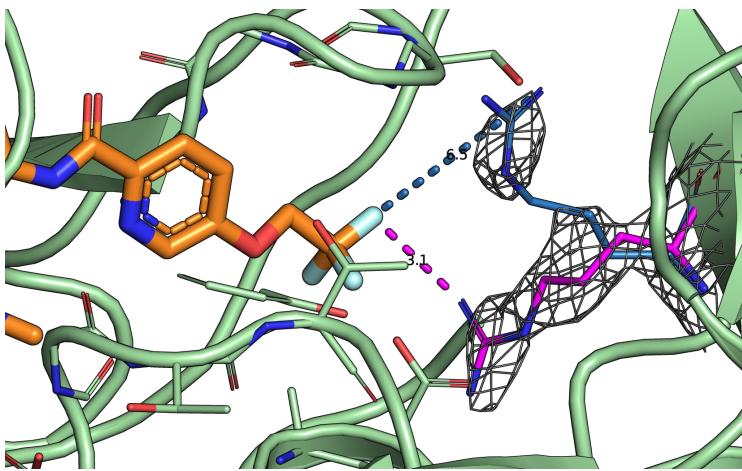


boat/chair?

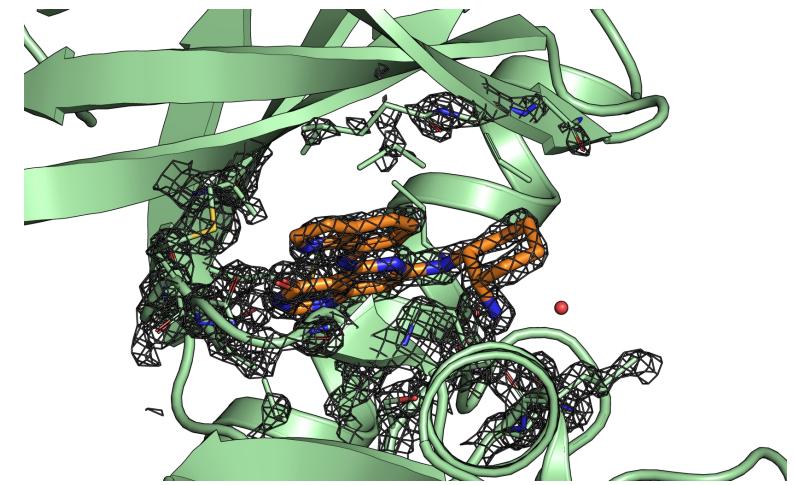
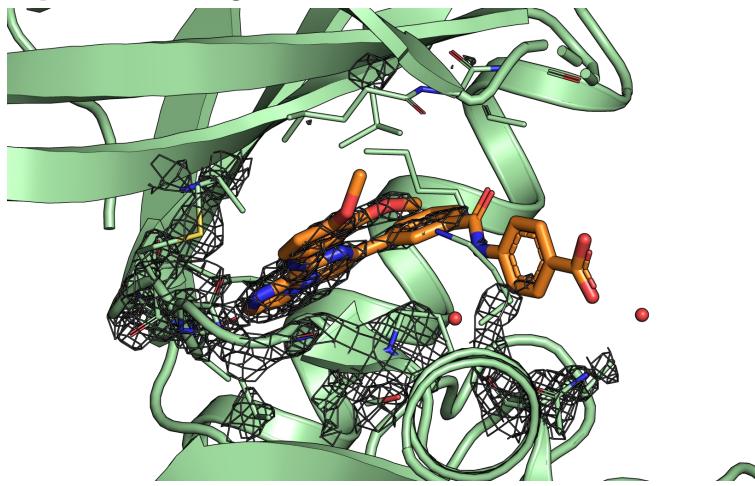


# Picking the best crystal structure requires care

BACE (Hunt)



spleen tyrosine kinase

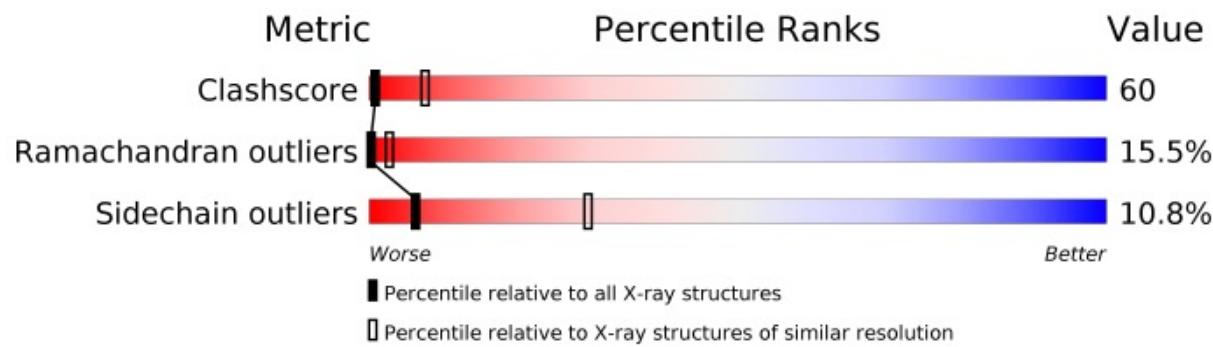


**Which protein structure?**

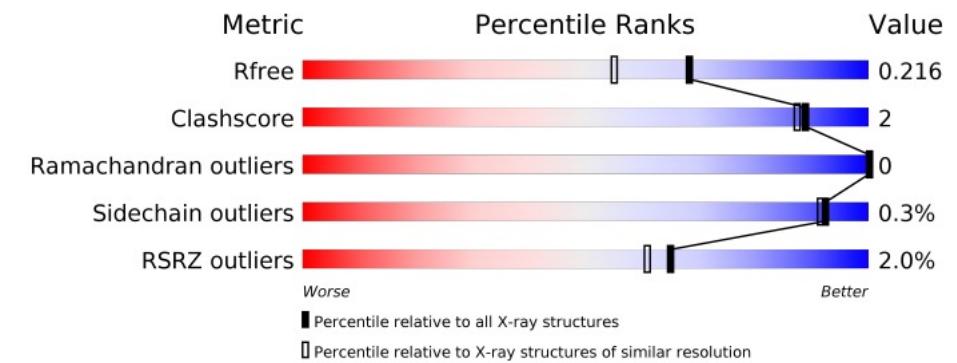
- Alternative side chain conformations need to be assessed carefully
- Active site residue densities are important for choosing the right crystal structure

# The RCSB PDB report can help with choosing structures

Jnk1 - 2GMX

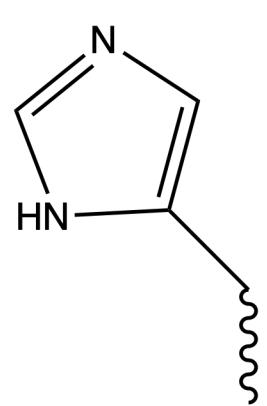


Jnk1 - 3ELJ

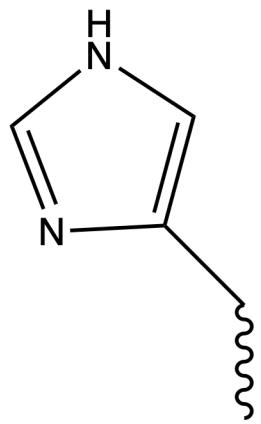


# How to choose the right tautomer?

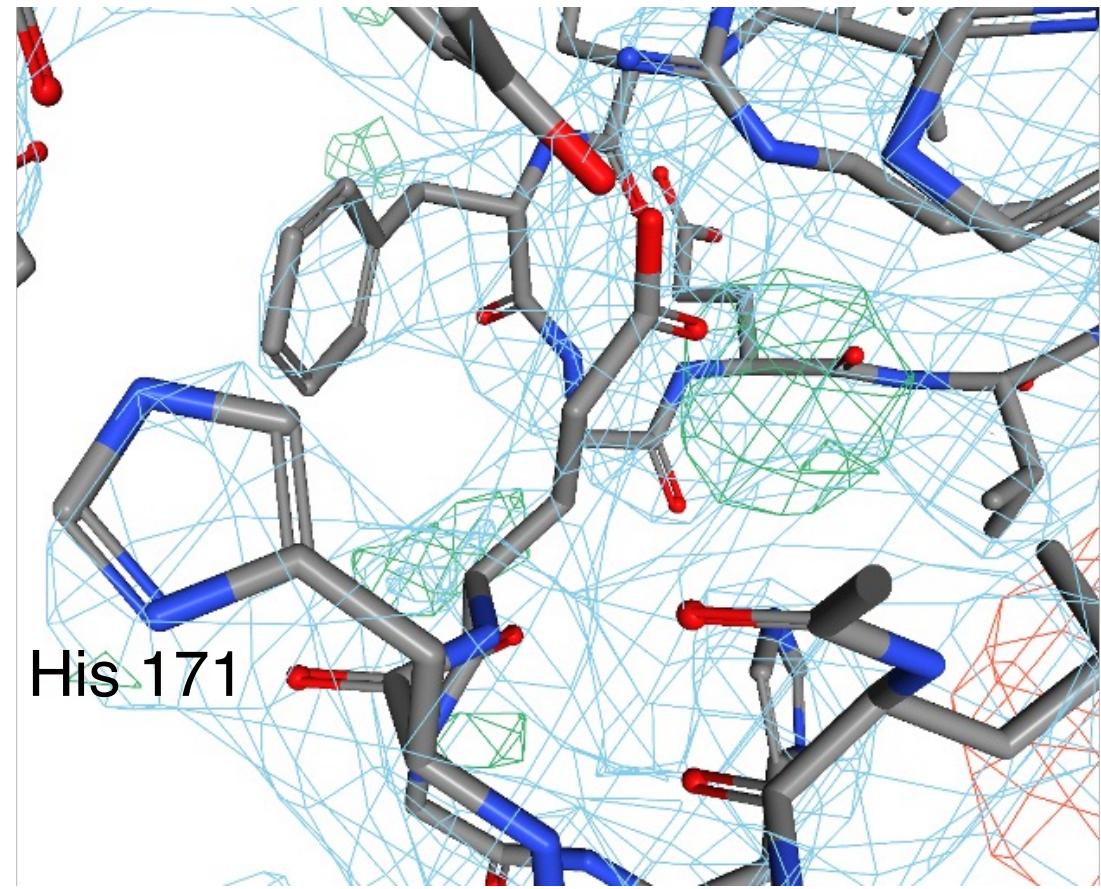
Which tautomer?



$\delta$ -tautomer

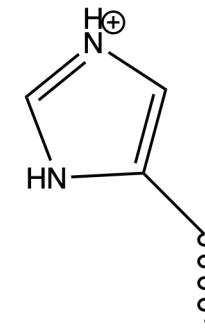
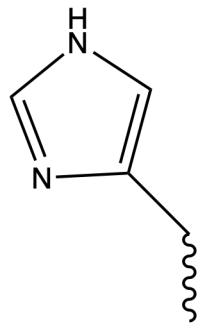
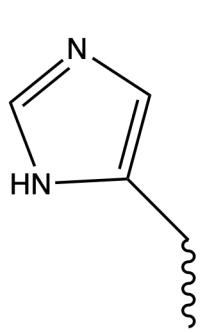


$\epsilon$ -tautomer



# How to choose the right tautomer and protonation state?

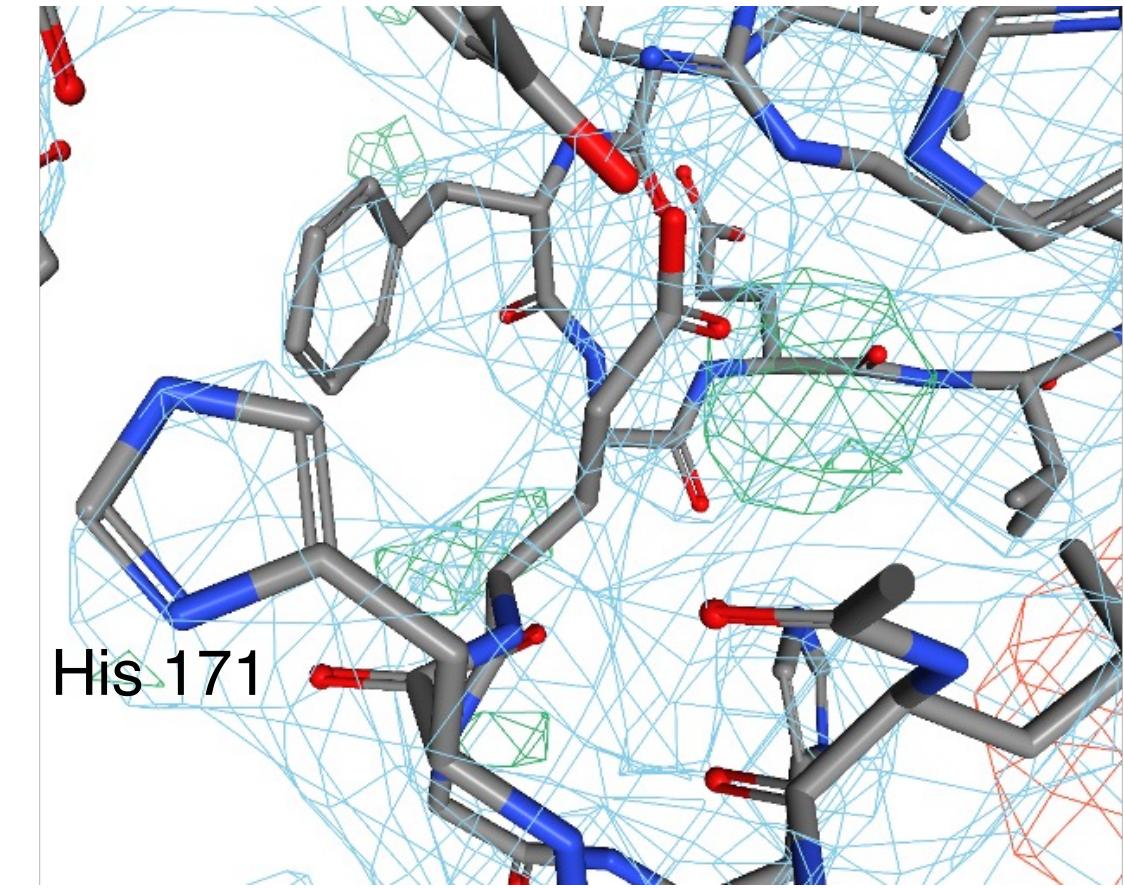
Which tautomer and protonation?



$\delta$ -tautomer

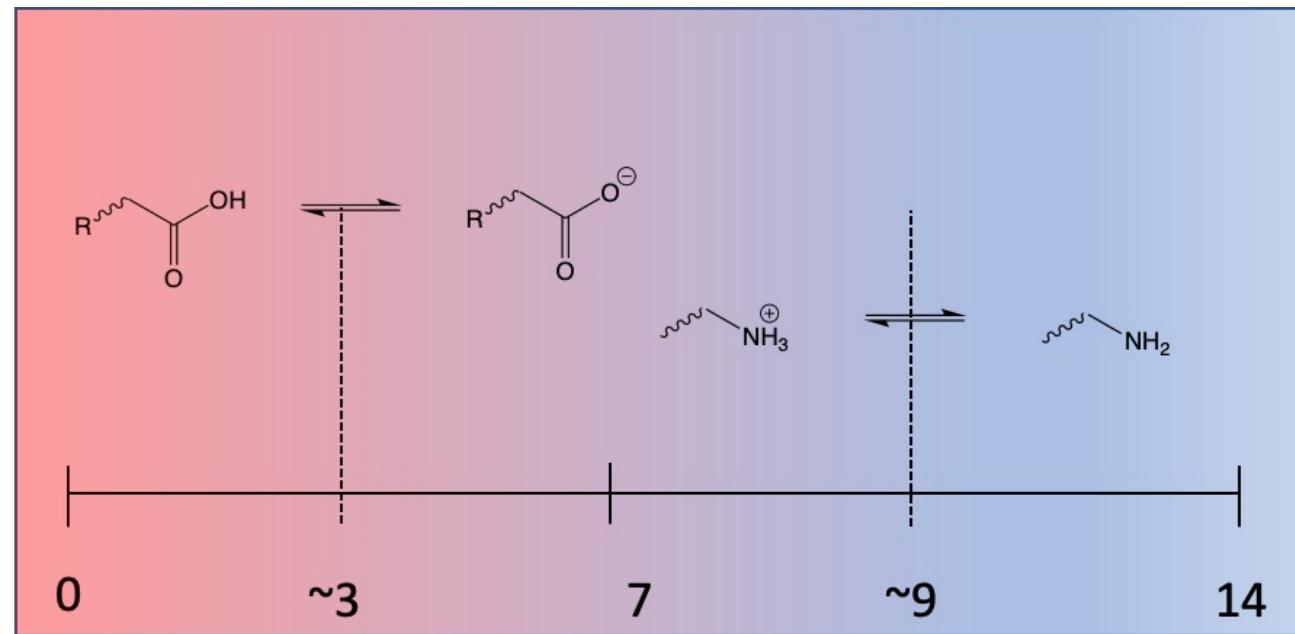
$\epsilon$ -tautomer

protonated



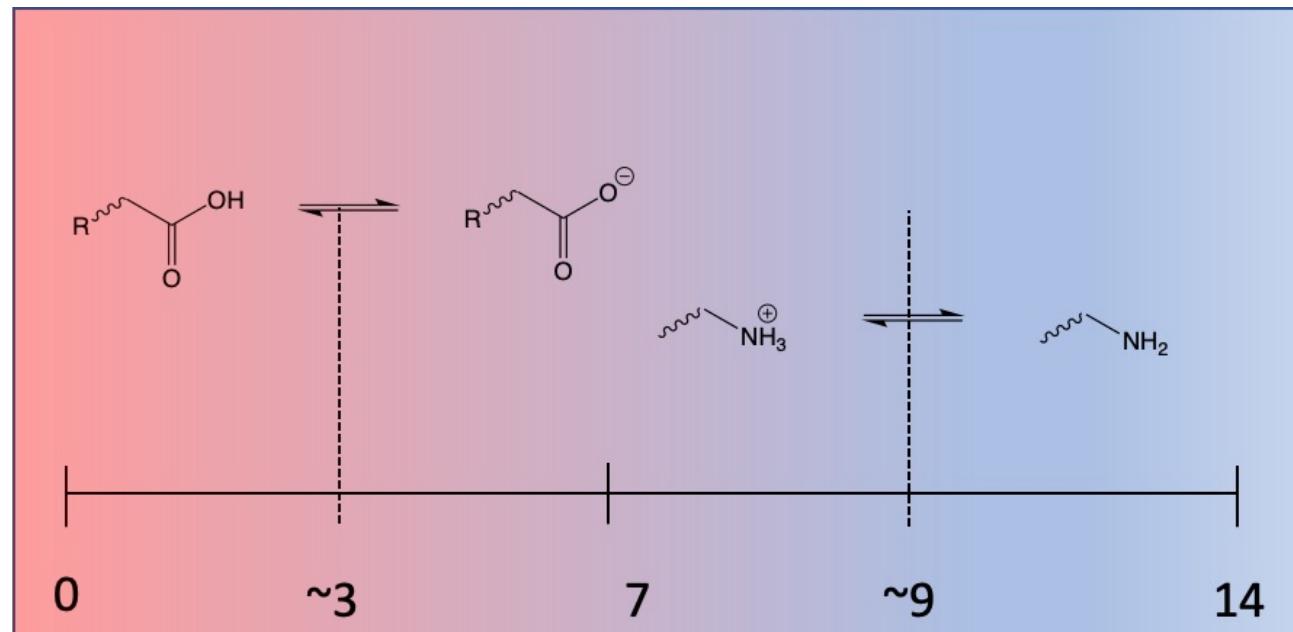
# We need to worry about pKa to decide protonation

- pKa: pH at which an acidic/basic group is 50% protonated/deprotonated.
- pKas are not fixed things!
- “Standard” values refer to the situation when the group is in dilute aqueous solution.
- Groups buried in the centre of hydrophobic proteins or close to other charged groups can show large pKa shifts.



# We need to worry about pKa to decide protonation

- Finding the right protonation state for proteins and ligands can be challenging!
- Poorly chosen protonation states
  - can lead to inaccurate simulations by (de)stabilising interactions
  - can lead to incorrectly estimating binding affinities or protein local or collective dynamics.



# Amino acids that need protonation state consideration

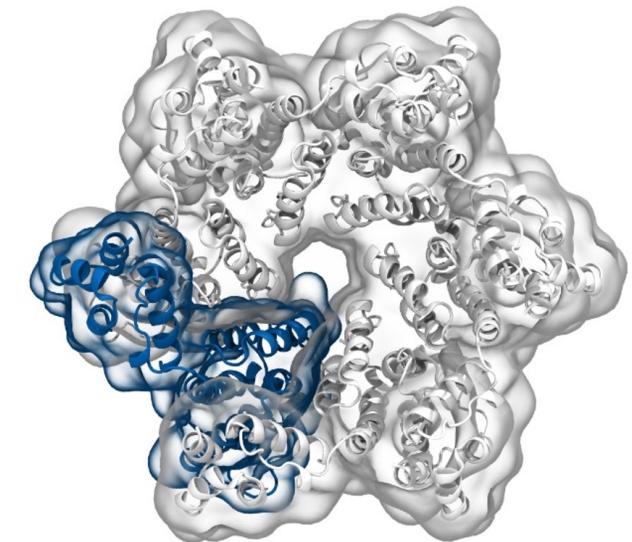
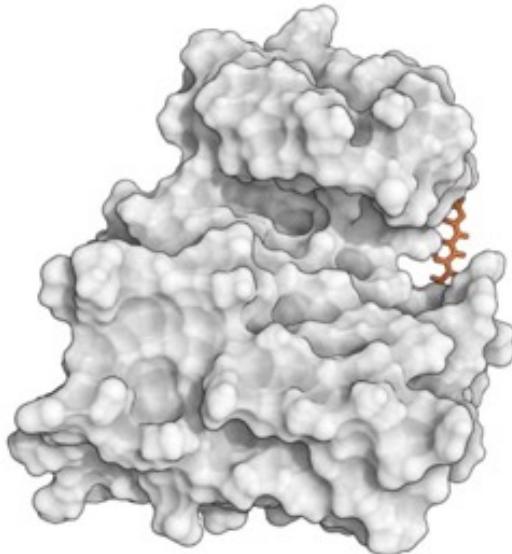
Amino acid	pKa	options	Significance*
Aspartic acid	3.65	-COOH instead of COO <sup>-</sup> ?	possible
Glutamic acid	4.25	-COOH instead of COO <sup>-</sup> ?	possible
Histidine	6.00	protonated instead of neutral?	very possible
Cysteine	8.18	-S <sup>-</sup> instead of SH?	very possible
Tyrosine	10.07	-O <sup>-</sup> instead of OH?	possible
Lysine	10.53	-NH <sub>2</sub> instead of NH <sub>3</sub> <sup>+</sup> ?	possible
Arginine	12.48	neutral instead of protonated?	unlikely

\* for a simulation around physiological pH

# Simulation of Biomolecules

## Setting up a protein simulation

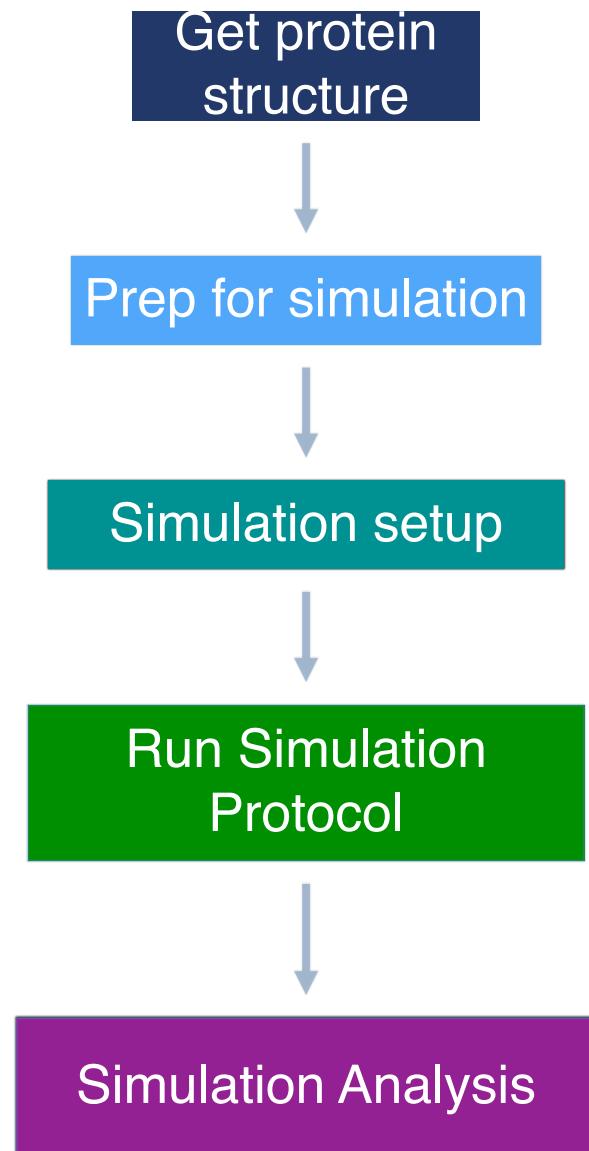
Delivered by **Tim Spankie**  
University of Edinburgh  
January 2025



Based on content from a ten-hour series

by Matteo Degiacomi ([matteo.degiacomi@ed.ac.uk](mailto:matteo.degiacomi@ed.ac.uk)) and Antonia Mey ([antonia.mey@ed.ac.uk](mailto:antonia.mey@ed.ac.uk))

# A typical workflow for molecular dynamics



Getting your protein structure

AlphaFold Protein Structure Database

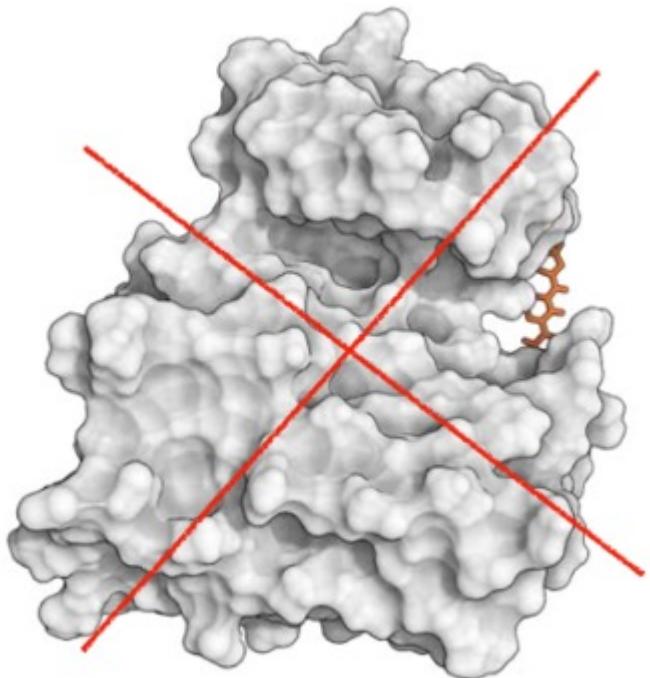


Getting ligands/co-factors

ZINC20



# Disclaimer!



Running biomolecular MD can take days on specialised hardware.

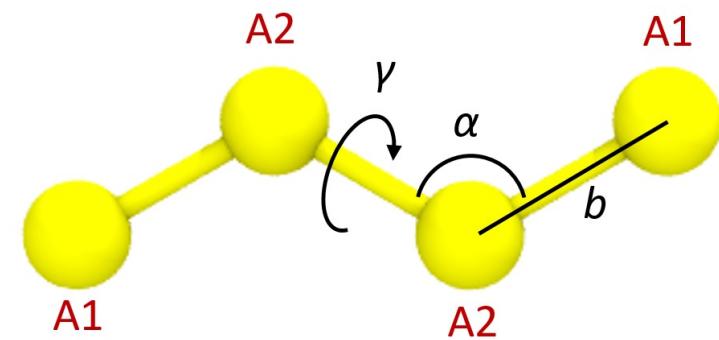
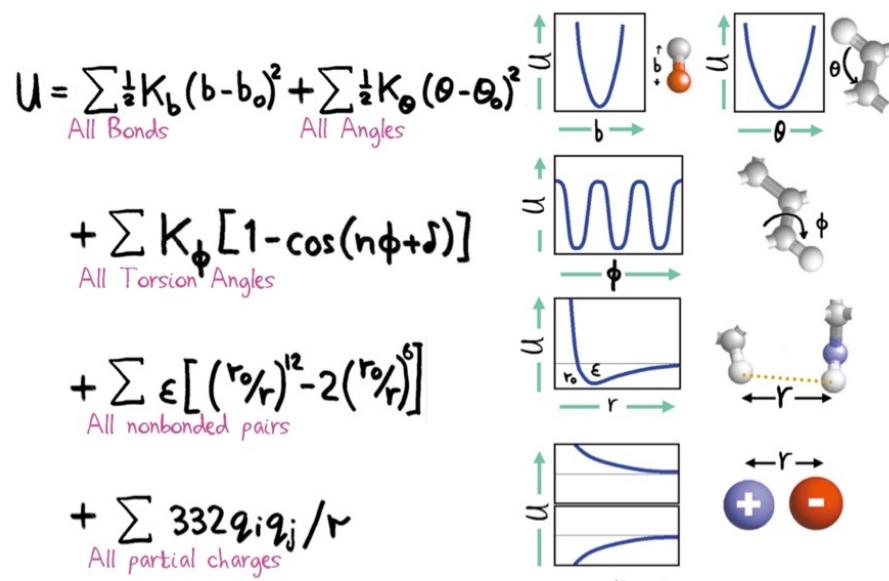
Today we will *not* run any of them, and instead will focus on fundamental principles using small molecules.

# Evaluating interatomic interactions

MD simulation require defining a system *topology*. This includes information on:

- *Connectivity* between atoms (covalent bond network)
- *Atom types* (depend on chemical environment, e.g., sp<sub>2</sub> or sp<sub>3</sub> Carbon)

$$\vec{F}_i = -\frac{dU}{d\vec{x}_i}$$



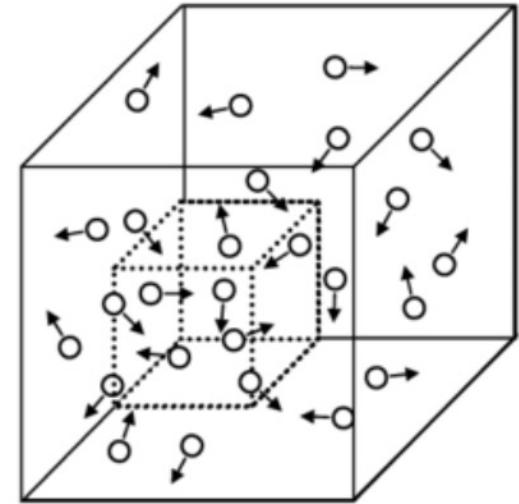
The functional form of interatomic interactions and their strengths is defined by a *force field*.

# Choosing your thermodynamic ensemble

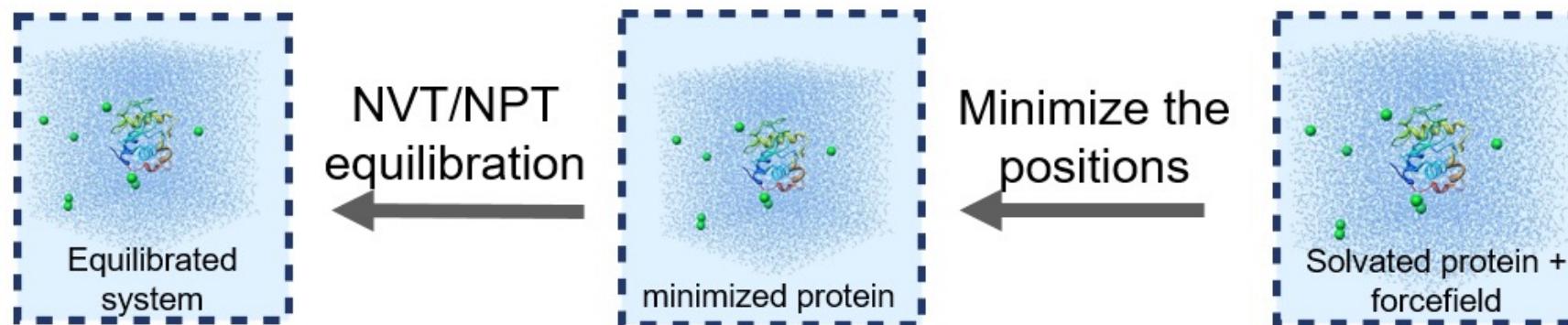
Simulations replicate a specific *thermodynamic ensemble* (typically NVT or NPT), or even grand canonical ( $\mu$ VT)

You will have different options to include *thermostats* (scaling atom velocities) and *barostats* (scaling positions) in your calculations:

- Nose-Hoover
- Berendsen
- Parrinello-Rahman
- Langevin piston
- ...

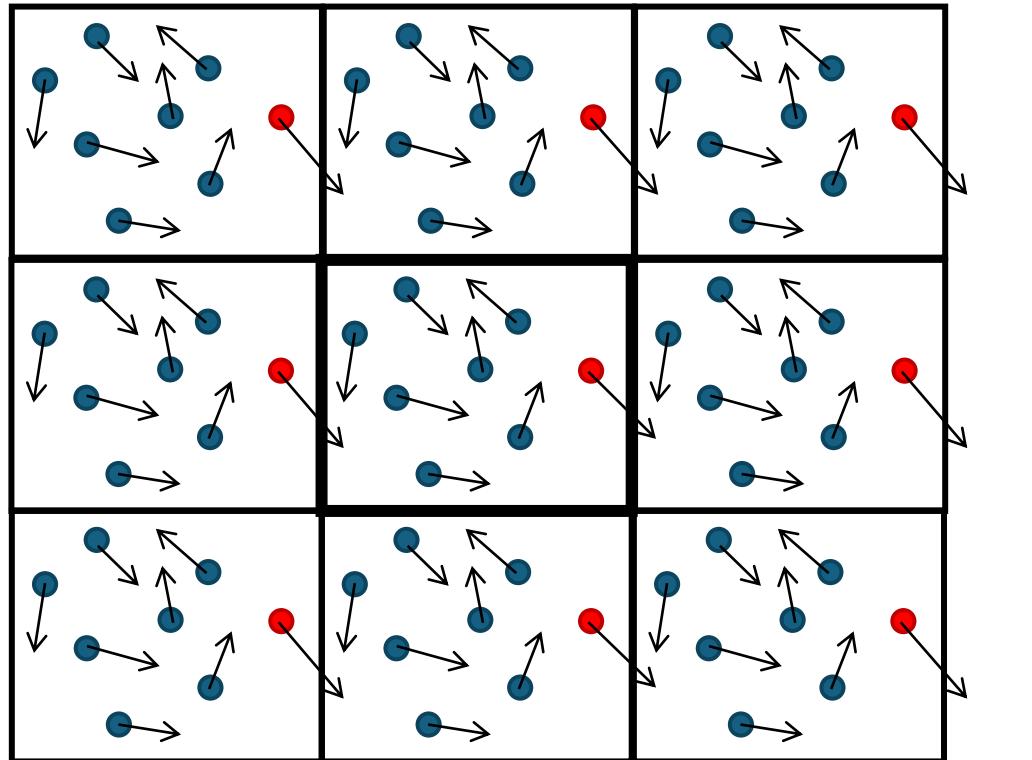


$\mu$ VT

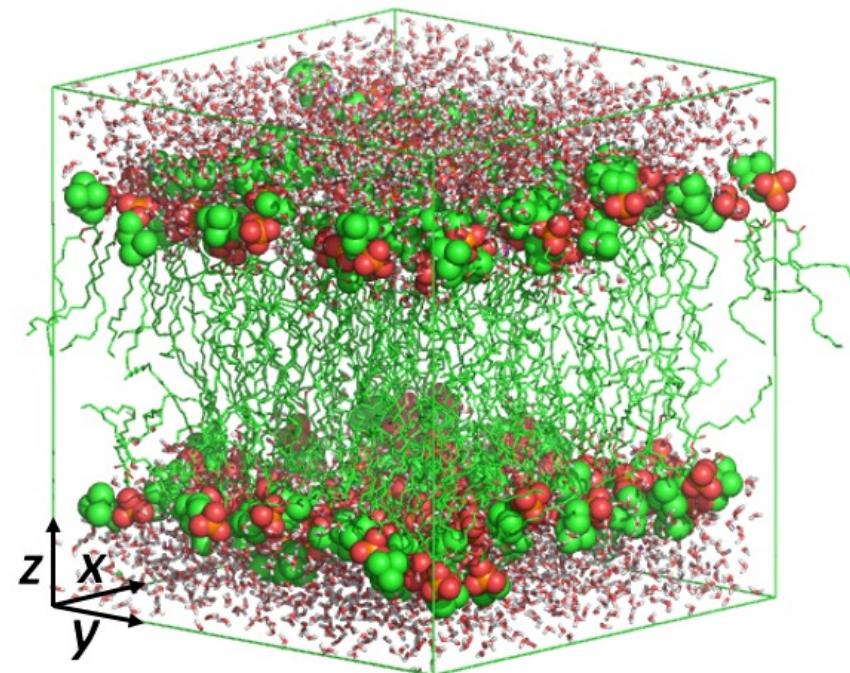


# Periodic boundary conditions (PBC) and pressure coupling

Useful to reduce finite-size effect and simulate bulk

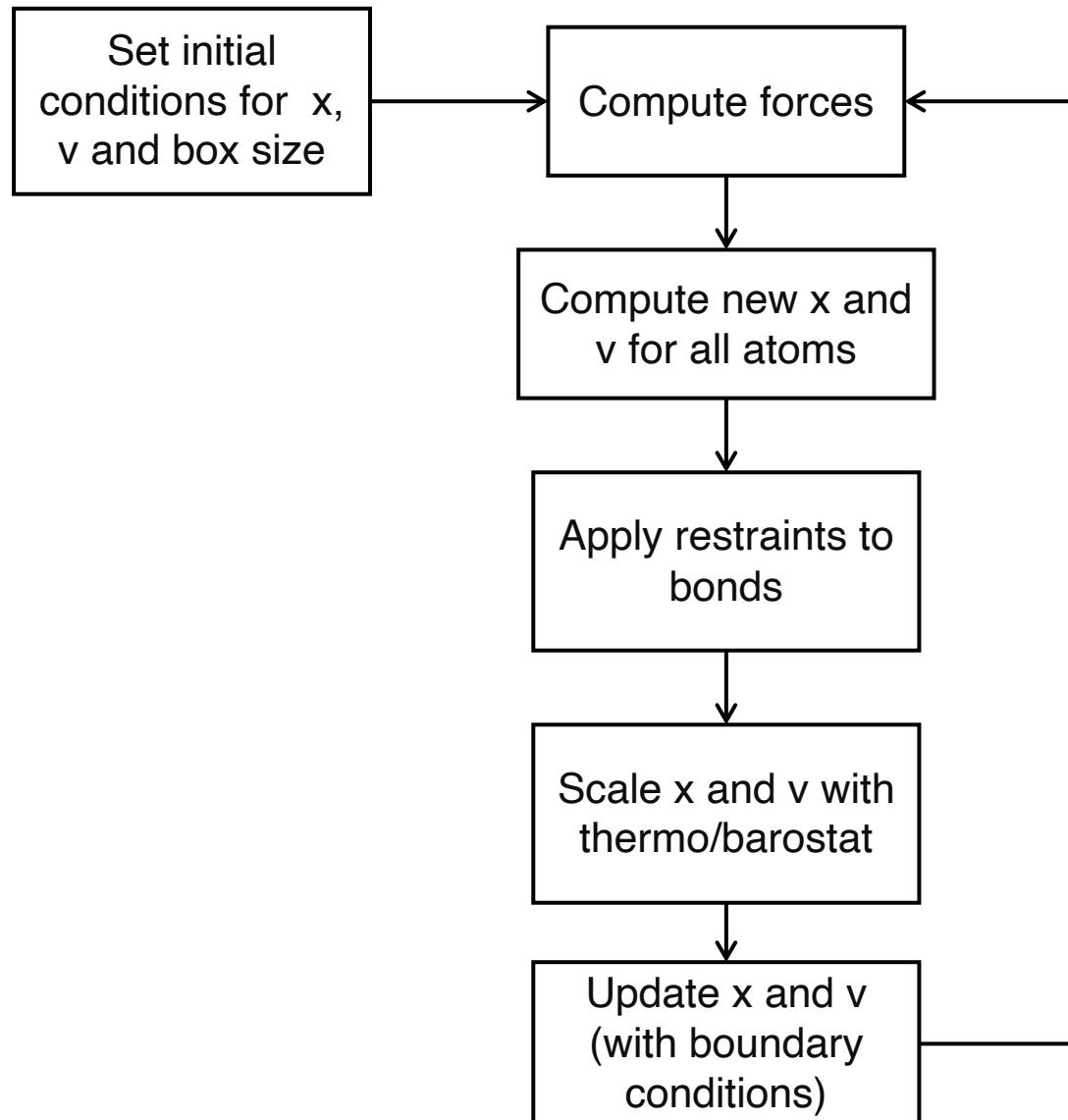


Typically, PBC applied in  $x$ ,  $y$  and  $z$  direction

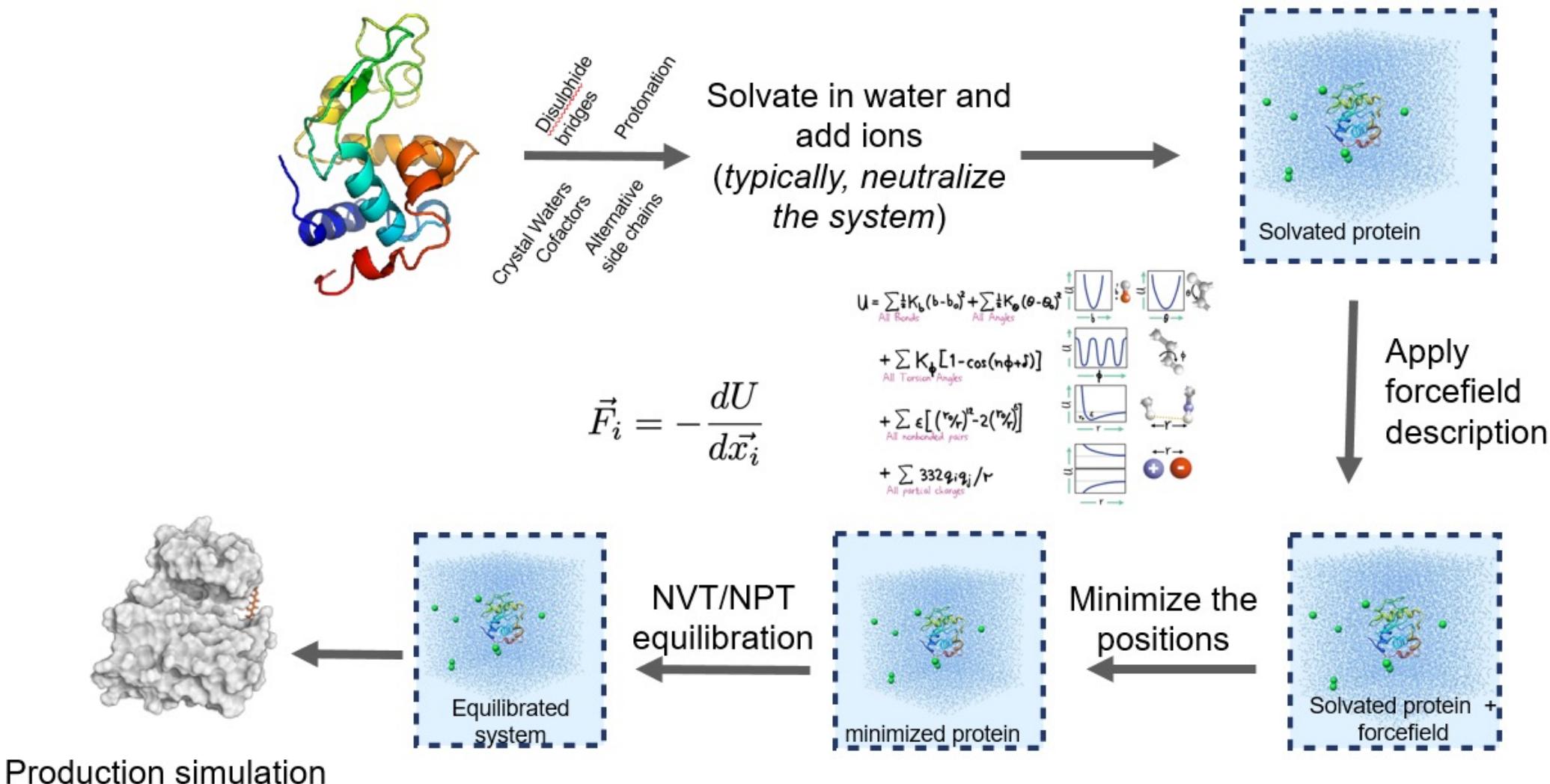


For membrane systems, use semi-isotropic pressure coupling ( $x, y \neq z$ ,  
lipids compressibility is direction-dependent)

# A Molecular Dynamics timestep



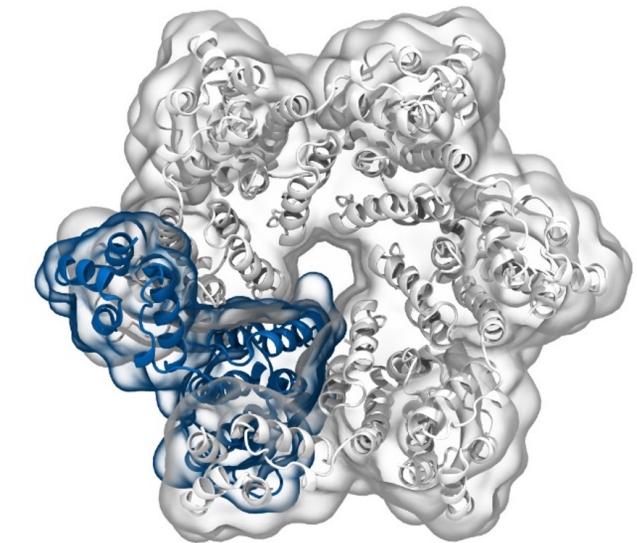
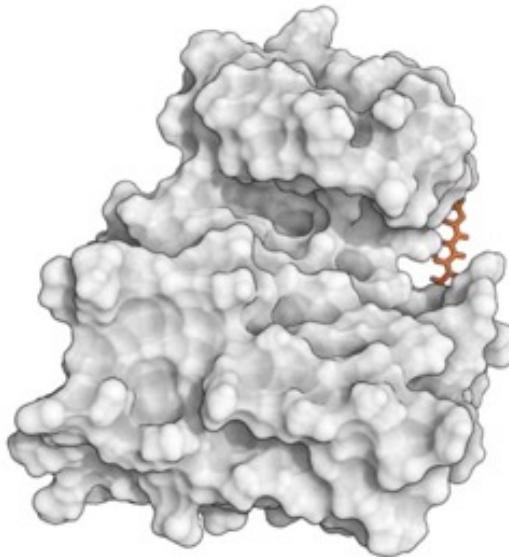
# Molecular dynamics require multiple steps for the setup of simulations



# Simulation of Biomolecules

## Basic Simulation Analysis

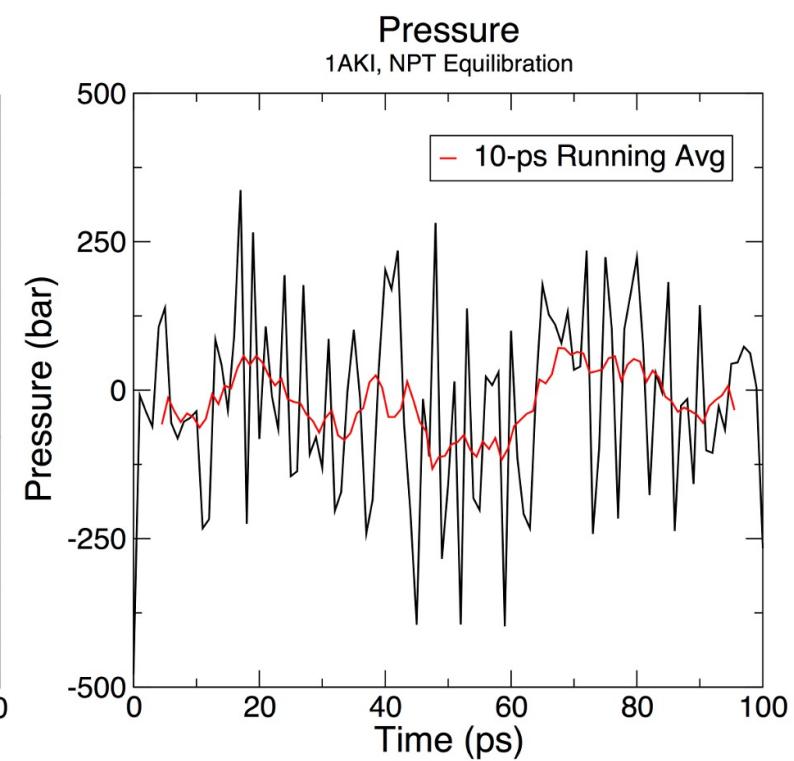
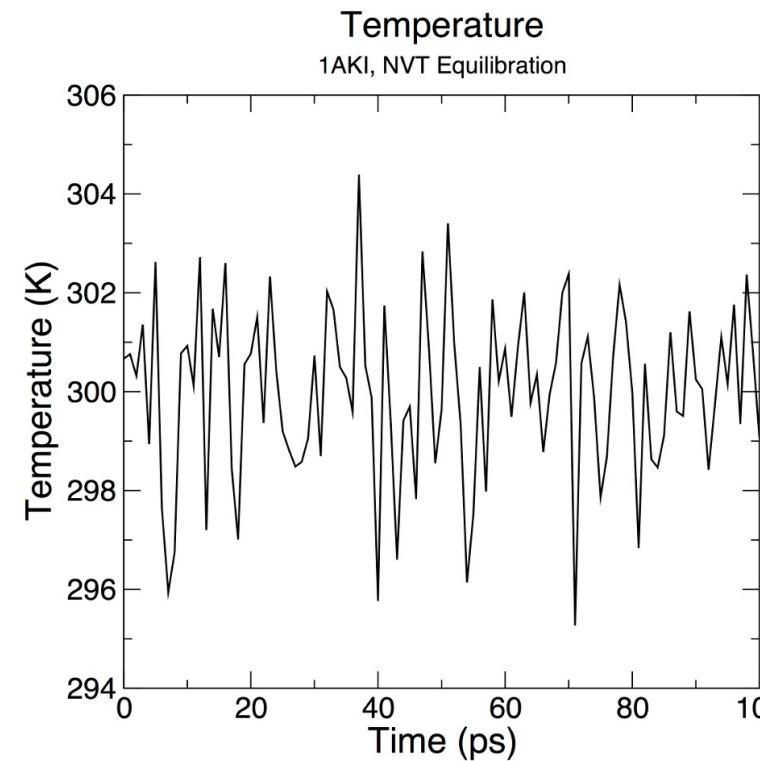
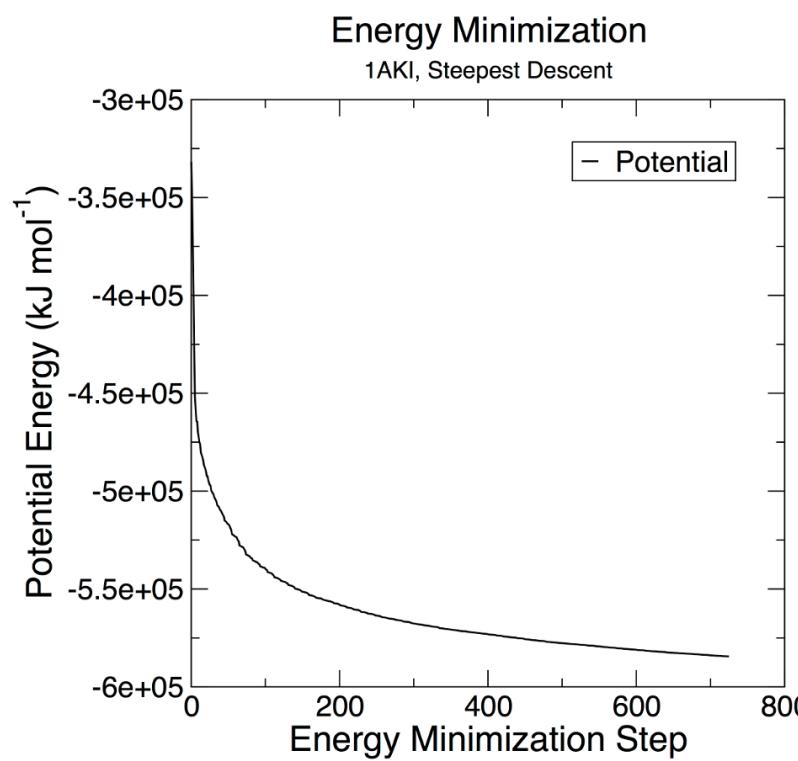
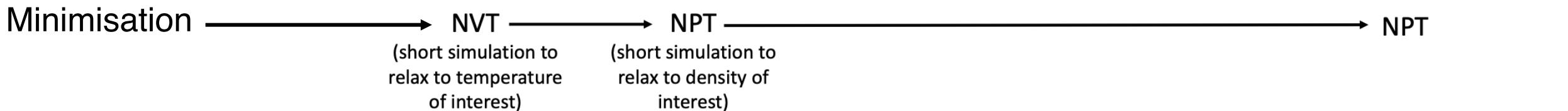
Delivered by **Tim Spankie**  
University of Edinburgh  
January 2025



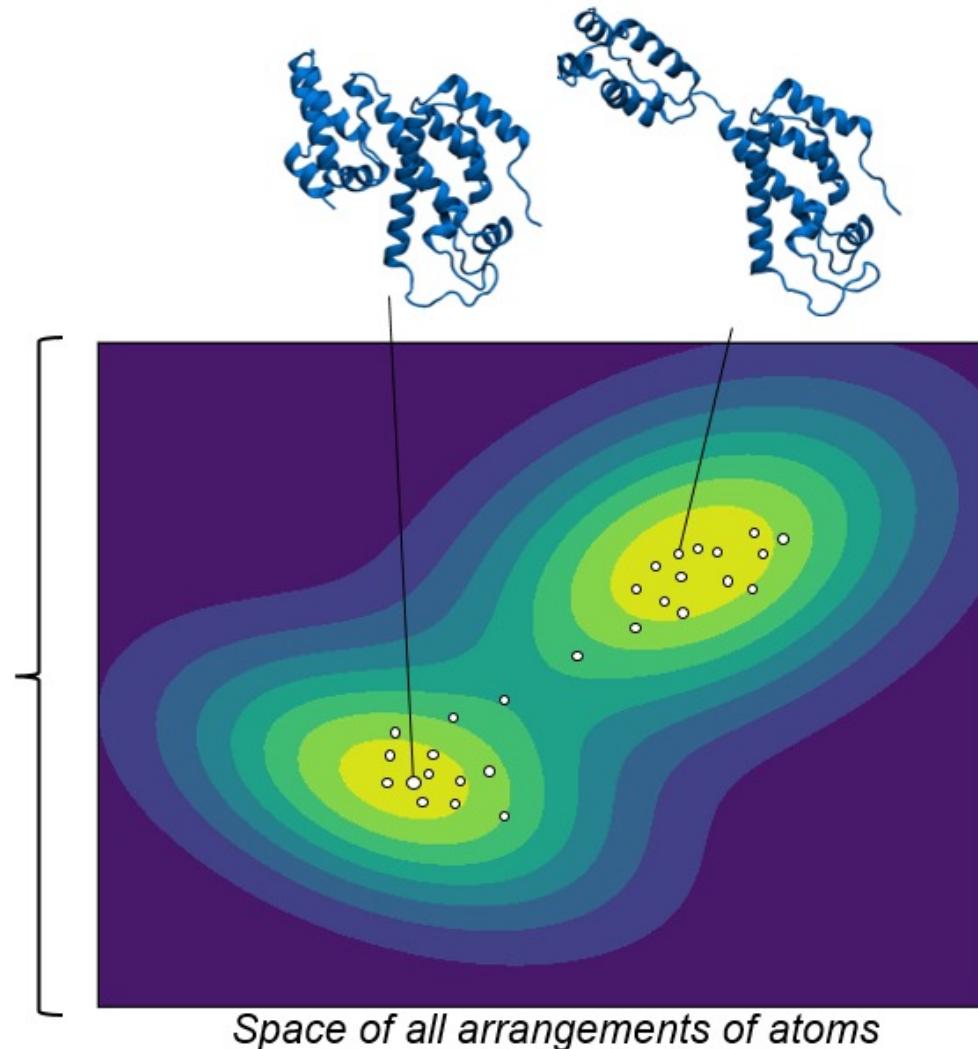
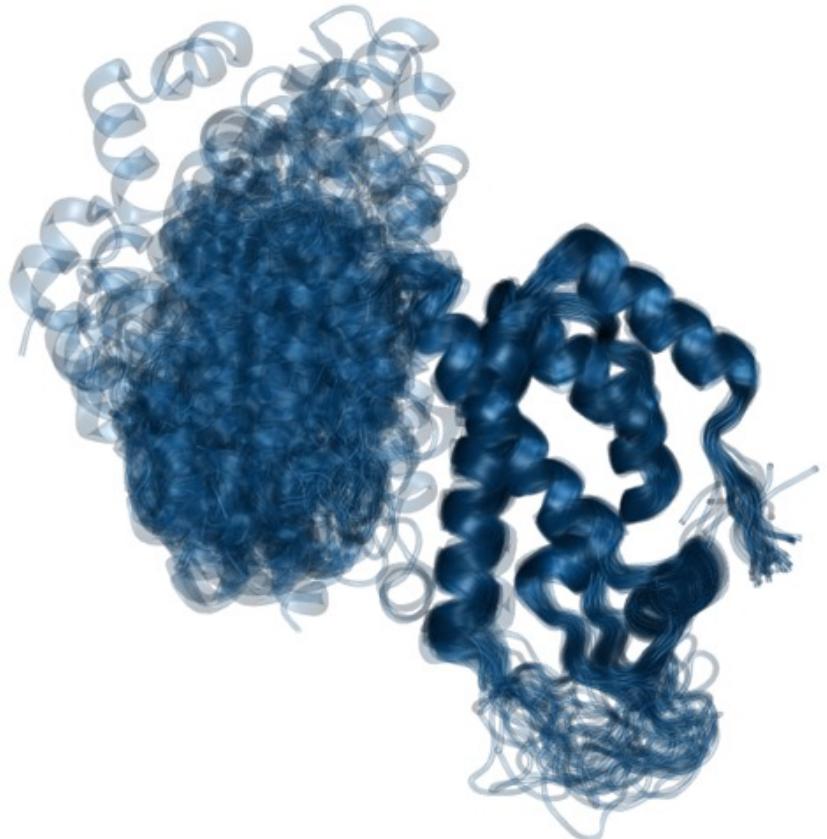
Based on content from a ten-hour series <https://github.com/CCPBioSim/BioSim-analysis-workshop>  
by Matteo Degiacomi ([matteo.degiacomi@ed.ac.uk](mailto:matteo.degiacomi@ed.ac.uk)) and Antonia Mey ([antonia.mey@ed.ac.uk](mailto:antonia.mey@ed.ac.uk))

# Volume and pressure equilibration

Steps until production:



# [Recall] Sampling the conformational space

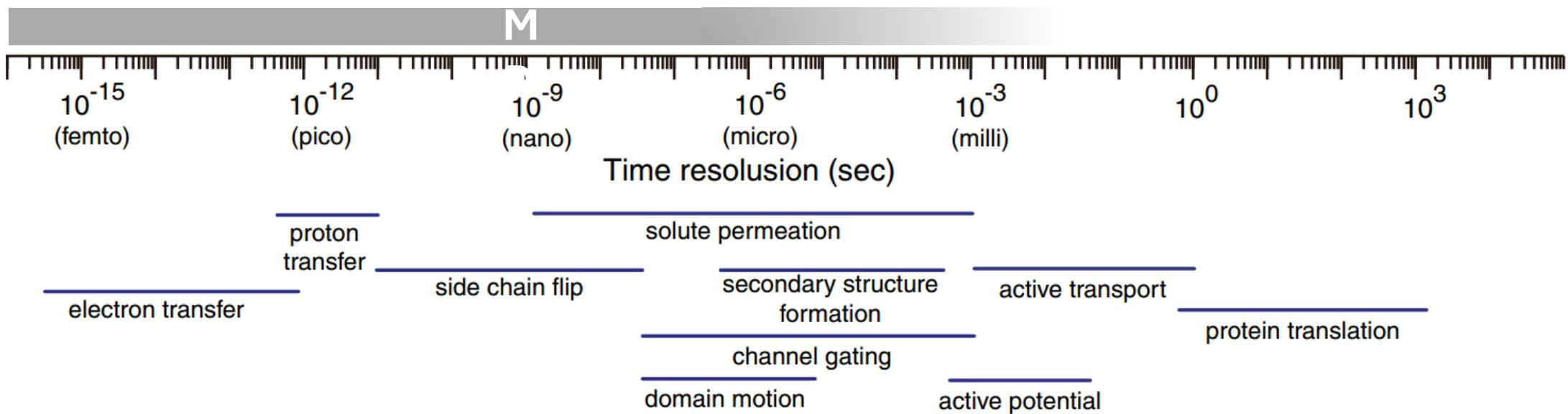
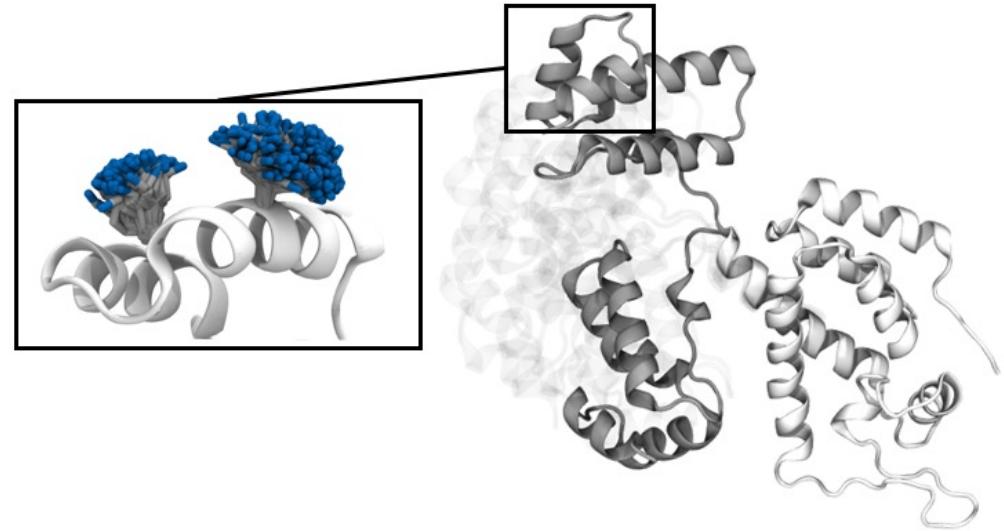


Probability of sampling a conformation is inversely proportional to its energy:  $p_i \propto e^{-\varepsilon_i/kT}$

# Timescales in biology

Different regions, different timescales:

- Side chains faster than backbone
- Loops faster than helices and sheets
- backbone faster than side chains
- Protein surface faster than core

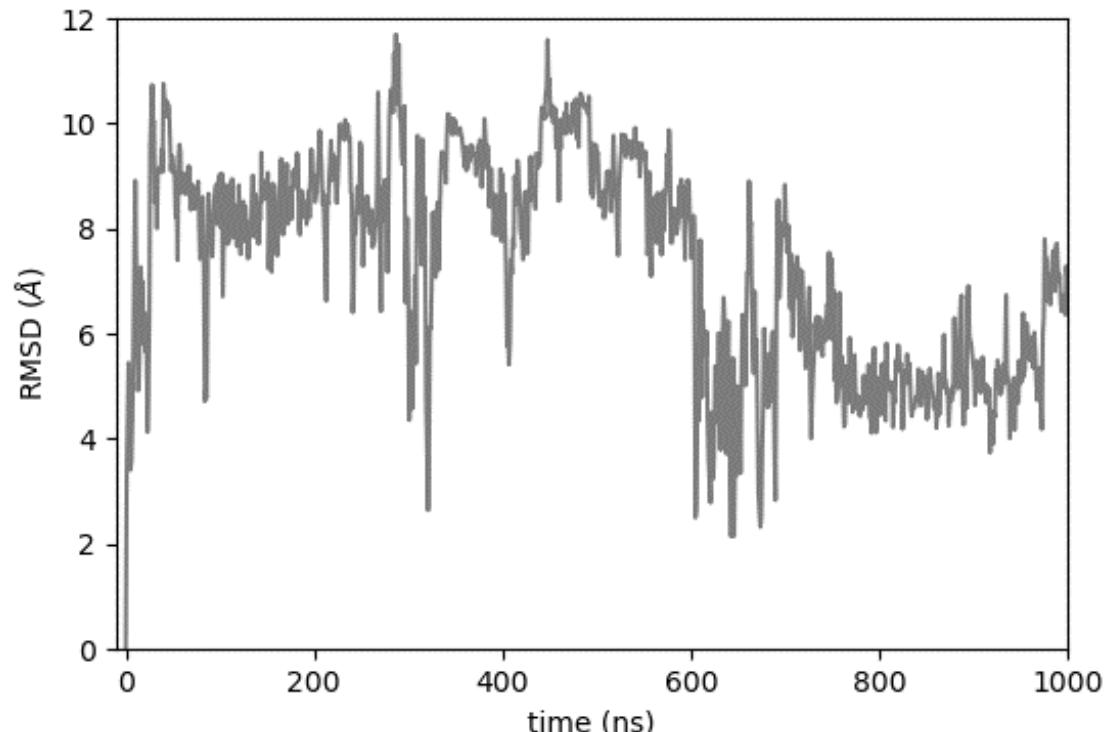
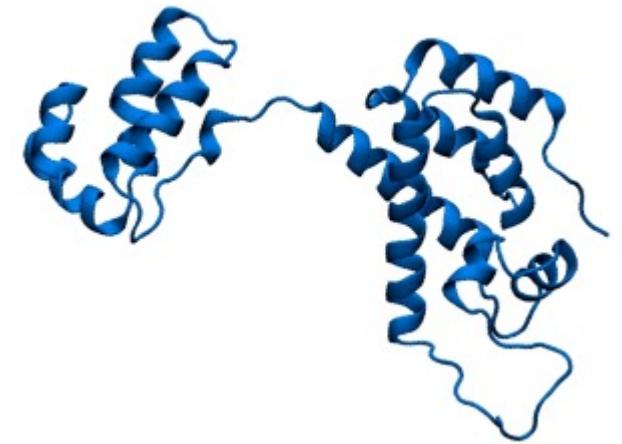


# Root Mean Square Deviation (RMSD)

Given a system with  $N$  atoms, and a reference arrangement  $x_0$ :

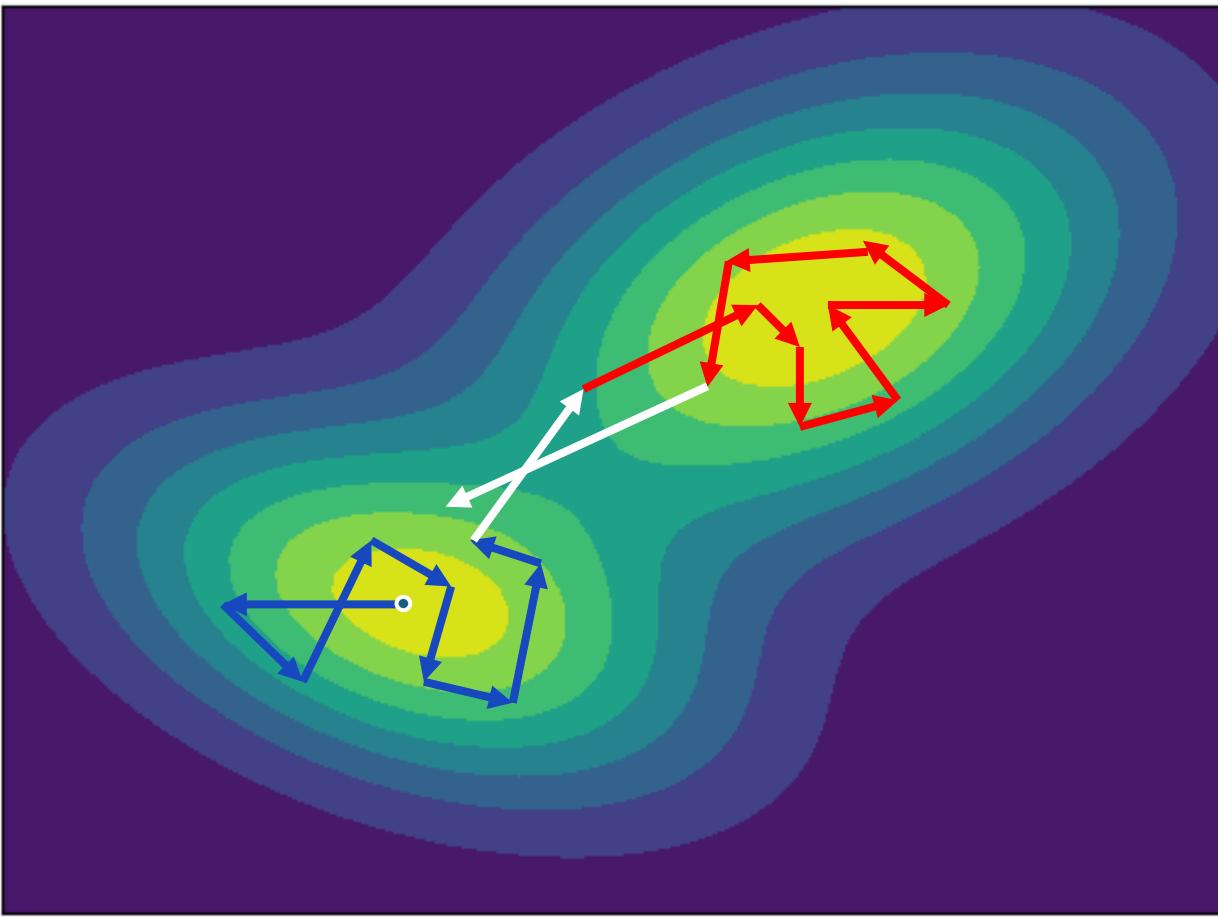
$$RMSD = \sqrt{\frac{1}{N} \sum_{i=0}^N (X_i - x_0)^2}$$

In MD,  $x_0$  is often the first conformation in the simulation.

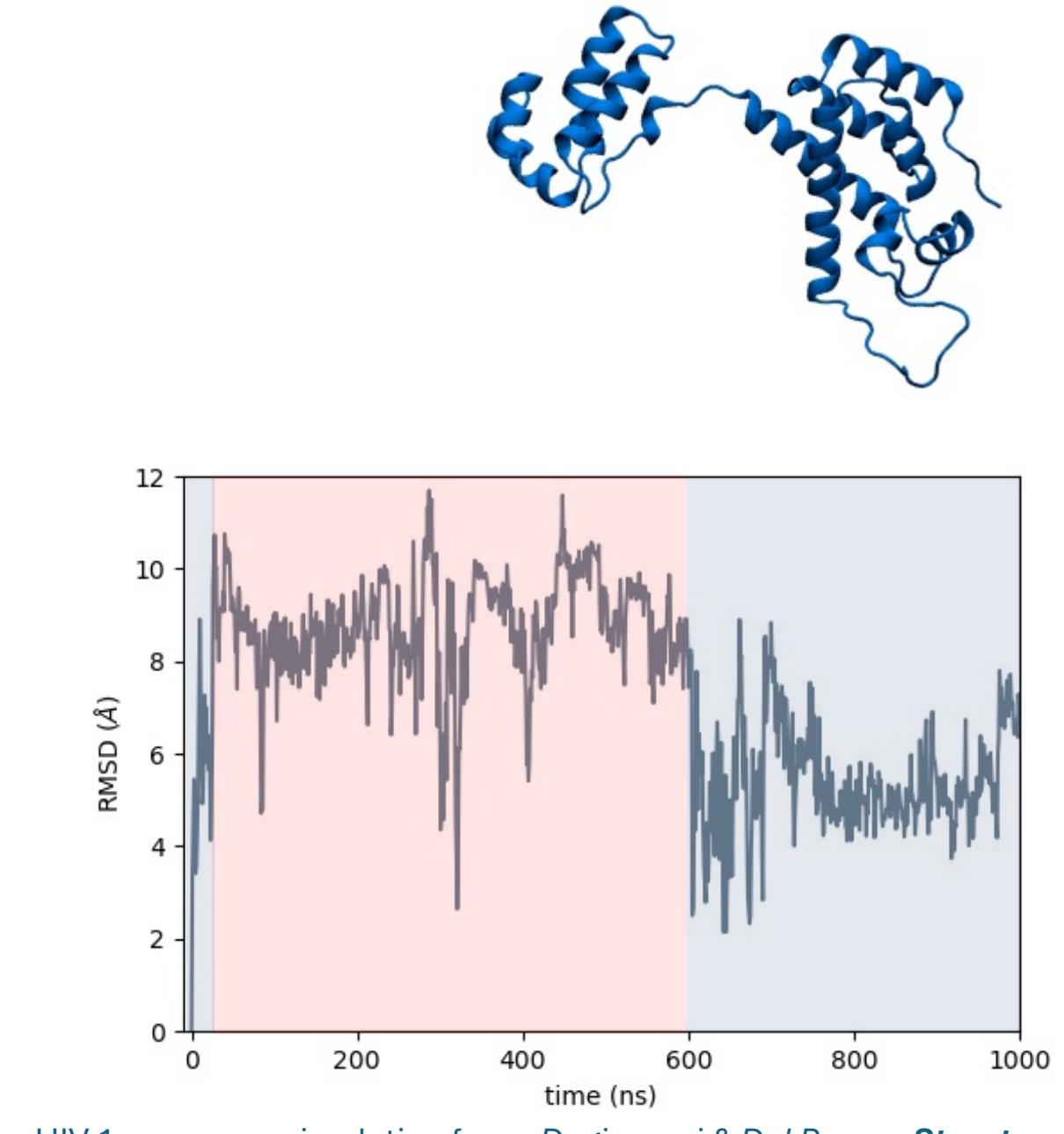


HIV-1 capsomer simulation from: *Degiacomi & Dal Peraro, Structure*,

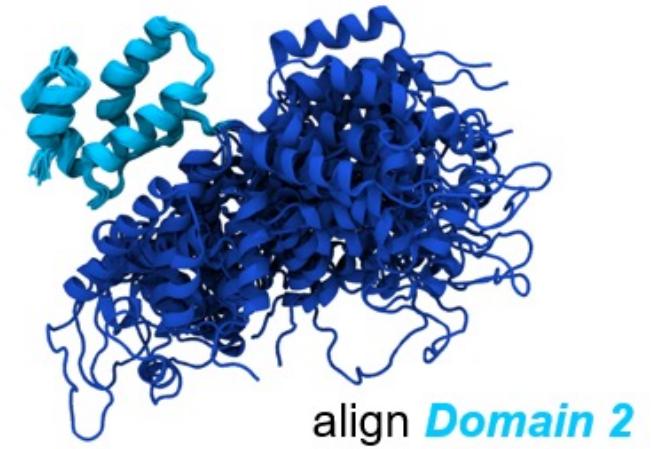
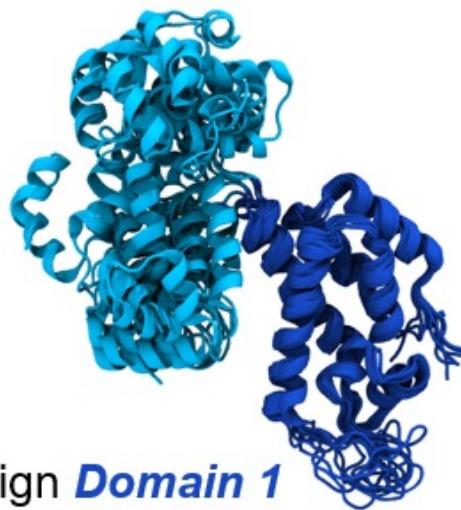
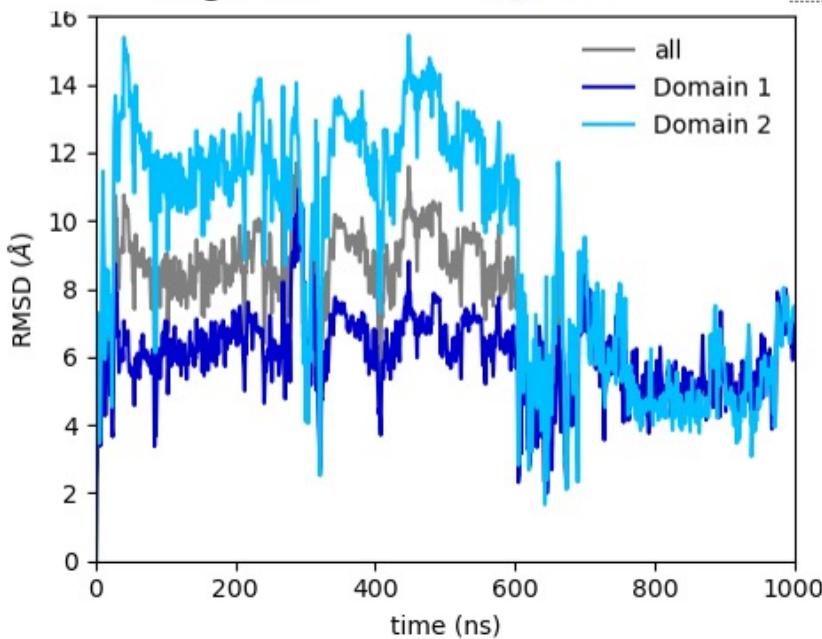
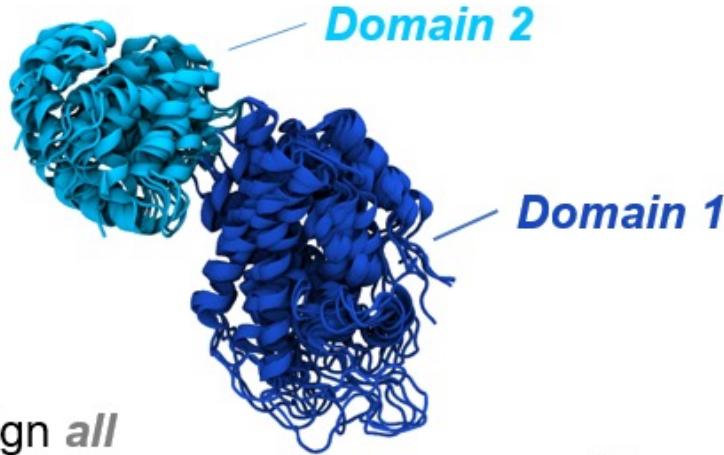
# Convergence?



*Refrain from using RMSD as a single indicator of simulation convergence.*

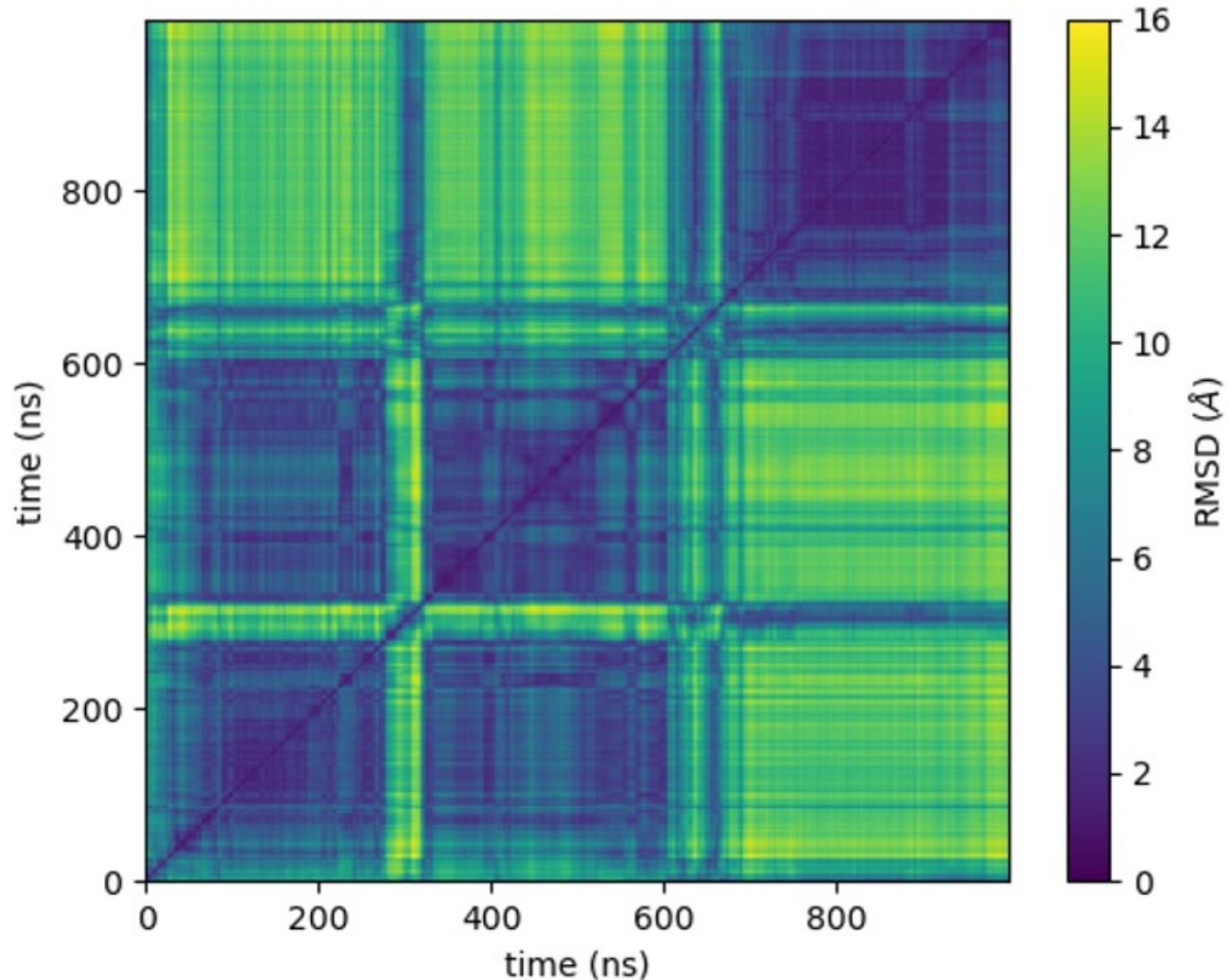


# RMSD is alignment- and selection-dependent



# Pairwise RMSD

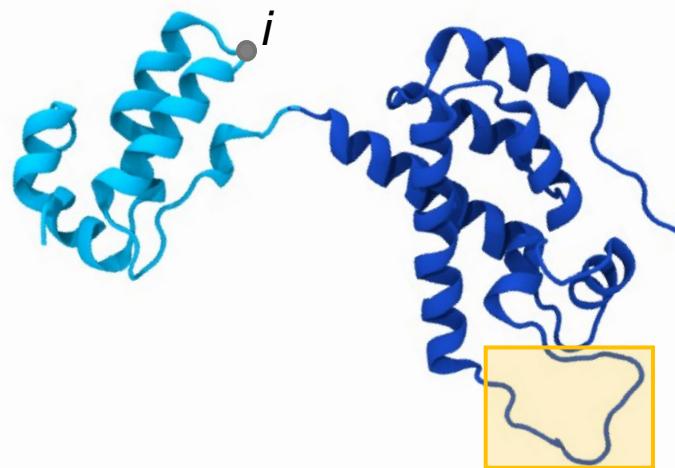
- Two structures with same RMSD from a reference are not forcefully similar to each other.
- Pairwise RMSD helps seeing if protein re-visits conformations throughout the simulation.



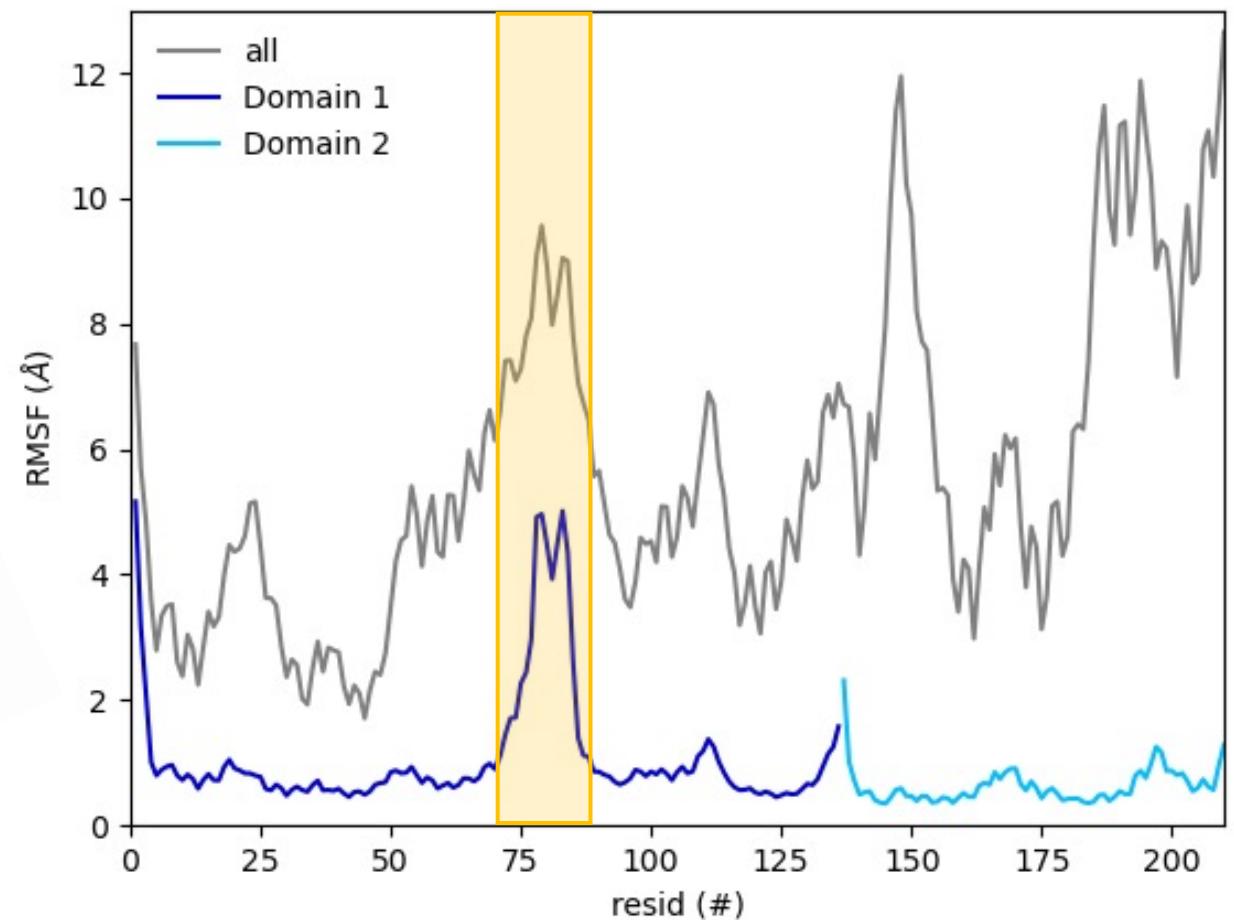
# Root Mean Square Fluctuation (RMSF)

The RMSF  $\sigma_i$  of atom  $i$  calculates how much it fluctuates around its mean position  $\langle X_i \rangle$ .

$$\sigma_i = \sqrt{\langle (X_i - \langle X_i \rangle)^2 \rangle}$$

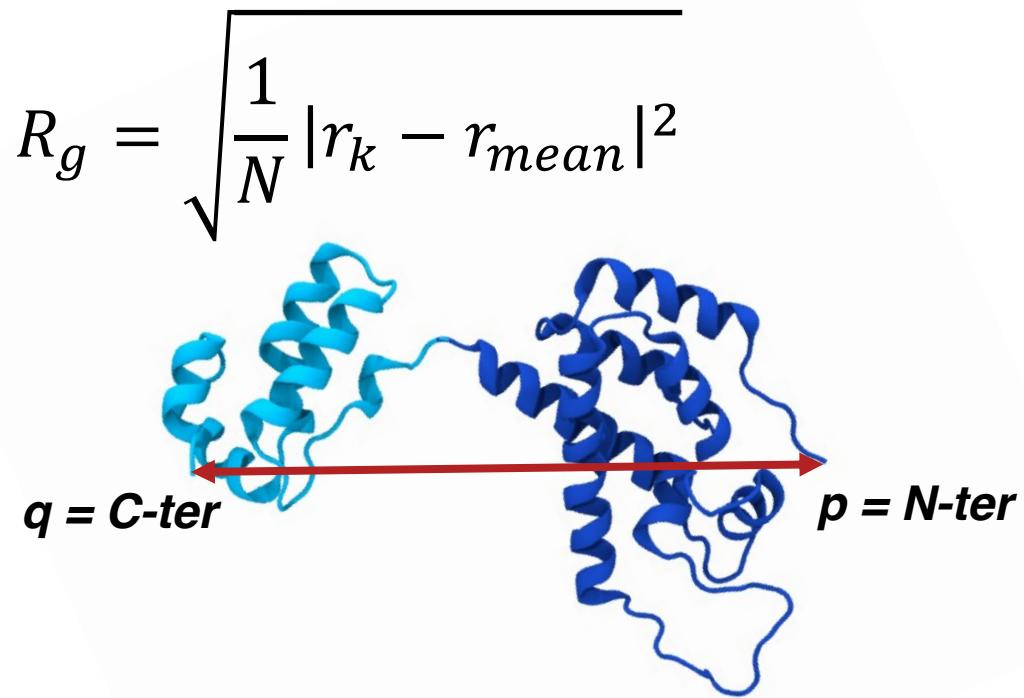


Helps identifying flexible/rigid regions.  
Typically done on  $C_\alpha$  atoms.  
Warning: result depends on alignment!

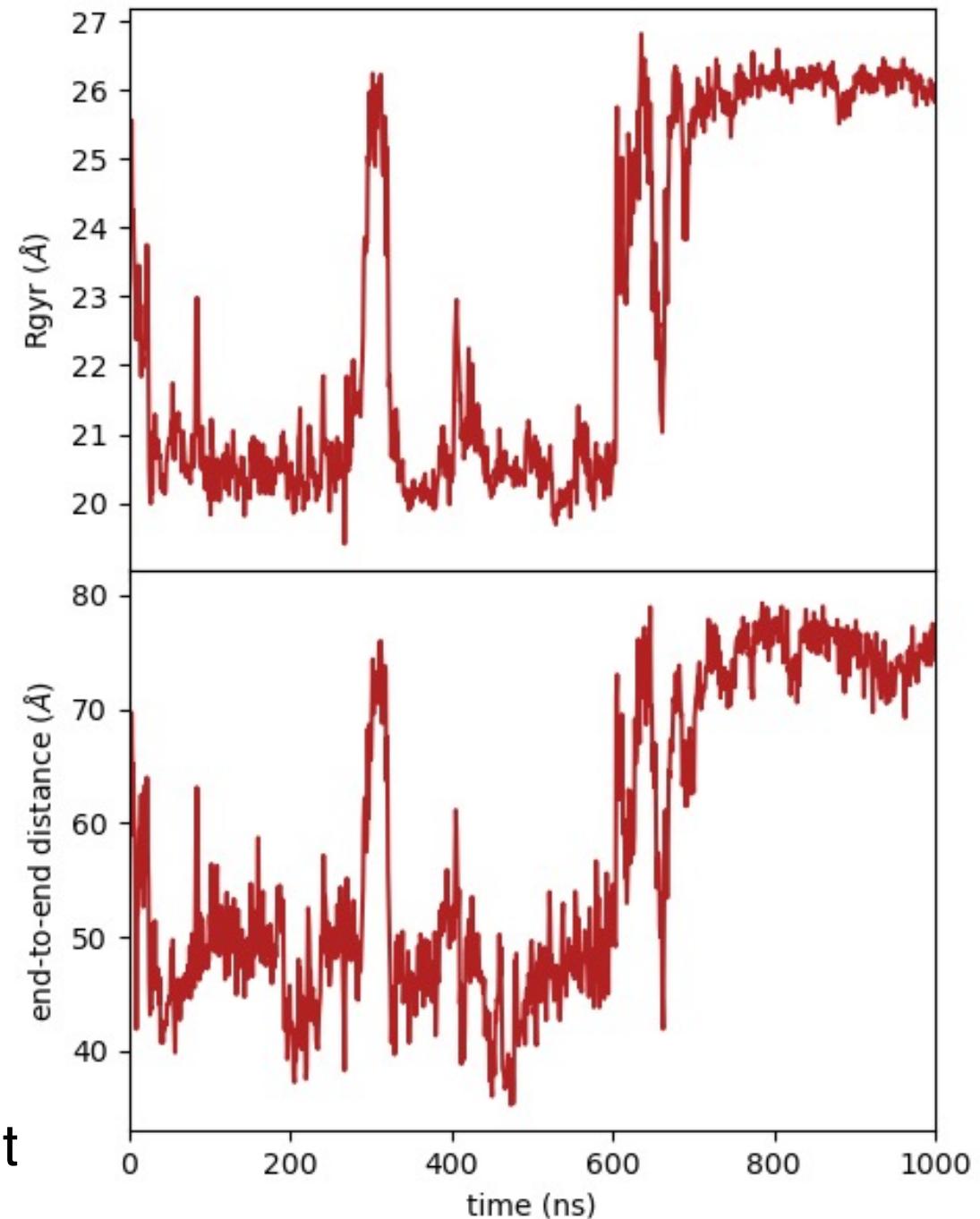


# end-to-end distance and Radius of Gyration ( $R_g$ )

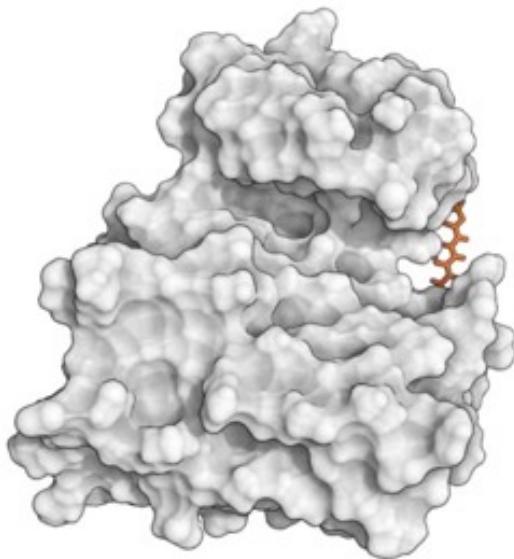
$$d(p, q) = \|p - q\|$$



Help quantifying protein compaction.  
Internal properties: do *not* depend on alignment

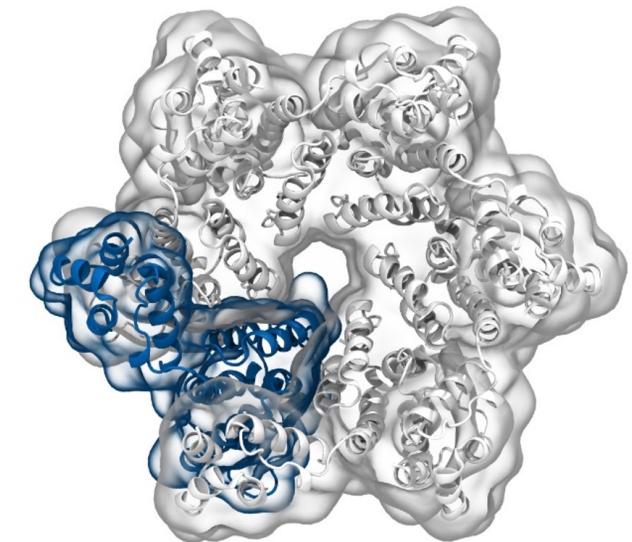


# Simulation of Biomolecules



## Dimensionality Reduction

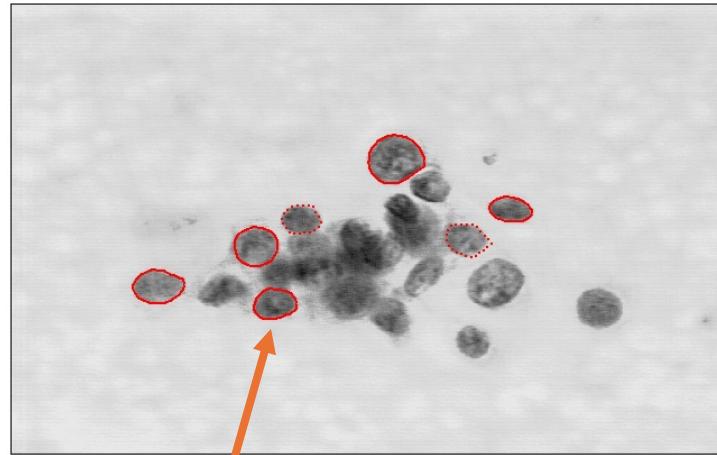
Delivered by **Tim Spankie**  
University of Edinburgh  
January 2025



Based on content from a ten-hour series <https://github.com/CCPBioSim/BioSim-analysis-workshop>  
by Matteo Degiacomi ([matteo.degiacomi@ed.ac.uk](mailto:matteo.degiacomi@ed.ac.uk)) and Antonia Mey ([antonia.mey@ed.ac.uk](mailto:antonia.mey@ed.ac.uk))

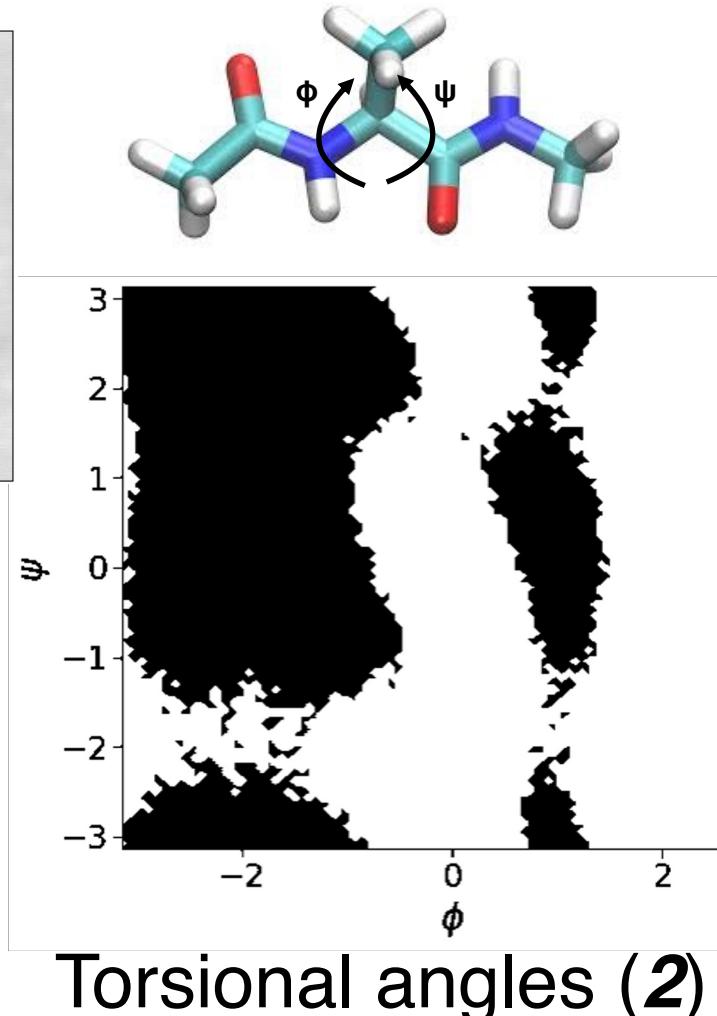
# *features* are possible ways to represent data

5  
0  
4  
1  
9  
2  
1  
3  
1  
4

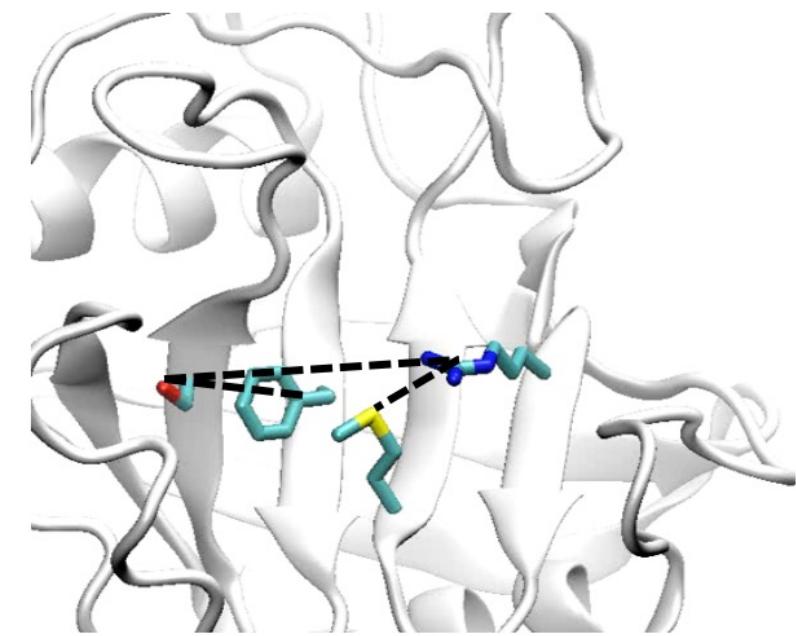


Sphericity (1)

Pixels colour  
( $28 \times 28 = 784$ )



Torsional angles (2)



atomic positions (459)  
atomic distances (3)

# Not all features are useful

Task: predict the weather in Edinburgh  
using historical data

data = { $X$ ,  $Y$ ,  $Z$ } → sun, rain, snow

{Temperature (C),  
~~Temperature (K)~~,  
Humidity (g/m<sup>3</sup>)}

2 decorrelated  
features

{Temperature (C),  
~~Swiss cheese export (£)~~,  
Humidity (g/m<sup>3</sup>)}

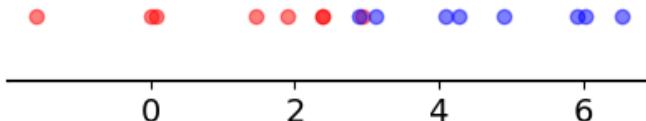
2 relevant  
features



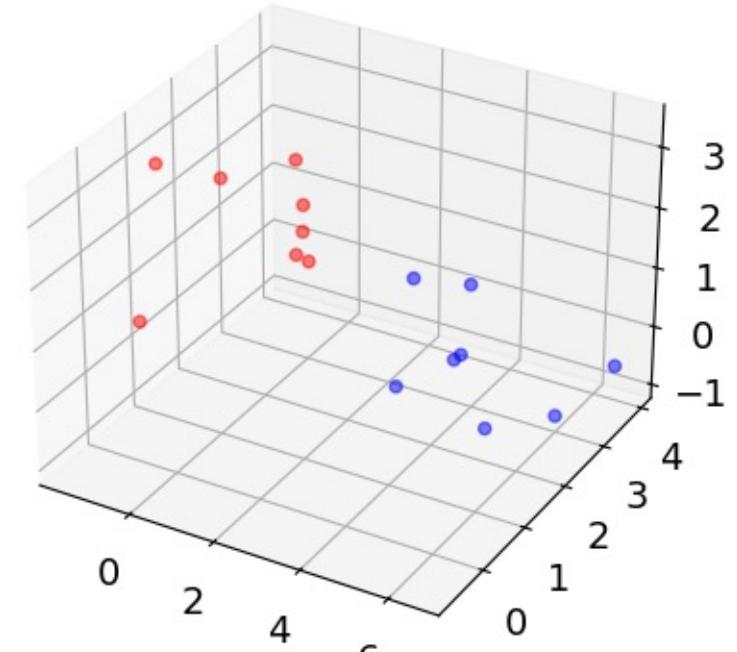
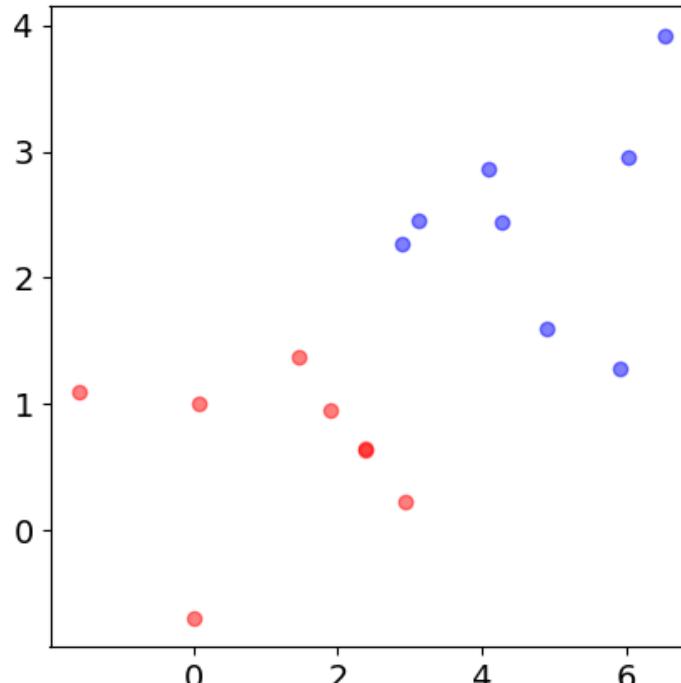
{Avg. gas expenditure (£),  
Heat strokes (#),  
Slipping accidents (#),  
Sunscreen sold (£)}

4 features connected to  
another quantity:  
temperature

# Curse of dimensionality



$$\|x\|_2 = \sqrt{\sum_{i=1}^N x_i^2}$$

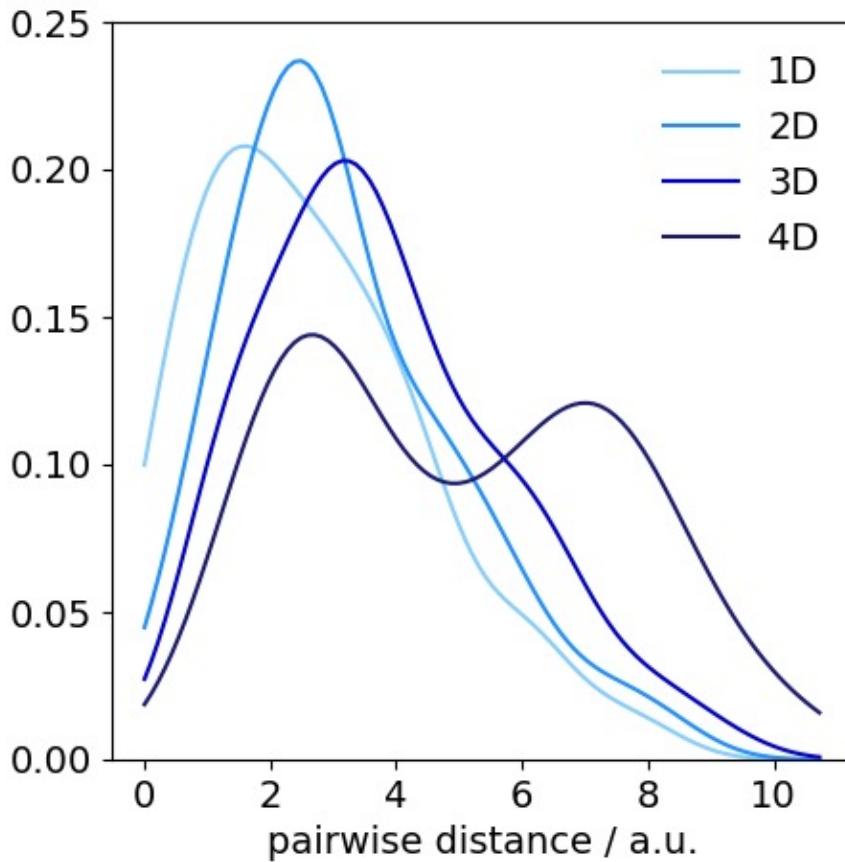


Distances between  $M$  data points  $x \in \mathbb{R}^N$  increase, when  $N$  increases

**Problem:** less data density increases uncertainty on underlying data structure

# [Extra] Curse of dimensionality

$$\|x\|_2 = \sqrt{\sum_{i=1}^N x_i^2}$$



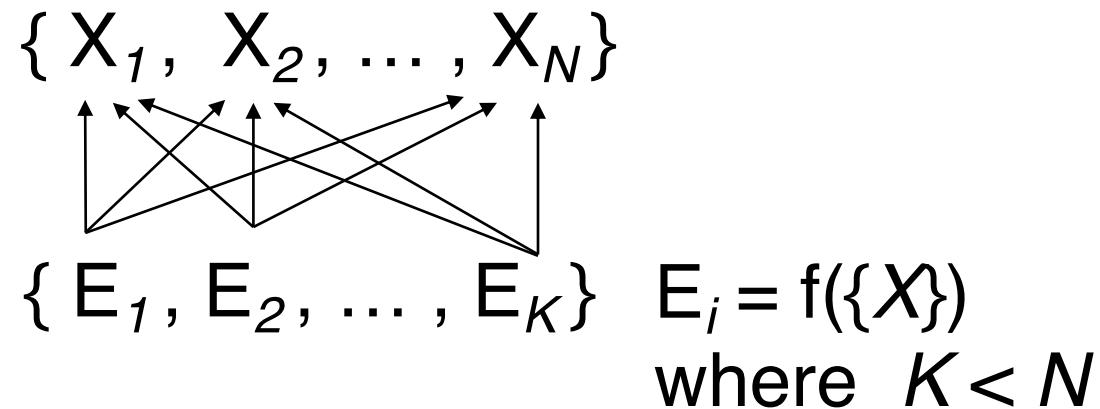
Distribution of pairwise distances between points shown in previous slide

Distances between  $M$  data points  $x \in \mathbb{R}^N$  increase, when  $N$  increases

**Problem:** less data density increases uncertainty on underlying data structure

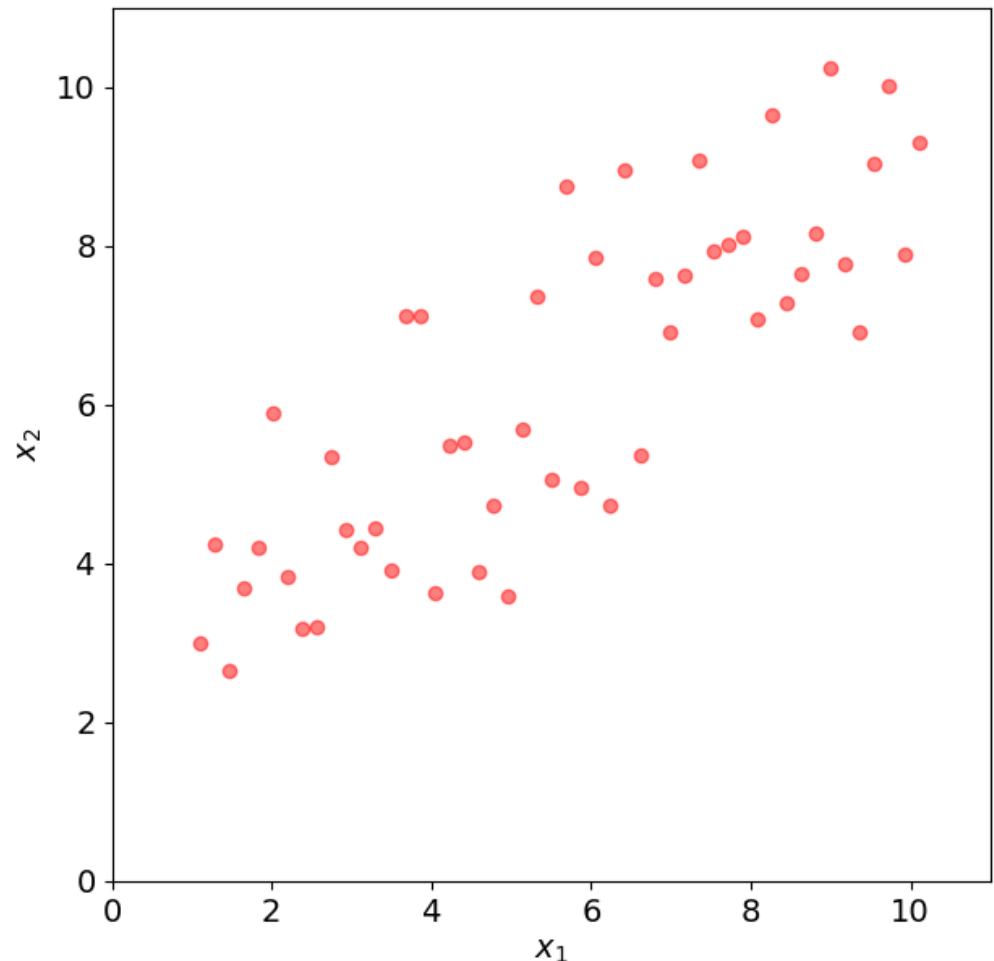
# Reducing features increases data density

- Choose appropriate features [expert user]
- Remove features
- Find a lower dimensional representation  $E$  of features  $X$



# Principal Components Analysis (PCA)

Let  $\mathbf{X}$  a dataset of  $M$  datapoints in  $N$  dimensions (here,  $M=50$  and  $N=2$ )

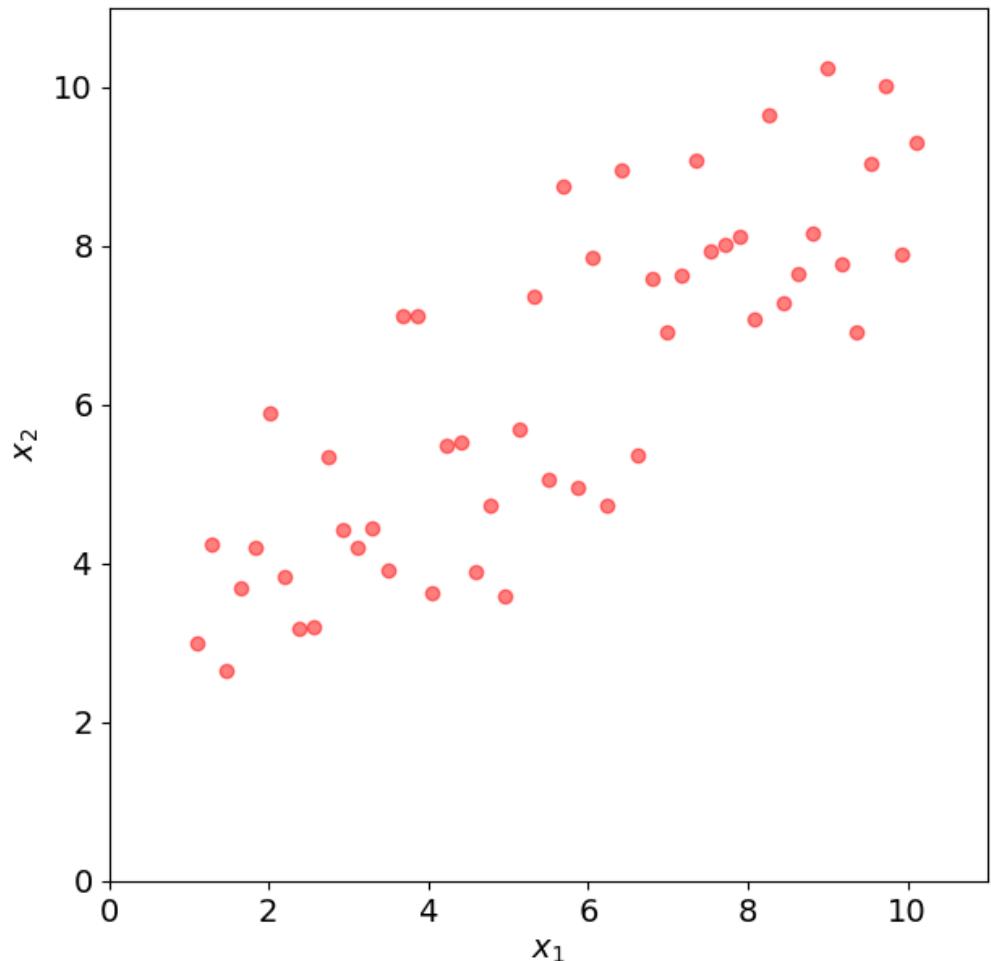


- Center data:  
$$\mathbf{X}' = \mathbf{X} - \boldsymbol{\mu}$$
- Compute data covariance matrix  $\mathbf{C}$ :  
$$c_{i,j} = \frac{1}{M} \sum_{k=1}^M \mathbf{x}'_i \mathbf{x}'_j$$
- Calculate eigenvalue decomposition:  
$$\mathbf{C} = \mathbf{V} \boldsymbol{\lambda} \mathbf{V}^{-1}$$

$\mathbf{V}$  is an  $N \times N$  matrix of **eigenvectors**  
 $\boldsymbol{\lambda}$  is an  $N \times N$  diagonal matrix of **eigenvalues**

# [Extra] Principal Components Analysis (PCA)

Let  $\mathbf{X}$  a dataset of  $M$  datapoints in  $N$  dimensions (here,  $M=50$  and  $N=2$ )



An eigenvector  $\mathbf{v}$  of  $\mathbf{C}$  respects:

$$\mathbf{C}\mathbf{v} = \lambda\mathbf{v}$$

Find eigenvalues as the roots of the characteristic polynomial:

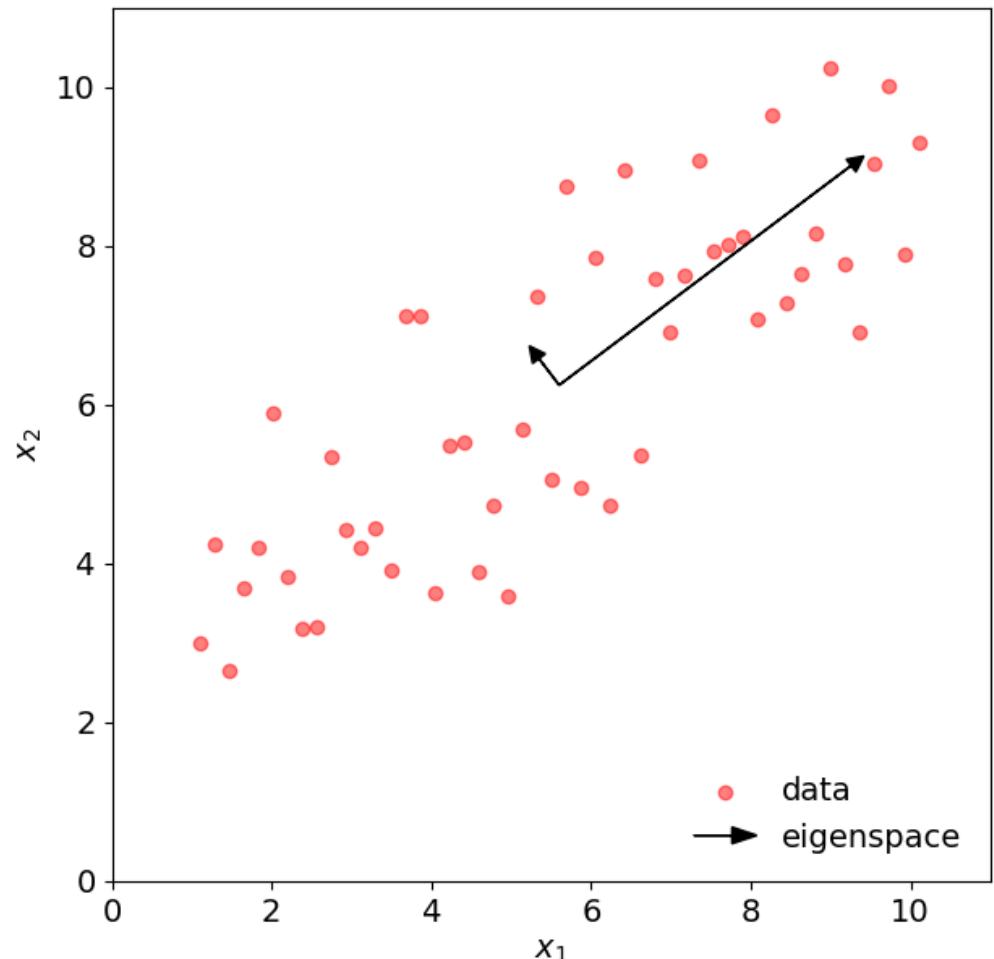
$$p(\lambda) = \det(\mathbf{C} - \lambda\mathbf{I}) = 0$$

The  $i$ -th eigenvector  $\mathbf{v}_i$  is found by solving:

$$\mathbf{C}\mathbf{v}_i = \lambda_i\mathbf{v}_i$$

# Principal Components Analysis (PCA)

Let  $\mathbf{X}$  a dataset of  $M$  datapoints in  $N$  dimensions (here,  $M=50$  and  $N=2$ )



$$\mathbf{C} = \mathbf{V}\boldsymbol{\lambda}\mathbf{V}^{-1}$$

$\mathbf{V} = [\mathbf{v}_1 \dots \mathbf{v}_N]$  eigenvectors:  
orthonormal base

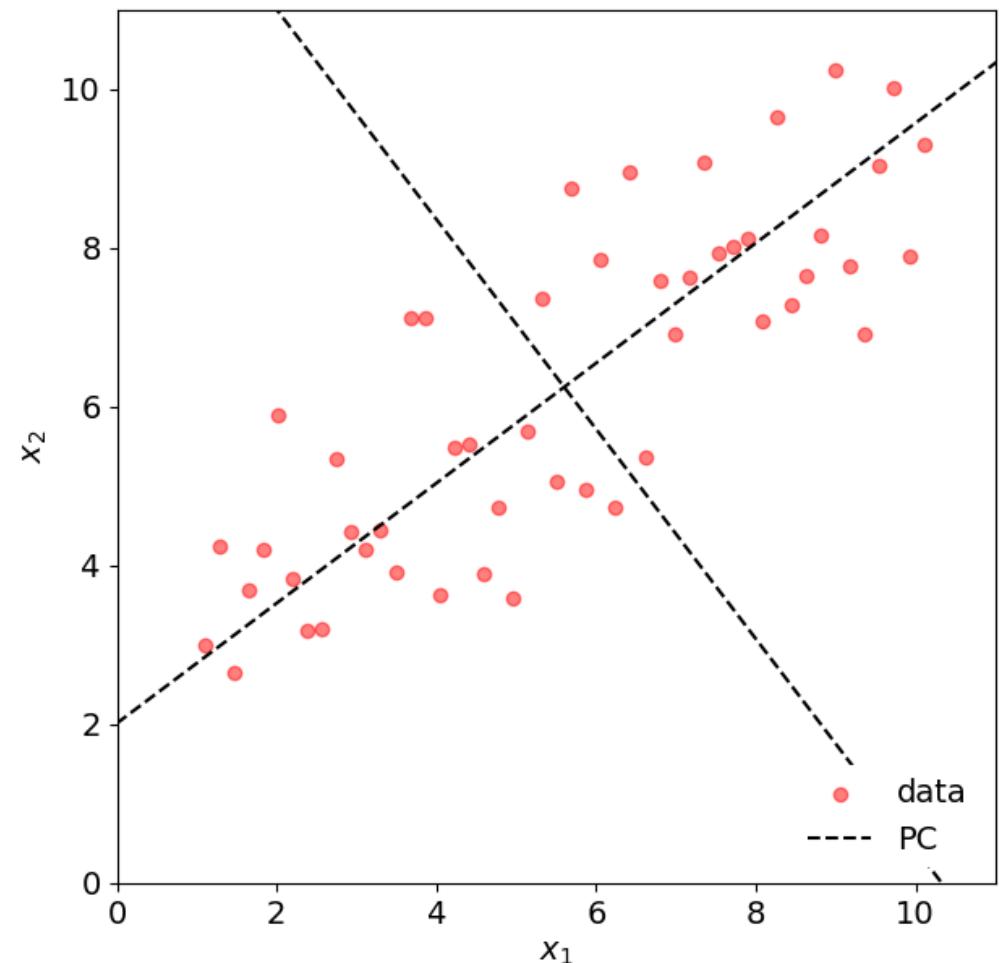
$\boldsymbol{\lambda}$  eigenvalues: scalars defining the  
importance of each eigenvector  
Importance  $r_i$  of each eigenvector  $\mathbf{v}_i$ :

$$r_i = \frac{\lambda_i}{\sum \lambda}$$

*Sort  $\mathbf{V}$  and  $\boldsymbol{\lambda}$  according to  $\lambda$*

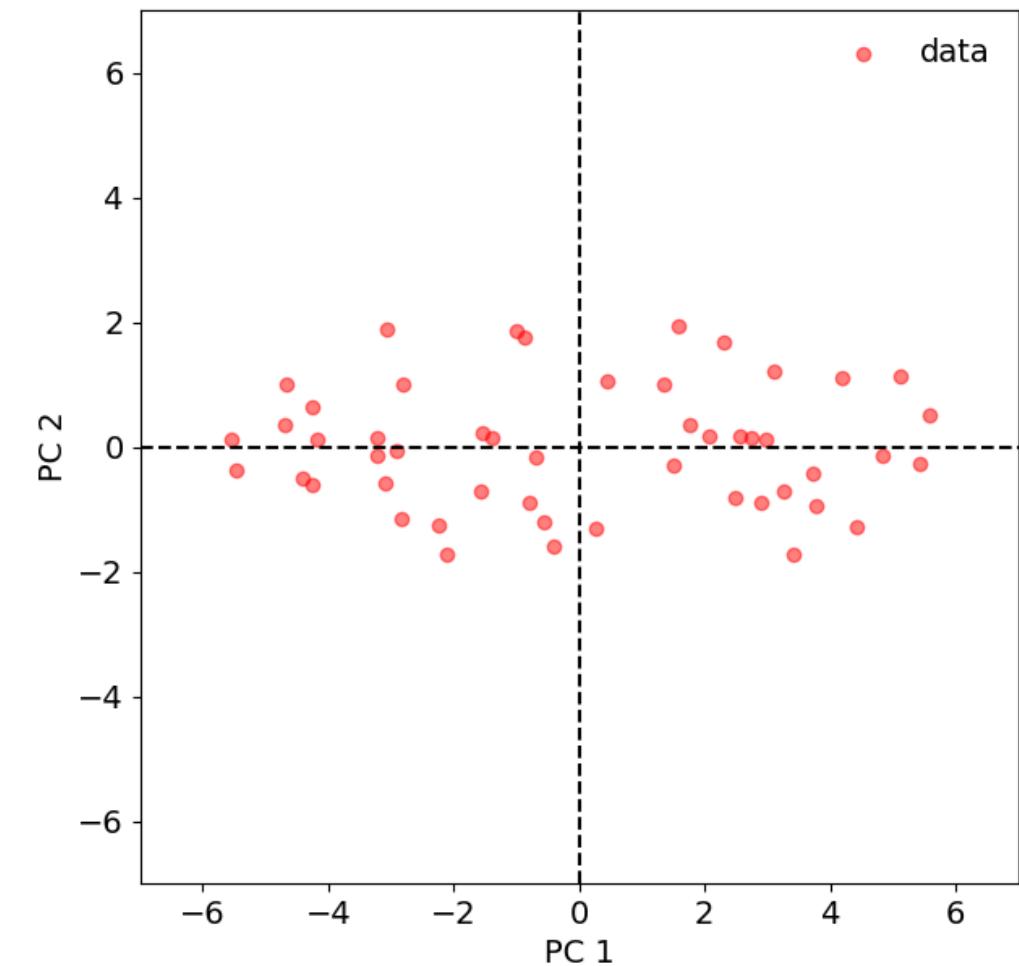
# Projection into the eigenspace

Let  $\mathbf{X}$  a dataset of  $M$  datapoints in  $N$  dimensions (here,  $M=50$  and  $N=2$ )



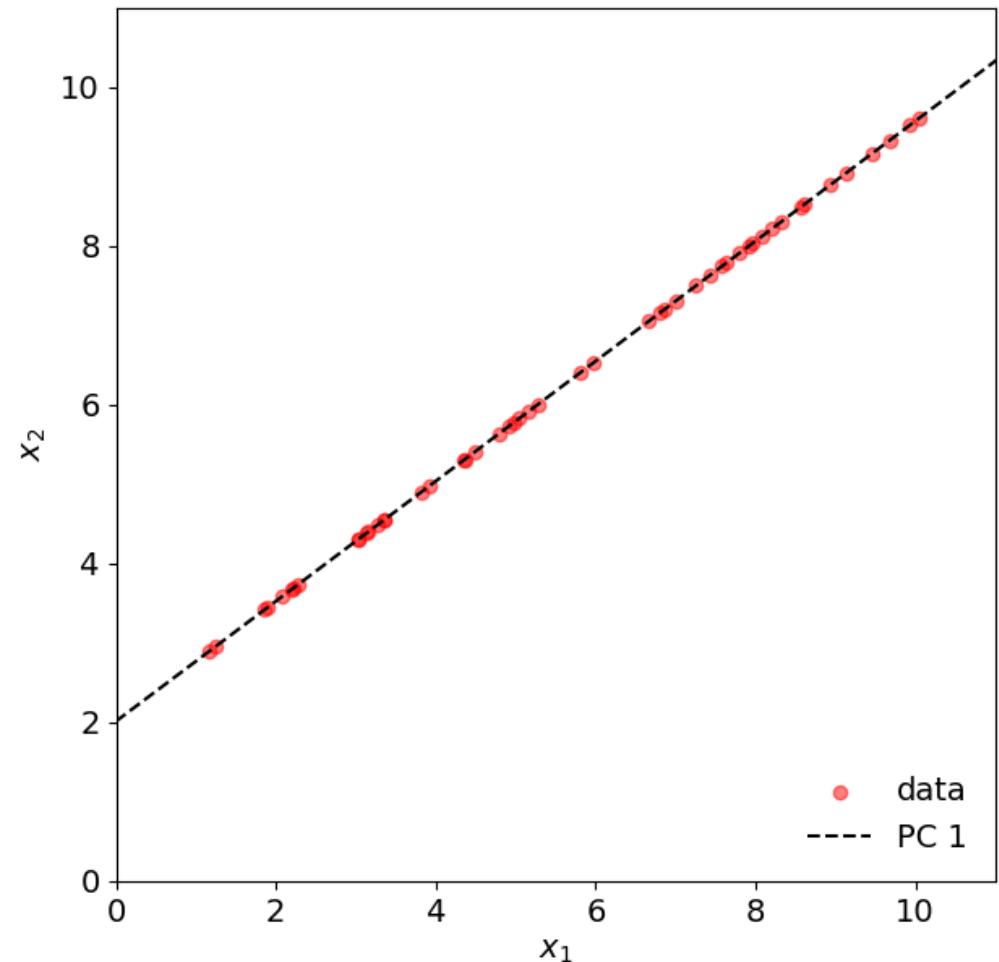
$$p = \mathbf{V}^T(\mathbf{x} - \boldsymbol{\mu})$$

transform



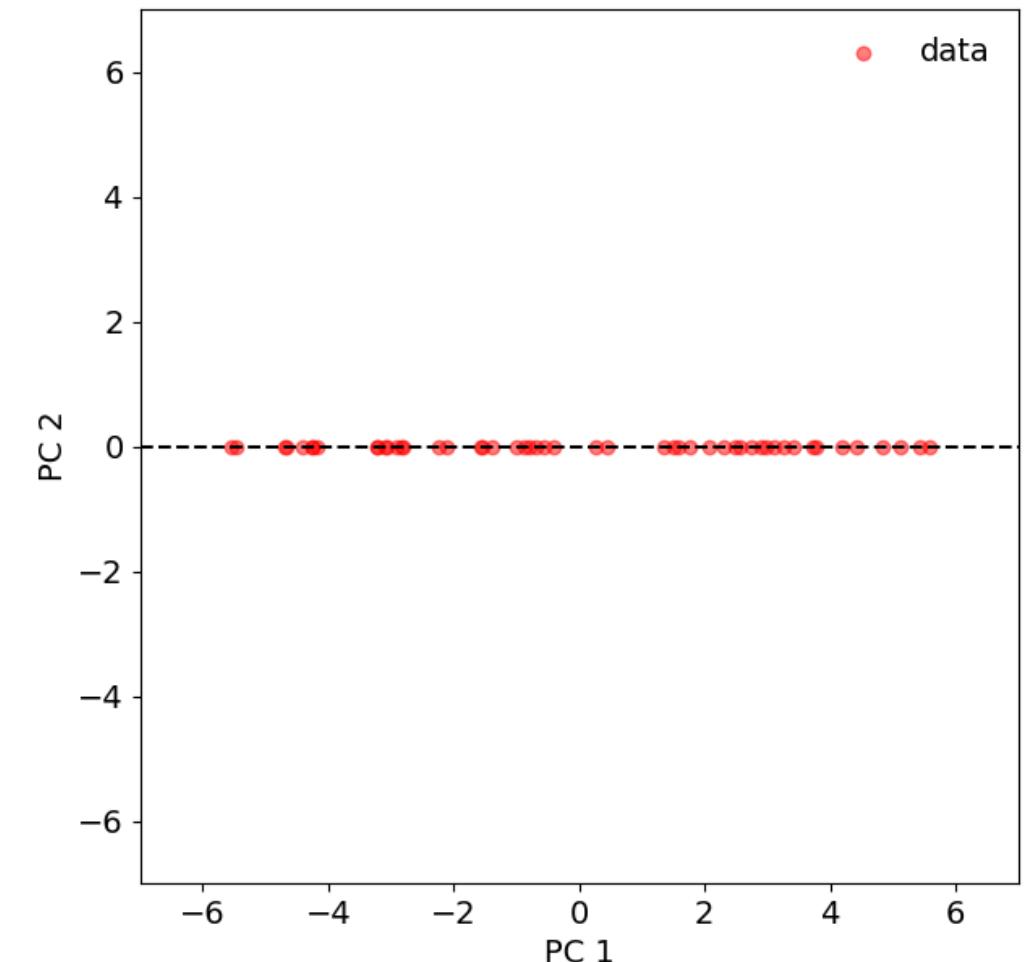
# Dimensionality reduction

Remove dimensions that least contribute to data variance



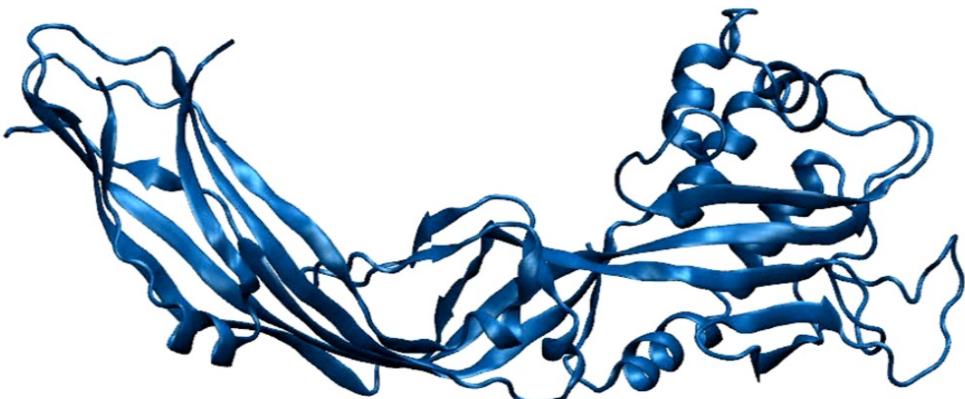
$$p = \mathbf{V}^T(\mathbf{x} - \boldsymbol{\mu})$$

transform

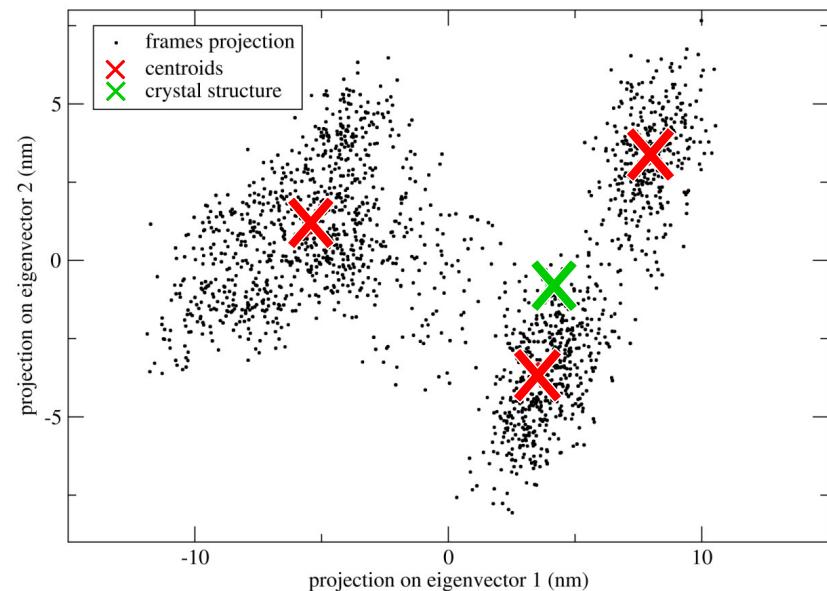


# [Example 2] Identifying dominant motions in proteins

Protein MD simulation



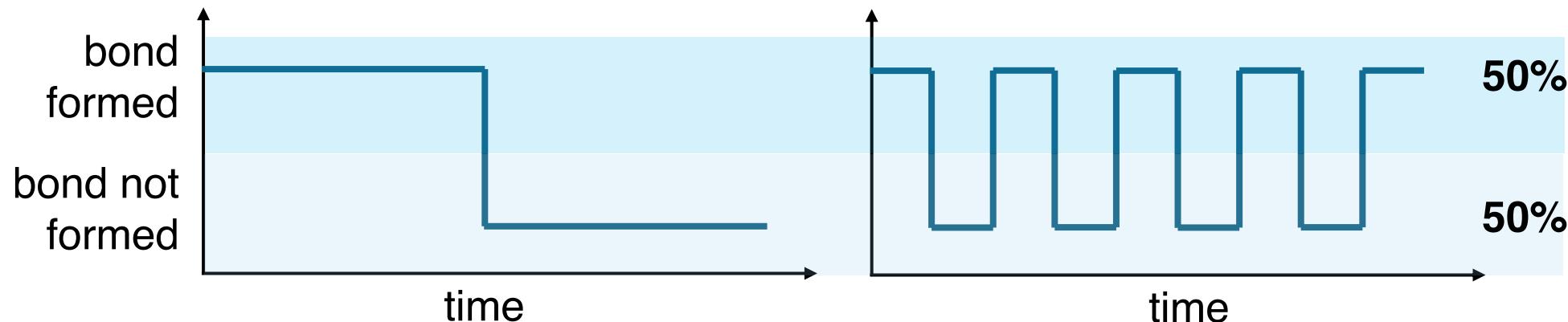
Eigenspace of  $C_\alpha$  coordinates



- Simulations are complex and noisy
- The first PCs capture large-scale collective motions, last ones capture noise

# largest-scale motion $\neq$ slowest motion

**Thought experiment reminder:** typically hydrogen bond is considered established if donor-acceptor distance  $<2.5 \text{ \AA}$ , and donor-acceptor-hydrogen angle  $<20^\circ$ .



- Biological function is often determined by the kinetics of a process (e.g., a specific conformational change).
- Two processes with different kinetics can have the same statistical distribution in terms of structure.
- PCA separates in terms of structure, not timescales!