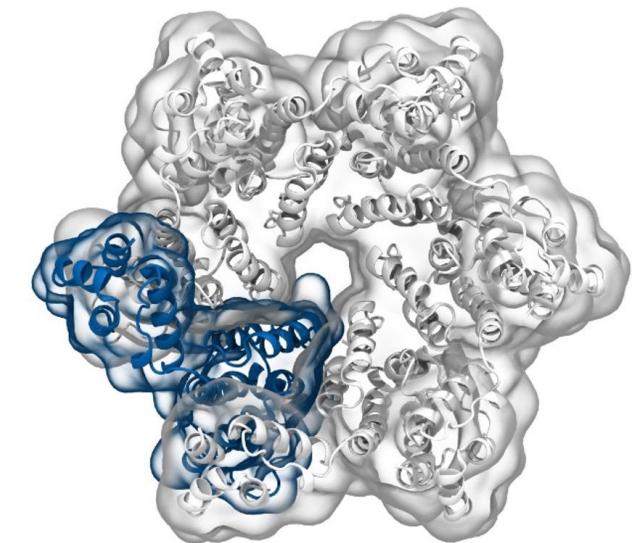
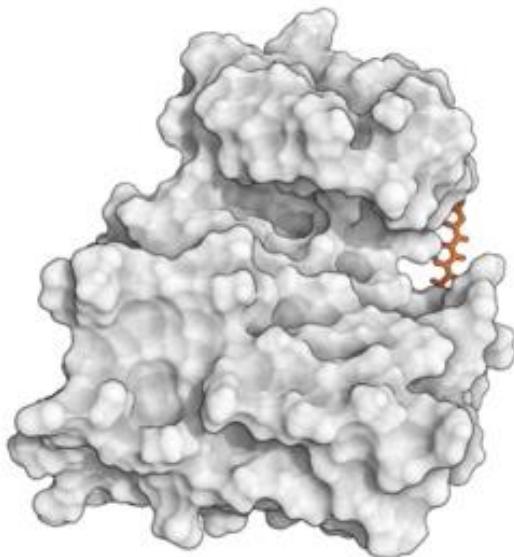


# Simulation of Biomolecules

## Introduction

**2024 CCP5 Summer School**



Dr Matteo Degiacomi

Durham University

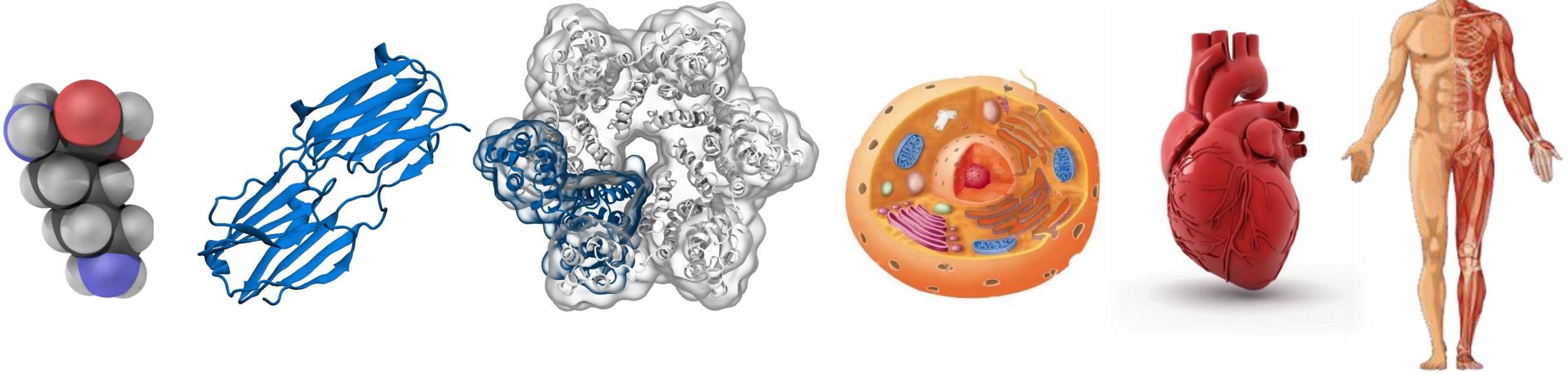
[matteo.t.degiacomi@durham.ac.uk](mailto:matteo.t.degiacomi@durham.ac.uk)

Dr Antonia Mey

University of Edinburgh

[antonia.mey@ed.ac.uk](mailto:antonia.mey@ed.ac.uk)

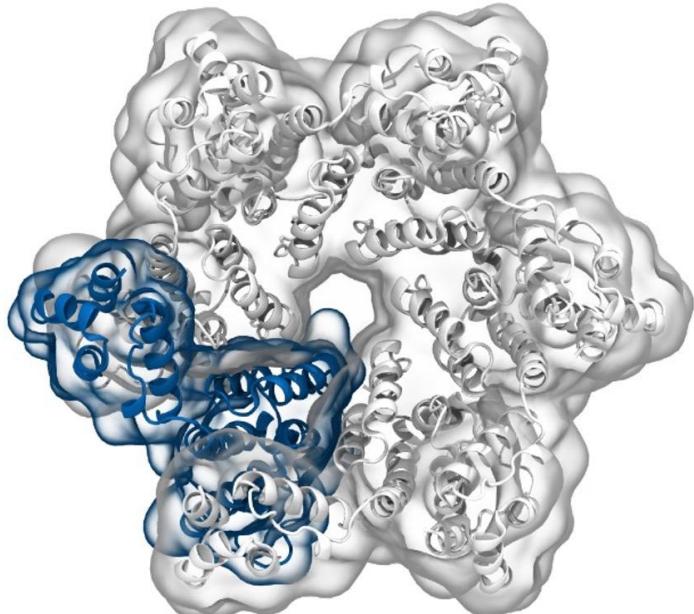
# Life emerges from molecular assembly



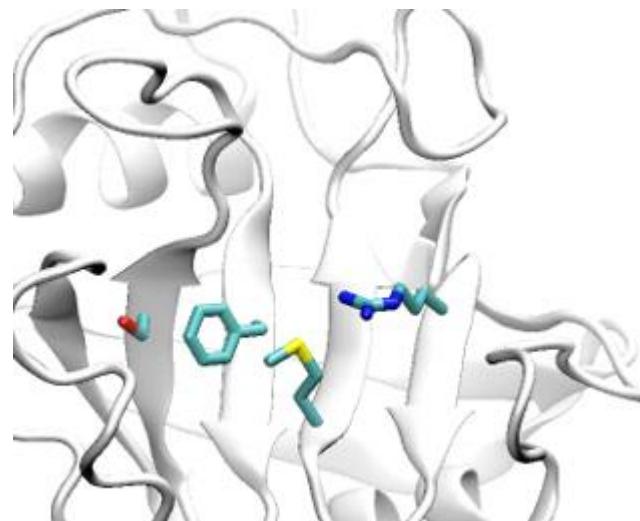
COMPLEXITY

# Structure and dynamics determine protein (mal)function

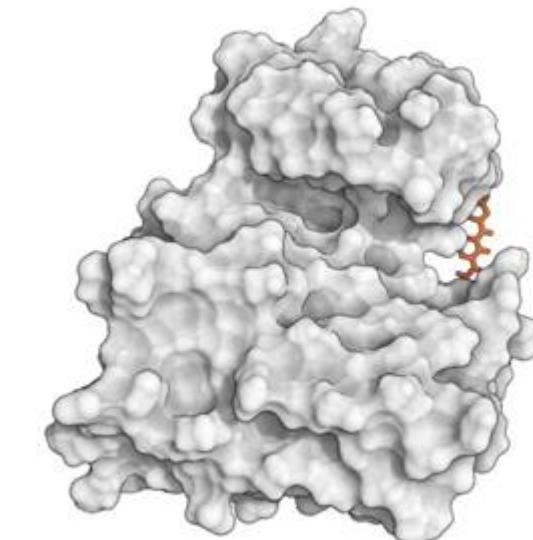
HIV Capsomer



Cyclophilin



Tyrosine kinase —dasatanib



M.T. Degiacomi, *Structure*, 2019

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

Y Shan et al. *JACS*, 2011

# General Information

[https://github.com/CCPBioSim/CCP5\\_Simulation\\_of\\_BioMolecules](https://github.com/CCPBioSim/CCP5_Simulation_of_BioMolecules)

**Tuesday 23/7**

What are Proteins?  
Protein preparation  
Protein-ligand docking

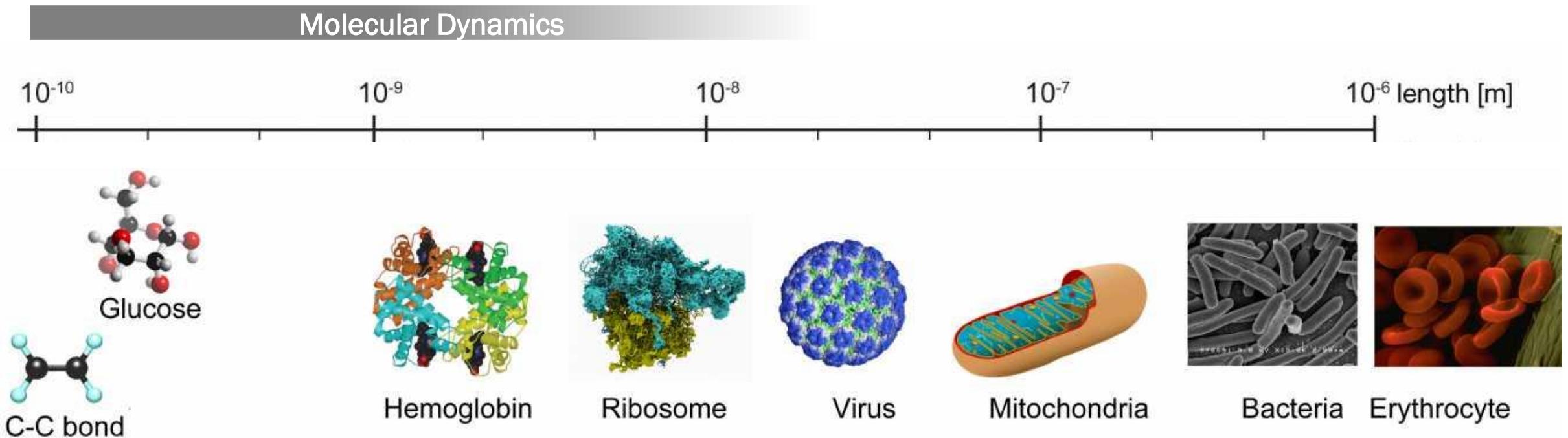
**Wednesday 24/7**

MD setup: BioSimSpace  
Data analysis: MDAnalysis  
Data analysis: ML, part 1

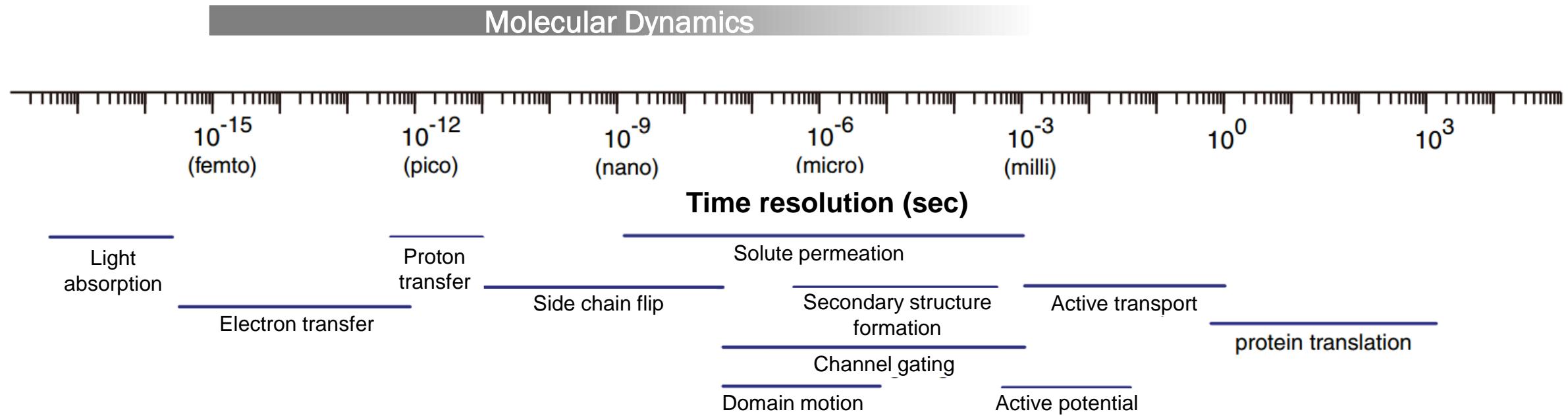
**Thursday 25/7**

Data analysis: ML, part 2  
Data analysis: MSM

# Sizes in biochemistry



# Timescales in biochemistry

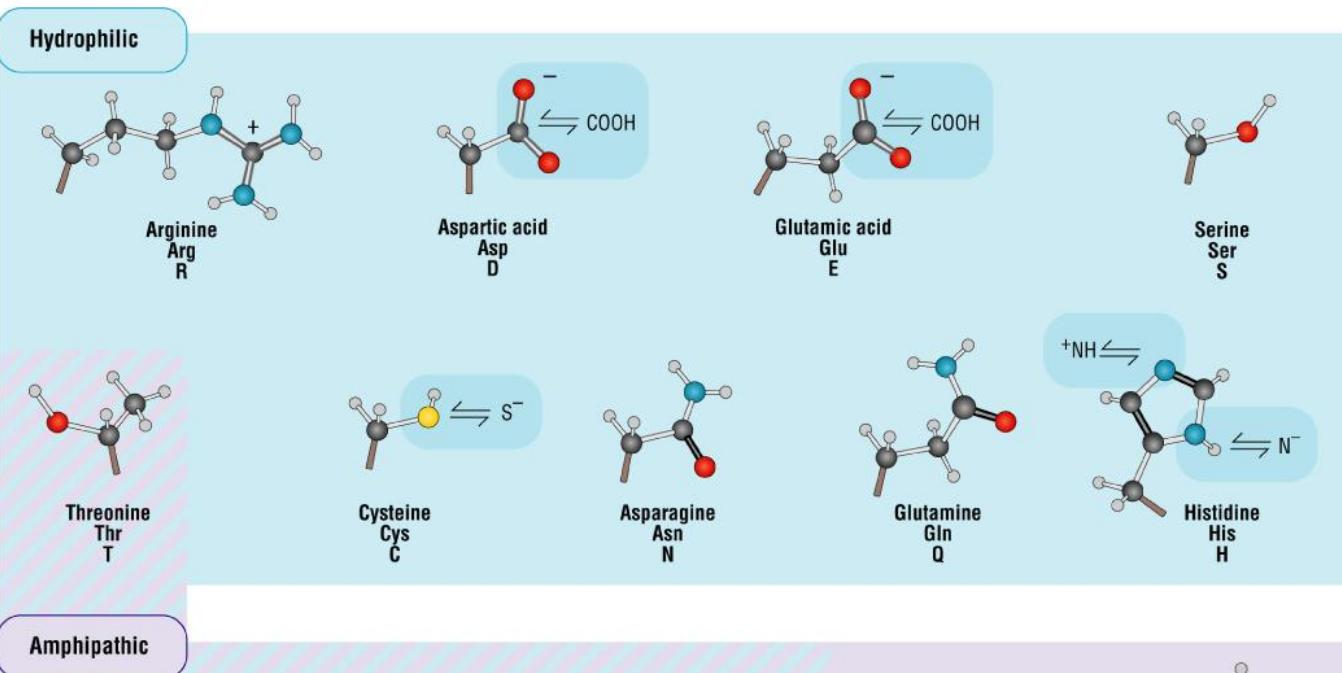
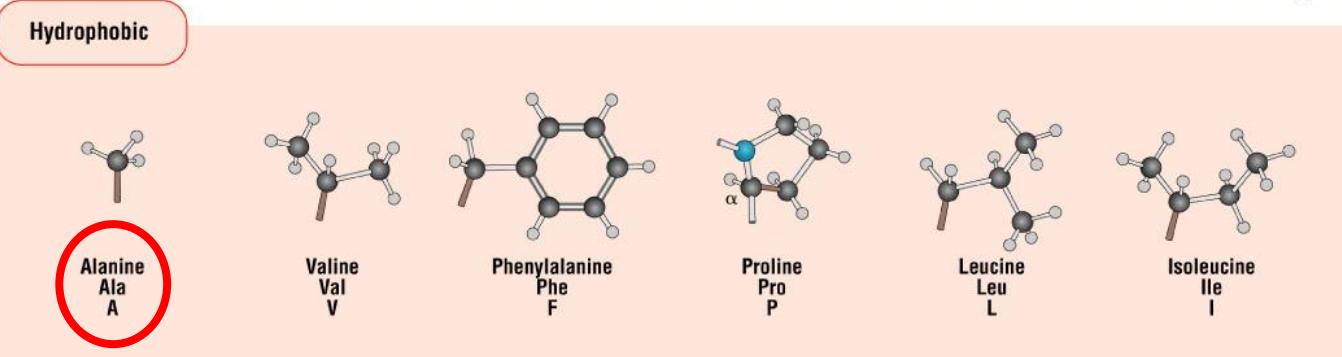
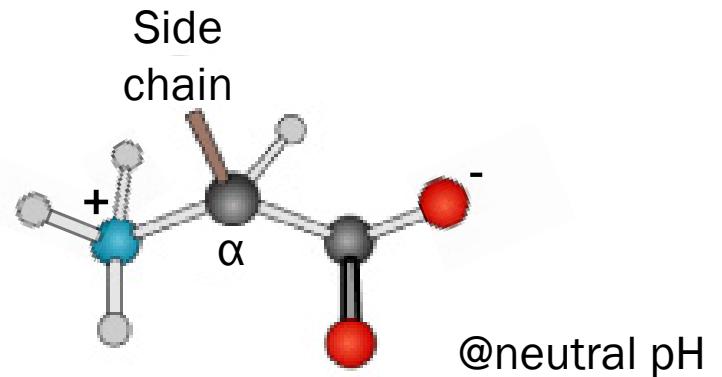


# **Part 1: what is a protein?**

# Proteins are amino acids polymers

Amino acids are composed of:

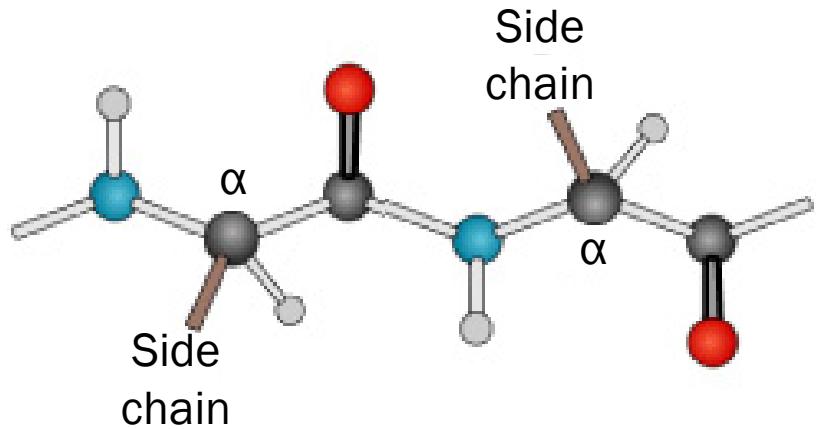
- **Backbone** (conserved)
- **Side chain** (variable)



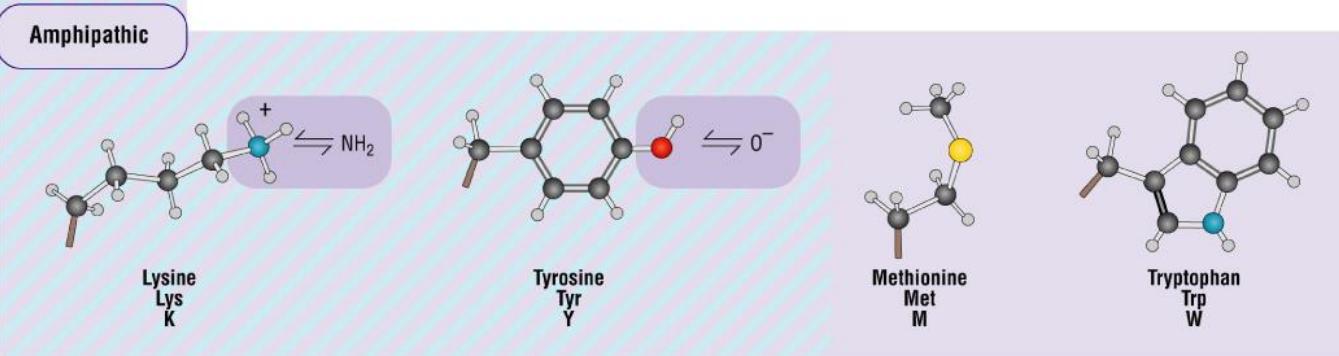
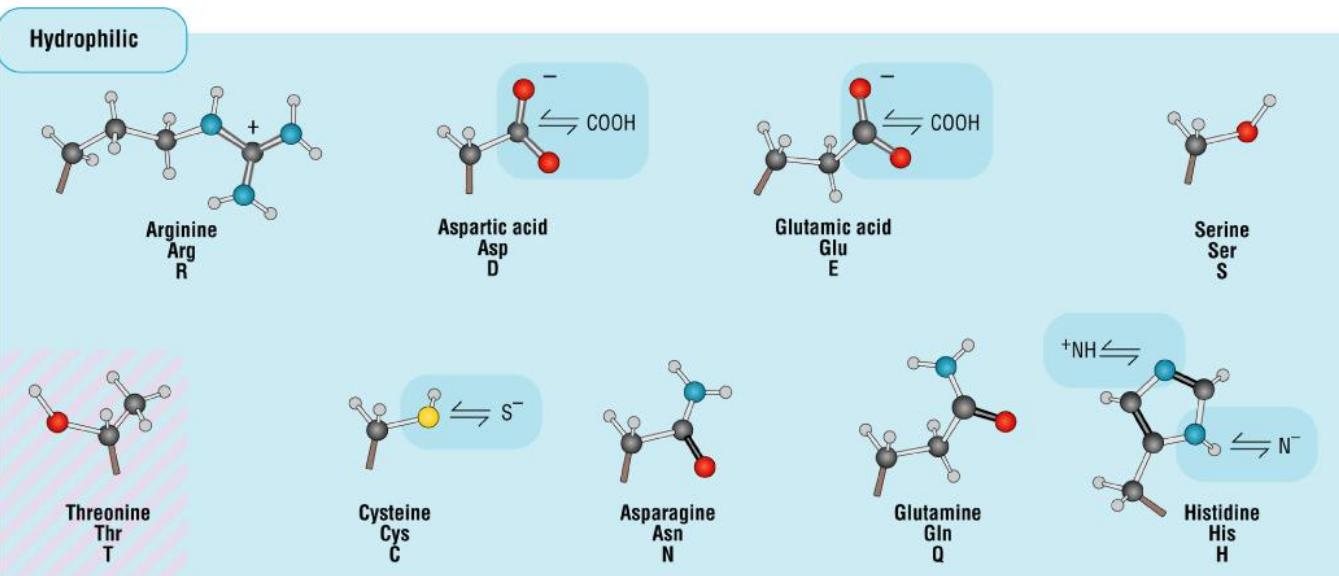
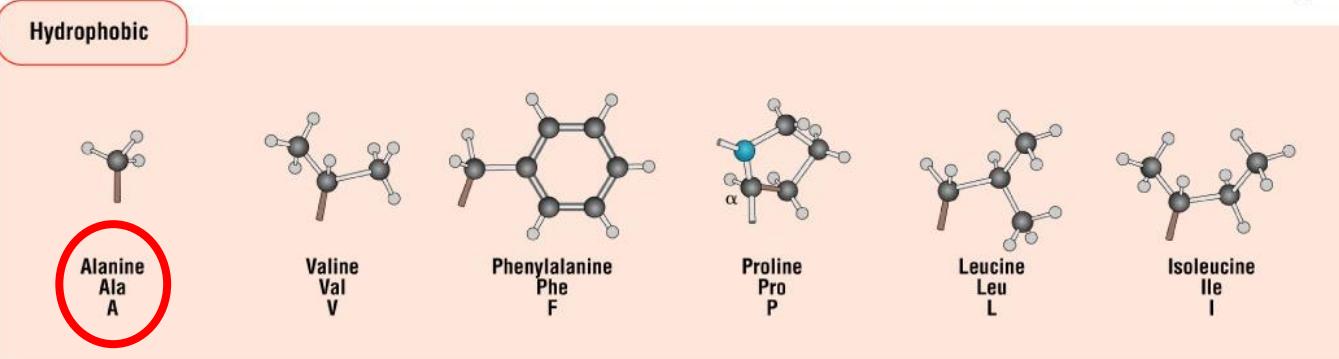
# Proteins are amino acids polymers

Amino acids are composed of:

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



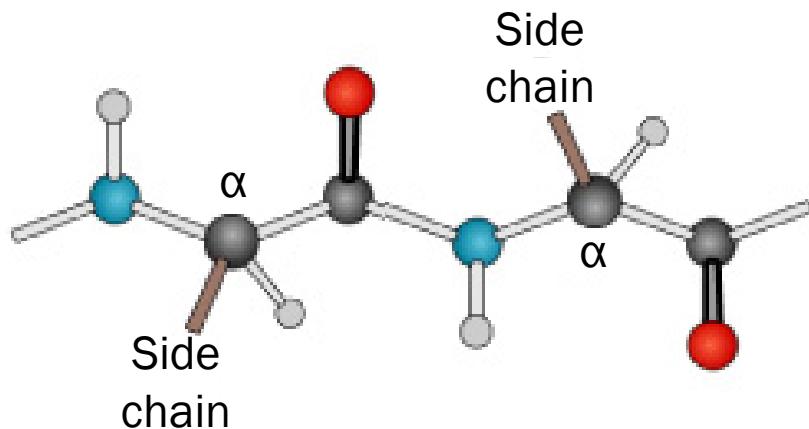
Amino acids polymerize forming a *peptidic bond* (condensation)



# Proteins are amino acids polymers

Amino acids are composed of:

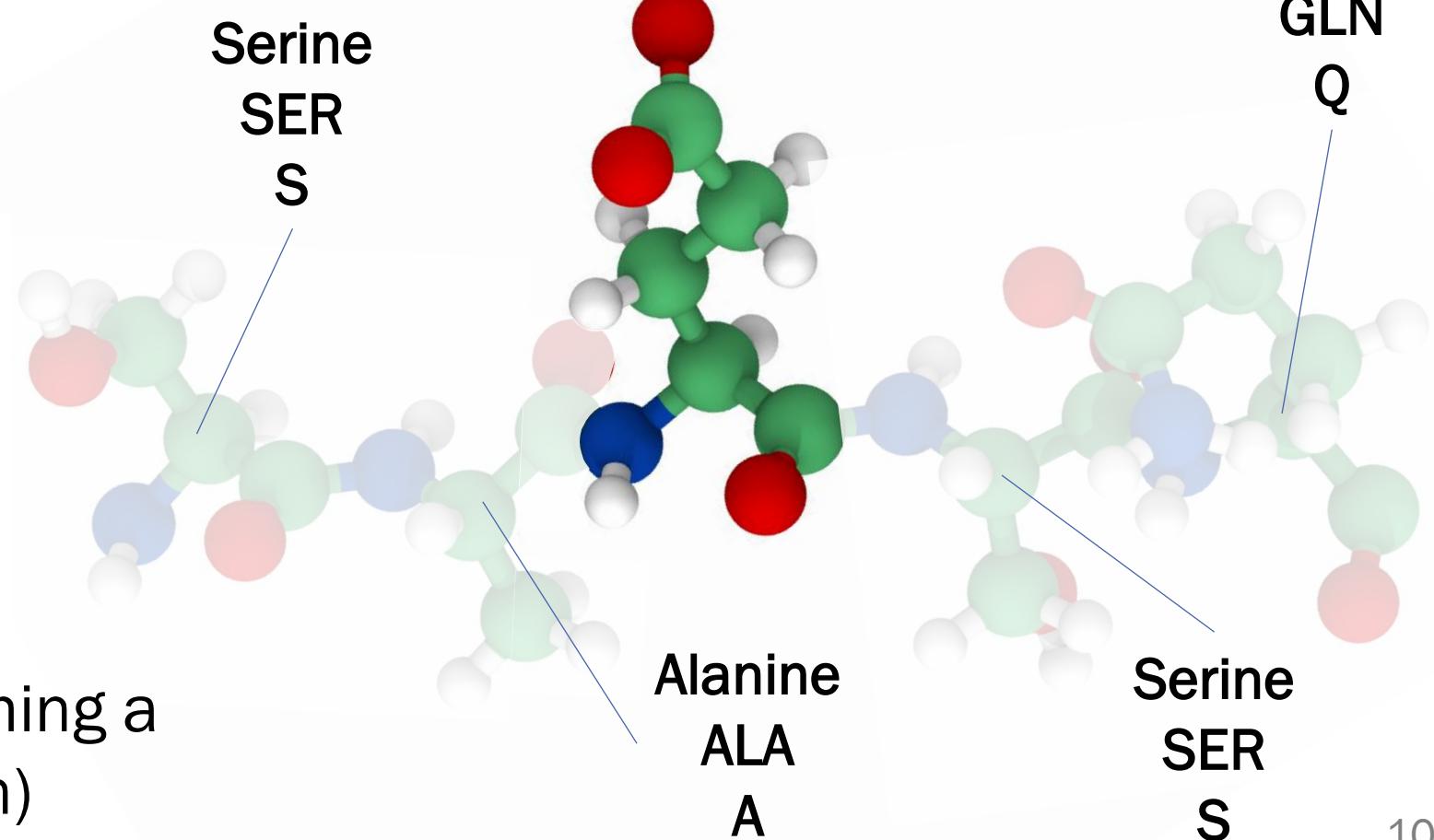
- **Backbone** (conserved)
- **Side chain** (variable)



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

Glutamic acid  
GLU  
E

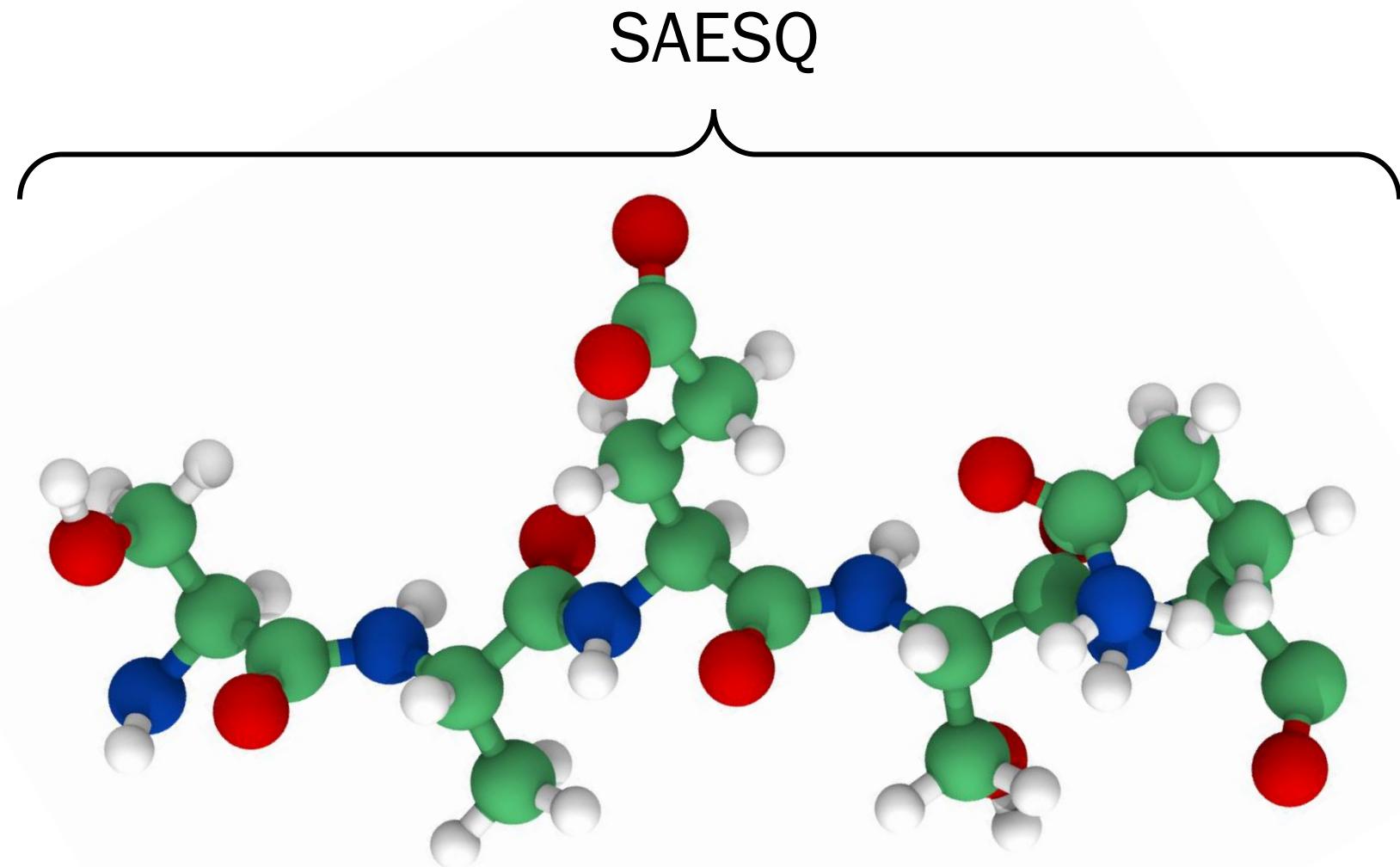
Glutamine  
GLN  
Q



# Protein Primary Structure: Sequence

Proteins with similar sequence, likely:

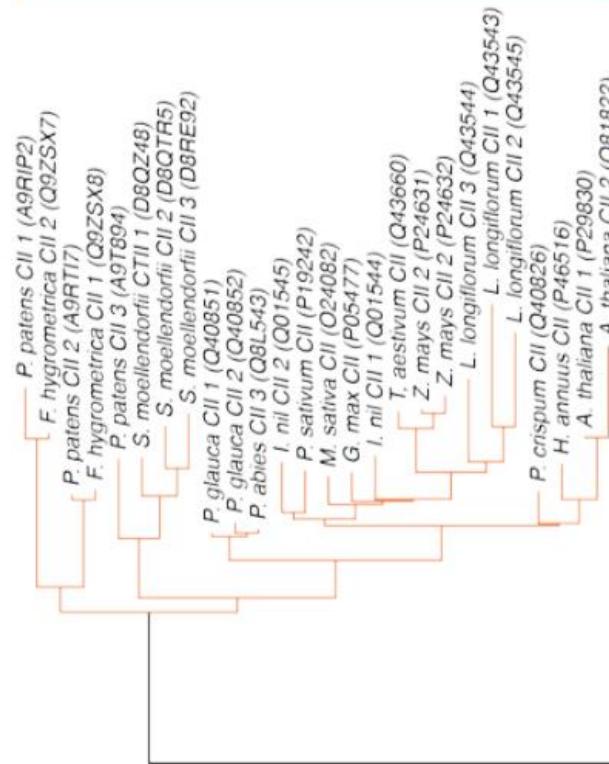
- have similar functions in an organism
- are evolutionarily related



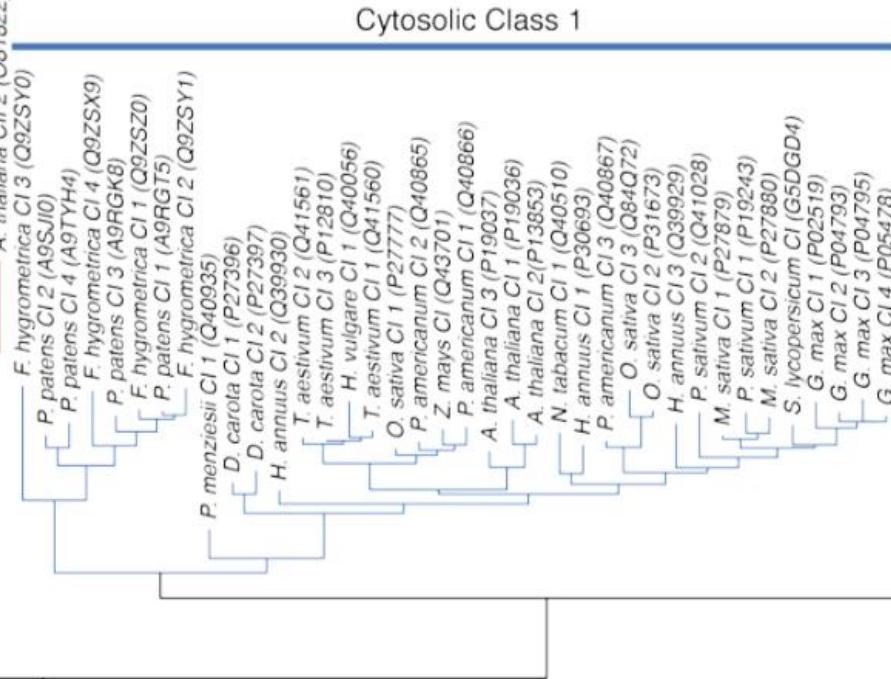
# Protein Primary Structure: Sequence

Angiosperm 1 MSLIPSFFSGRRSNVFD-PF--SL-DVWDPLKD-FPFSNSSPSASPRENPAFV-STRVDWKETPEAHVFKA  
DLPGLKKEEVKVEVE  
Gymnosperm 1 MSIIPSFFGRRSSSAFD-PF--SL-DVWDPFRAFTDLGGGPGSQFVNEASAVA-NTQIDWKETPEAHIFKA  
DLPGLKKEEVKIELE  
Bryophyte 1 MAL--SLFGSRGNGVFD-PF--EFGSVWDPFSA---PESGLSRKLAGDAHAGA-NTRIDWRETPEAHIFKA  
DLPGLRKEEVKIQVV  
Angiosperm 2 -----MDLDSPLFNTLHHIMDLTDDTTEKNLNAPTRTYVRDAKAMA-ATPADVKEHPNSYVFMVDMPGVKSGDIKVQVE  
Gymnosperm 2 -----MAMD-PSLITVQHLLGVPDD-LEKLLNAPTHSYMRDTKAMA-STPVDVKEYPNSYVFIIDMPGLKSNDIKVQVE  
Bryophyte 2 -----MEFVVFDTD-PFLTSLHQHVHEPESDLERKIKRKRRSQHDEPRHVTIATPVDVKEKKDAYLFIA  
DVPGLQKTDIEVQIE

Cytosolic Class 2



Cytosolic Class 1

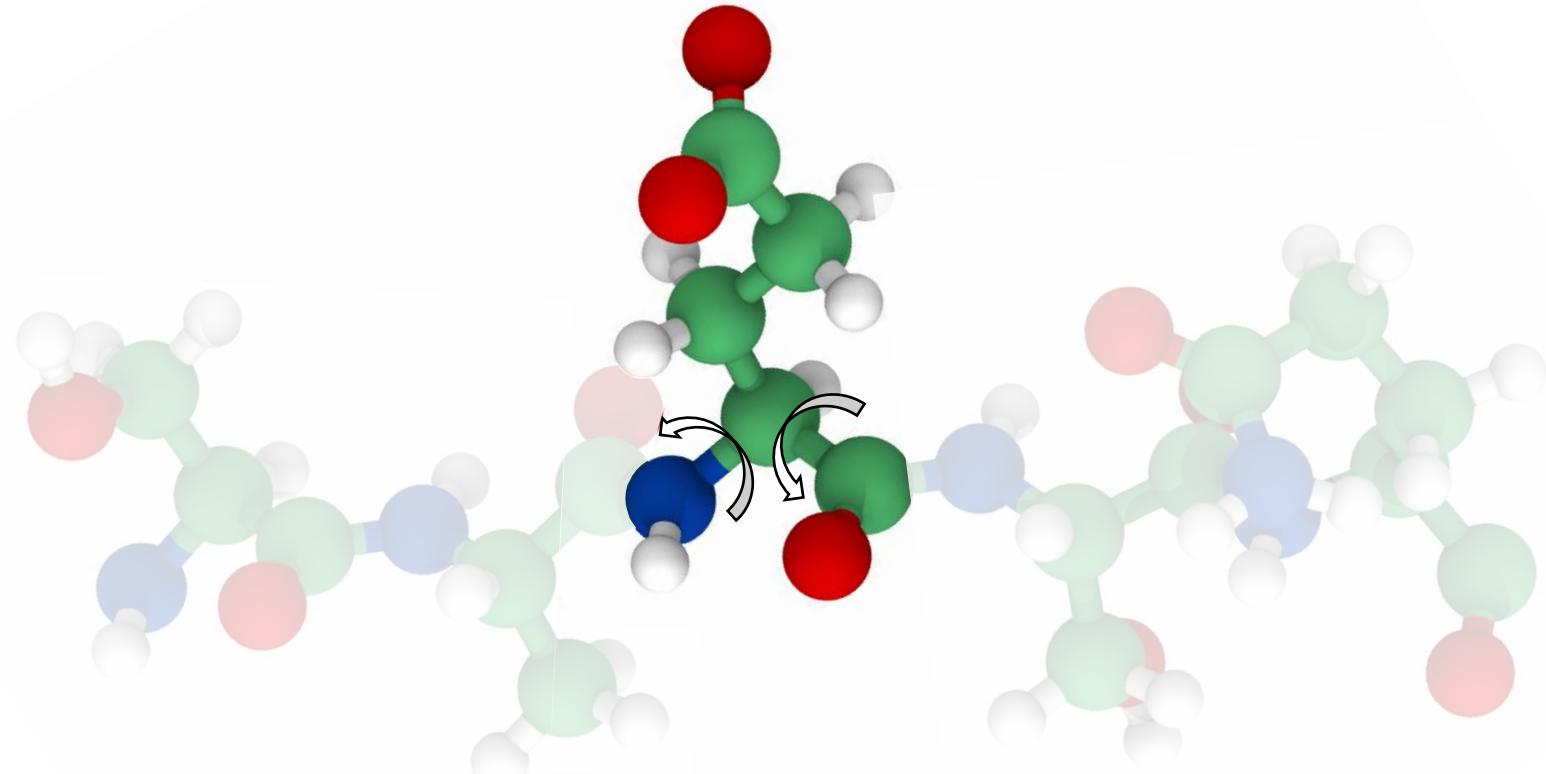


Proteins with similar sequence, likely:

- have similar functions in an organism
- are evolutionarily related

# Protein Secondary Structure

The amino acid chain path is determined by backbone torsional angles

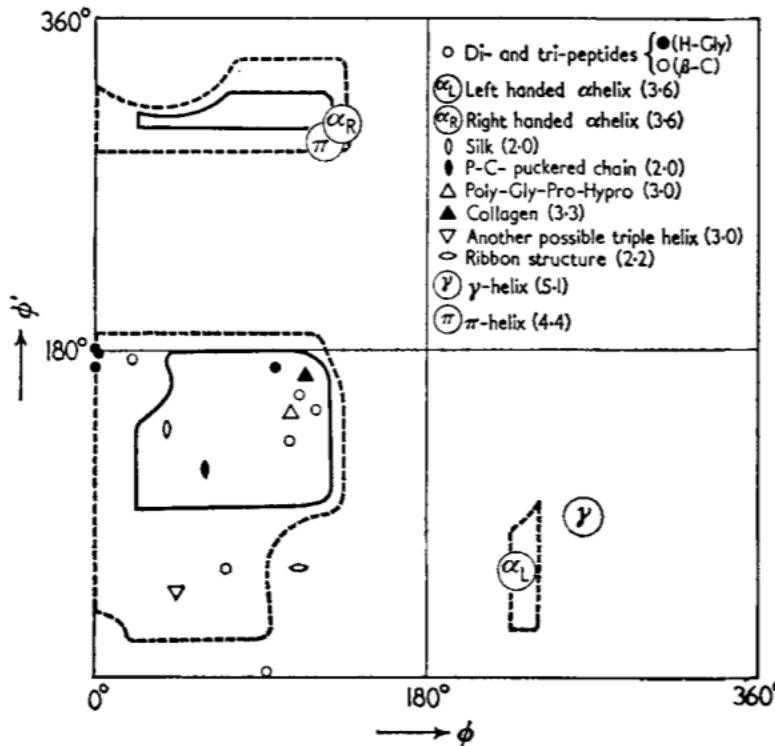
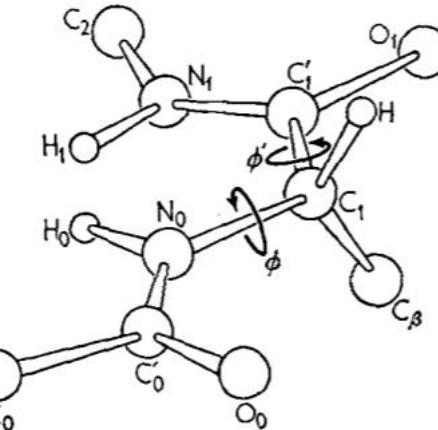


# Protein Secondary Structure

The amino acid chain path is determined by backbone torsional angles

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

G.N. Ramachandran, C. Ramakrishnan, V. Sasisekharan. *Stereochemistry of polypeptide chain configurations*. Journal of Molecular Biology, 1963



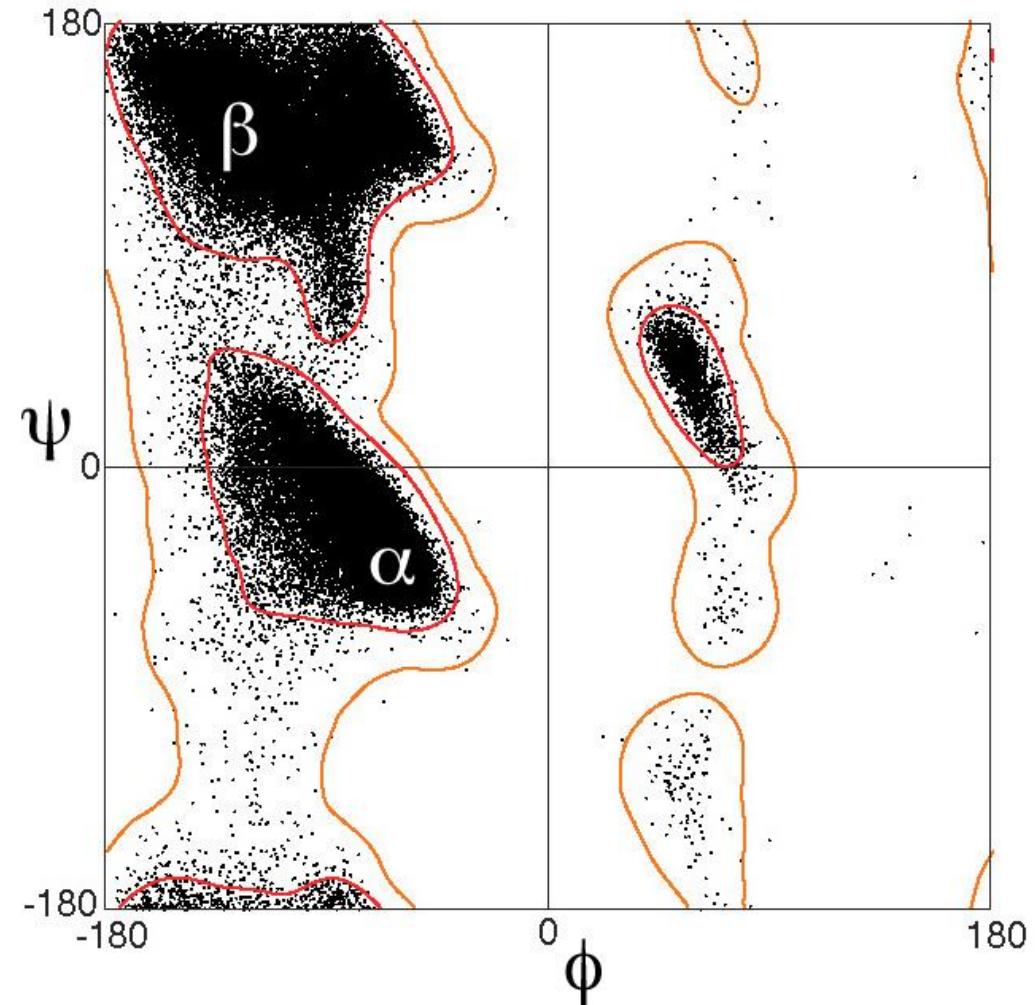
# Protein Secondary Structure

The amino acid chain path is determined by backbone torsional angles

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

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



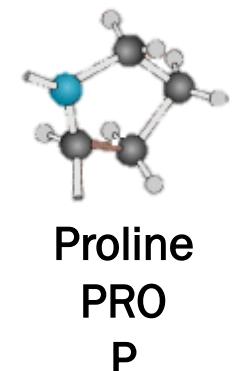
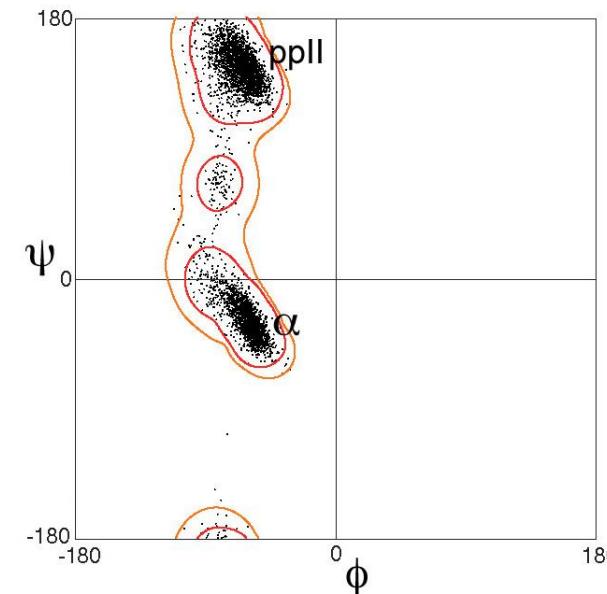
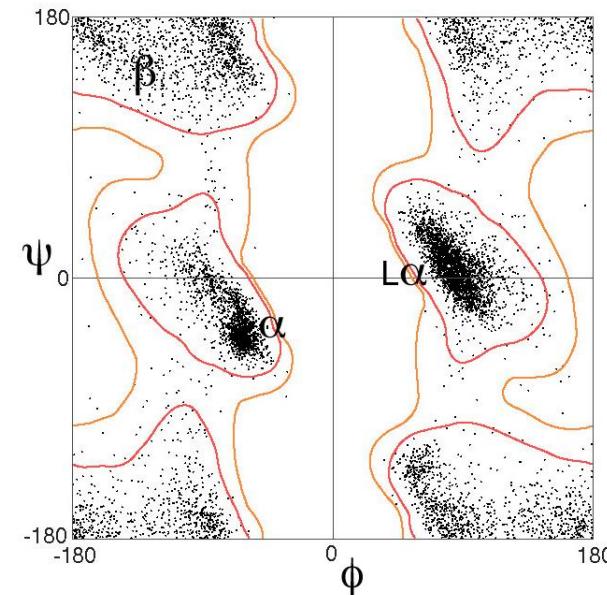
# [Extra] Protein Secondary Structure

The amino acid chain path is determined by backbone torsional angles

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

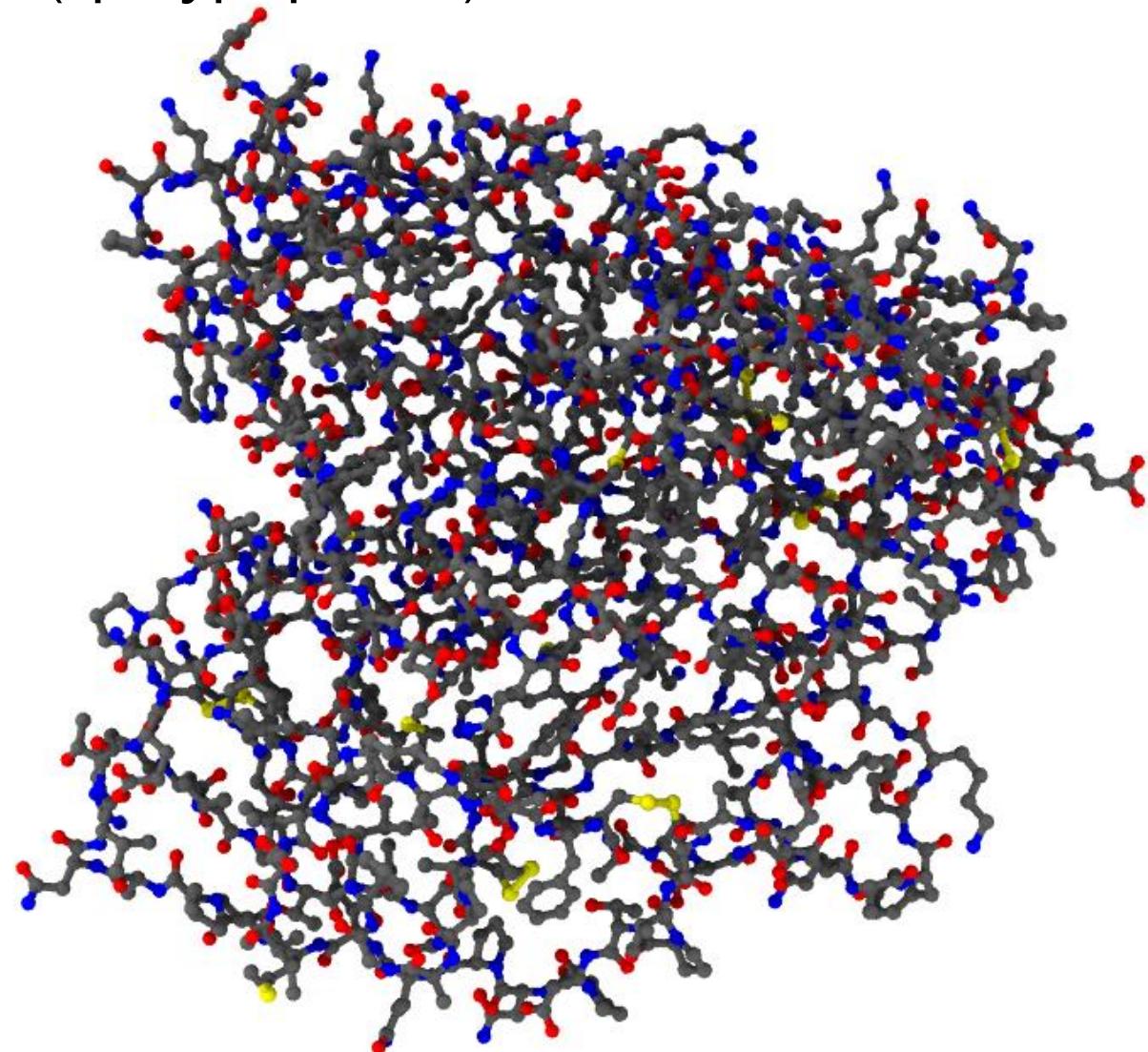
DSSP classification defines 7 secondary structure elements (regions in the plot):

- 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



# 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

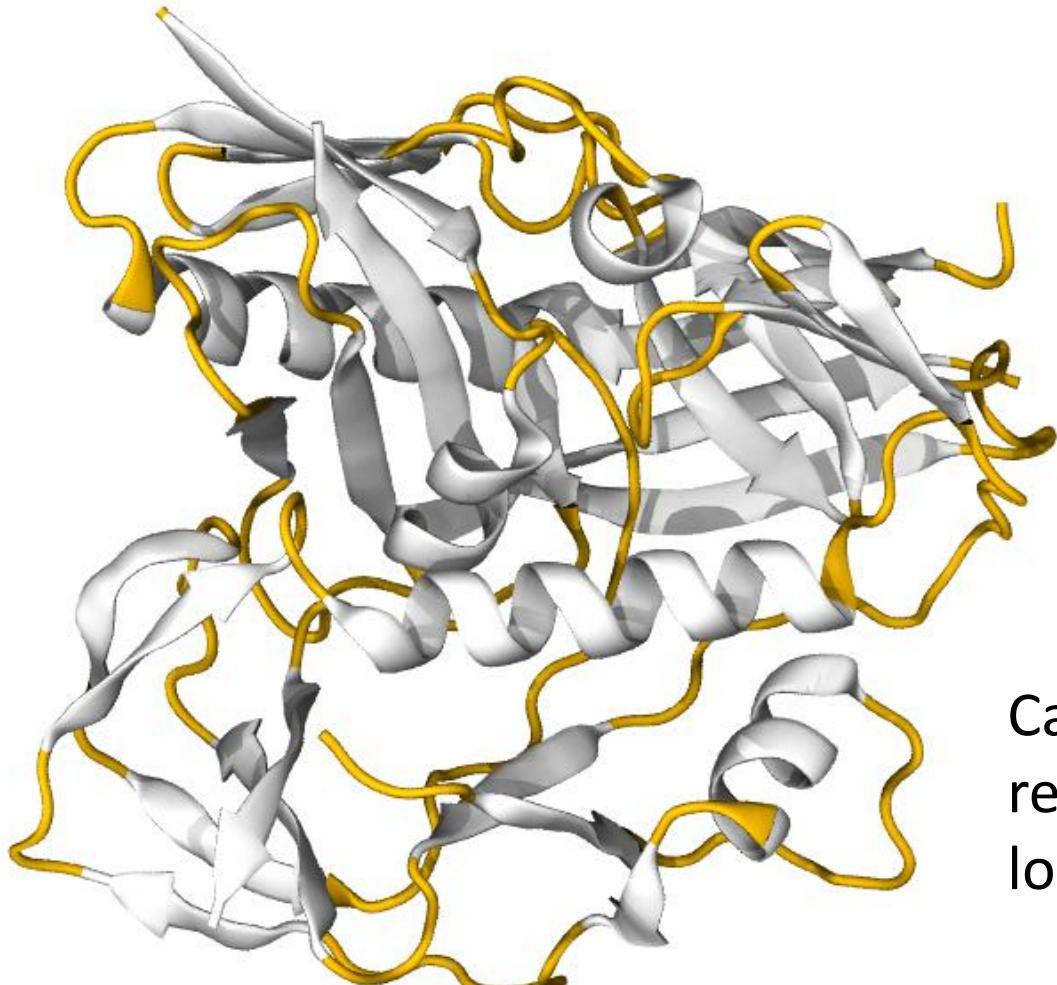
Protein («polypeptide»): 10 to >1000 amino acids



Cartoon  
representation:  
sheets

# Protein tertiary structure: folding

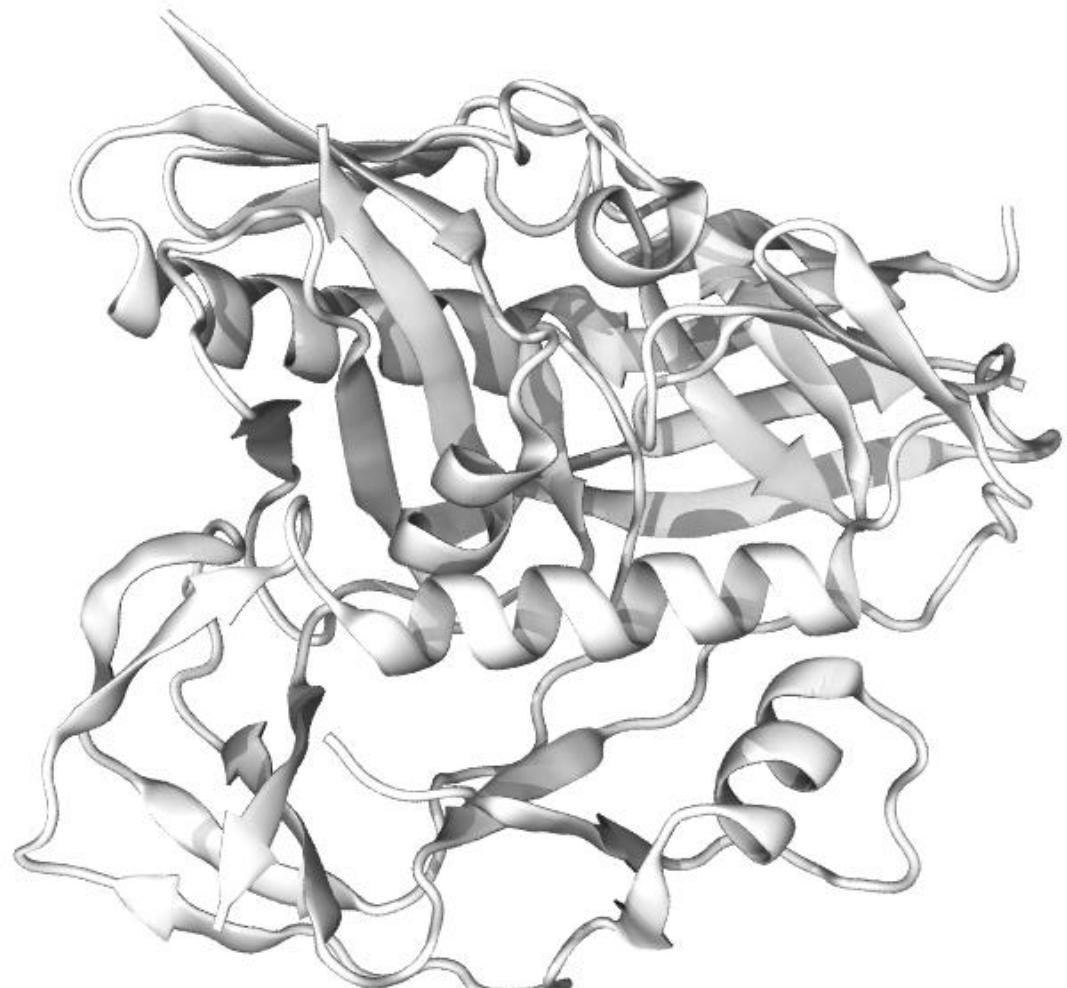
Protein («polypeptide»): 10 to >1000 amino acids



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:

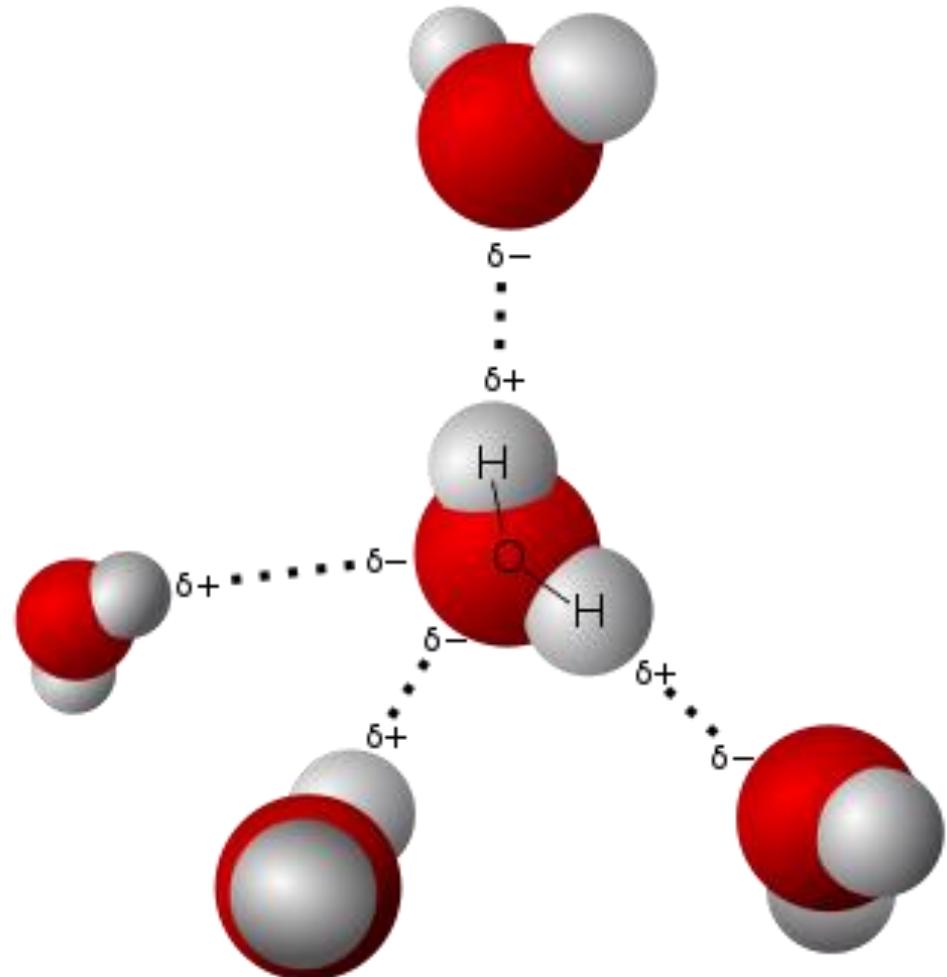
- donor-acceptor distance typically 1.6-2 Å
- donor-acceptor-hydrogen angle must be small

Hydrogen bond energy in biomolecular

systems typically 5-25 kJ/mol, e.g.:

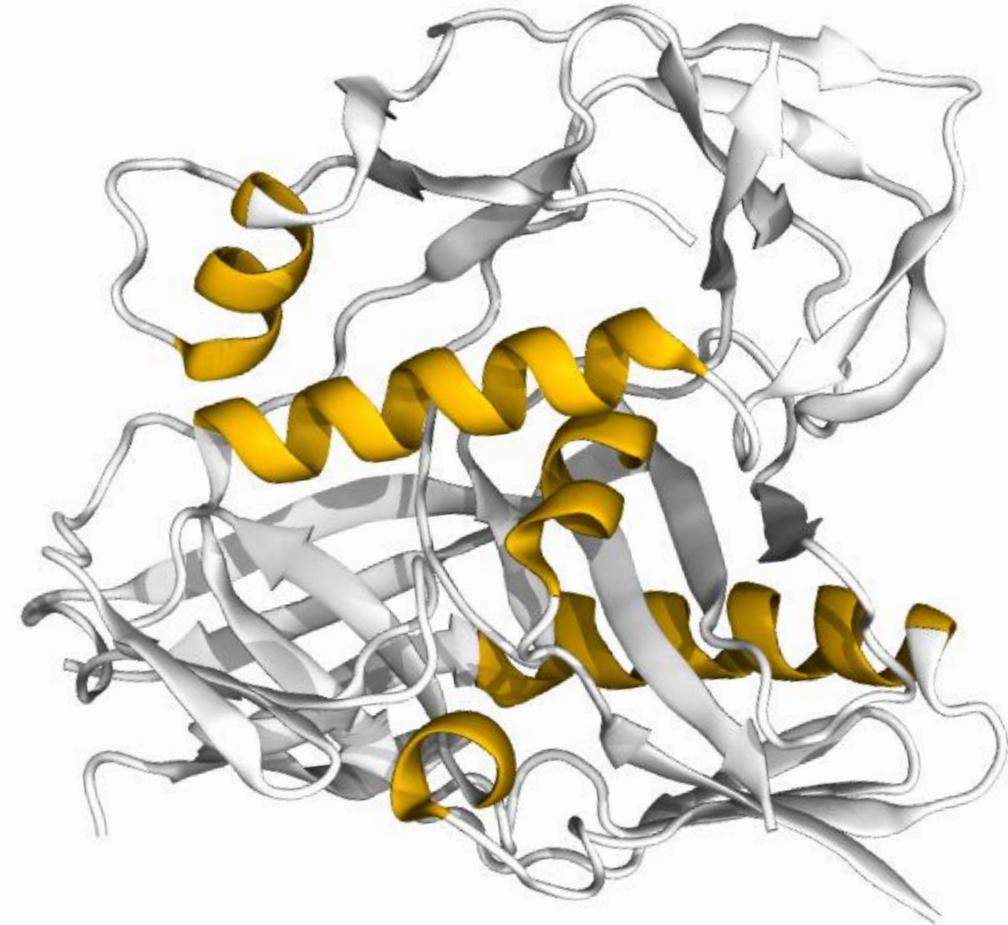
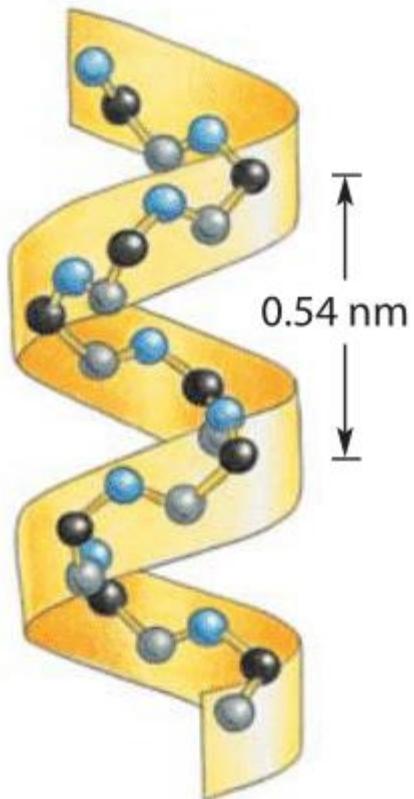
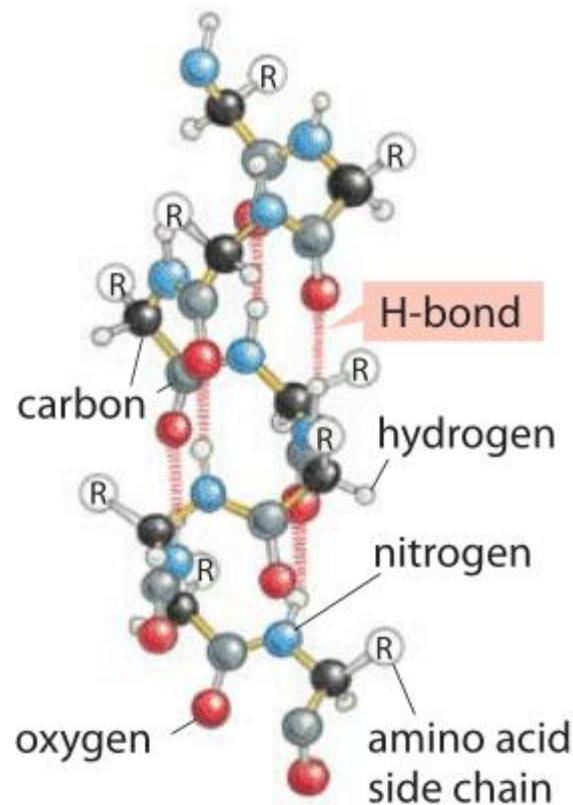
- O-H $\cdots$ O , 21 kJ/mol (5.0 kcal/mol)
- N-H $\cdots$ O , 8 kJ/mol (1.9 kcal/mol)

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



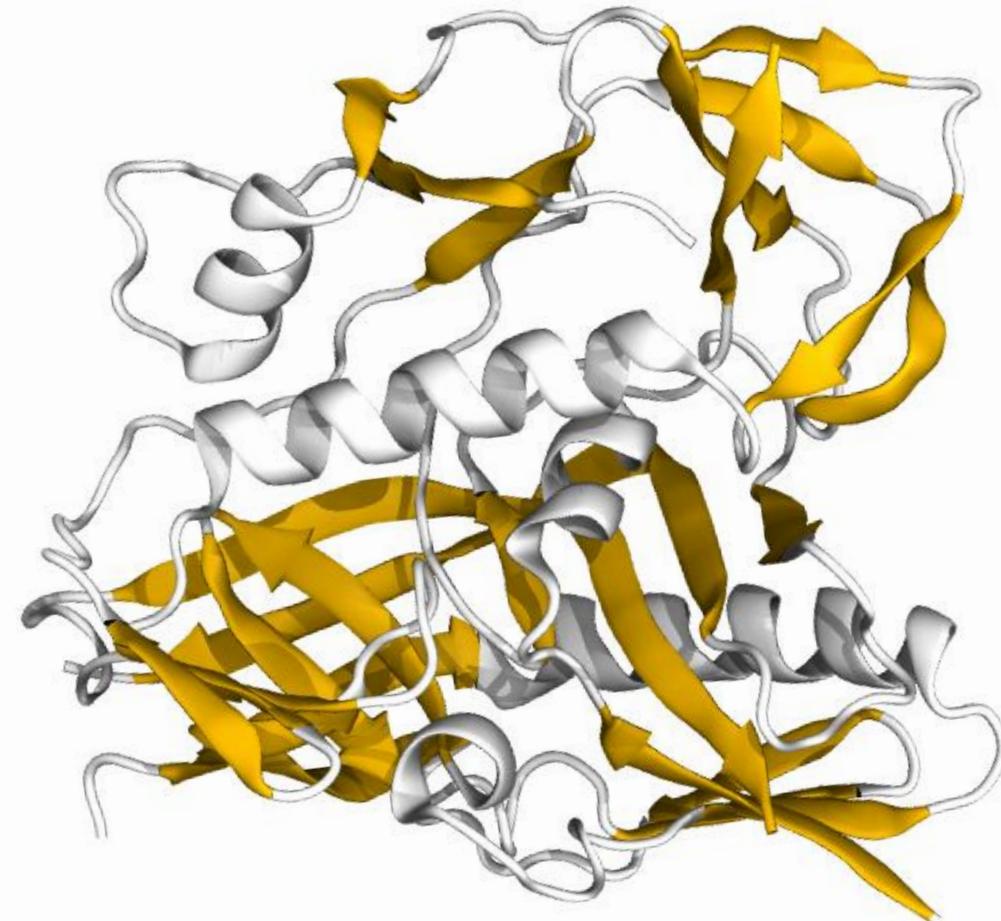
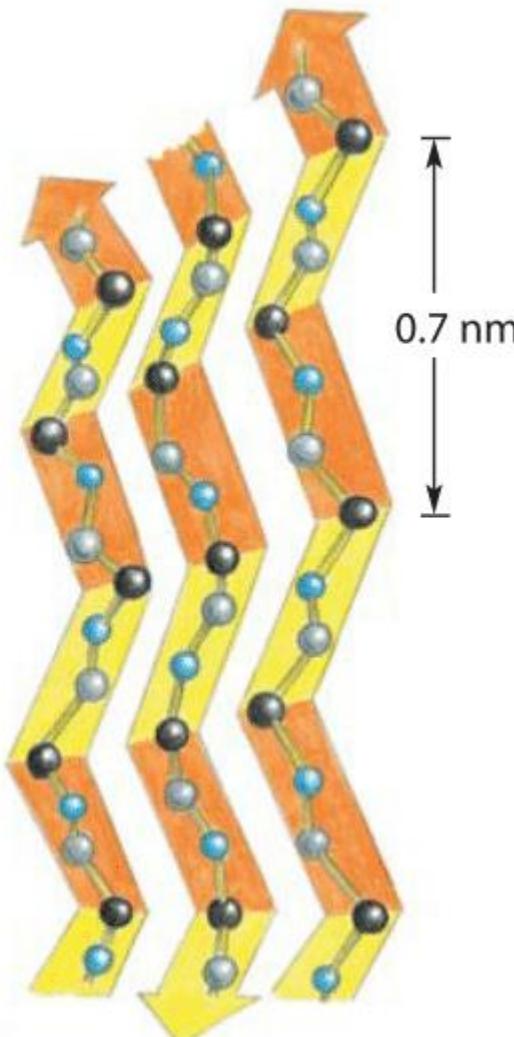
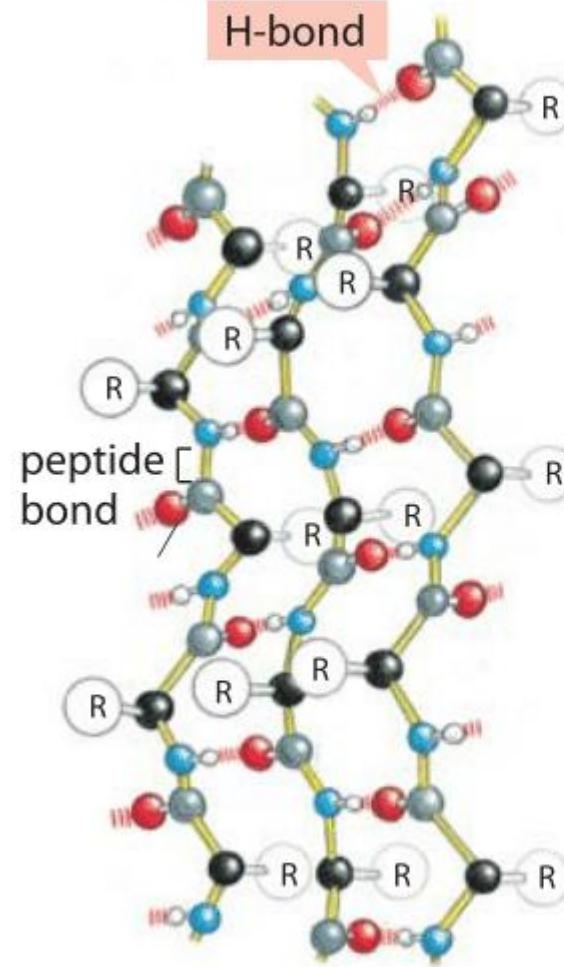
# Hydrogen bonding on protein backbone

alpha helix

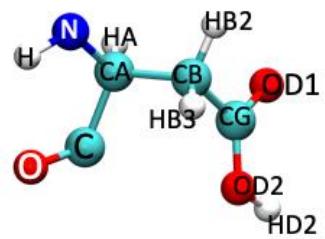


# Hydrogen bonding on protein backbone

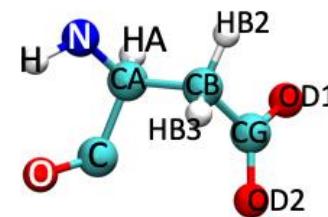
beta sheet



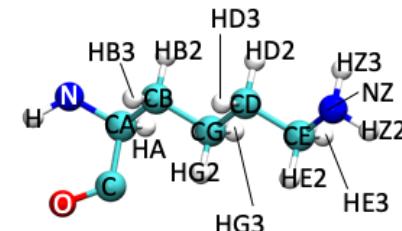
# Hydrogen bonding: effect of *local* pH



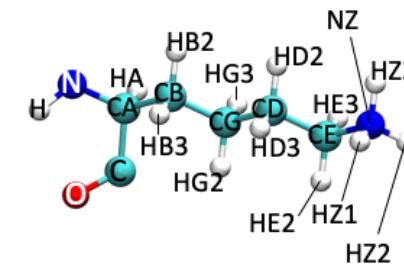
ASH, Aspartic acid\*, D



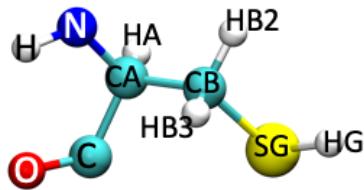
ASP, Aspartate, D



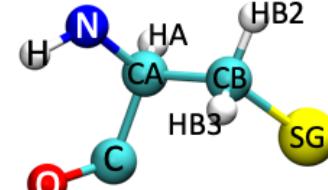
LYN, Lysine\*, K



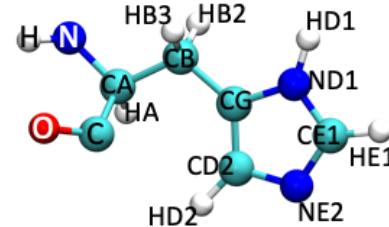
LYS, Lysine, K



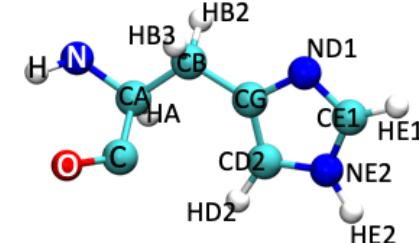
CYS, Cysteine, C



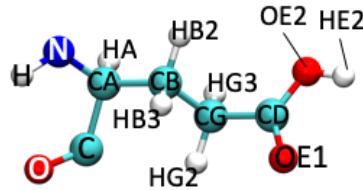
CYX, Cysteine\*\*, C



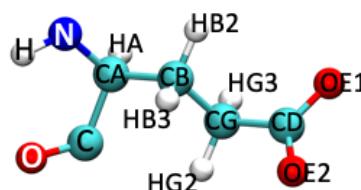
Histidine\*, H



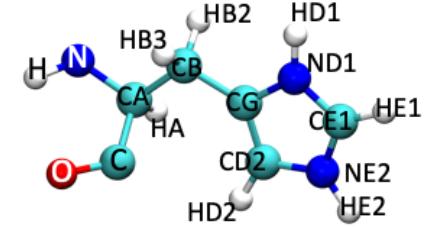
HIE, Histidine, H



GLH, Glutamic Acid\*, E



GLU, Glutamate, E

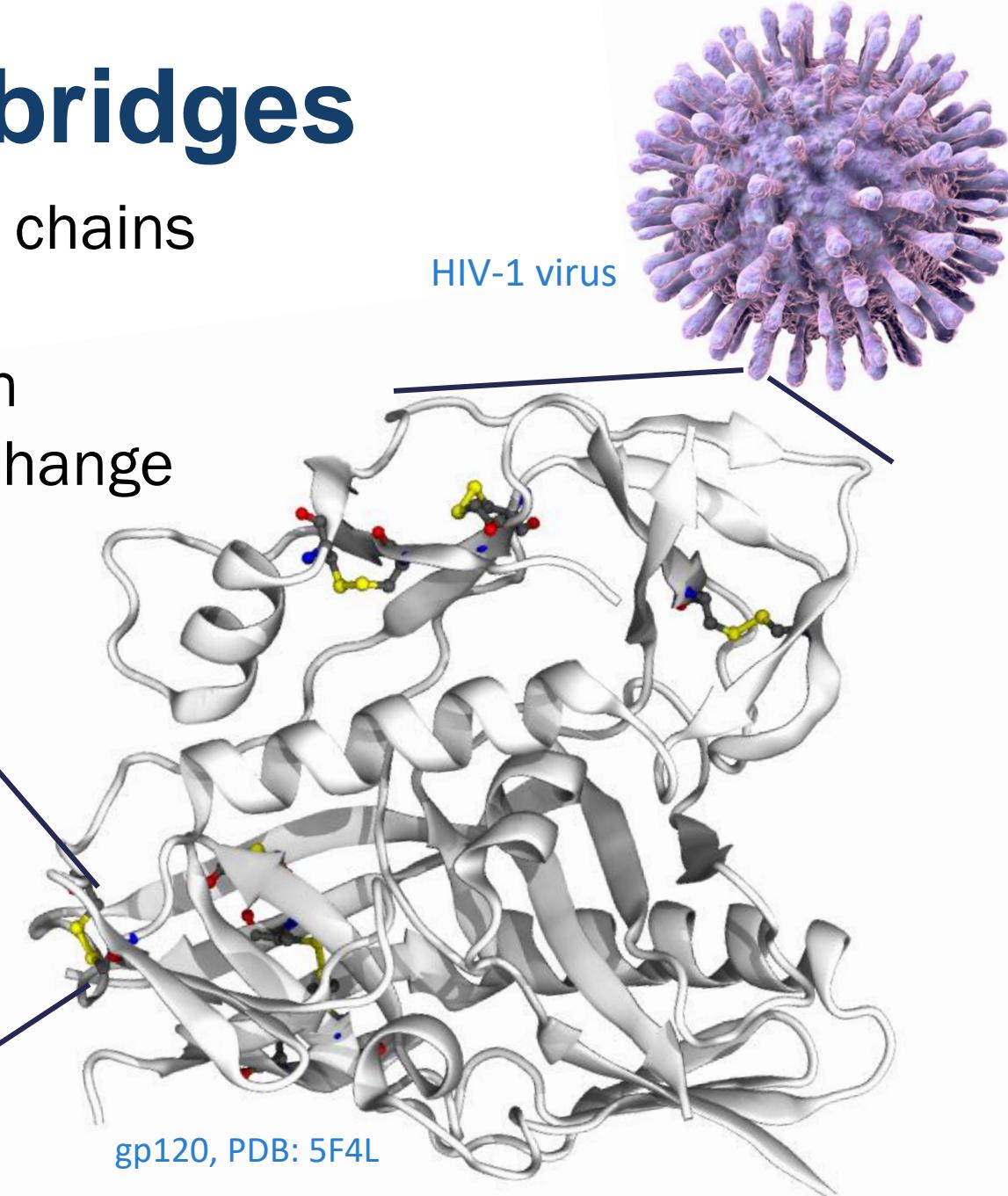
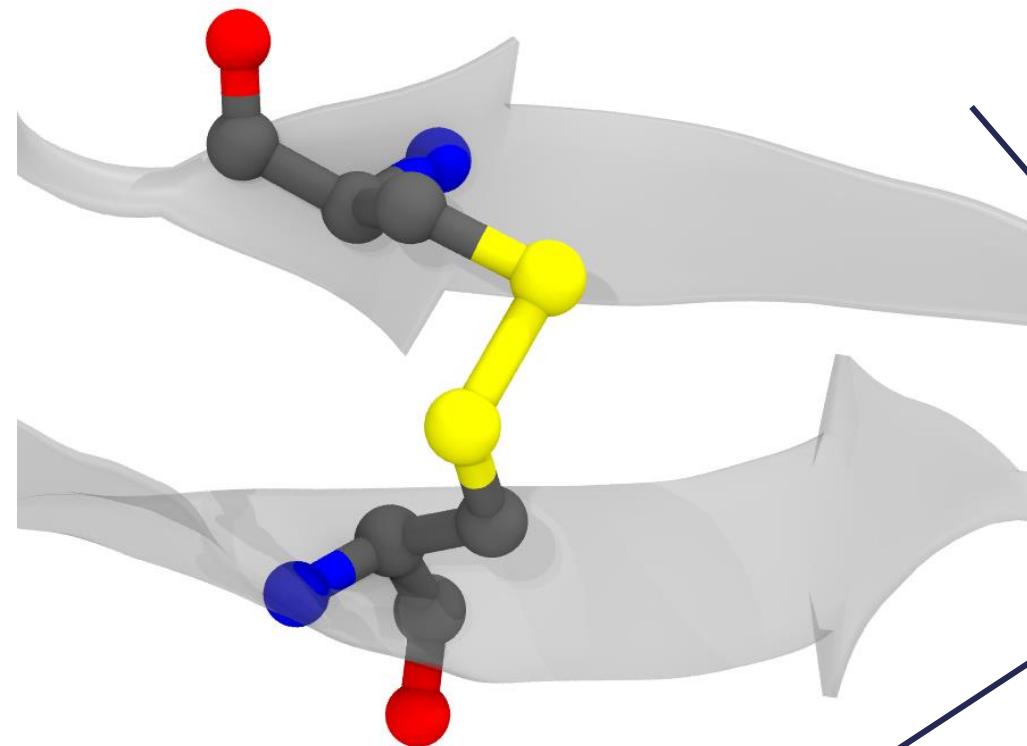


HIP, Histidine\*, H

# Disulfide bridges

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

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



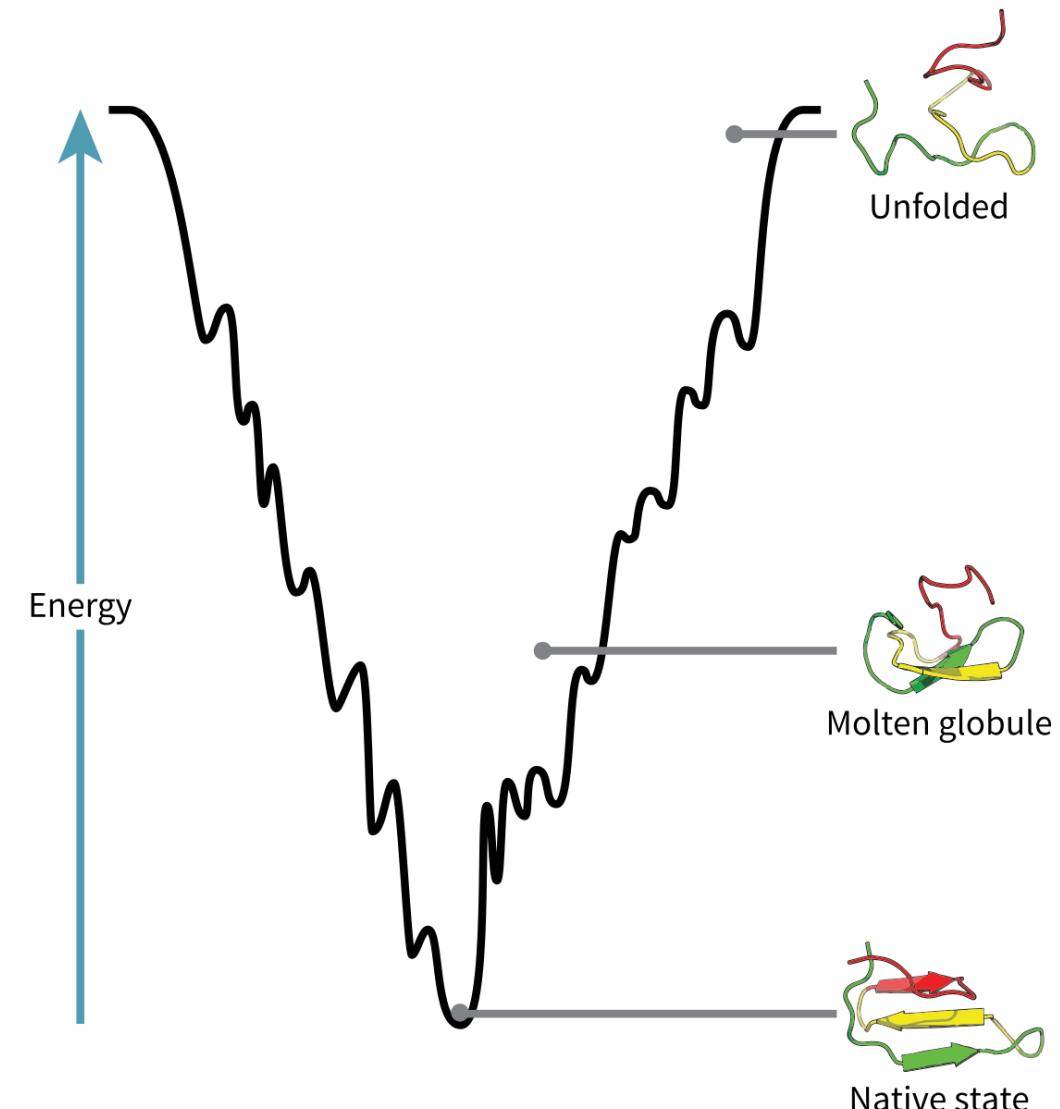
# [Extra] Proteins fold in low energy structures

$$\Delta G = \Delta H - T\Delta S$$

Proteins fold spontaneously, i.e.  $\Delta G < 0$

$\Delta S_{\text{protein}}$  is  $< 0$

$\Delta S = \Delta S_{\text{protein}} + \Delta S_{\text{water}} > 0$



# [Extra] Proteins fold in low energy structures

$$\Delta G = \Delta H - T\Delta S$$

Proteins fold spontaneously, i.e.  $\Delta G < 0$

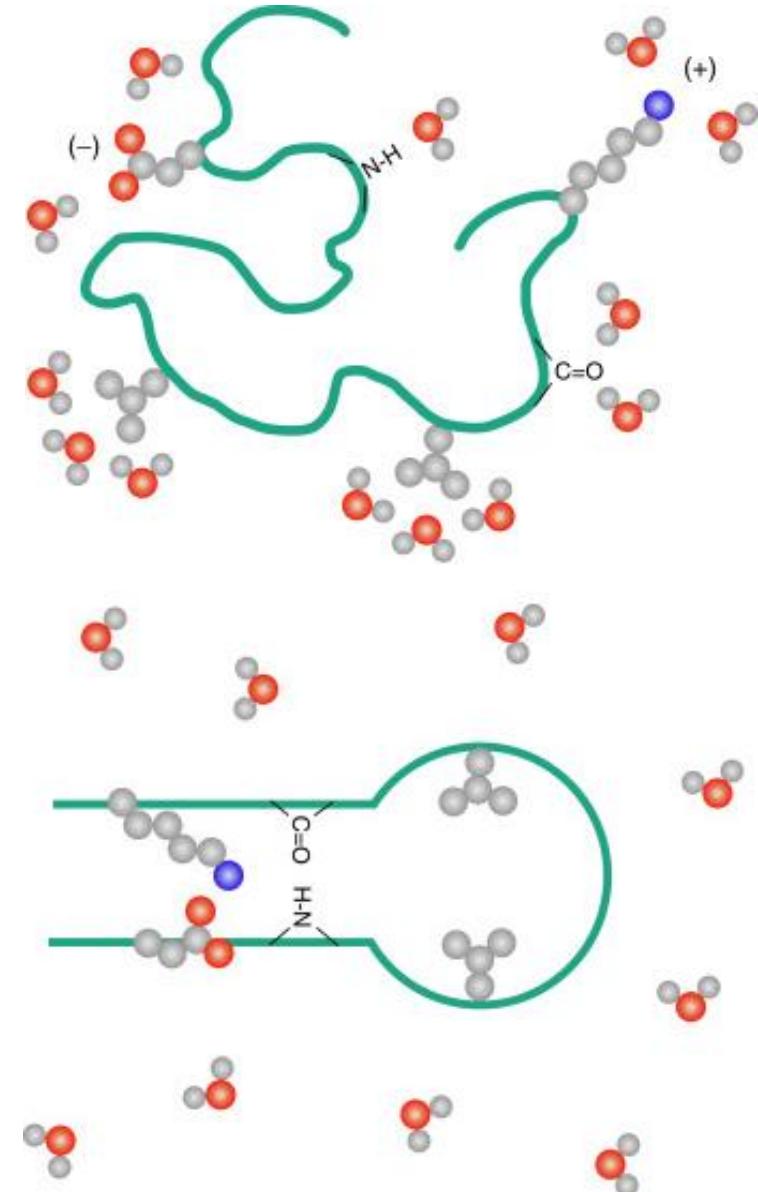
$\Delta S_{\text{protein}}$  is  $< 0$

$$\Delta S = \Delta S_{\text{protein}} + \Delta S_{\text{water}} > 0$$

$H = U + PV$ ; at physiological conditions  $\Delta H \approx \Delta U$

**$\Delta H$  is small:** a folded protein forms bonds with itself, an unfolded one forms bonds with water.

Hydrophobic collapse drives folding!



# [Extra] Levinthal's paradox

How can proteins find their unique fold in sub-second timescales?

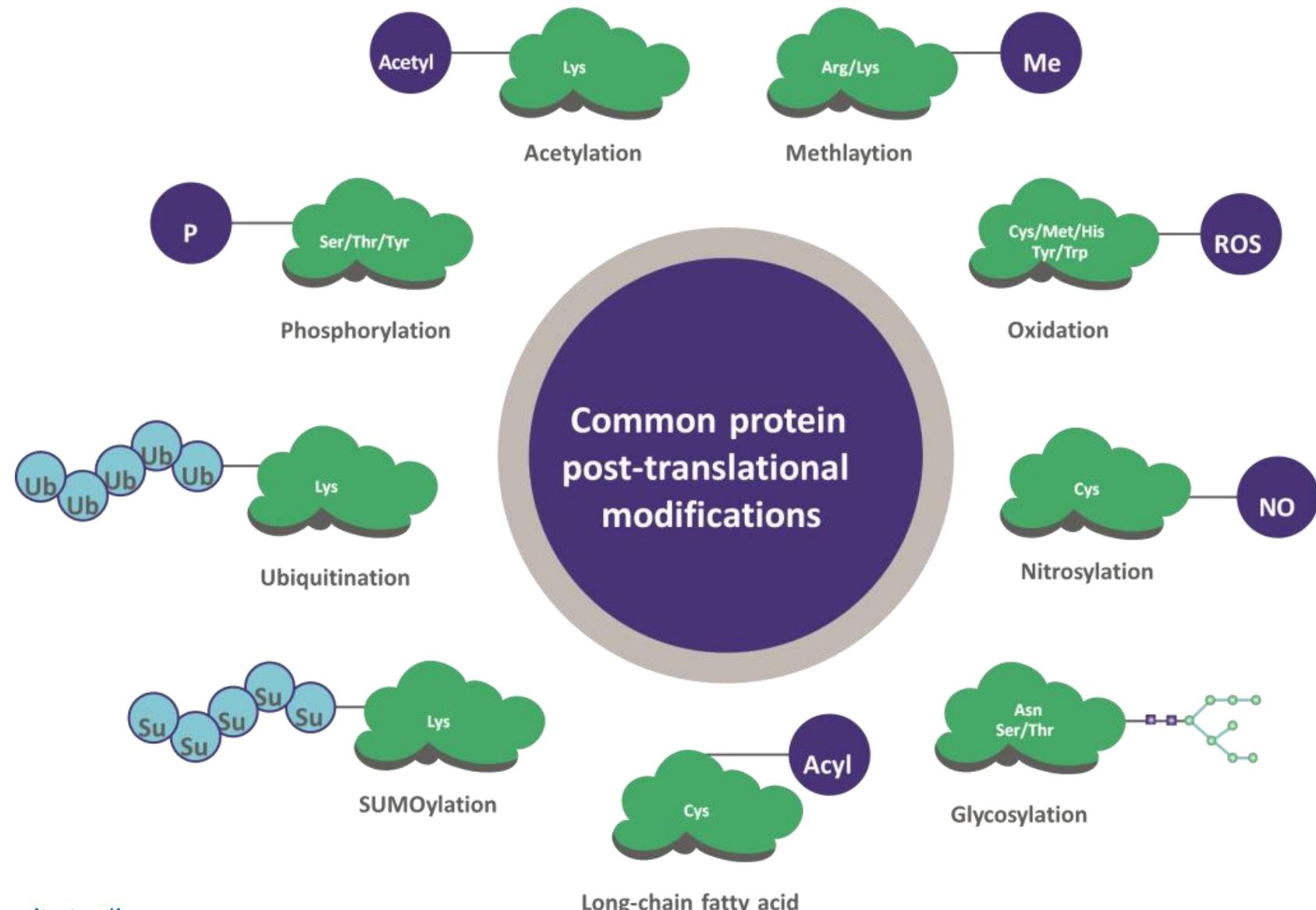
- 100 amino acids-long protein
- $2 \times 100 - 2 = 198$  backbone torsional angles
- each torsional angle has on average 3 stable positions
- $3^{198}$  different configurations
- side chain rotation in the fs timescale ( $10^{-12}$  s)
- random search worst case scenario:  $10^{-12} \times 3^{198} \approx 10^{-12} \times 10^{99} = 10^{87}$  s
- age of the universe  $\approx 13.7$  billion years =  $4.32 \times 10^{17}$  s

**Protein folding cannot be a random process**

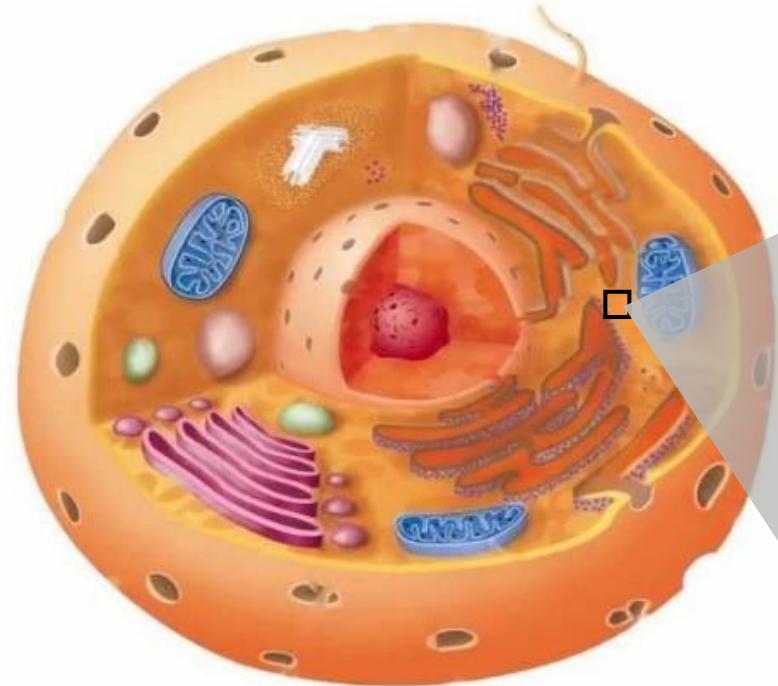
# Post-translational modifications (PTMs)

Once expressed, proteins can be modified by the covalent binding of other molecules.

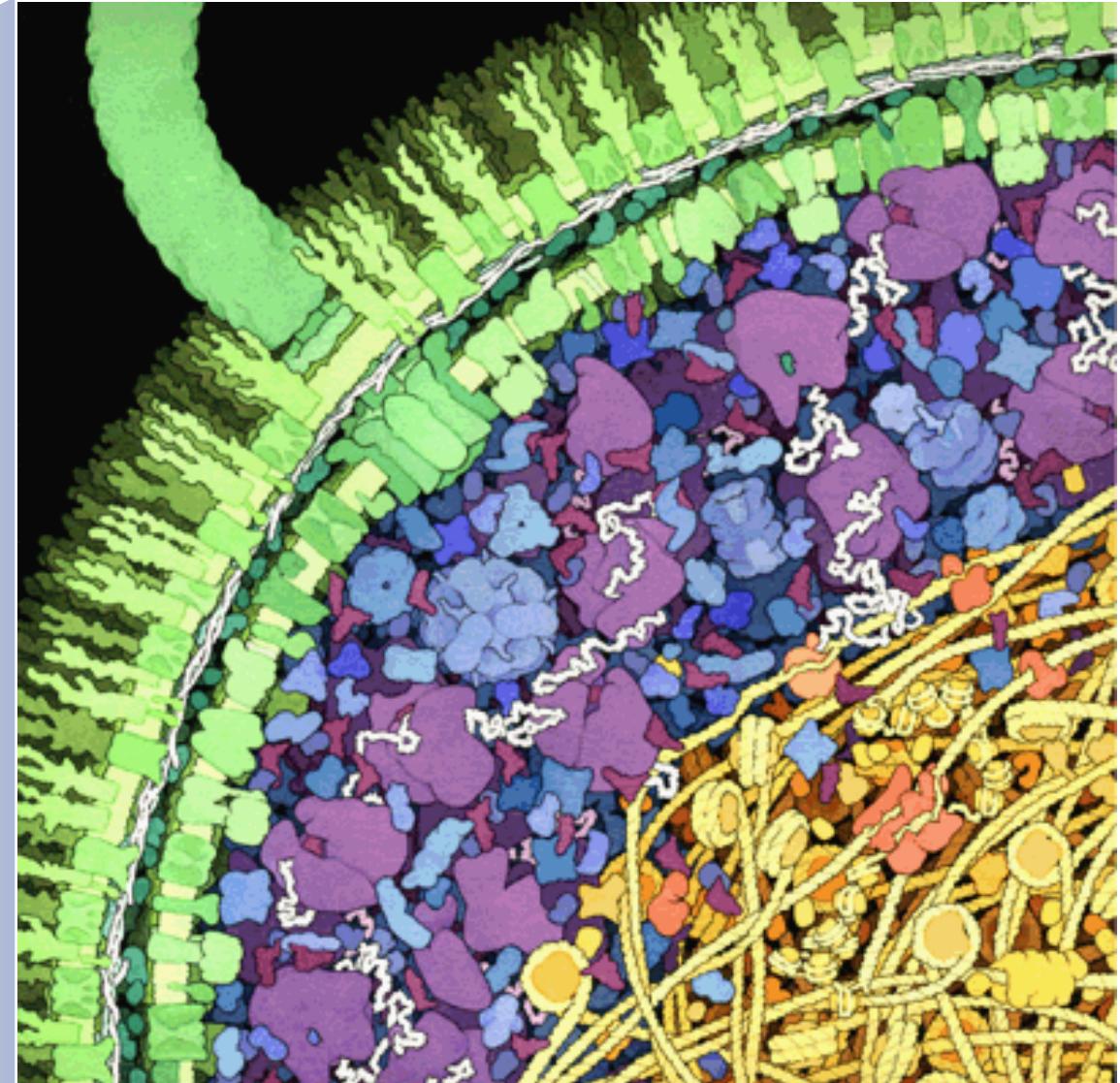
The addition/removal of most PTMs is controlled by the organism, to modulate protein function.



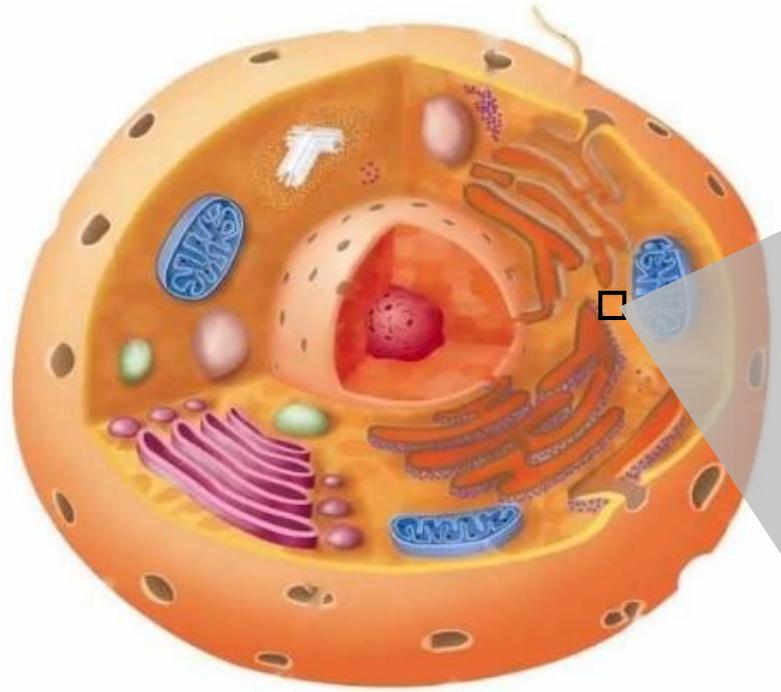
# The intracellular space



Crowded environment!



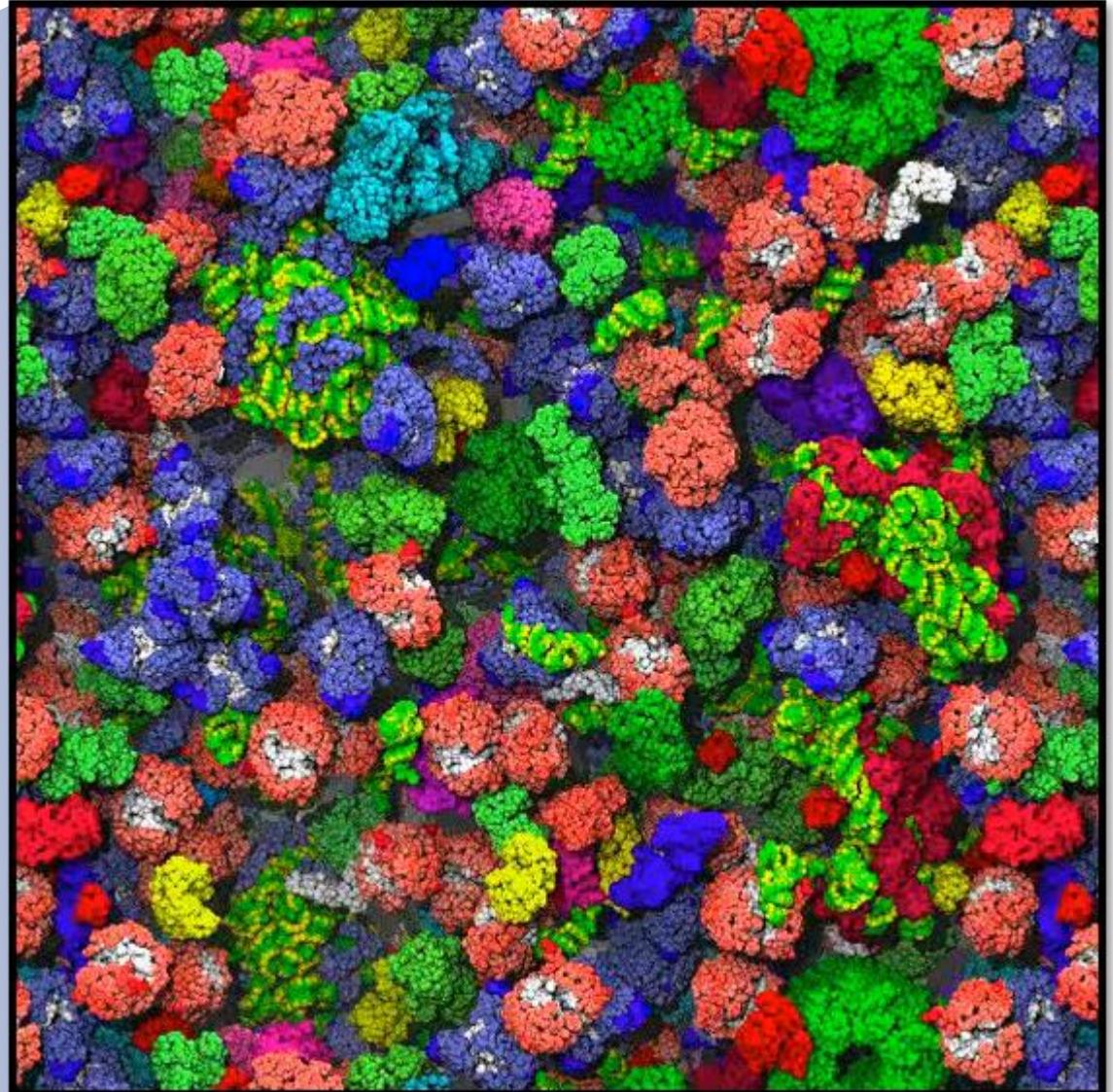
# The intracellular space



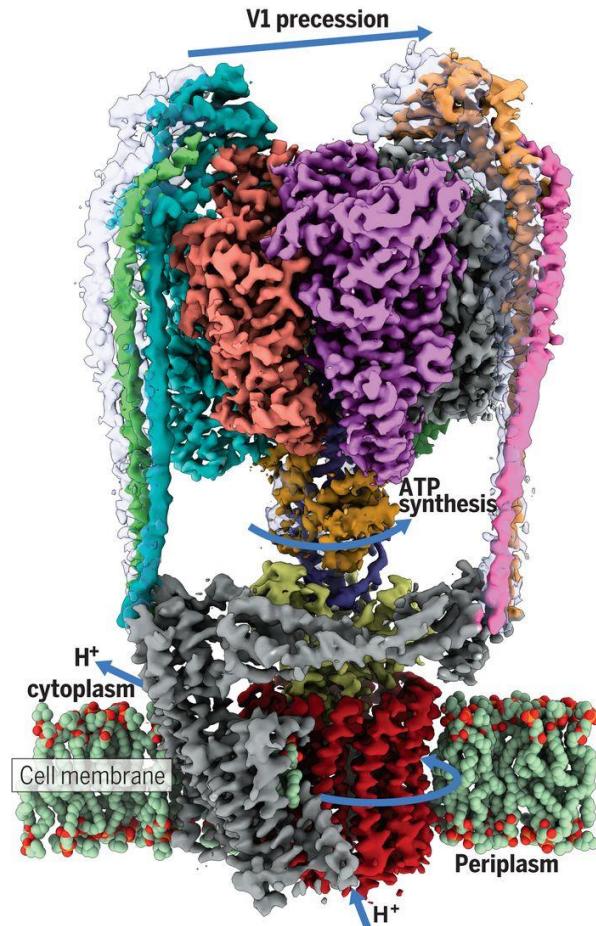
Crowded environment!

Brownian motion: proteins bump into other molecules all the time!

Most contacts are short-lived

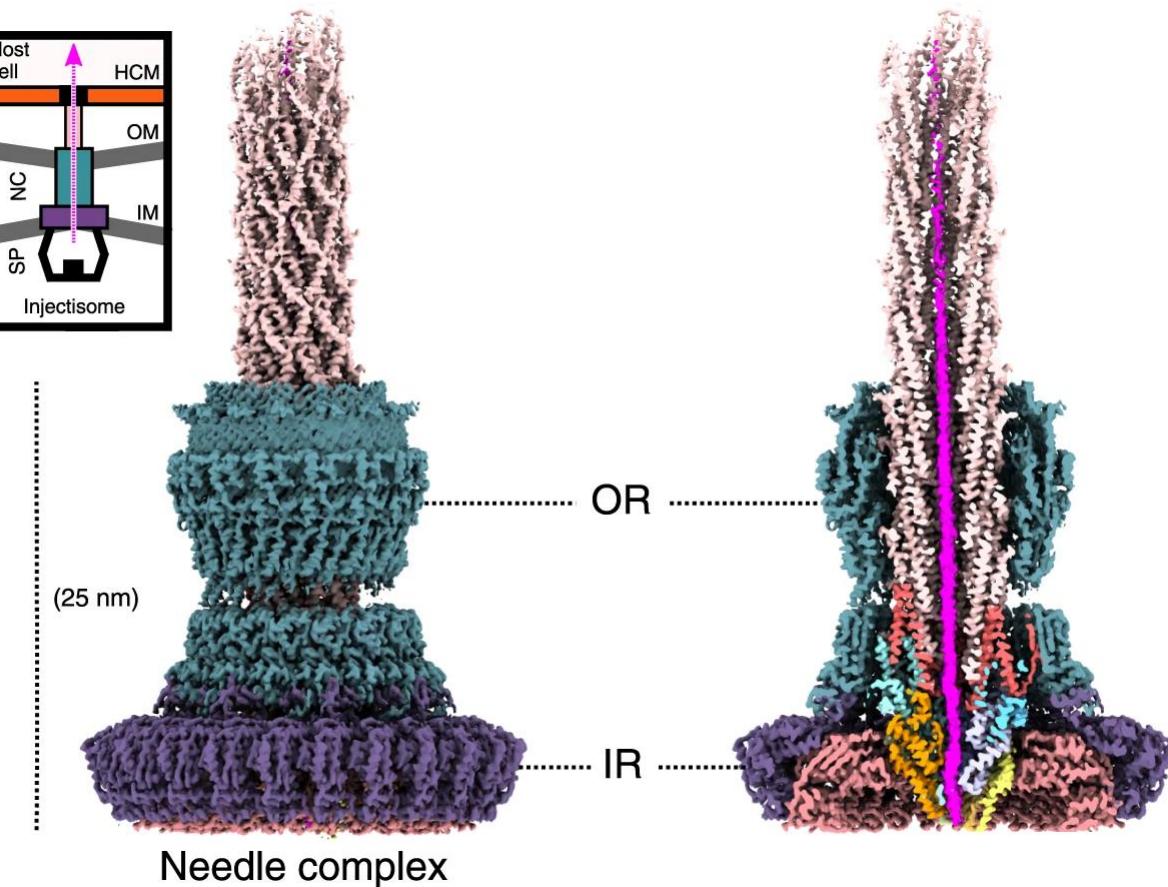
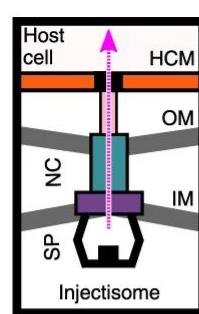


# Protein quaternary structure: assembly



ATP syntase

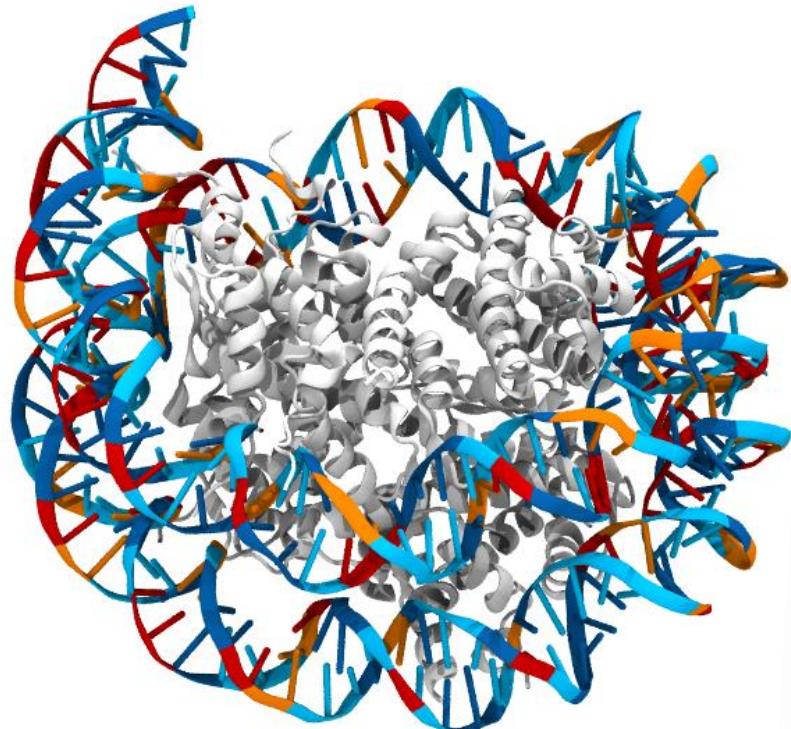
L. Zhou and L. A. Sazanov, *Structure and conformational plasticity of the intact Thermus thermophilus V/A-type ATPase*, *Science*, 2019



Injectisome

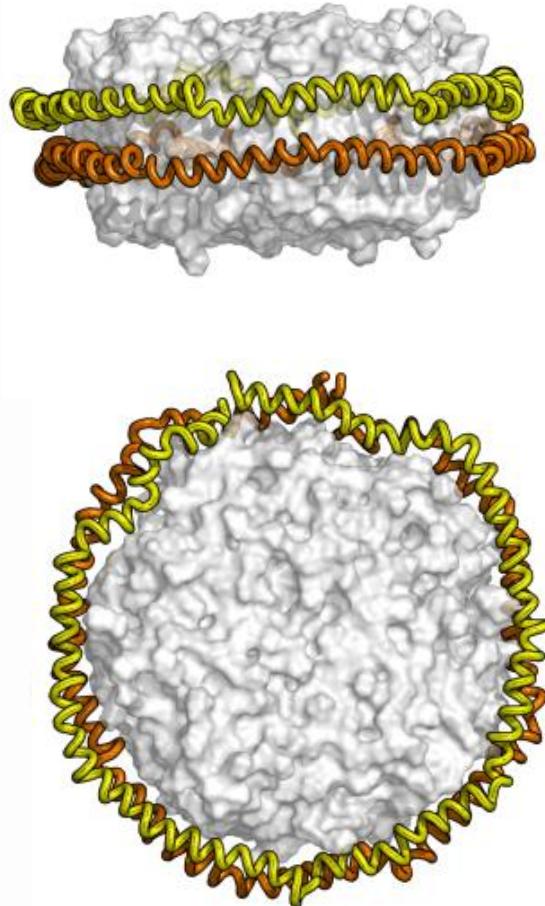
S. Miletic et al., *Substrate-engaged type III secretion system structures reveal gating mechanism for unfolded protein translocation*, *Nature Comms*, 2021

# The structure determines the function



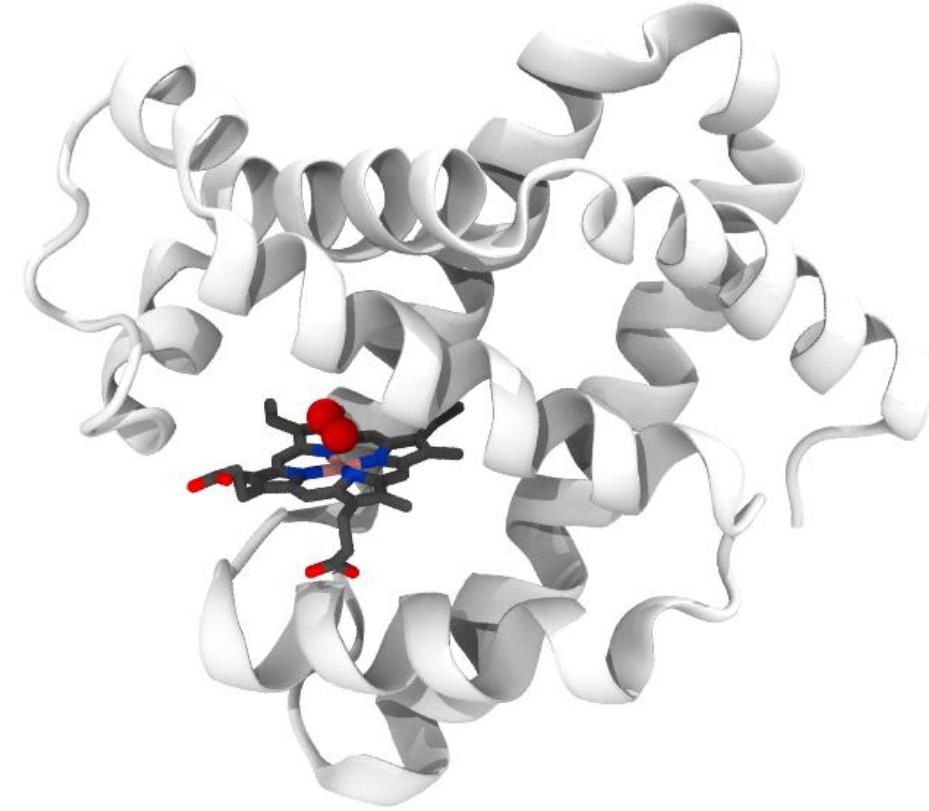
Nucleosome (PDB: 5CPI)

K. Luger et al., *Crystal Structure of the nucleosome core particle at 2.8 Å resolution*, Nature, 1997



Lipoprotein (PDB: 1AV1)

D.W. Bohrani et al., *Crystal structure of truncated human apolipoprotein A-I suggests a lipid-bound conformation*, PNAS, 1997



Myoglobin (PDB: 1MBO)

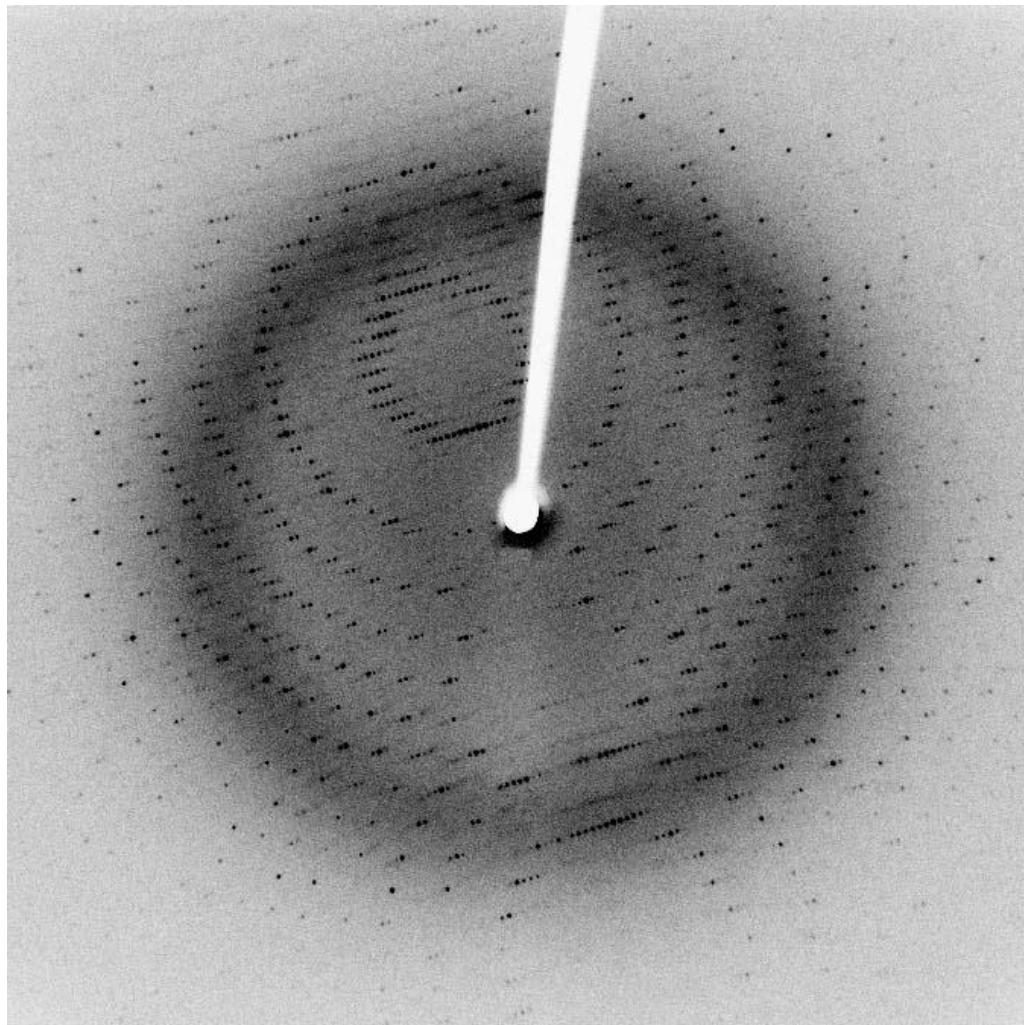
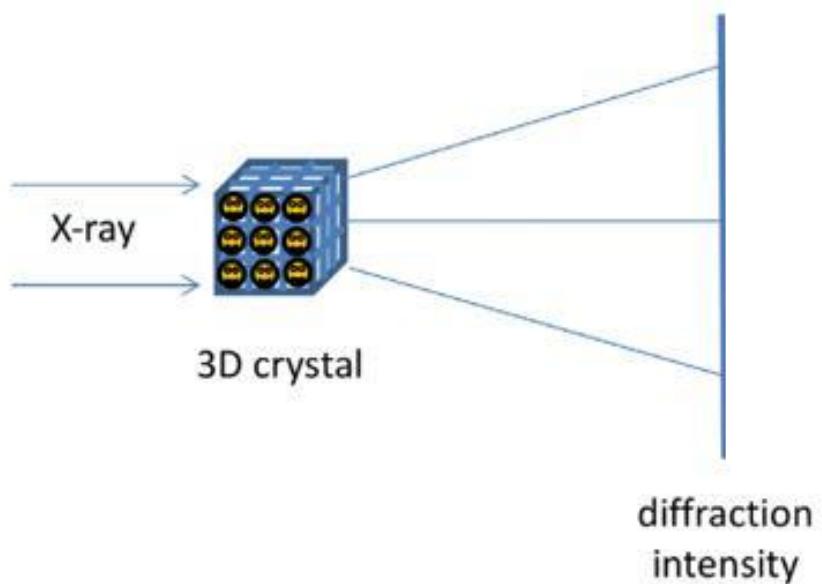
J.C. Kendrew et al., *A three-dimensional Model of the Myoglobin Molecule obtained by X-Ray Analysis*, Nature, 1958

# Proteins, illness, and drug design

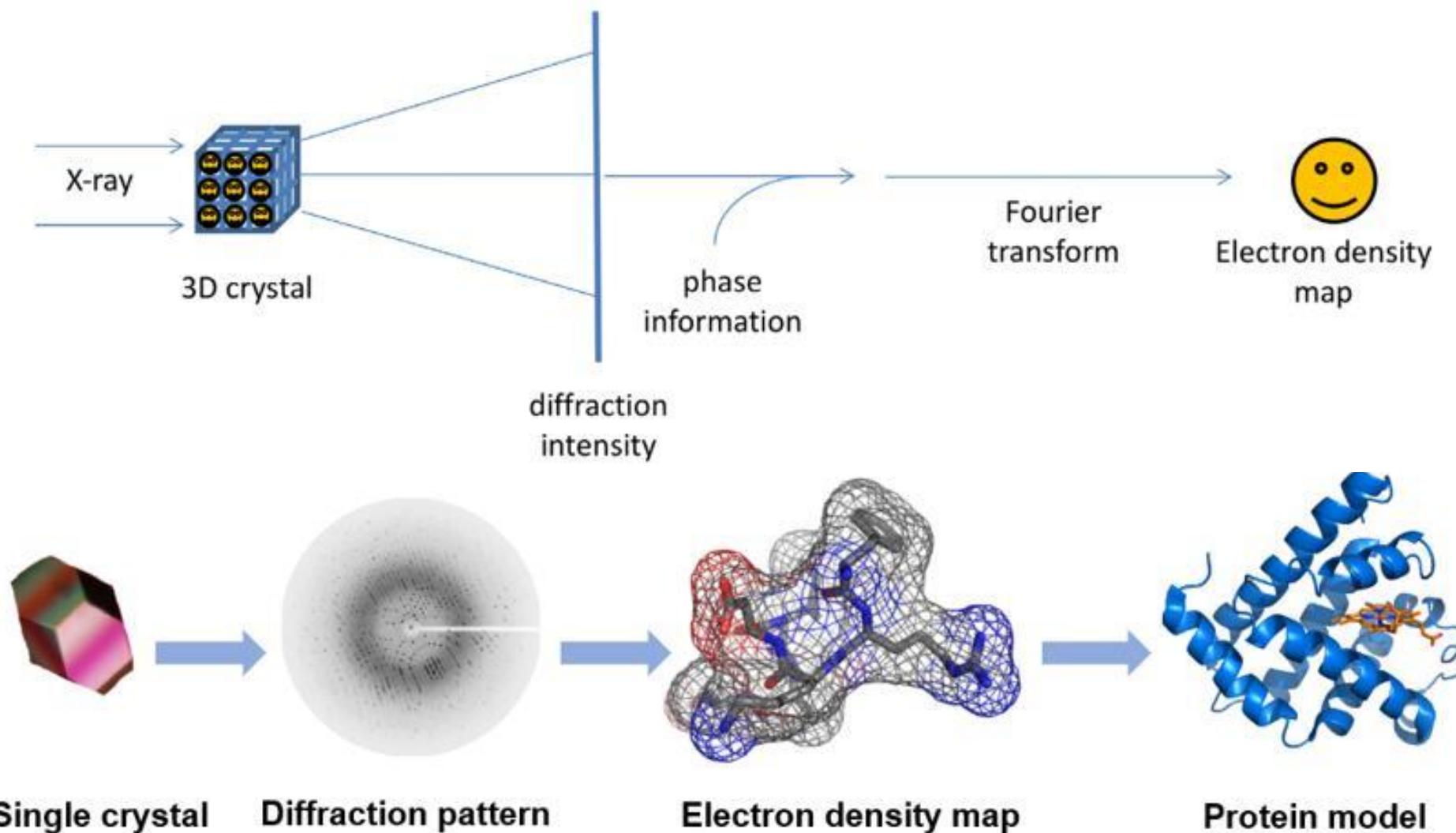
- **Proteins and diseases** (e.g. covid-19, salmonella, flu, ...)
  - pathogen's own metabolism/structure
  - pathogen's weapon
- **Proteins and disorders** (e.g. Cancer, Alzheimer, ...)
  - own protein misfolds
  - own protein folds, but has different dynamics
- **Drug**
  - Small molecule designed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

## **Part 2: structure determination**

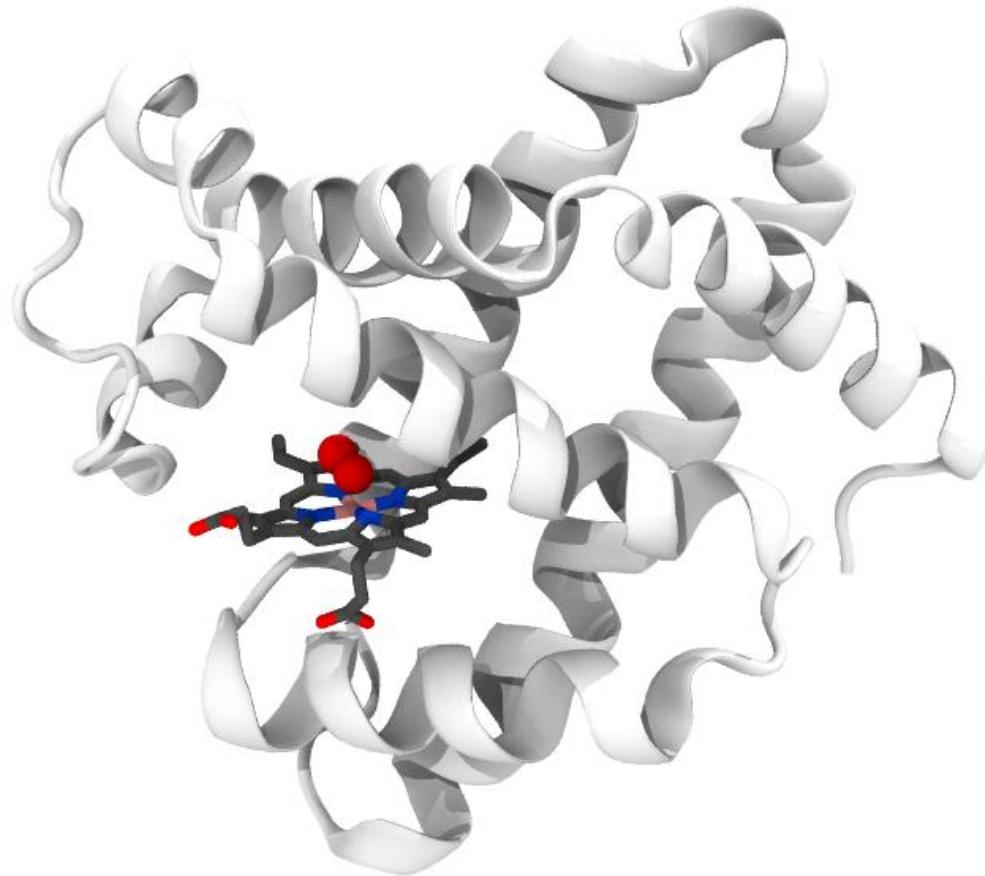
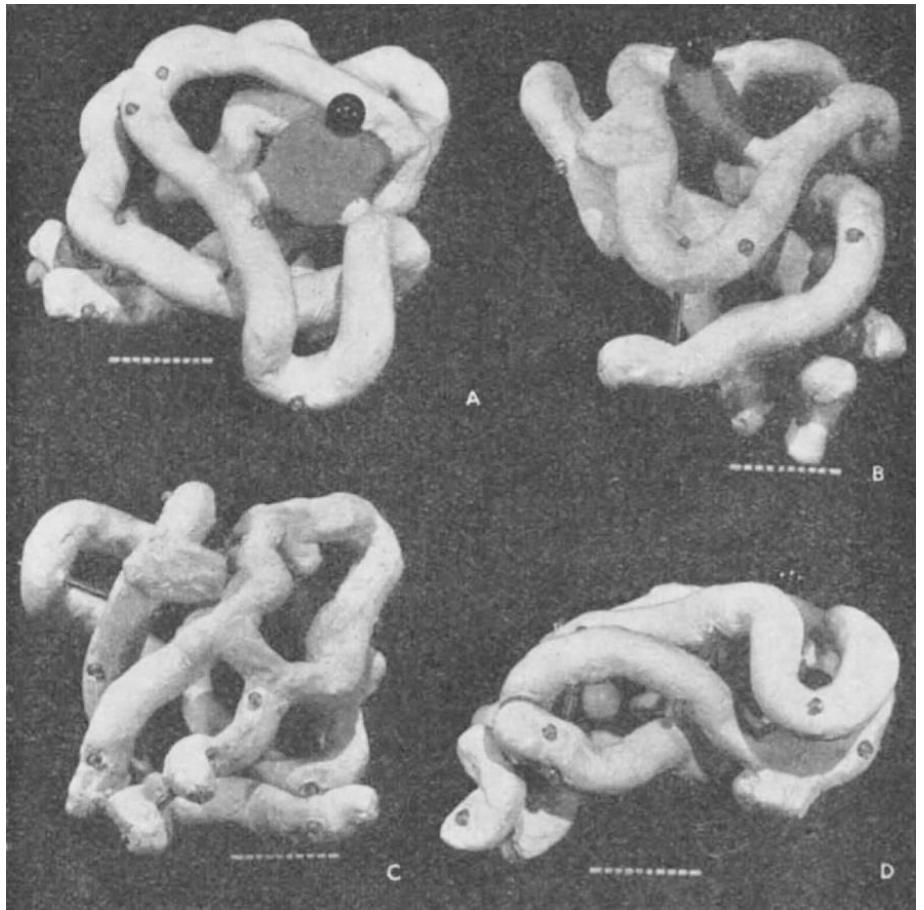
# Structure determination: X-ray crystallography



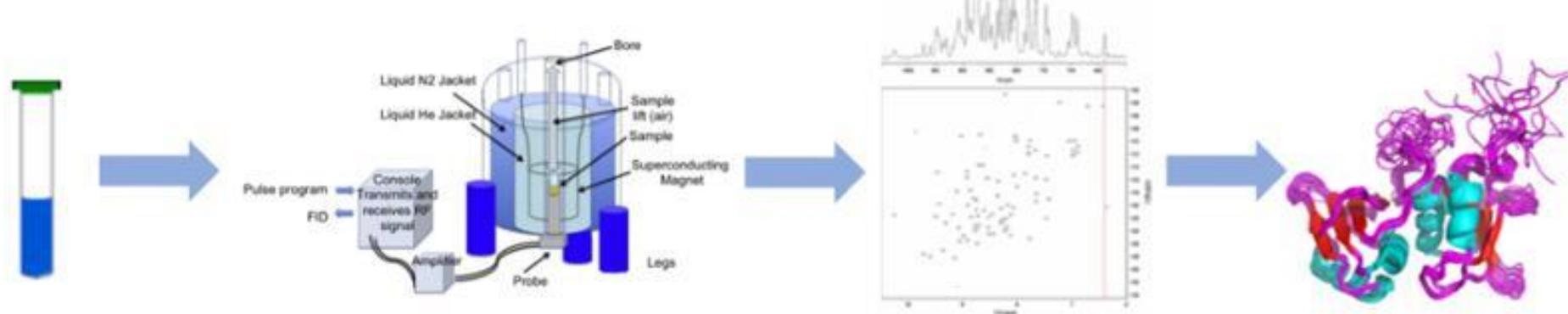
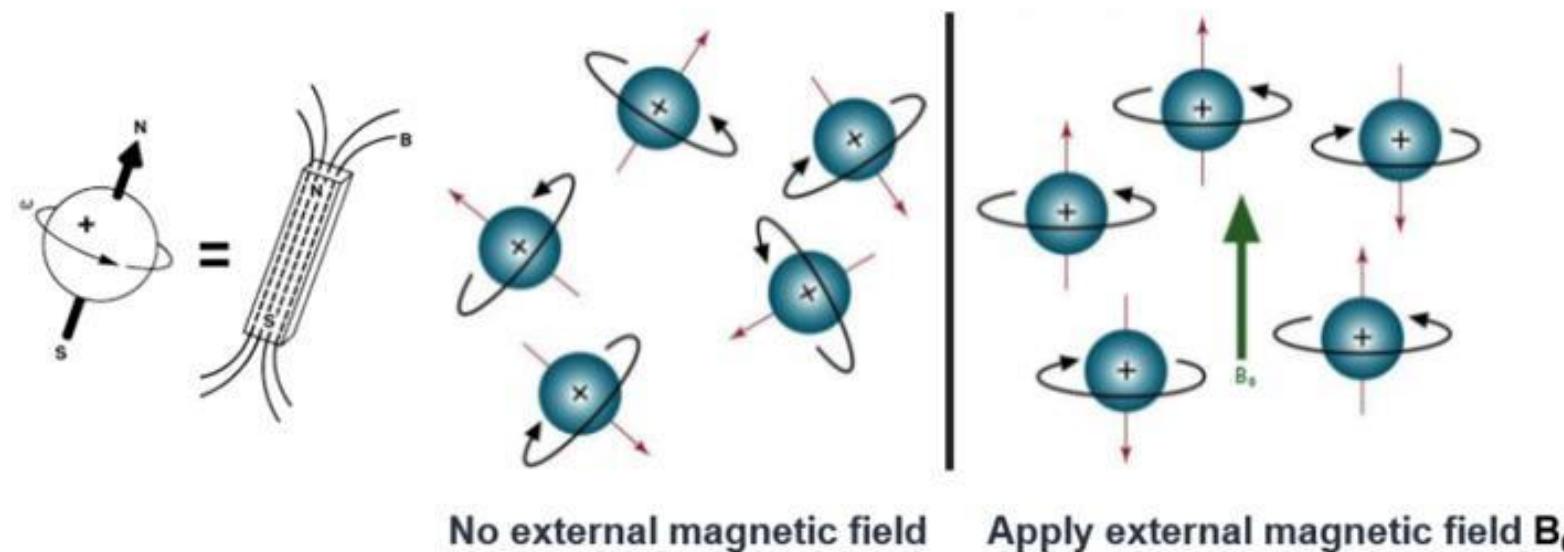
# Structure determination: X-ray crystallography



# Structure determination: X-ray crystallography



# Structure determination: Nuclear Magnetic Resonance (NMR)



**Sample preparation**

**Data acquisition**

**Spectral processing**

**Structural analysis**

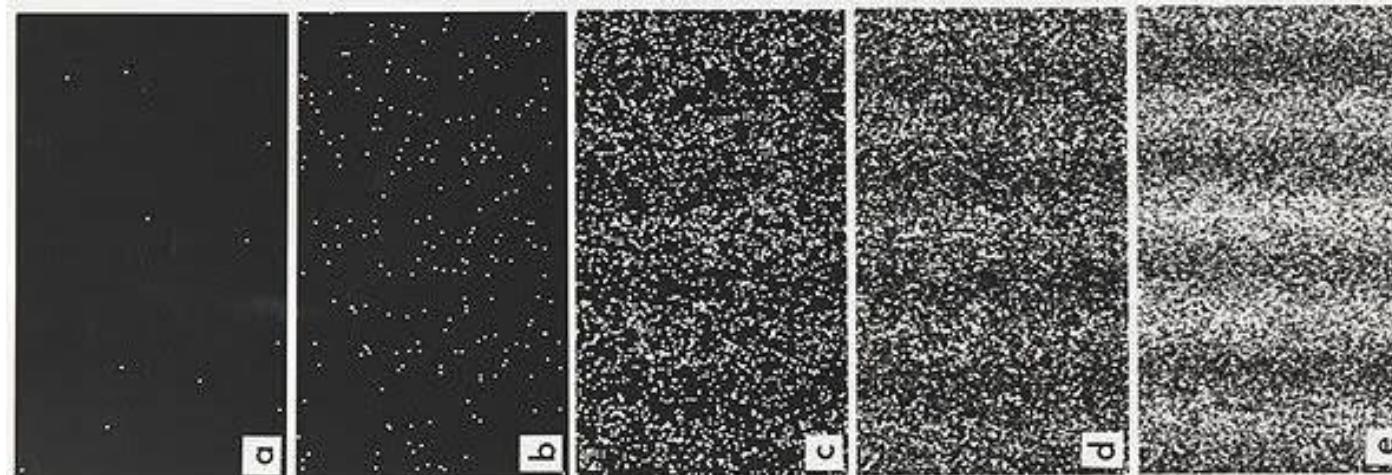
# [Extra] Structure determination: Electron Microscopy (EM)

- Particles as «waves that transfers energy and momentum»

$$\lambda = \frac{h}{p}$$

$\lambda$  : wavelength  
 $p$  : momentum  
 $h$  : Planck constant

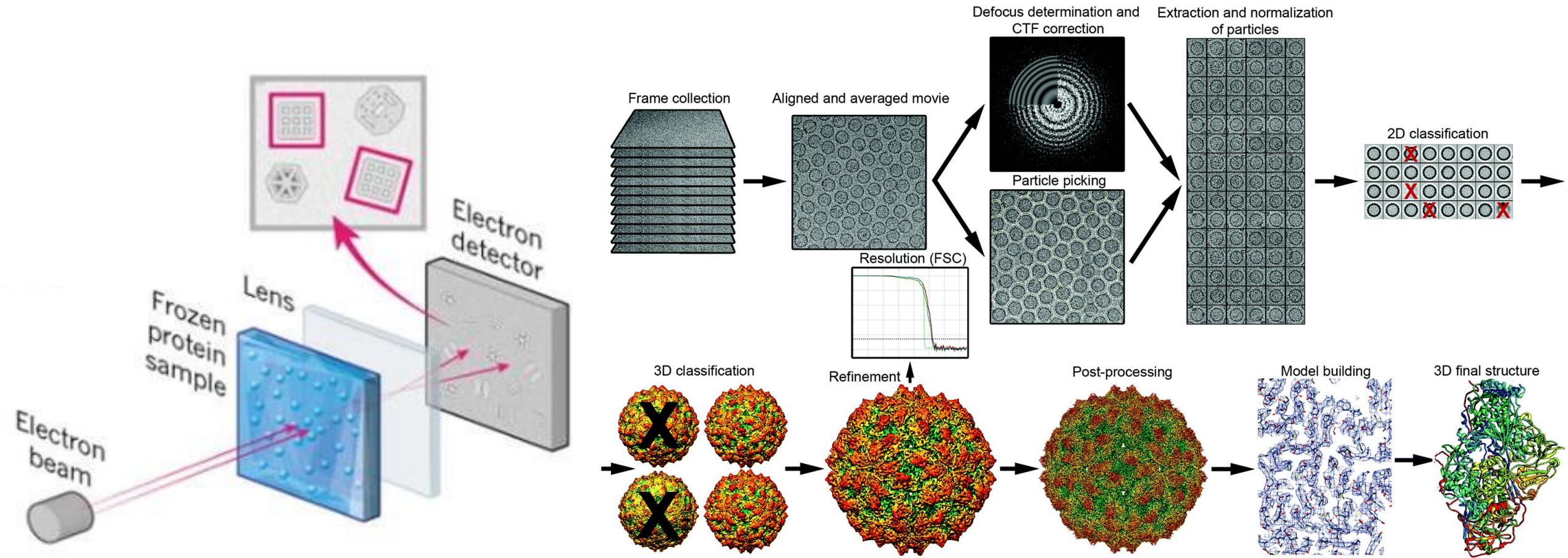
- Davisson–Germer experiment: electrons diffract too!



Louis De Broglie  
1892-1987

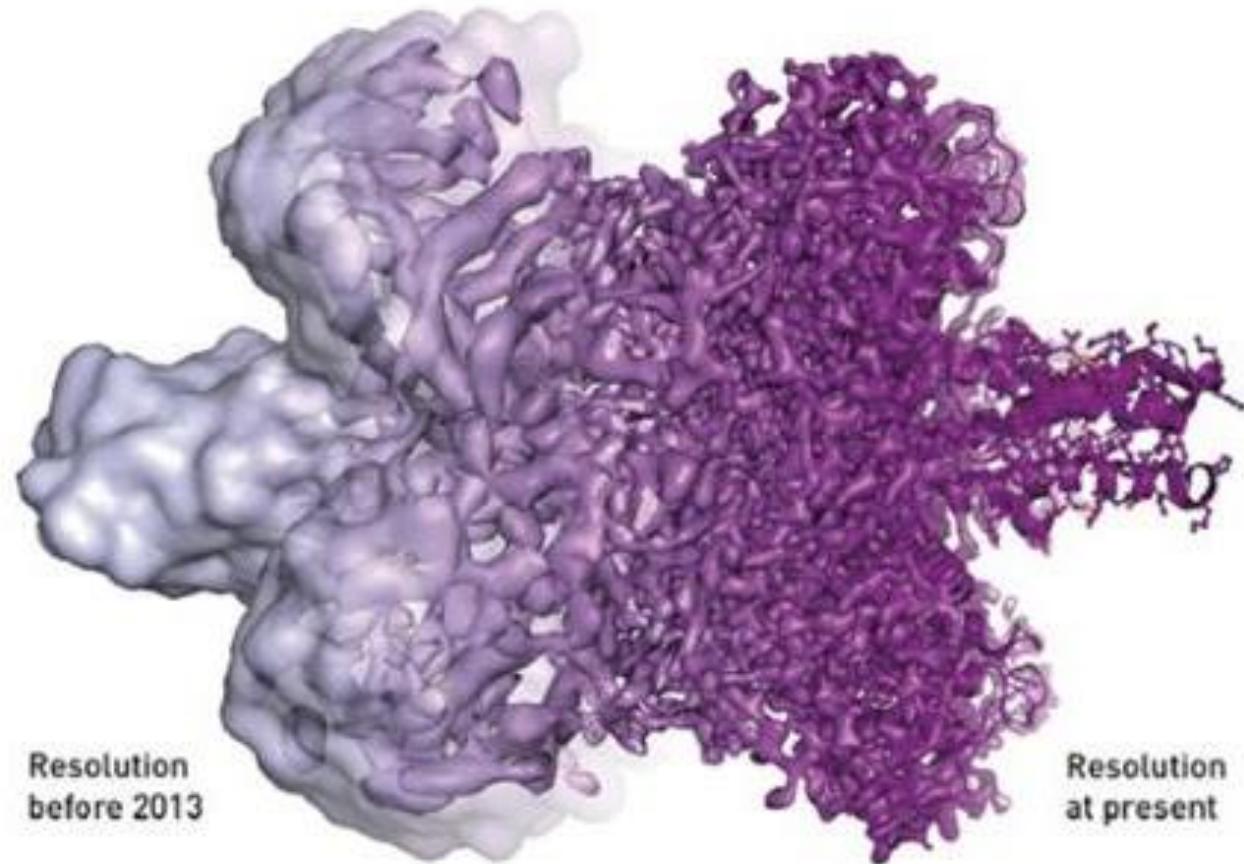
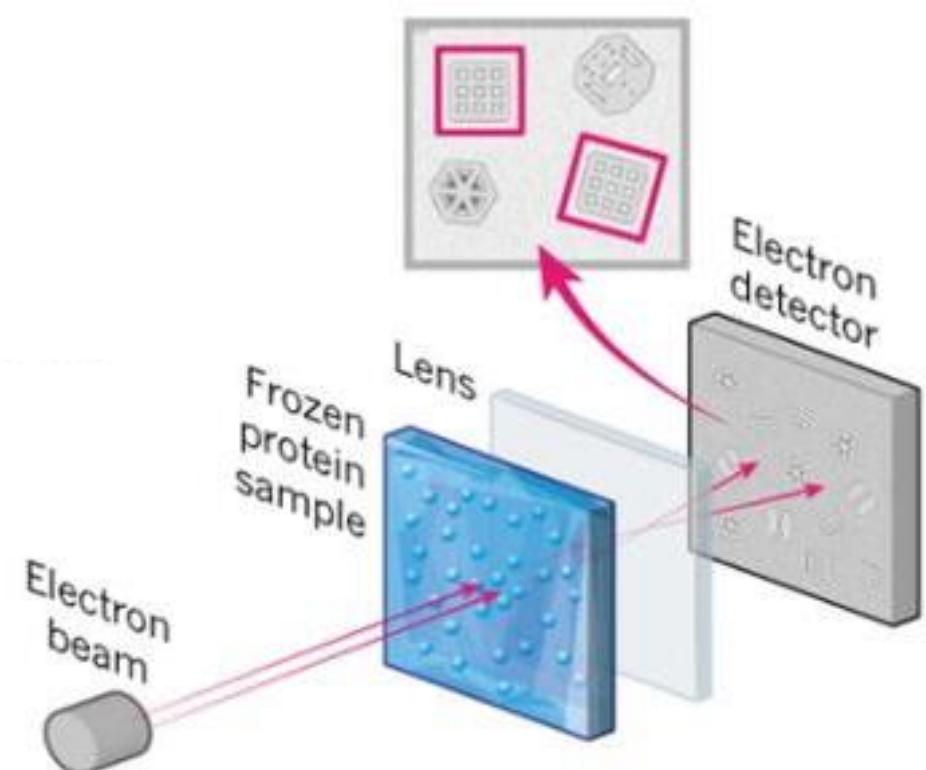
A. Tonomura *et al.*, *Demonstration of single-electron buildup of an interference pattern*, American Journal of Physics, 1989

# Structure determination: Electron Microscopy (EM)



# Structure determination: Electron Microscopy (EM)

## Resolution Revolution



# The Protein Data Bank (PDB)

- Molecular structures are deposited in the Protein Data Bank (PDB)
  - 1971: foundation of PDB at Brookhaven National Laboratory

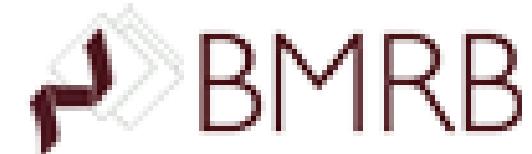


A yellow rectangular card with black printed text and data. At the top, it shows "ATOM C 20 -0.03426 0.25407 -0.11938 1 1.0000". Below this is a grid of numbers representing a protein's backbone. The grid has 20 columns and 20 rows, with each cell containing a number from 1 to 5. The bottom of the card has the text "Printed in the U.K." and "CDL 5081".



# The Protein Data Bank (PDB)

- Molecular structures are deposited in the Protein Data Bank (PDB)
  - 1971: foundation of PDB at Brookhaven National Laboratory

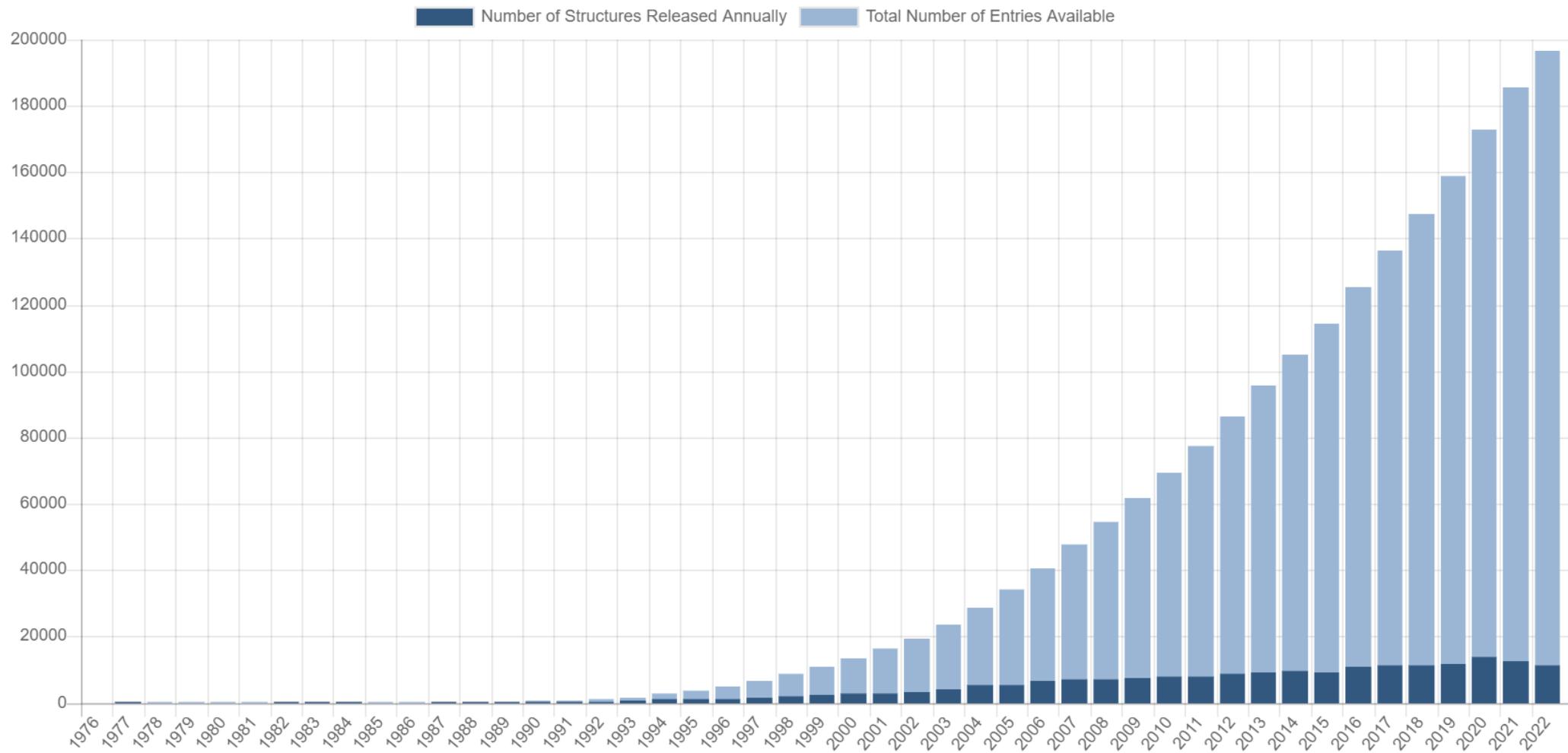


- 2003: wwPDB founded
  - now with four deposition centres
- Each molecule assigned a unique **4-characters code** (e.g. 1MBO, 1AV1, ...)

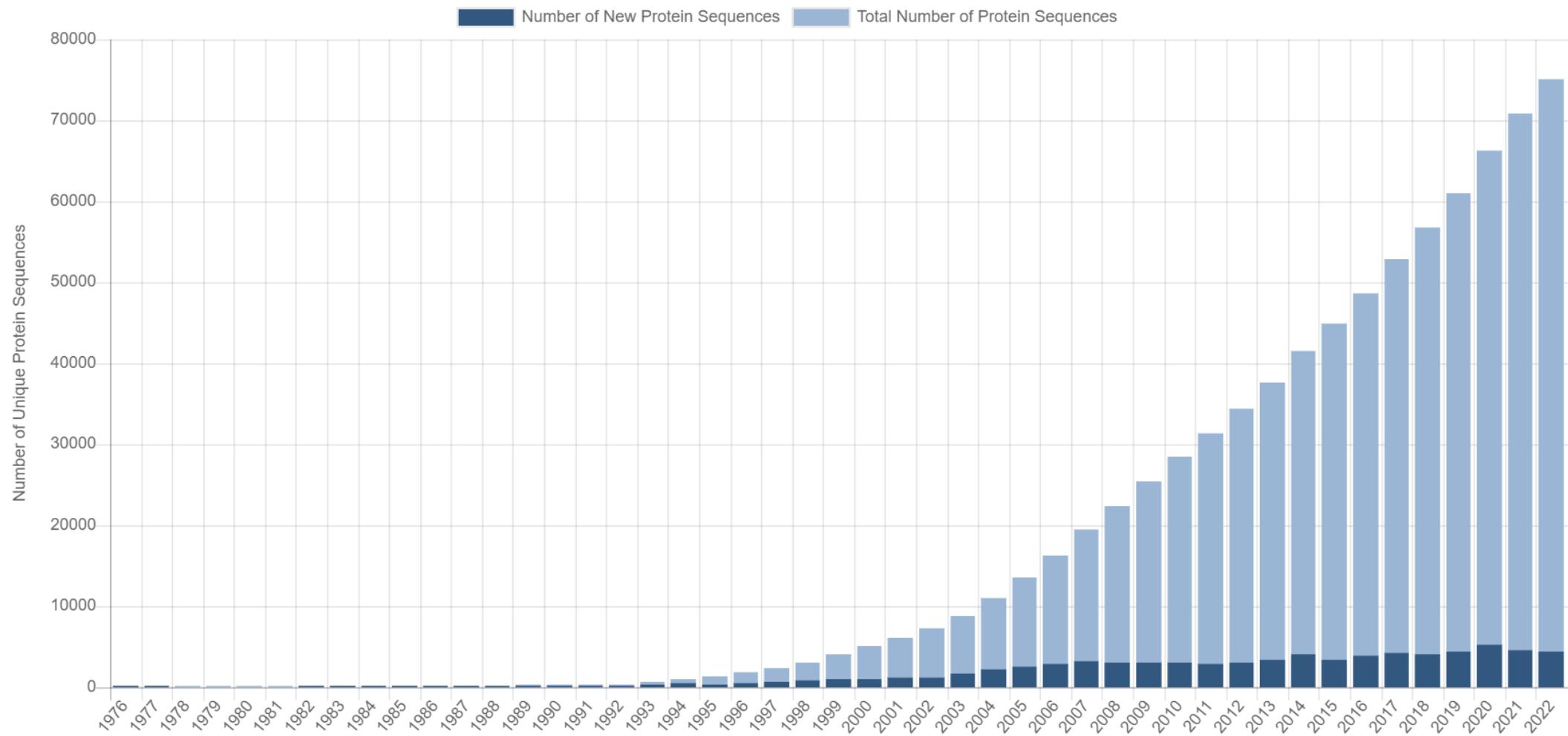
H.Berman, K.Henrick and H. Nakamura, *Announcing the worldwide Protein Data Bank*. Nature Structural & Molecular Biology, 2003

wwPDB consortium , *Protein Data Bank: the single global archive for 3D macromolecular structure data*, Nucleic Acids Research, 2019

# 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

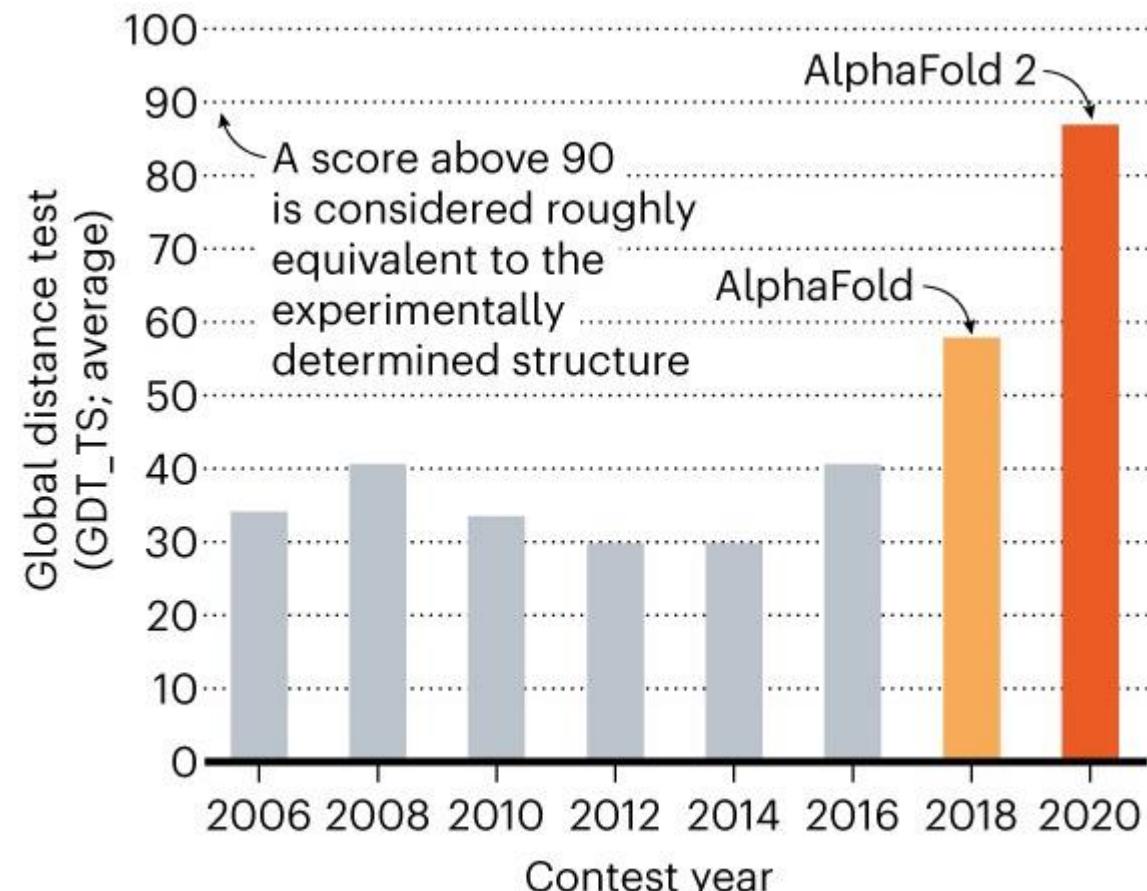
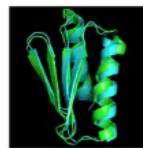
Protein Sequence

SQETRKKCTEMKKFKNCEVRCDESNHCVRCSDTKYTLC



prediction

Structure

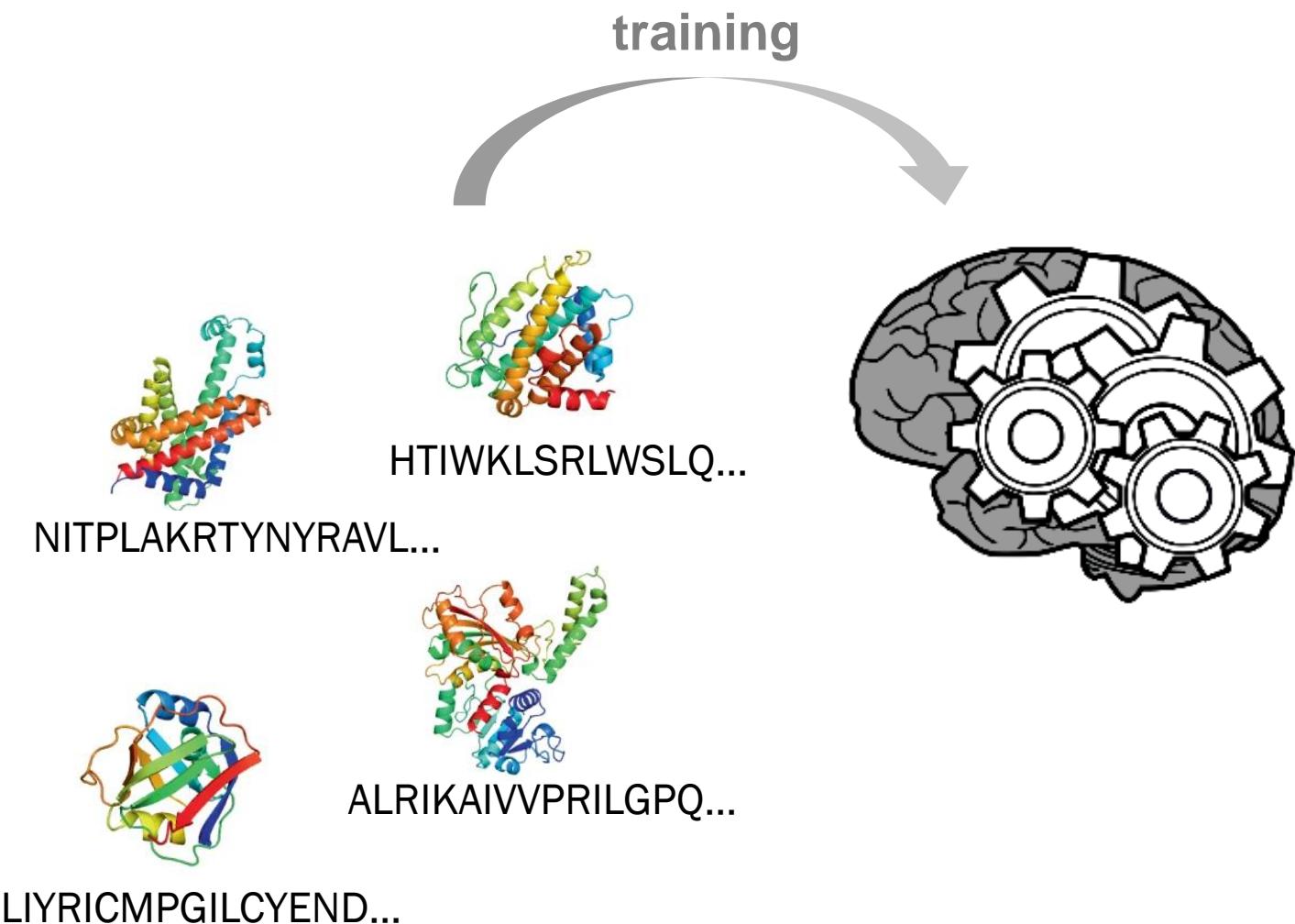


A.W. Senior et al., *Improved protein structure prediction using potentials from deep learning*, Nature, 2020

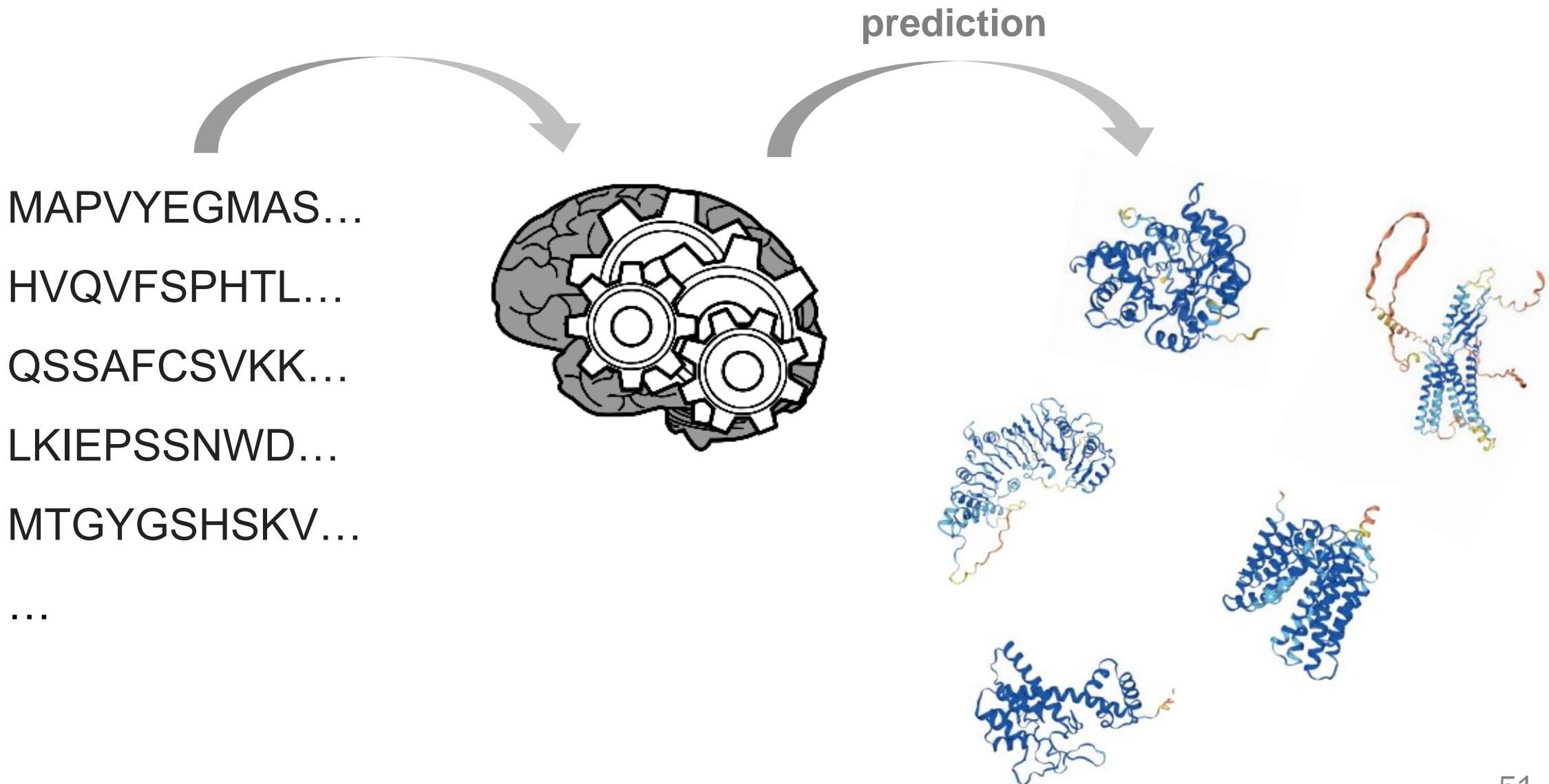
J. Jumper et al., *Highly accurate structure prediction with AlphaFold*, Nature, 2021

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

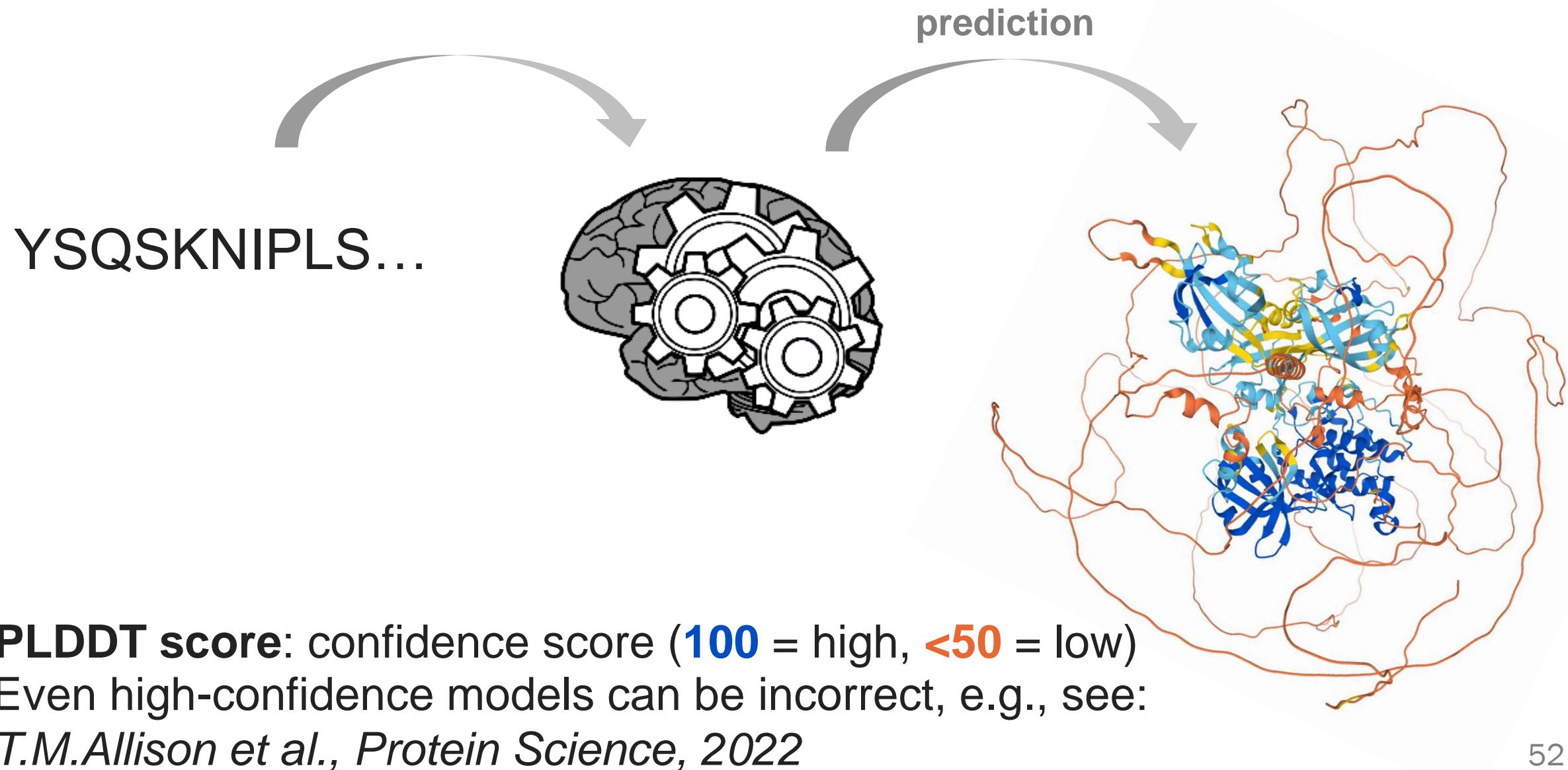
# Protein fold prediction: AlphaFold2



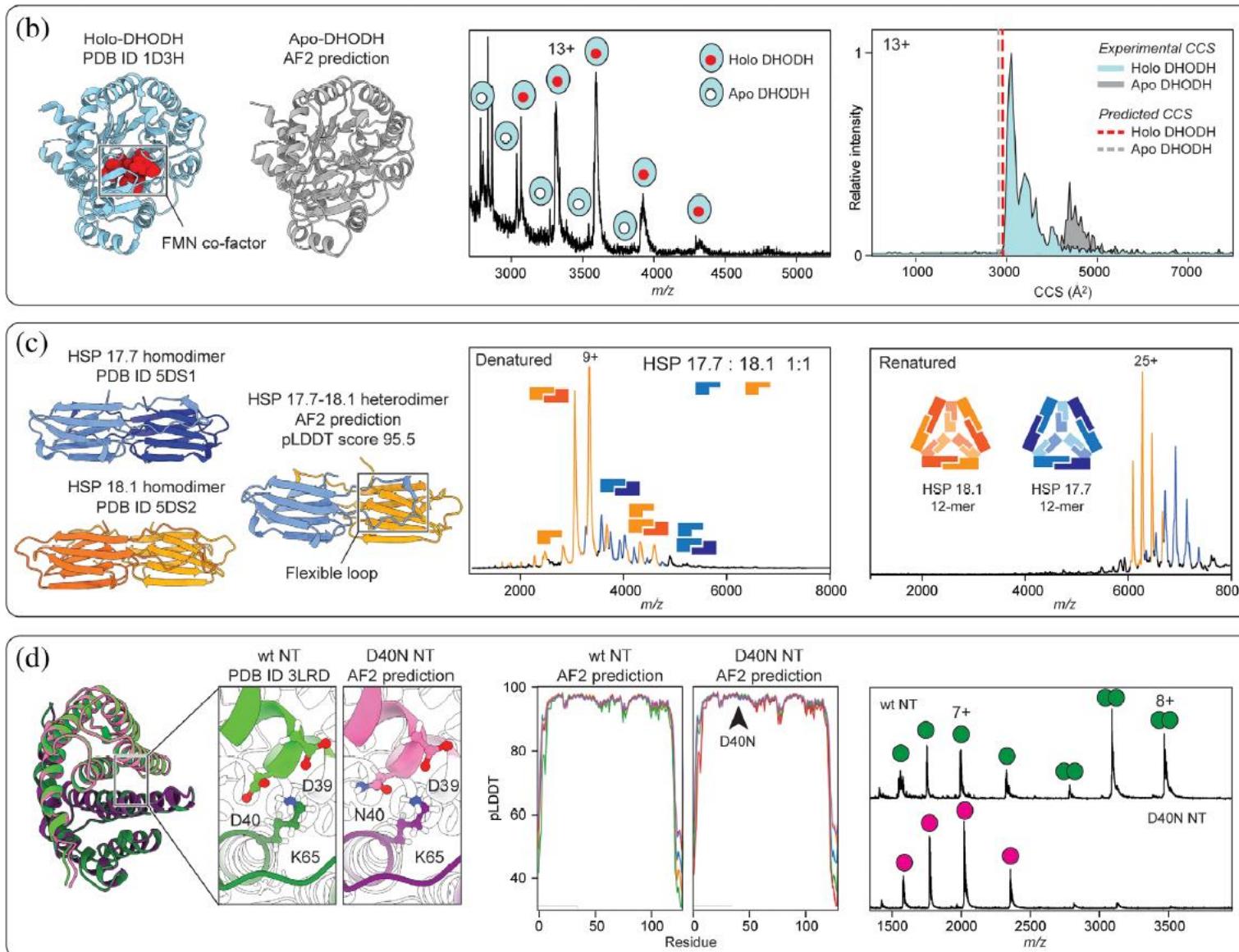
# Protein fold prediction: AlphaFold2



# Protein fold prediction: AlphaFold2



# Protein fold prediction: warning!



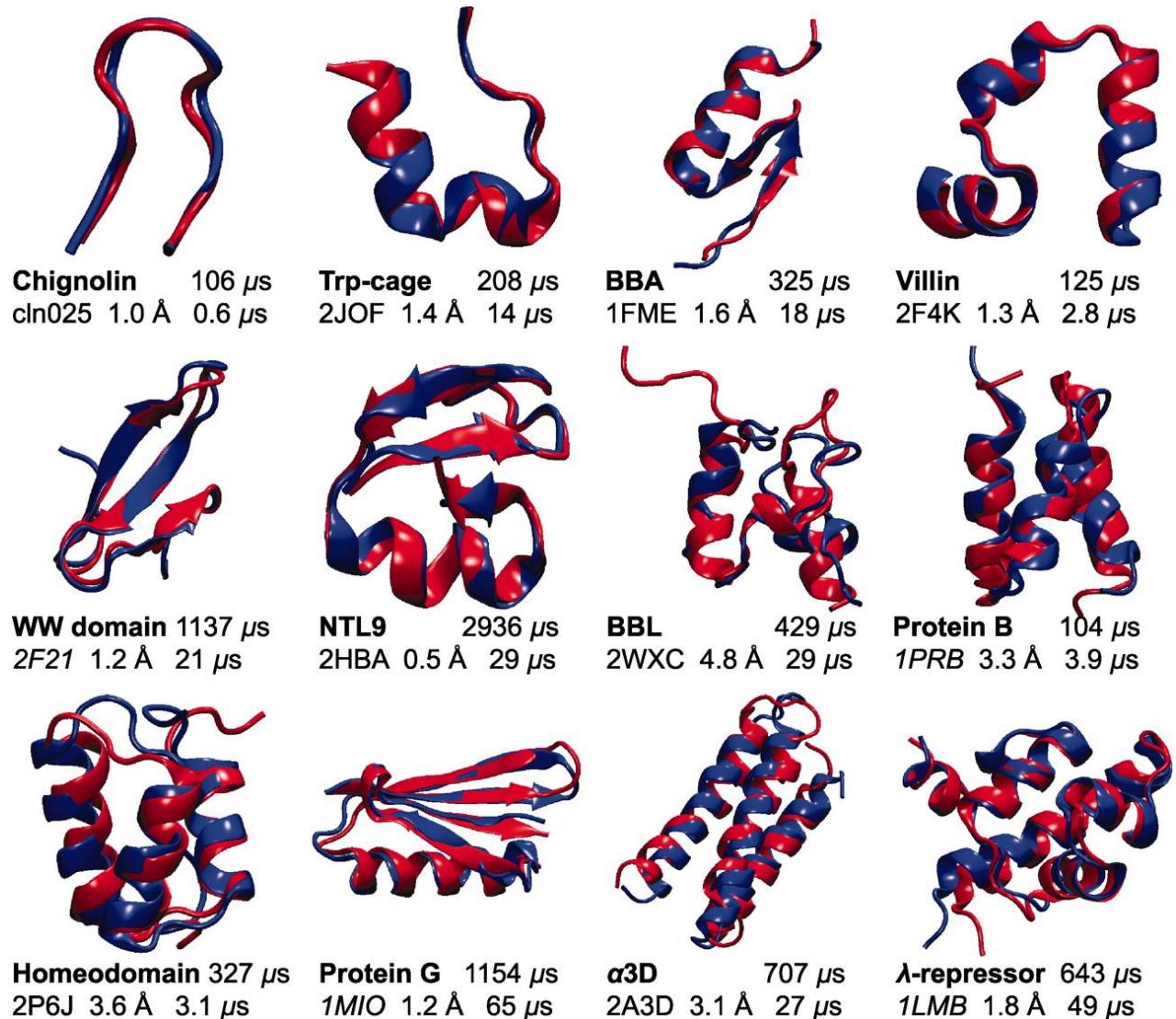
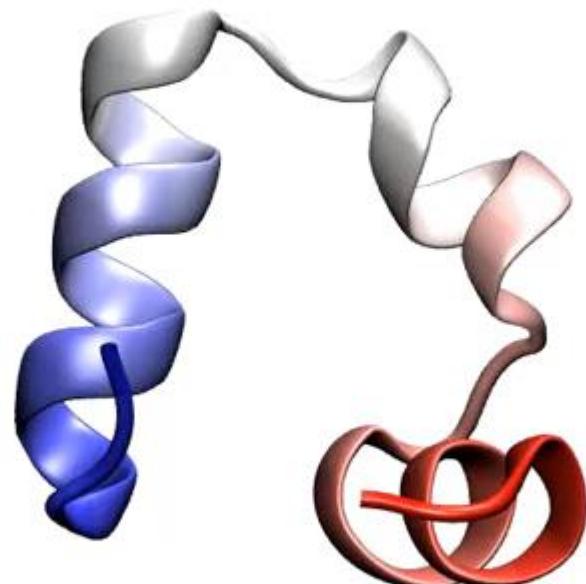
Apo protein predicted folded like holo state, but it should be unfolded

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 ( $<100 \mu\text{s}$ ) proteins can be studied via simulation

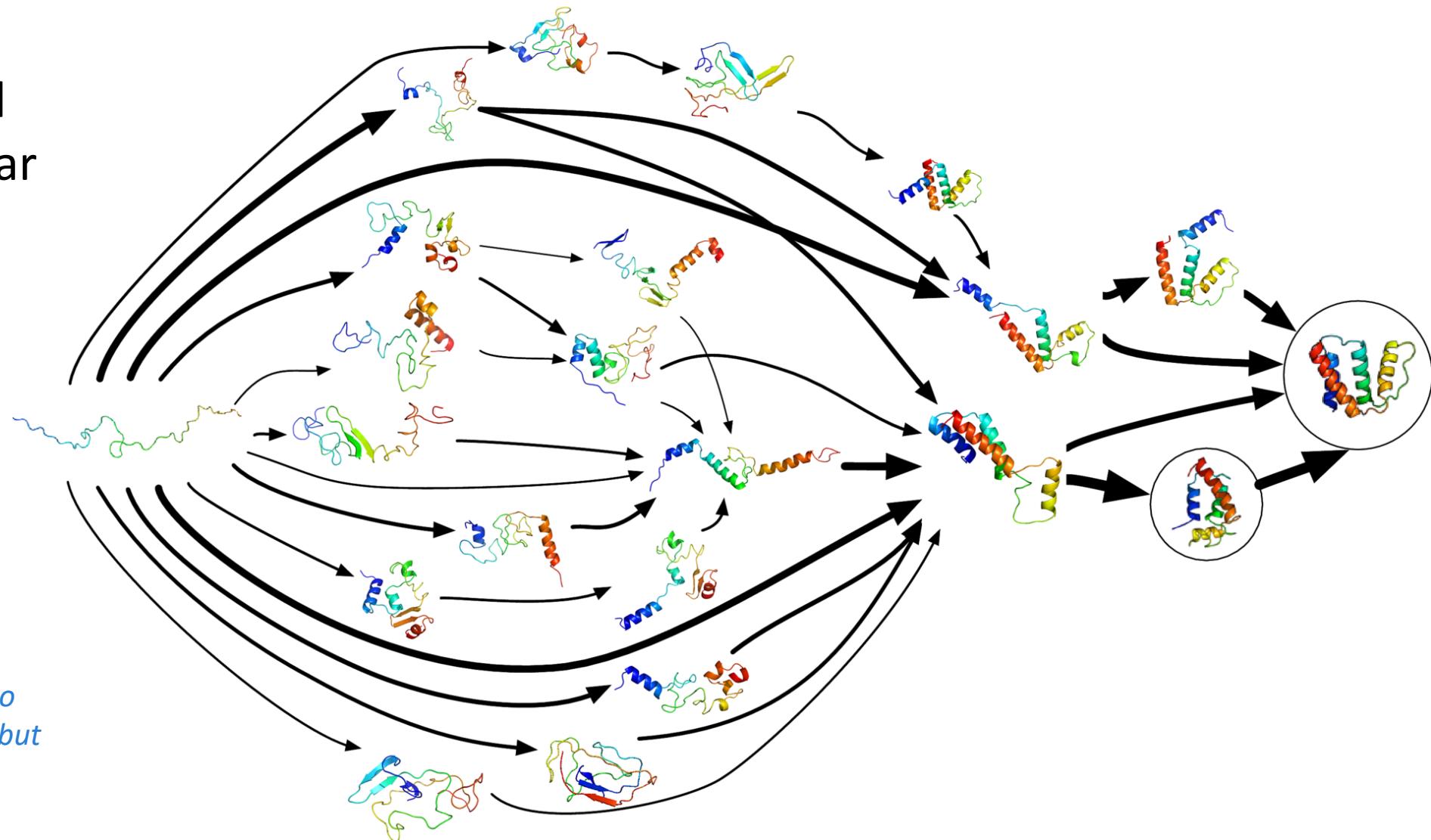


# Folding@Home

Combine distributed computing, molecular simulation and Markov State Modelling (MSM) to predict protein folding pathways.

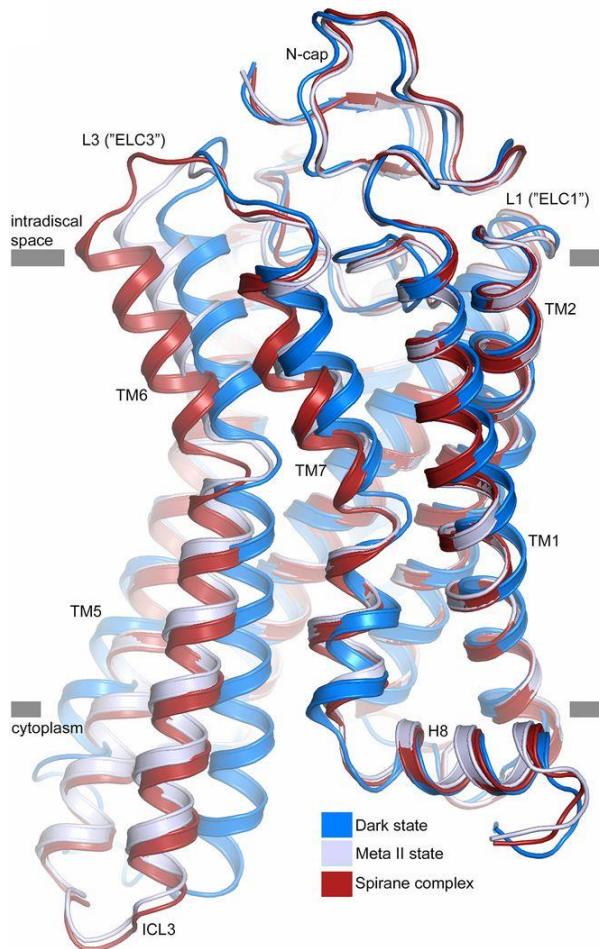
[www.foldingathome.org](http://www.foldingathome.org)

V. S. Pande, K. Beauchamp, G. R. Bowman, *Everything you wanted to know about Markov State Models but were afraid to ask. Methods*, 2010

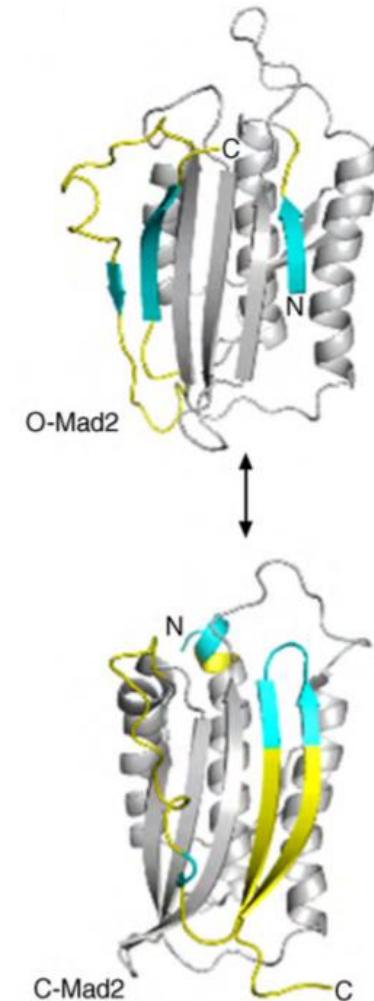


# Protein dynamics

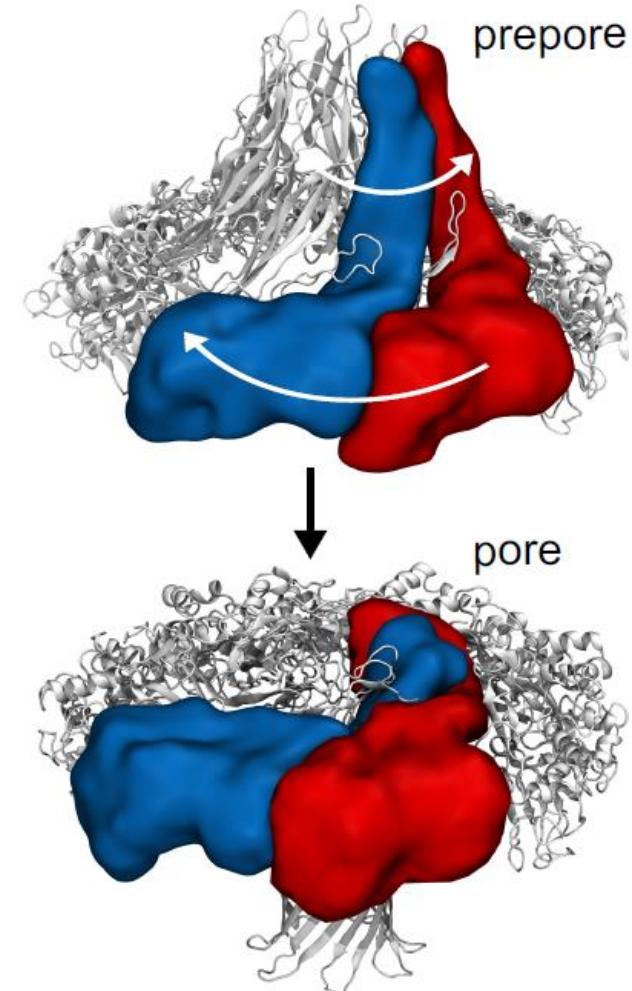
sequence + interaction with environment = conformational space



D. Mattle et al., *Ligand channel in pharmacologically stabilized rhodopsin*, PNAS, 2018



P.N. Bryan and J. Orban, *Proteins that switch folds*, Curr. Op. Struct. Biol., 2010



M.T. Degiacomi et al., *Molecular assembly of the aerolysin pore reveals a swirling membrane-insertion mechanism*, Nat. Chem. Biol., 2013

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