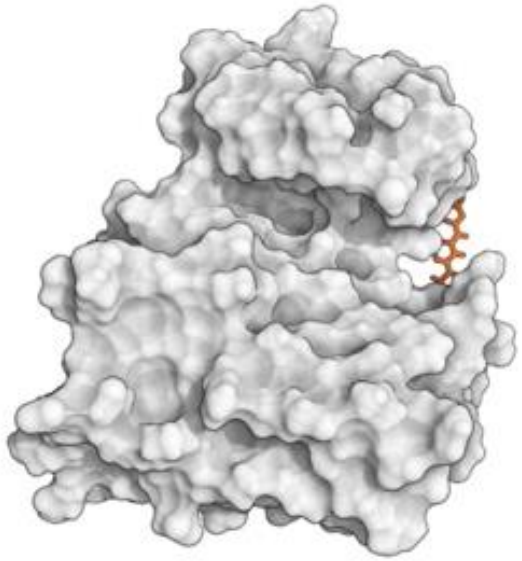
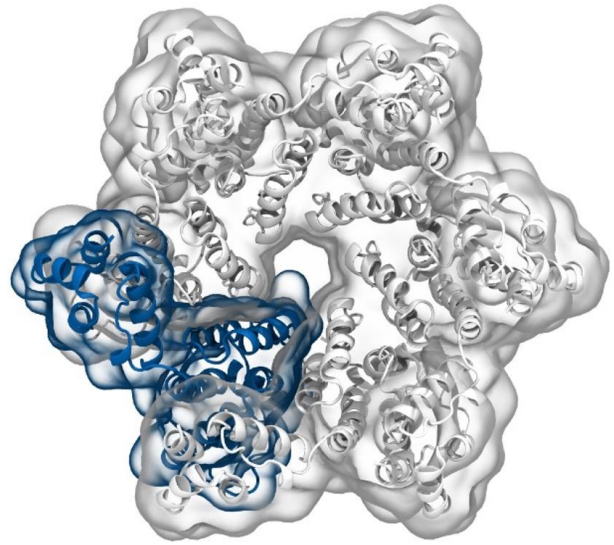
Simulation of Biomolecules



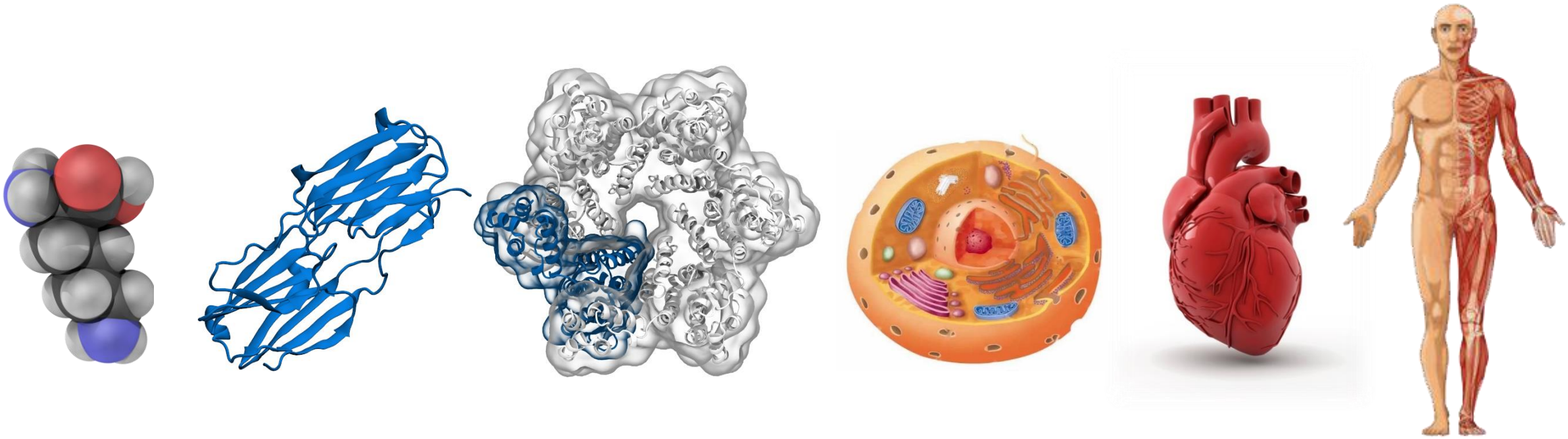
Introduction



Dr Matteo Degiacomi
University of Edinburgh
matteo.degiacomi@ed.ac.uk

Dr Antonia Mey
University of Edinburgh
antonia.mey@ed.ac.uk

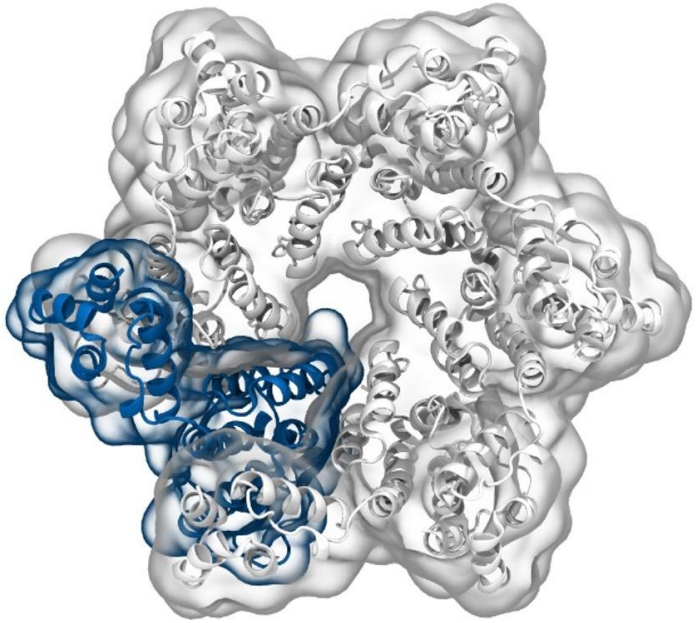
Life emerges from molecular assembly



COMPLEXITY

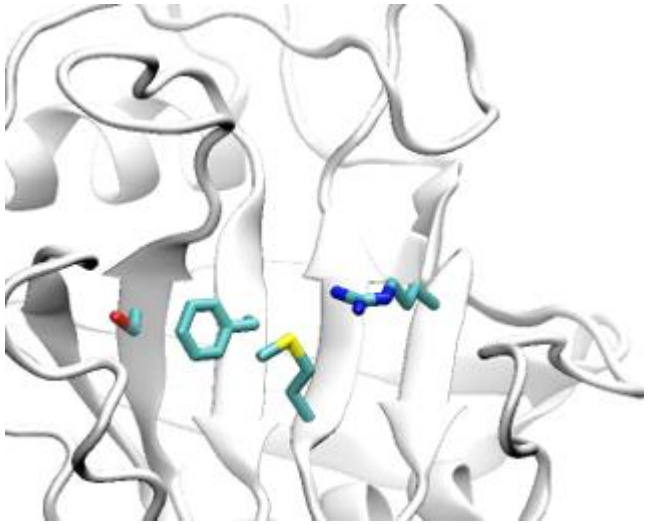
Structure and dynamics determine protein (mal)function

HIV Capsomer



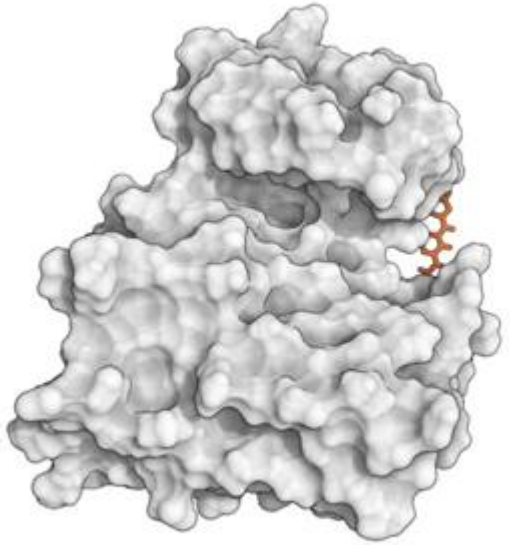
M.T. Degiacomi, ***Structure***, 2019

Cyclophilin



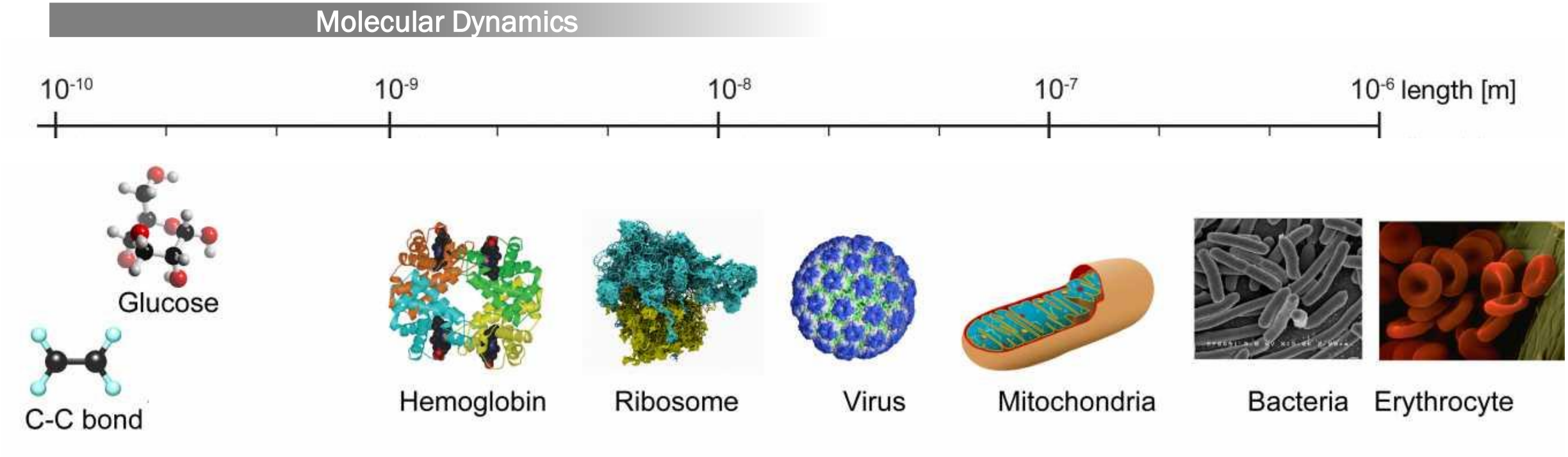
Wapeesittipan, Mey, et al., ***Comms. Chem.***, 2019

Tyrosine kinase —dasatanib

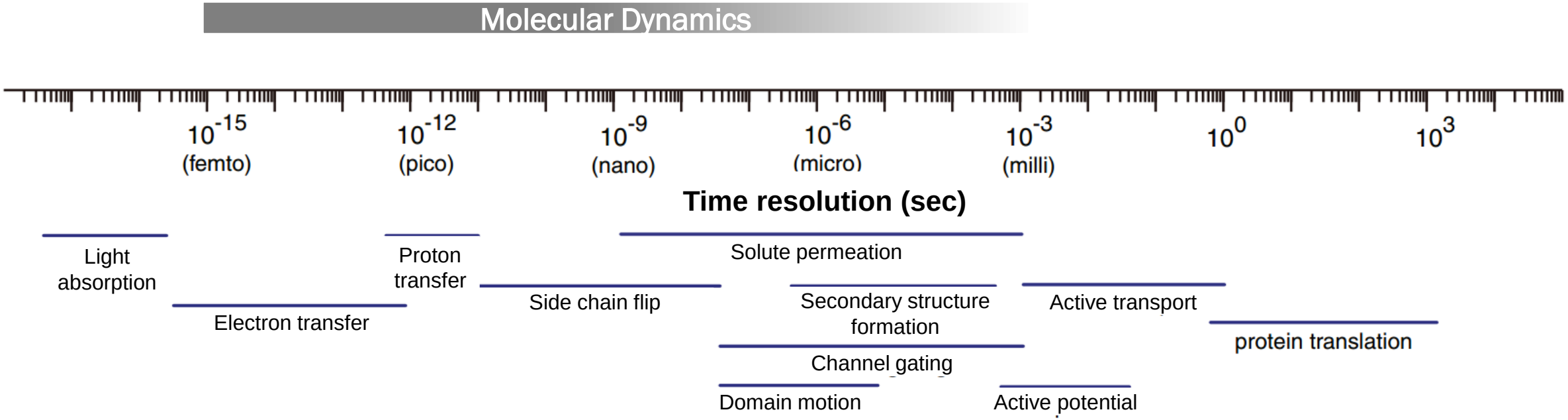


Y Shan et al. ***JACS***, 2011

Sizes in biochemistry



Timescales in biochemistry

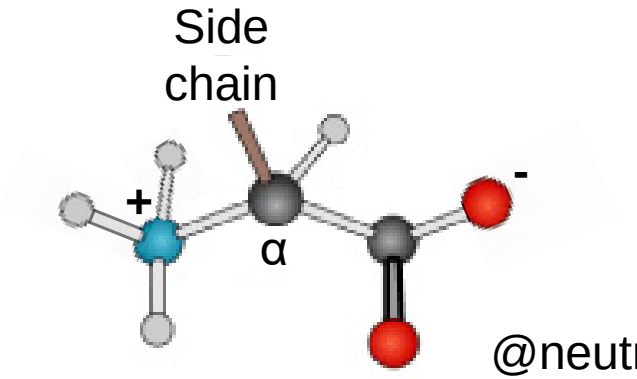


Part 1: What is a protein?

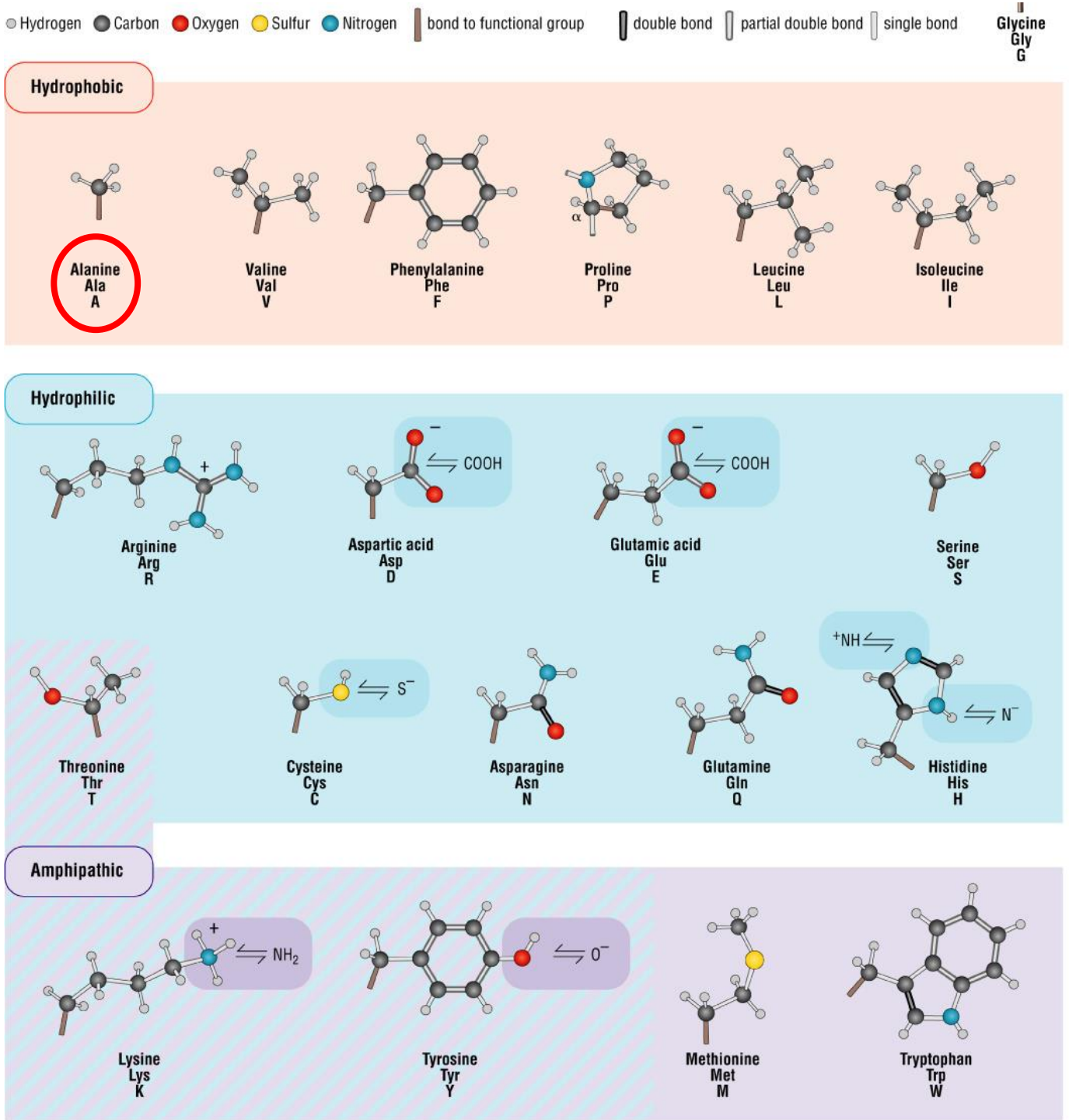
Proteins are amino acids polymers

Amino acids are composed of:

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



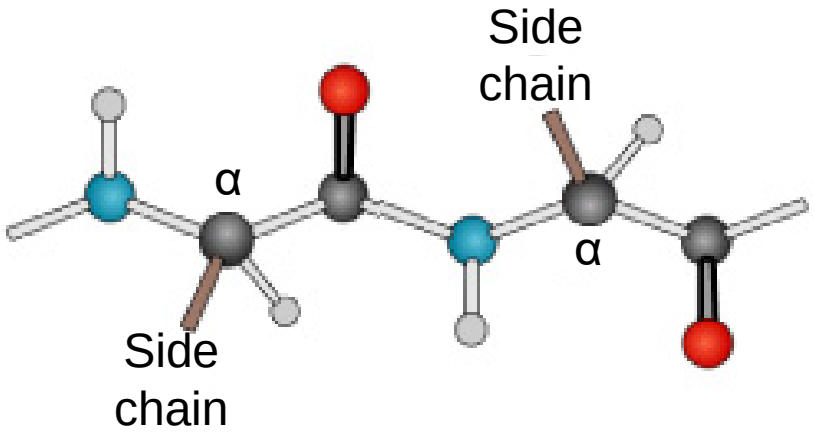
@neutral pH



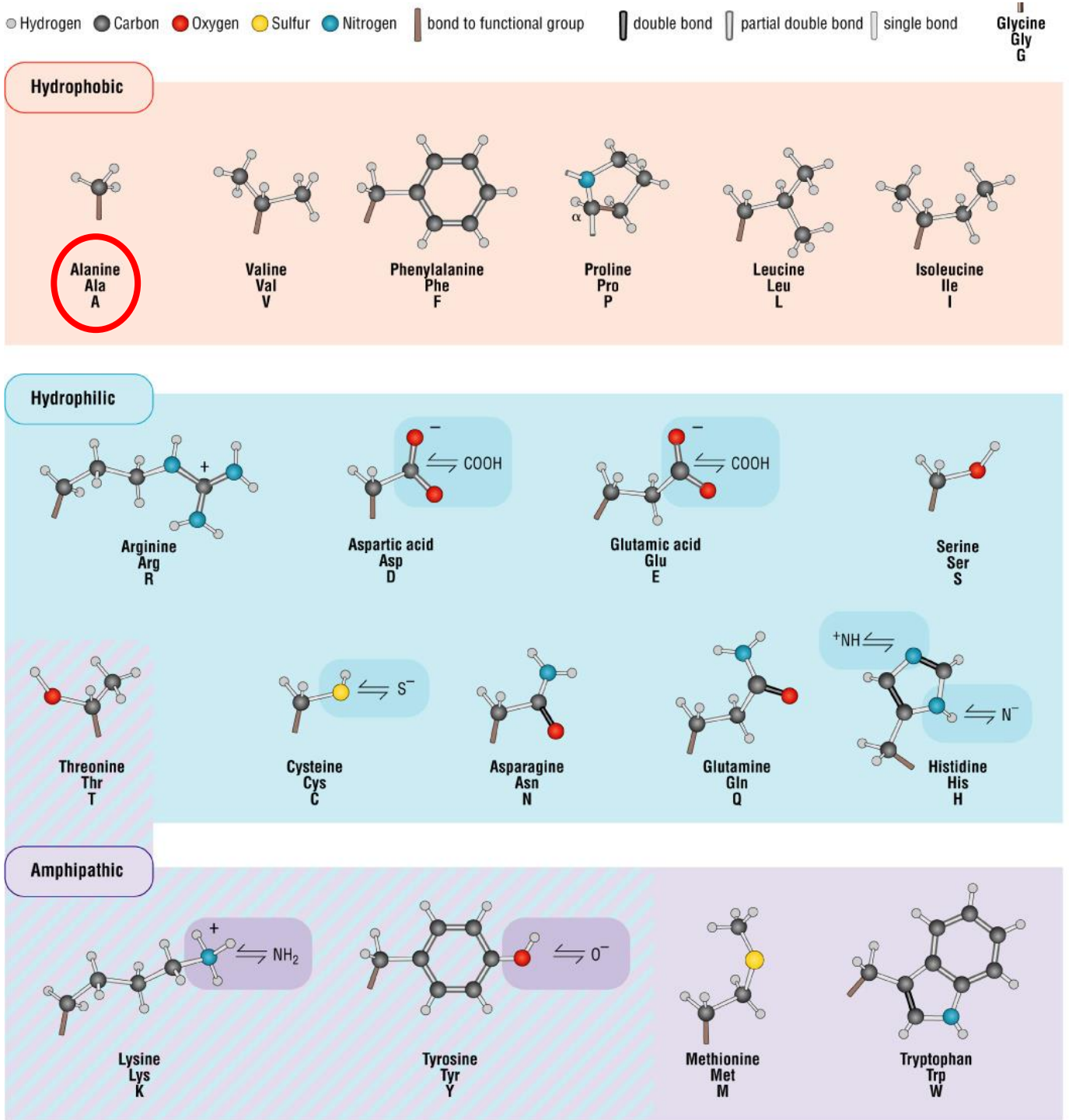
Proteins are amino acids polymers

Amino acids are composed of:

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



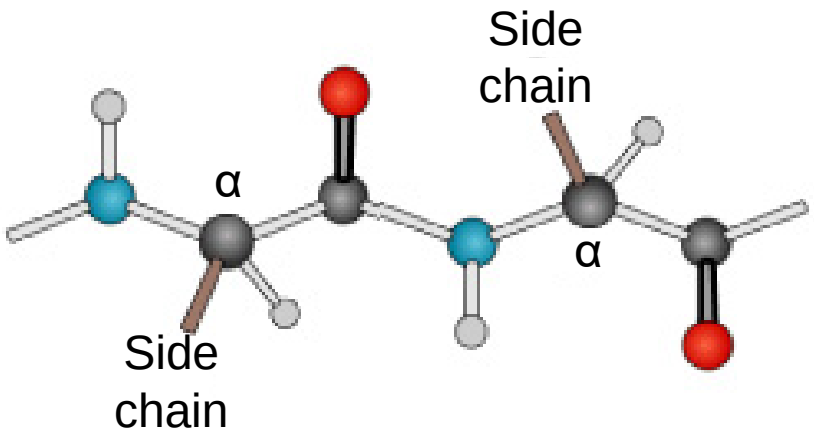
Amino acids polymerize forming a *peptidic bond* (condensation)



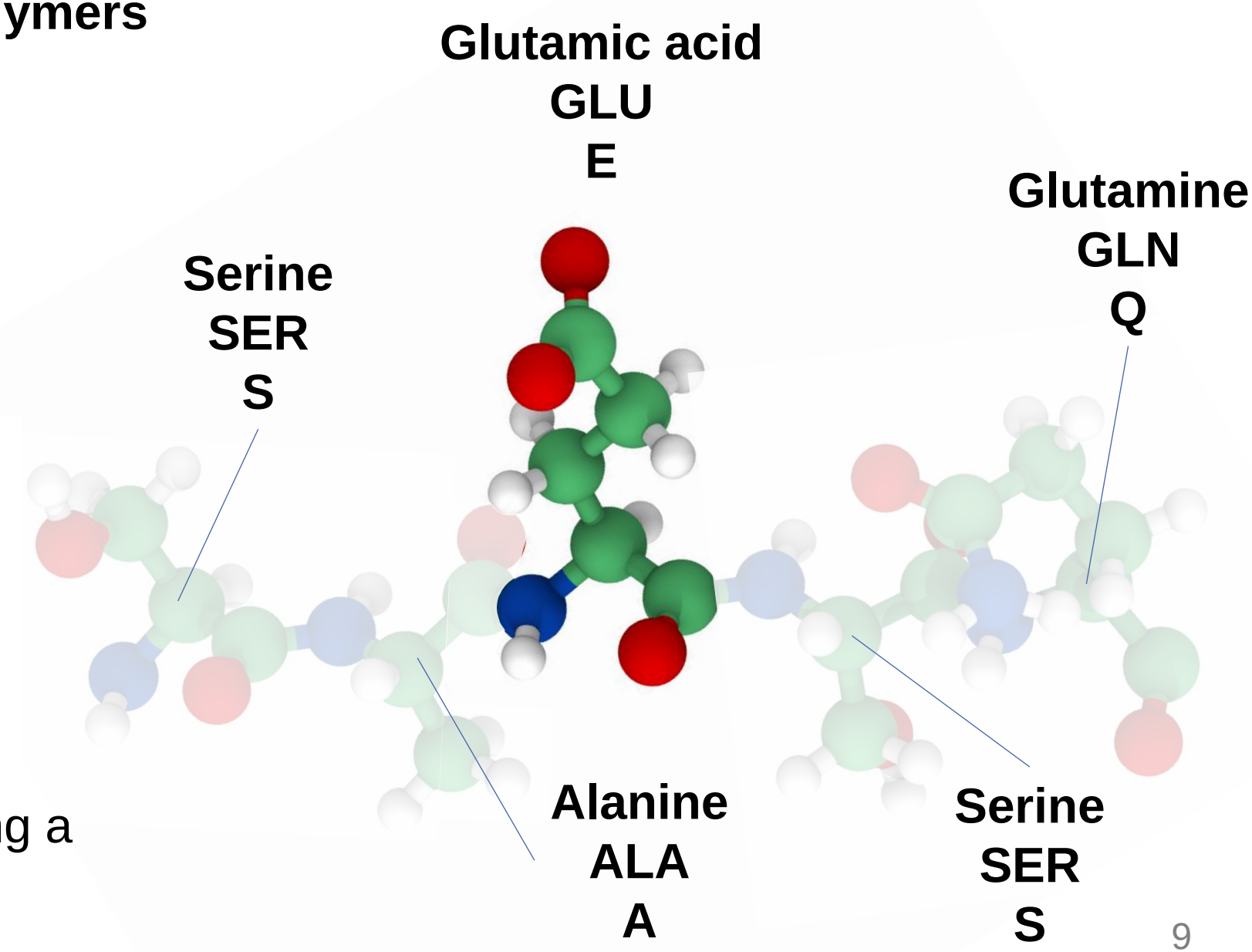
Proteins are amino acids polymers

Amino acids are composed of:

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

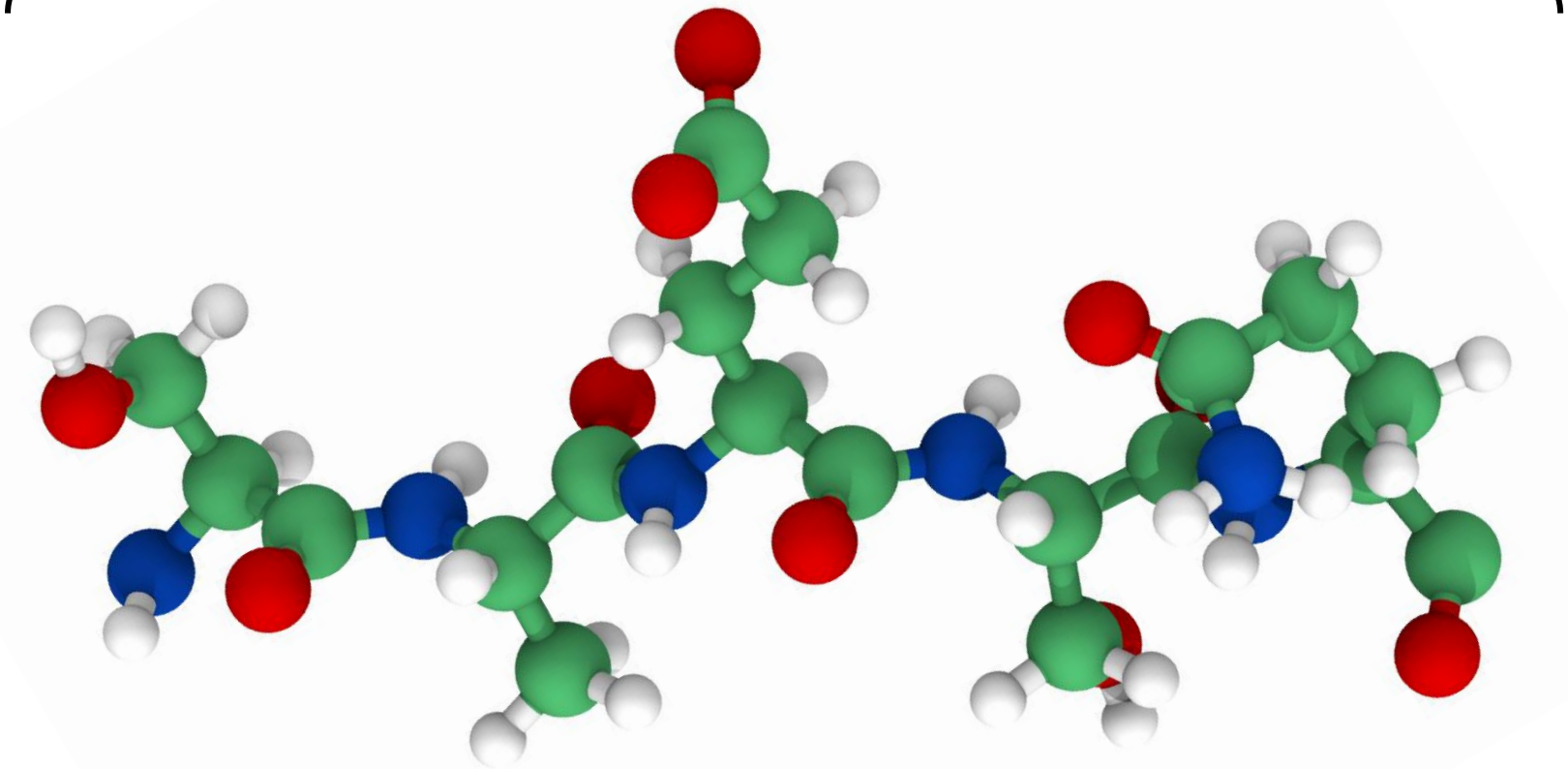


Amino acids polymerize forming a *peptidic bond* (condensation)



Protein Primary Structure: Sequence

SAESQ



Proteins with similar sequence, likely:

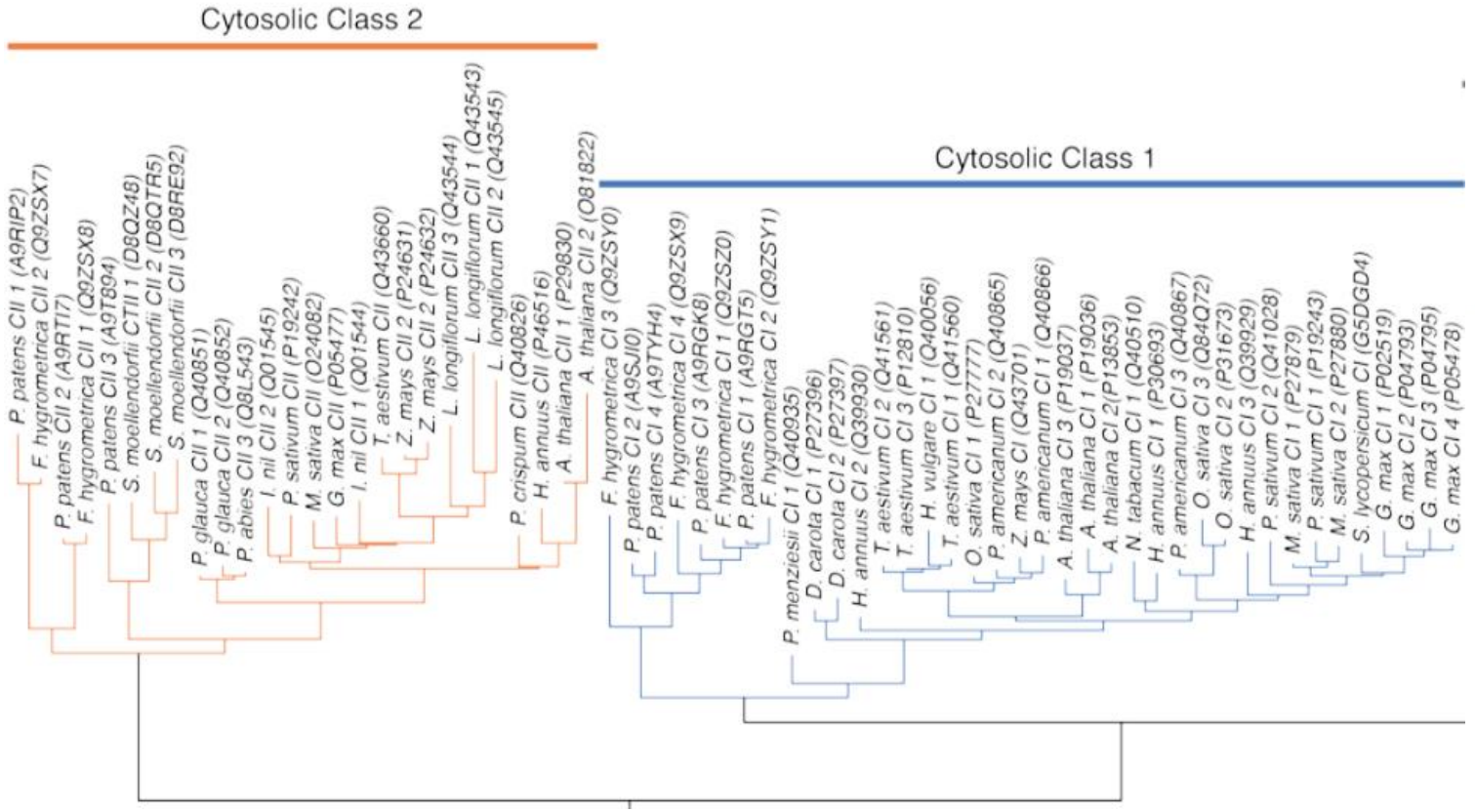
- have similar functions in an organism
- are evolutionarily related

Protein Primary Structure: Sequence

```
Angiosperm 1   MSLIPSFFSGRRSNVFD-PF--SL-DVWDPLKD-FPFSNSSPSASFPRENPAFV-STRVDWKETPEAHVFKADLPGLKKEEVKVEVE
Gymnosperm 1   MSIIPSFFGRRSSSAFD-PF--SL-DVWDPFRAFTDLSGGGPSGQFVNEASAVA-NTQIDWKETPEAHIFKADLPGLKKEEVKIELE
Bryophyte 1    MAL--SLFGSRGNGVFD-PF--EFGSVWDPFSA----PESGLSRKLAGDAHAGA-NTRIDWRETPEAHIFKADLPGLRKEEVKIQVV
Angiosperm 2   ------------MDLDSPLFNTLHHIMDLTDDTTEKNLNAPTRTYVRDAKAMA-ATPADVKEHPNSYVFMVDMPGVKSGDIKVQVE
Gymnosperm 2   ------------MAMD-PSLITVQHLLGVPDD-LEKLLNAPTHSYMRDTKAMA-STPVDVKEYPNSYVFIIDMPGLKSNDIKVQVE
Bryophyte 2    --------MEFVVFDTD-PFLTSLHQLVHEPESDLERKIKRKRRSQHDEPRHVTIATPVDVKEKKDAYLFIADVPGLQKTDIEVQIE
```

Proteins with similar sequence, likely:

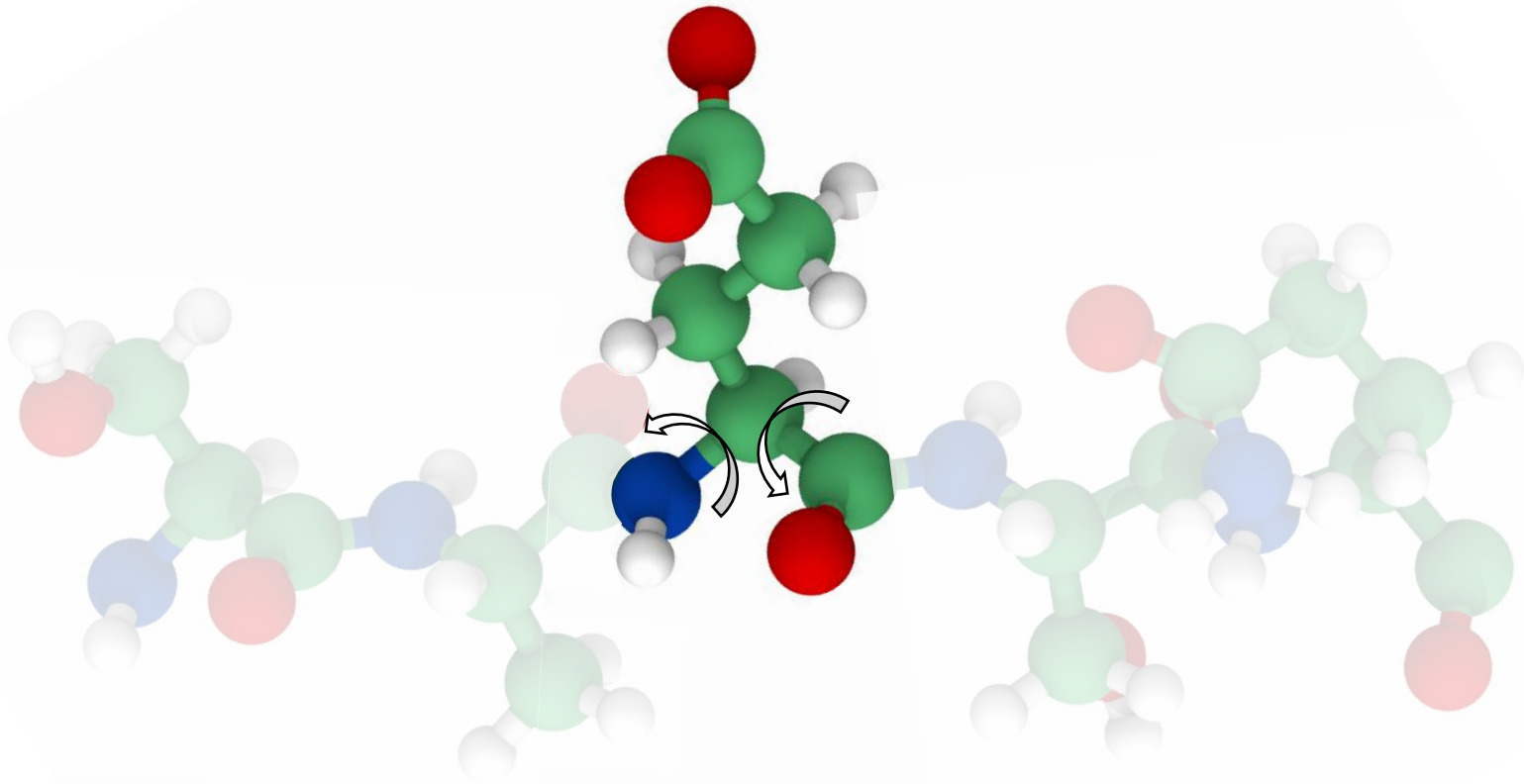
- have similar functions in an organism
- are evolutionarily related



G.K. Hochberg et al. *Structural principles that enable oligomeric small heat-shock protein paralogs to evolve distinct functions.* ***Science***, 2018 [SI]

Protein Secondary Structure

The amino acid chain path is determined by backbone torsional angles

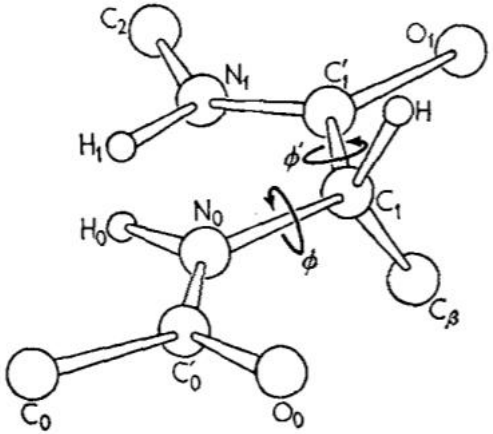


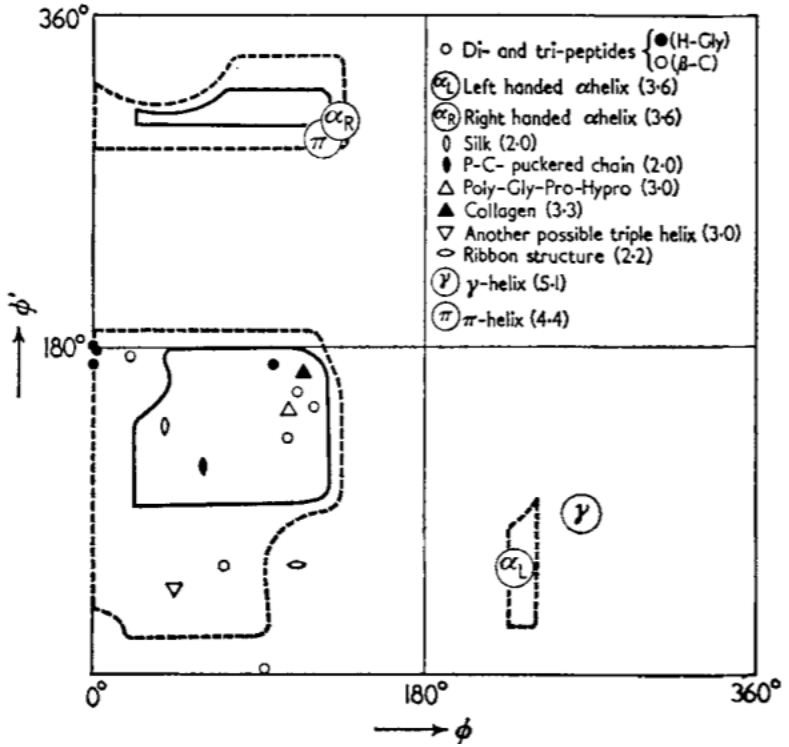
Protein Secondary Structure

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





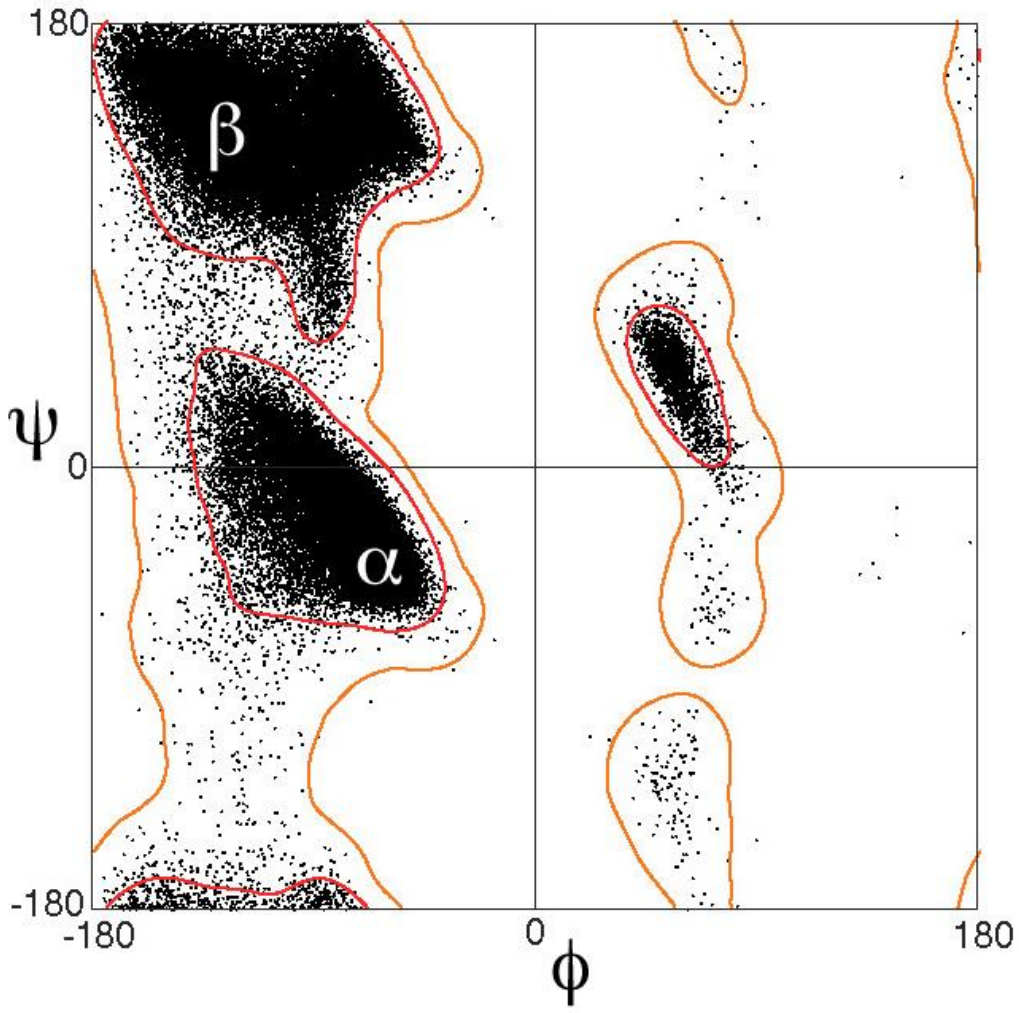
Protein Secondary Structure

The amino acid chain path is determined by backbone torsional angles

Ramachandran plot: scatter plot of amino acids backbone torsional angles

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

- H = α-helix
- G = 3_{10} helix
- I = π-helix
- B = residue in isolated β-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

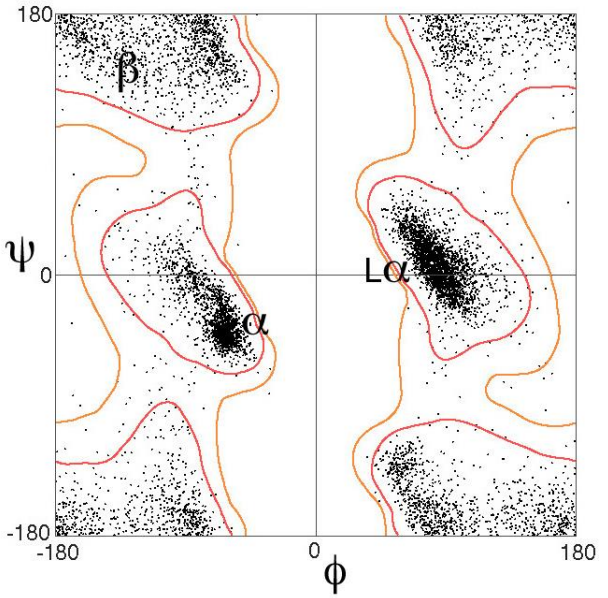
[Extra] Protein Secondary Structure

The amino acid chain path is determined by backbone torsional angles

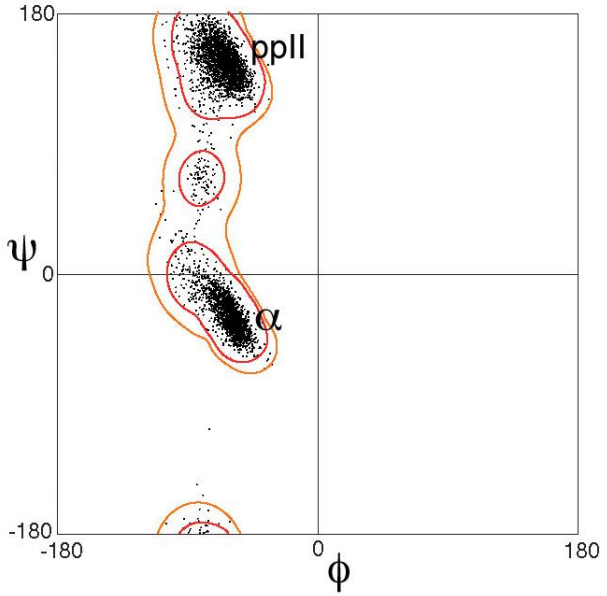
Ramachandran plot: scatter plot of amino acids backbone torsional angles

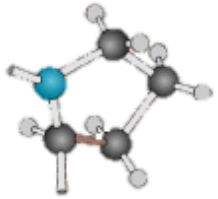
Dictionary of Secondary Structure in Proteins (DSSP) classification defines 7 secondary structure elements (regions in the plot):

- H = α-helix
- G = 3_{10} helix
- I = π-helix
- B = residue in isolated β-bridge
- E = extended strand
- T = hydrogen bonded turn
- S = bend



Glycine
GLY
G



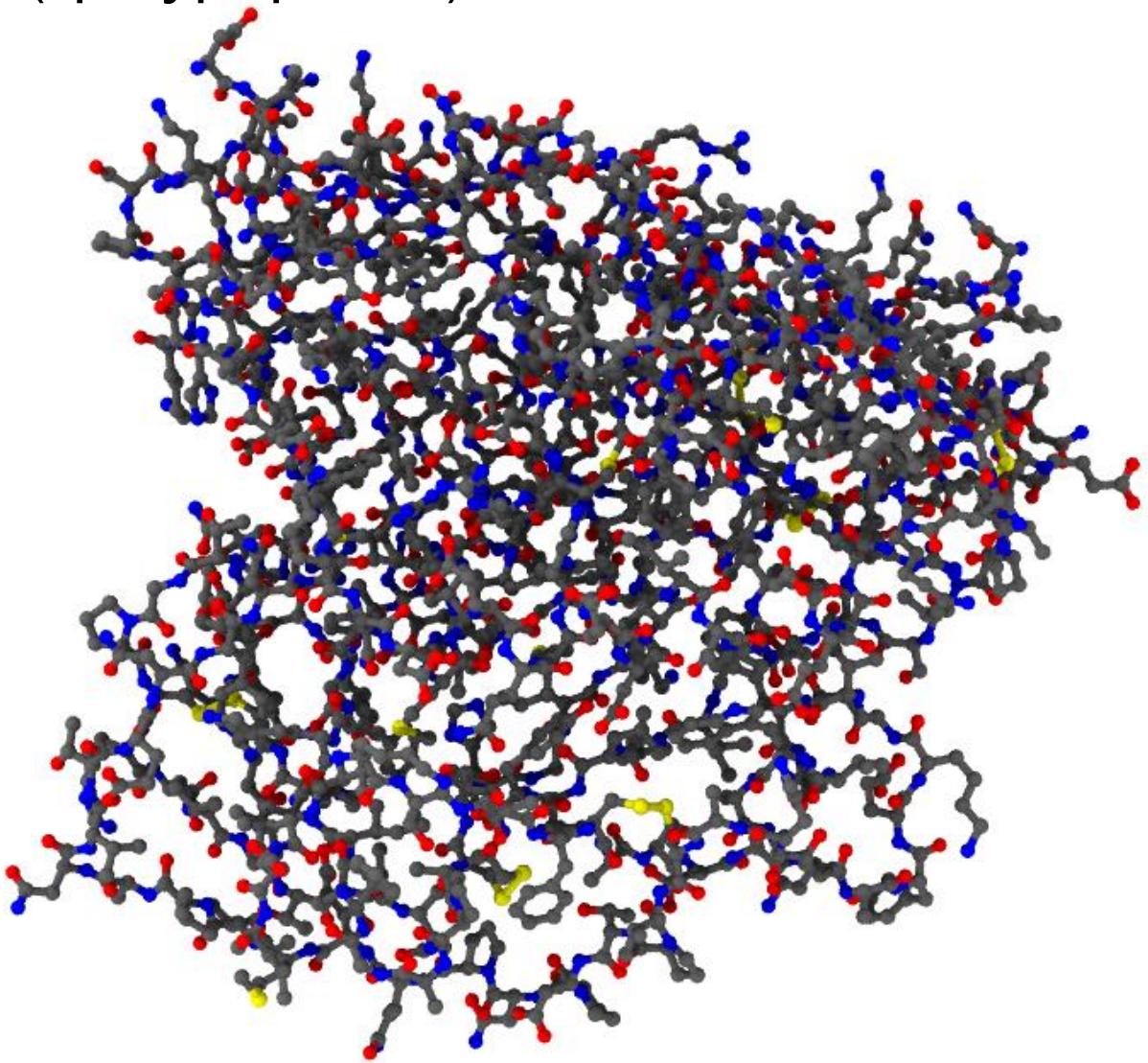


Proline
PRO
P

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

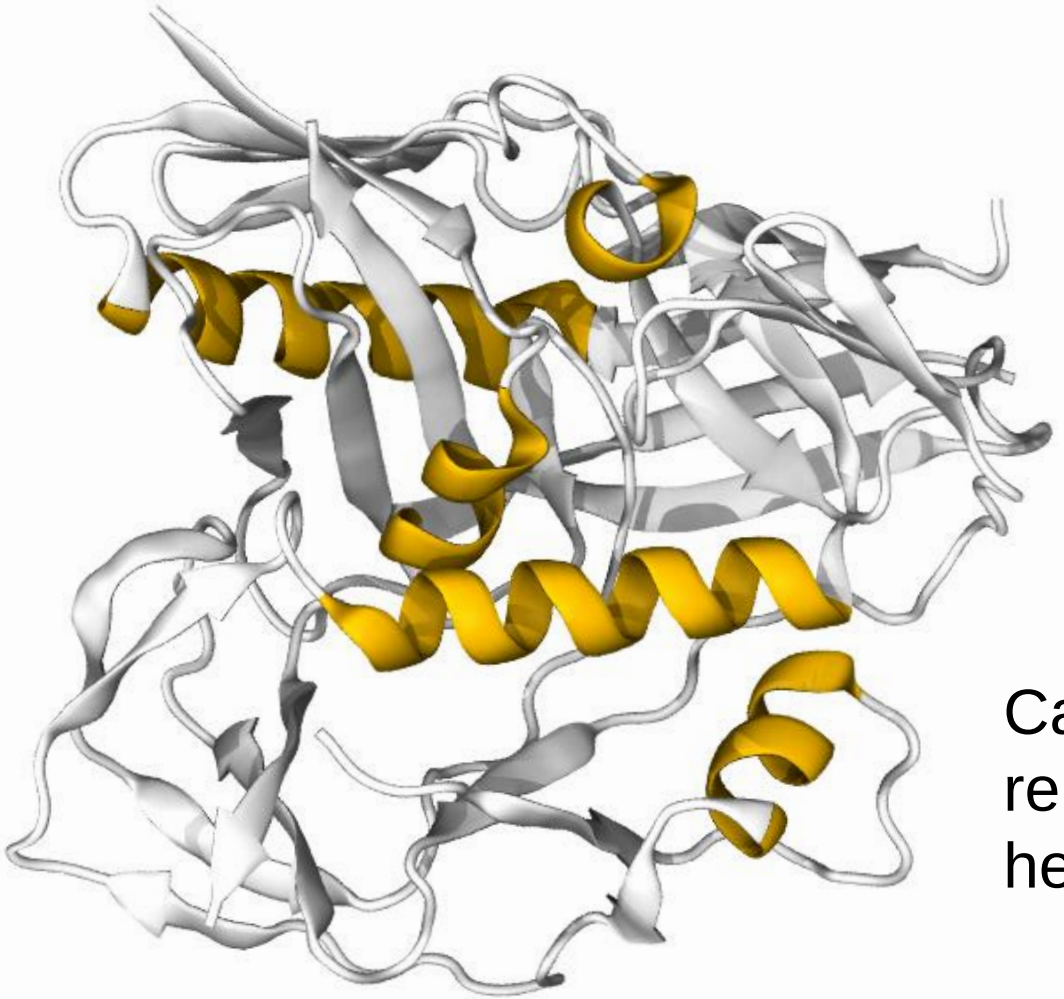
Protein tertiary structure: folding

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



Protein tertiary structure: folding

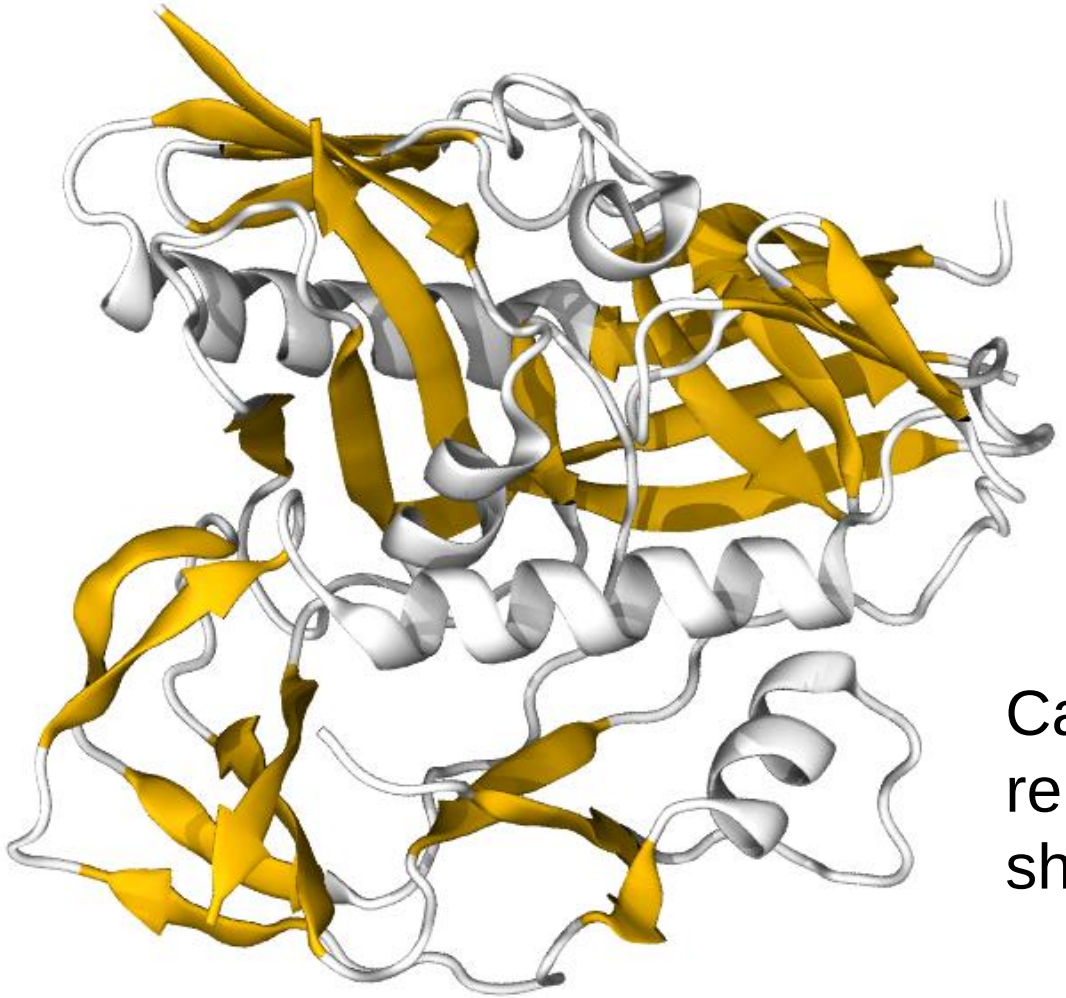
Protein («polypeptide»): 10 to >1000 amino acids



Cartoon representation: helices

Protein tertiary structure: folding

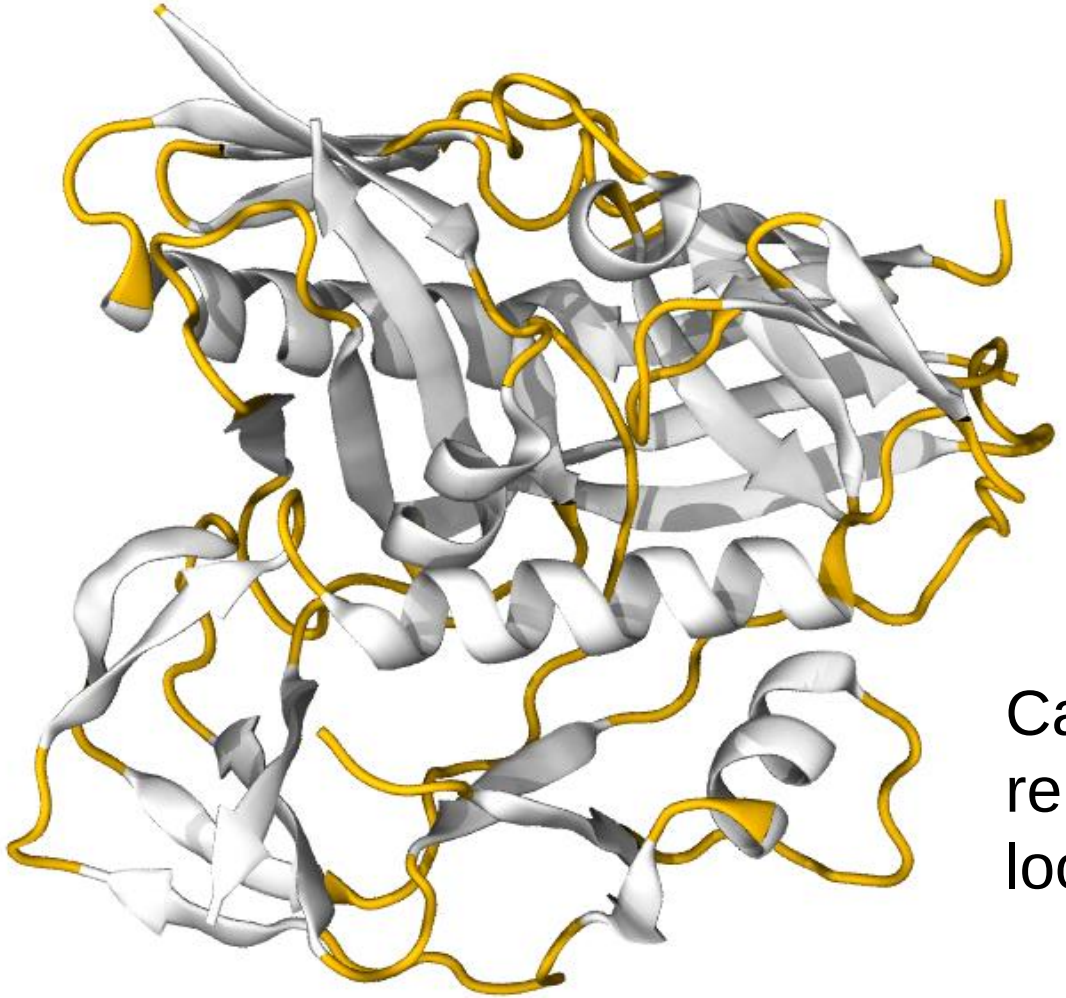
Protein («polypeptide»): 10 to >1000 amino acids



Cartoon representation: sheets

Protein tertiary structure: folding

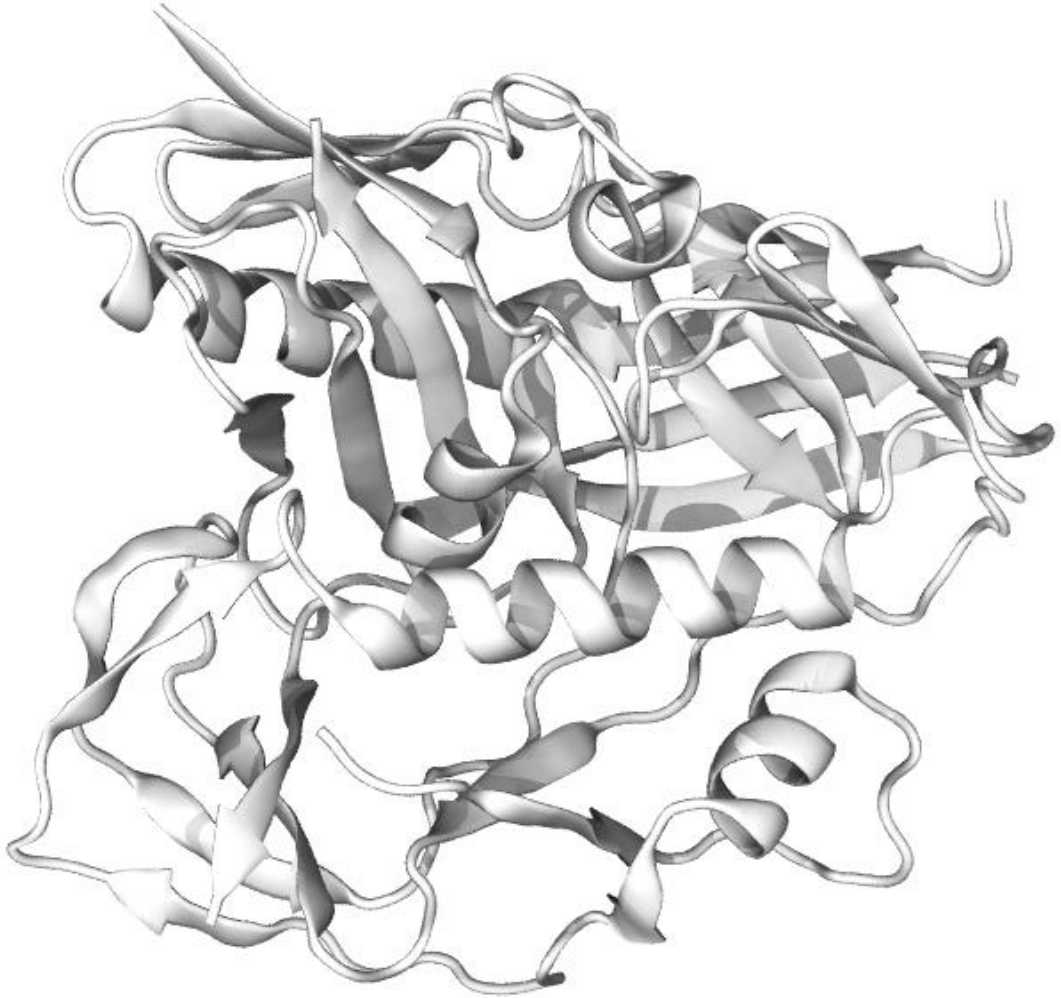
Protein («polypeptide»): 10 to >1000 amino acids



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

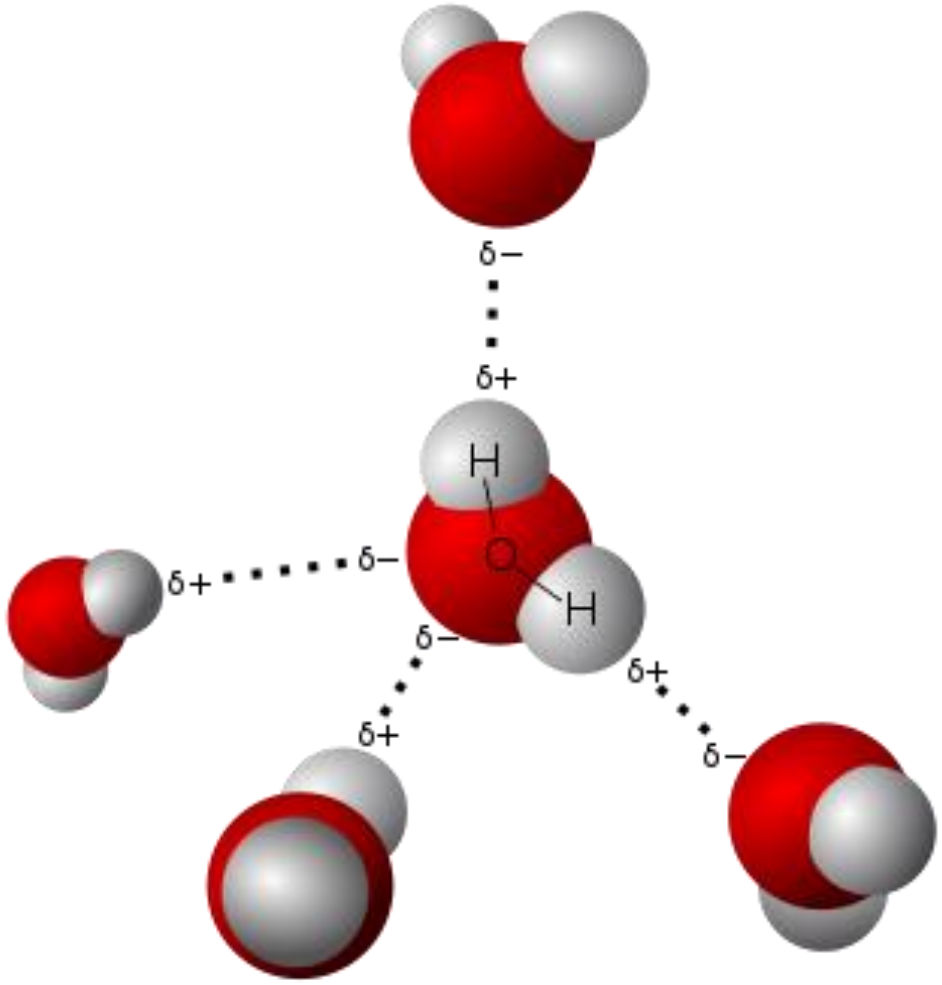
Electrostatic interaction. Structure:

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

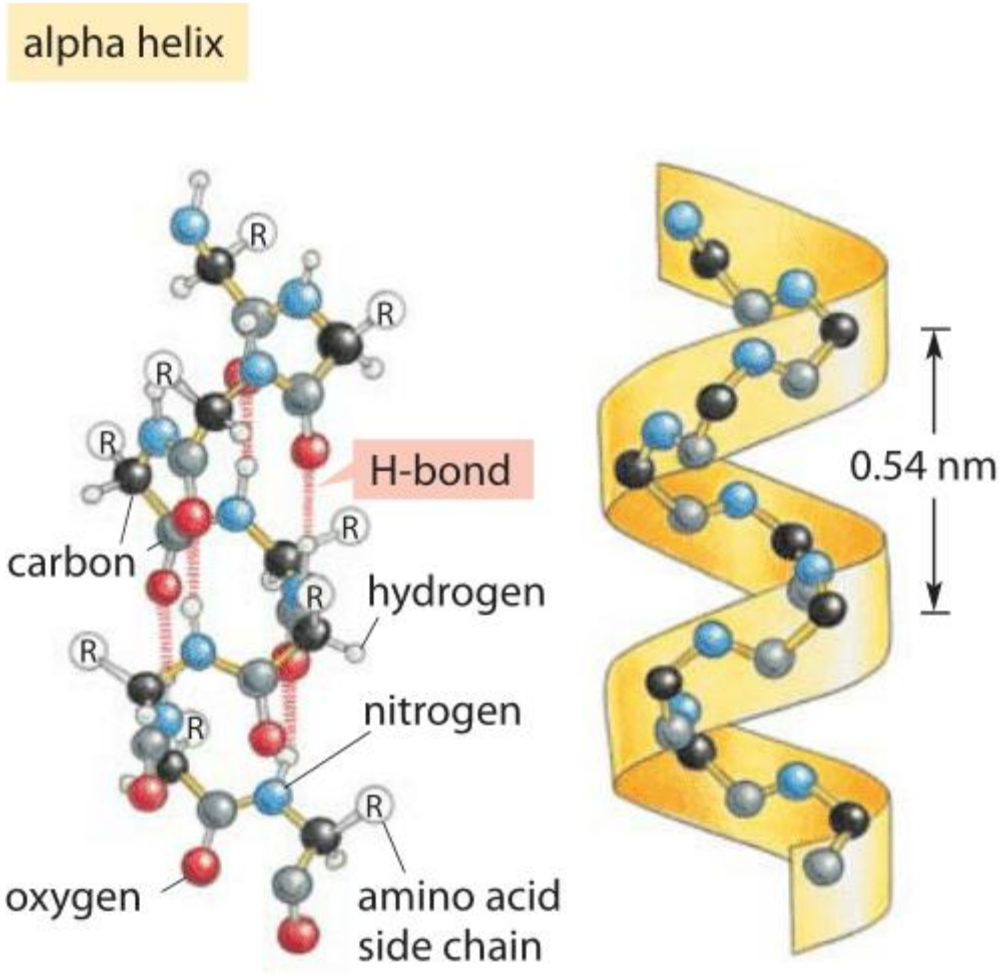
Hydrogen bond energy in biomolecular systems typically 5-25 kJ/mol, e.g.:

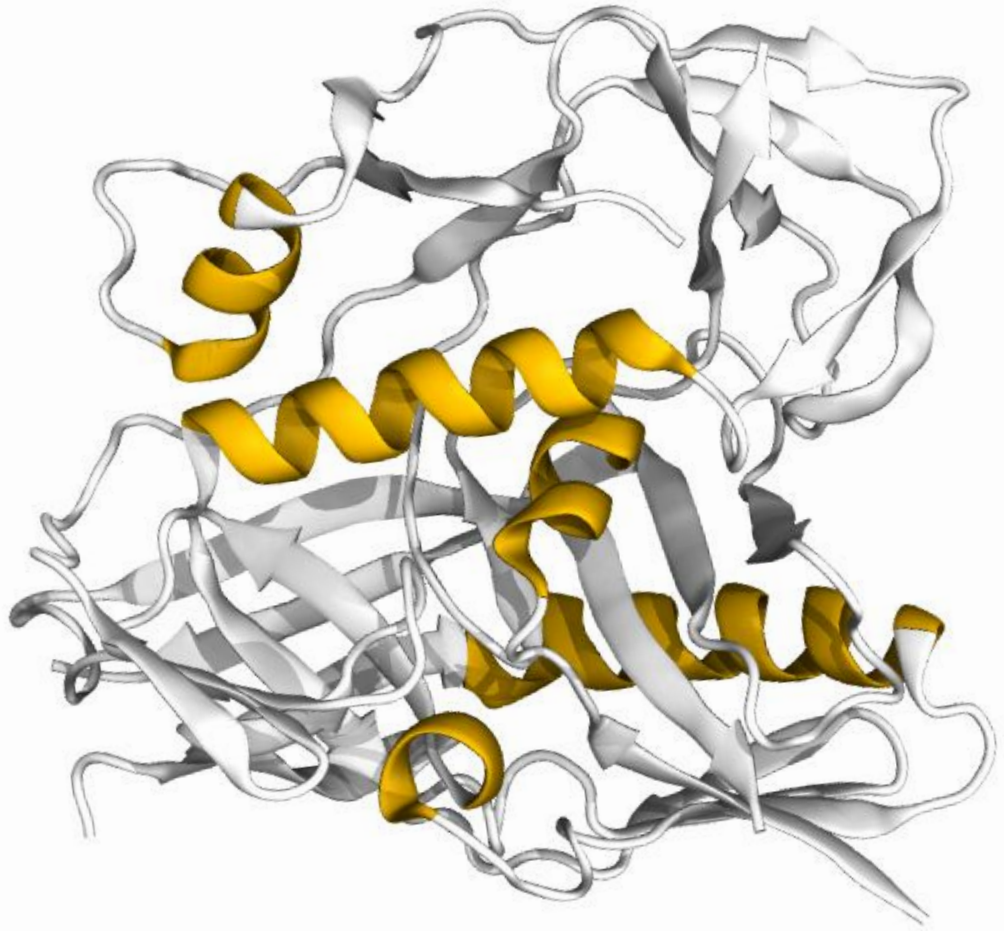
- O−H···:O , 21 kJ/mol (5.0 kcal/mol)
- N−H···:O , 8 kJ/mol (1.9 kcal/mol)

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



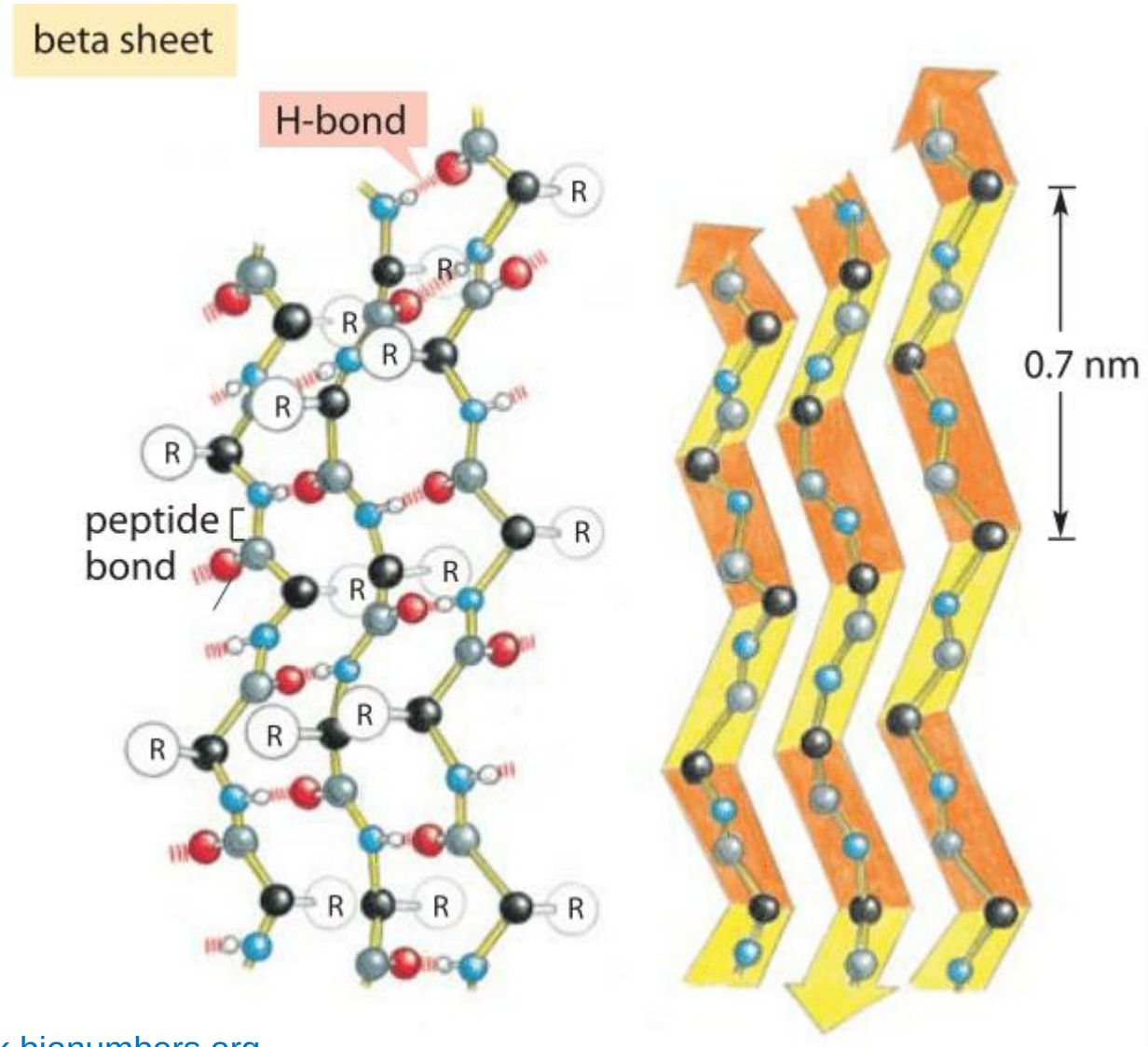
Hydrogen bonding on protein backbone





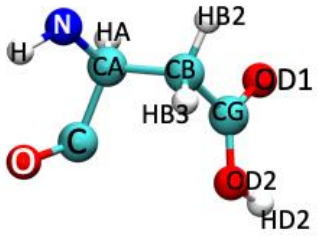
book.bionumbers.org

Hydrogen bonding on protein backbone

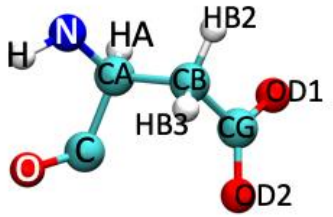


book.bionumbers.org

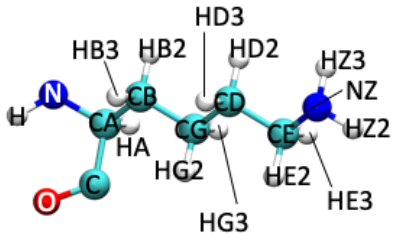
Hydrogen bonding: effect of *local* pH



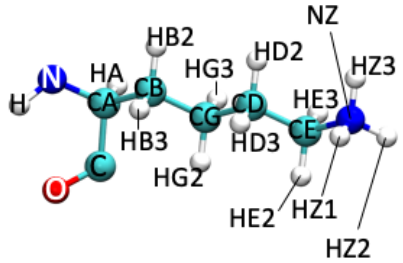
ASH, Aspartic acid*, D



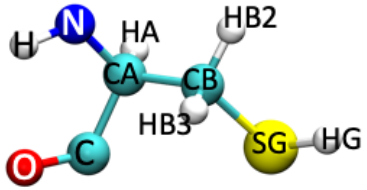
ASP, Aspartate, D



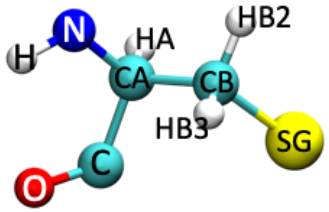
LYN, Lysine*, K



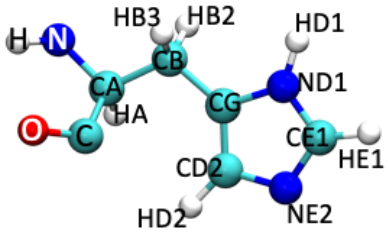
LYS, Lysine, K



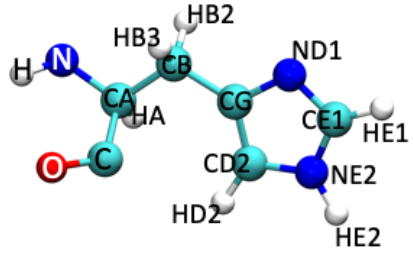
CYS, Cysteine, C



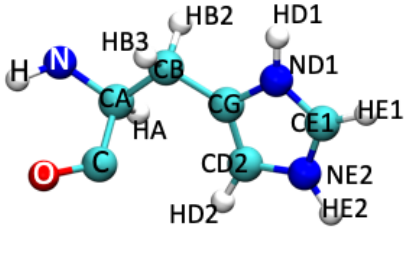
CYX, Cysteine**, C



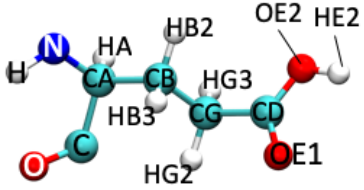
Histidine*, H



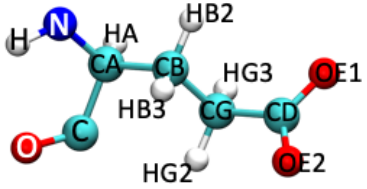
HIE, Histidine, H



HIP, Histidine*, H



GLH, Glutamic Acid*, E



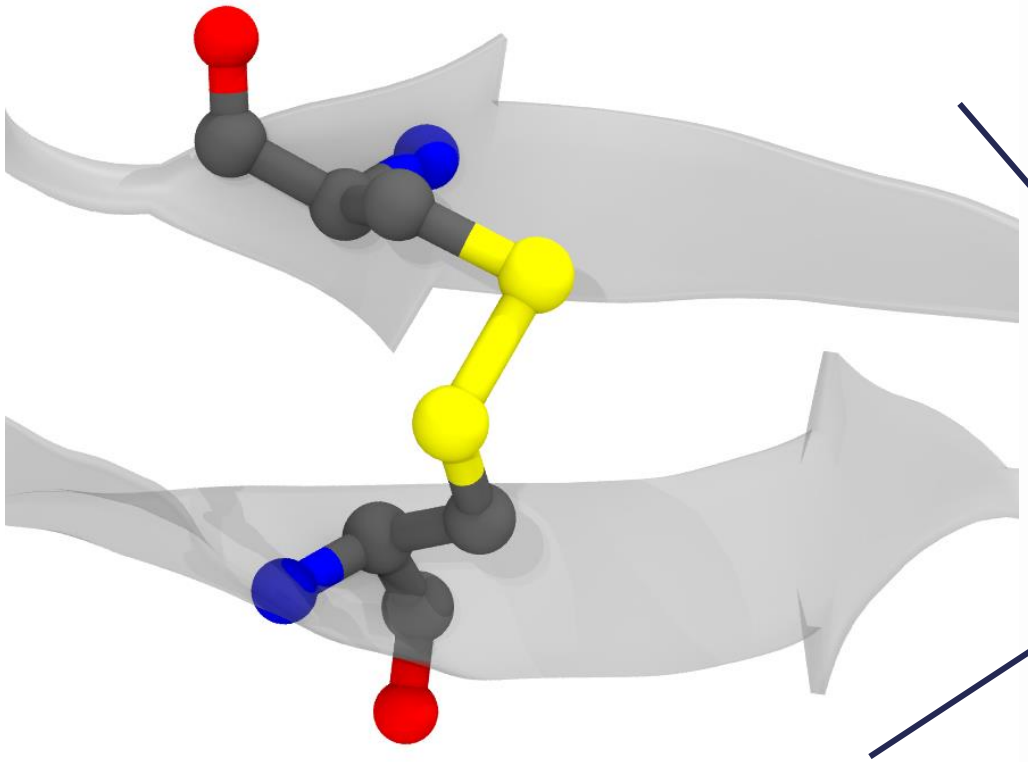
GLU, Glutamate, E

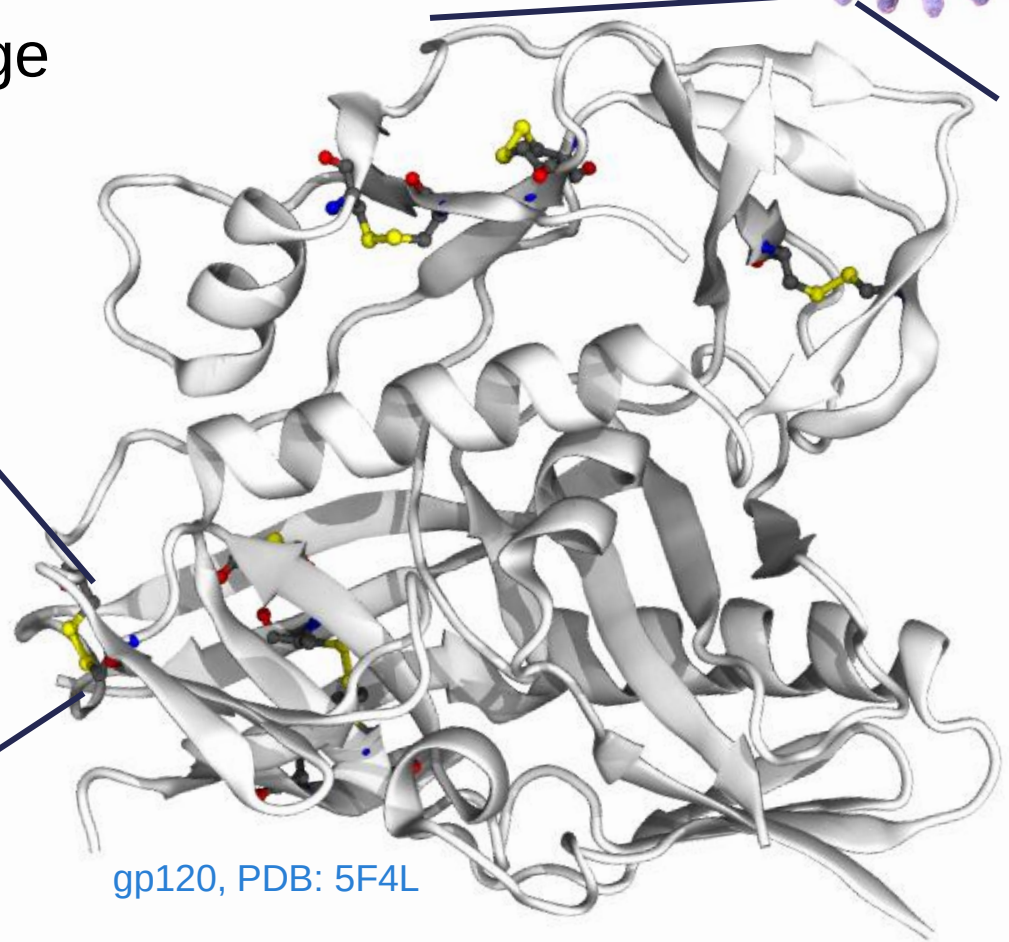
Disulfide bridges

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

- Protein more resistent to denaturation
- Changes in conditions ➡ structural change

HIV-1 virus





gp120, PDB: 5F4L

T.J.Bechtel and E.Weerapana, *From structure to redox: the diverse functional roles of disulfides and implications in disease*, ***Proteomics***, 2017

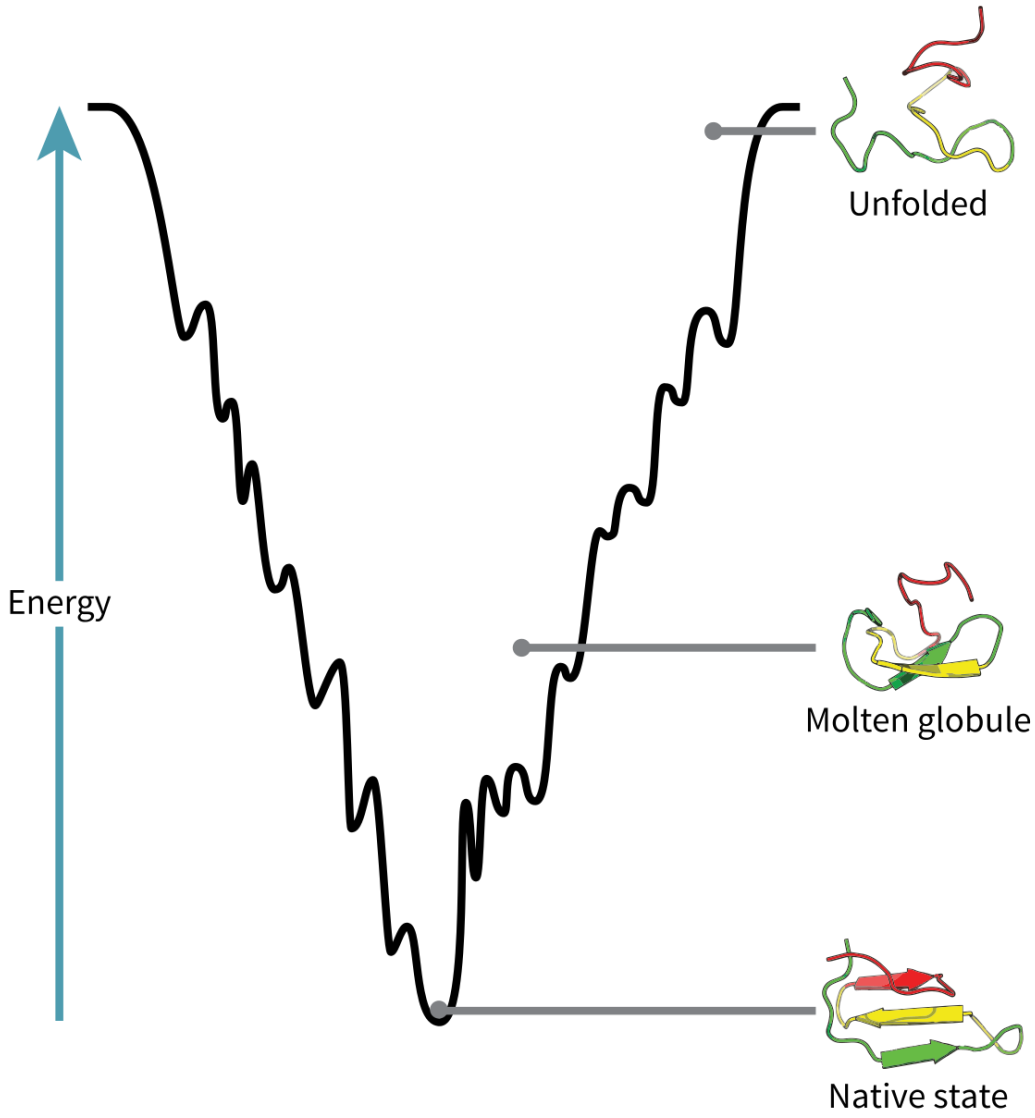
Proteins fold in low energy structures

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

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

$\Delta S_{protein}$ is < 0

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



Proteins fold in low energy structures

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

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

$\Delta S_{protein}$ is < 0

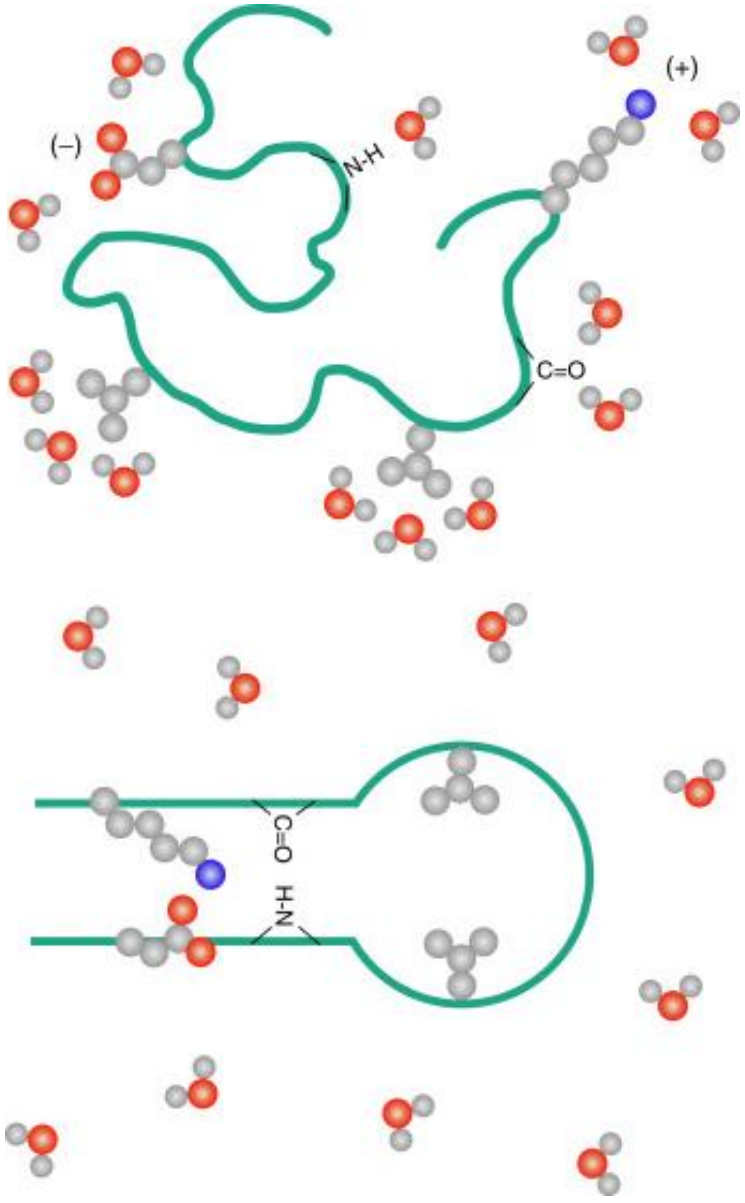
$\Delta S = \Delta S_{protein} + \Delta S_{water} > 0$

H = U + PV; at physiological conditions

$\Delta H \approx \Delta U$

ΔH 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
- 2x100 - 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 ≈ 13.7 billion years = **4.32×10^{17} s**

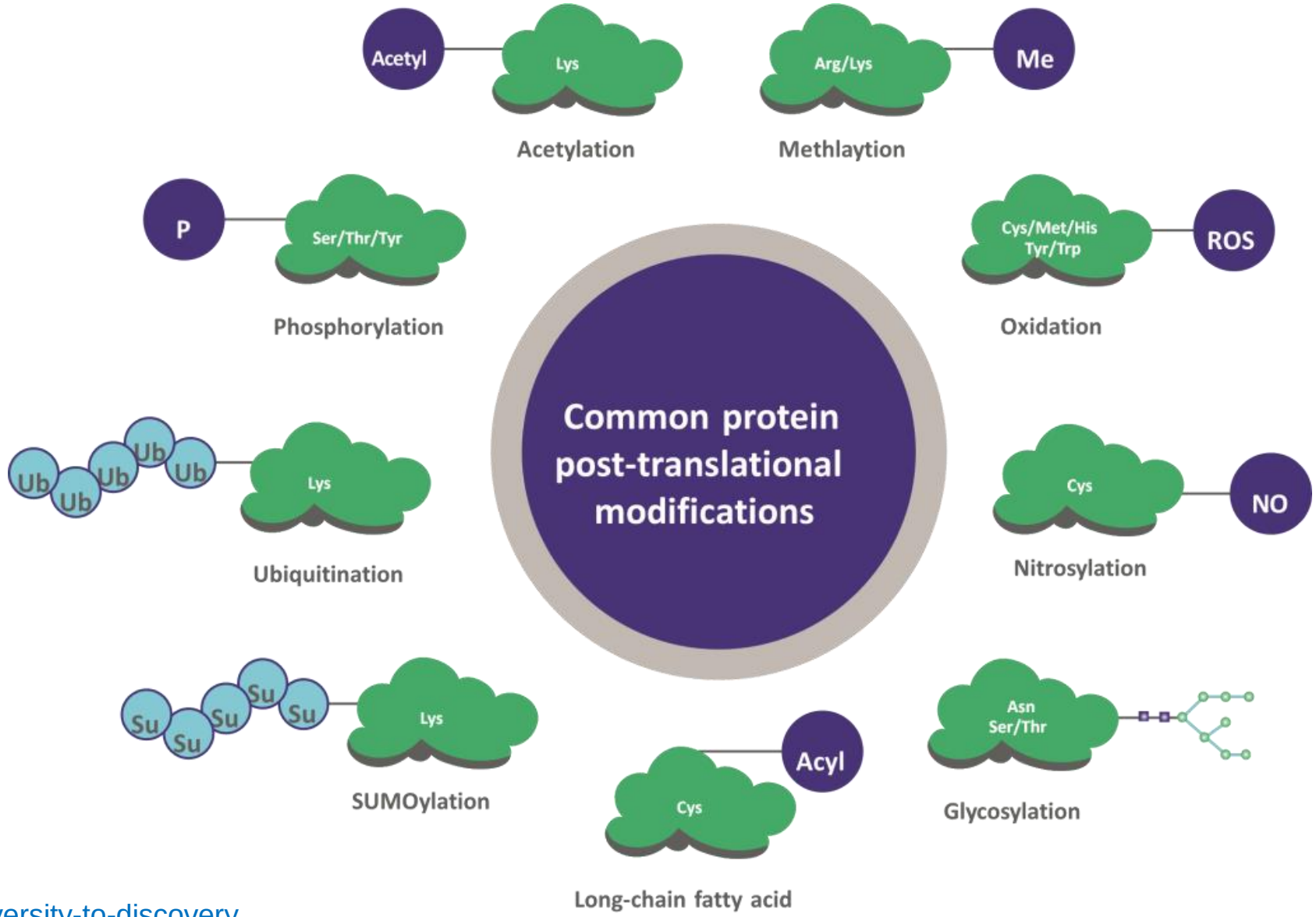
Protein folding cannot be a random process

Cyrus Levinthal. *How to Fold Graciously*. Mossbauer Spectroscopy in Biological Systems: Proceedings of a meeting held at Allerton House, Monticello, Illinois, 1969

Post-translational modifications (PTMs)

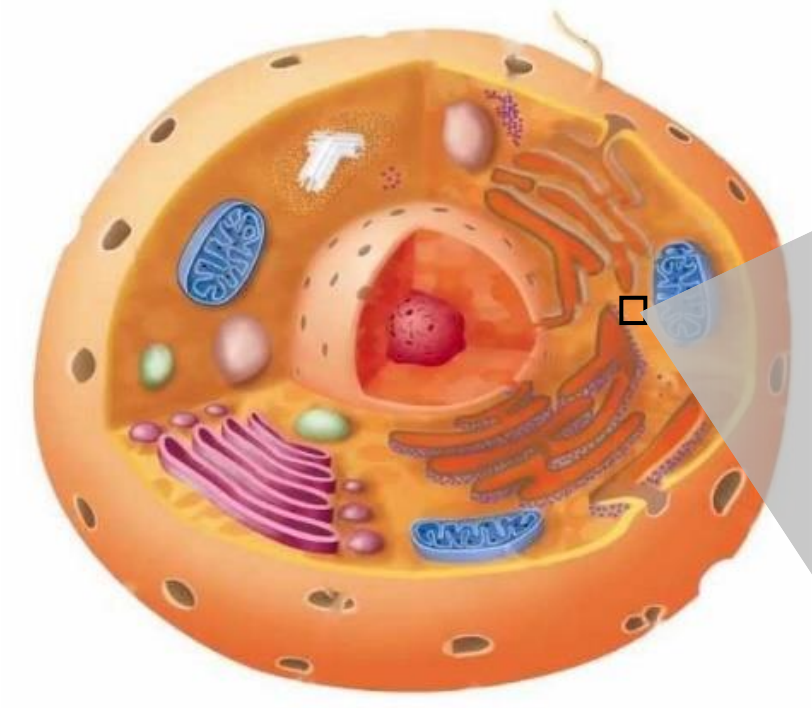
Once expressed, proteins can be modified by the covalent binding of other molecules.

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

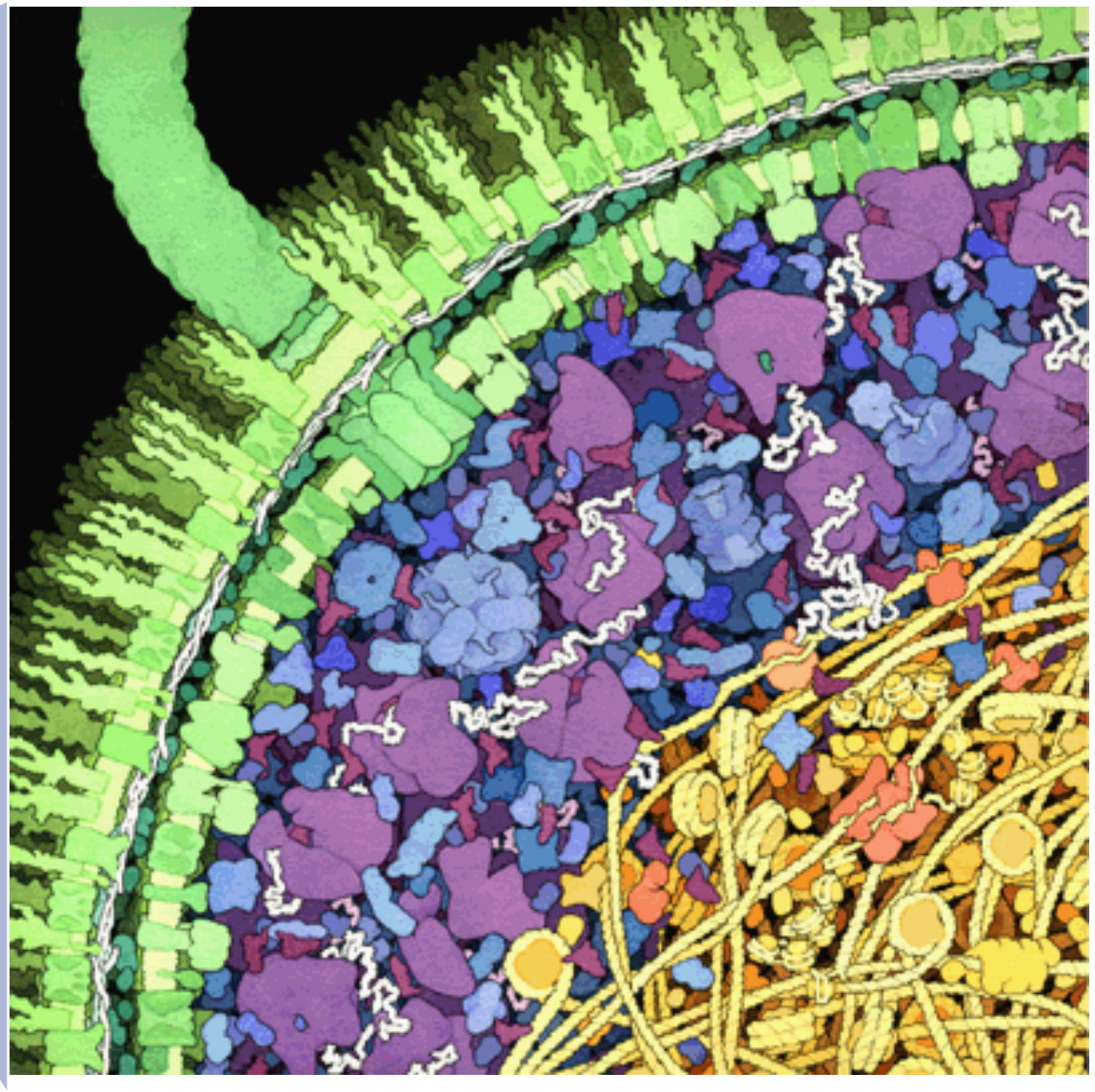


www.raybiotech.com/learning-center/translating-diversity-to-discovery

The intracellular space

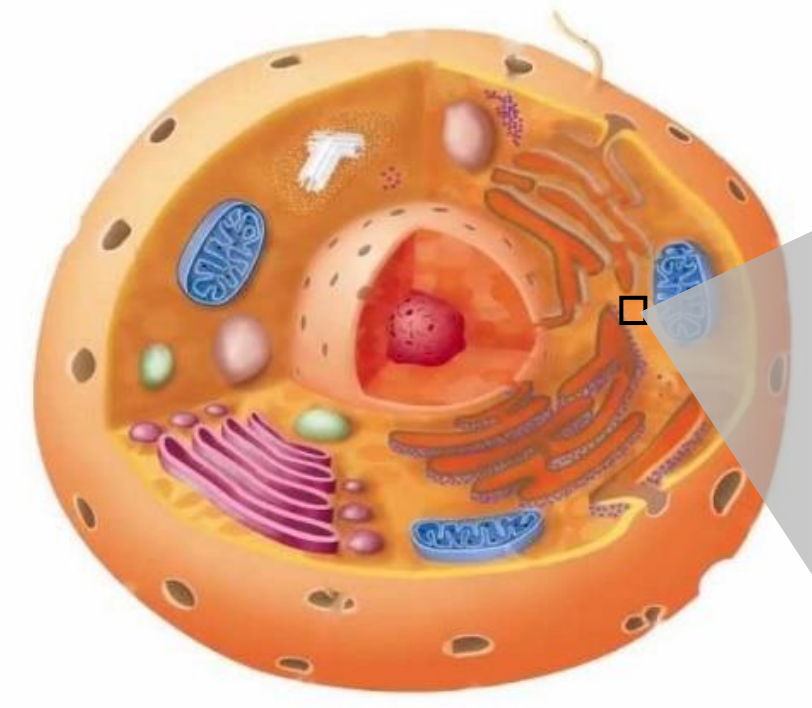


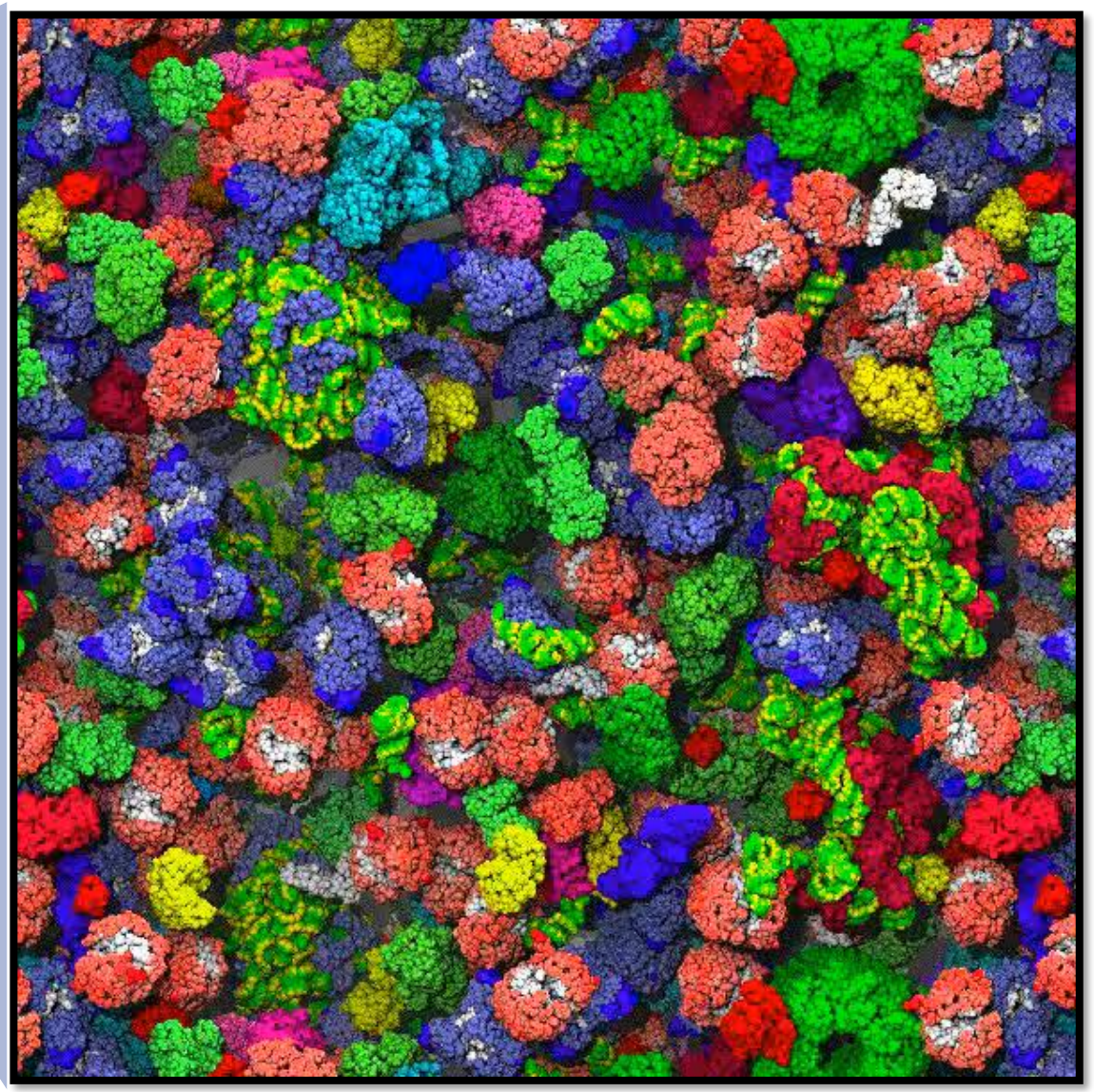
Crowded environment!



David G. Goodsell, *The machinery of Life*, Springer

The intracellular space





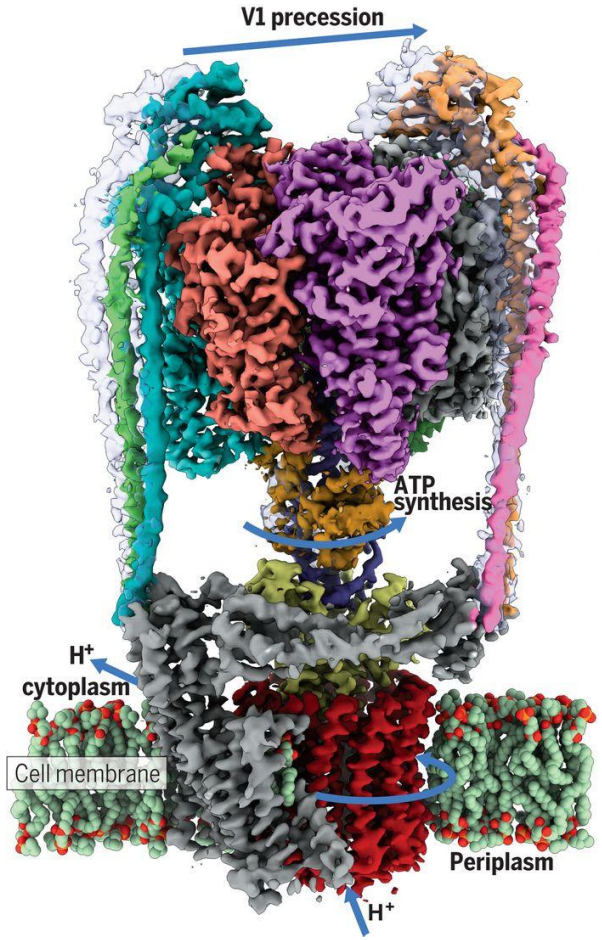
Crowded environment!

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

Most contacts are short-lived (ns-μs)

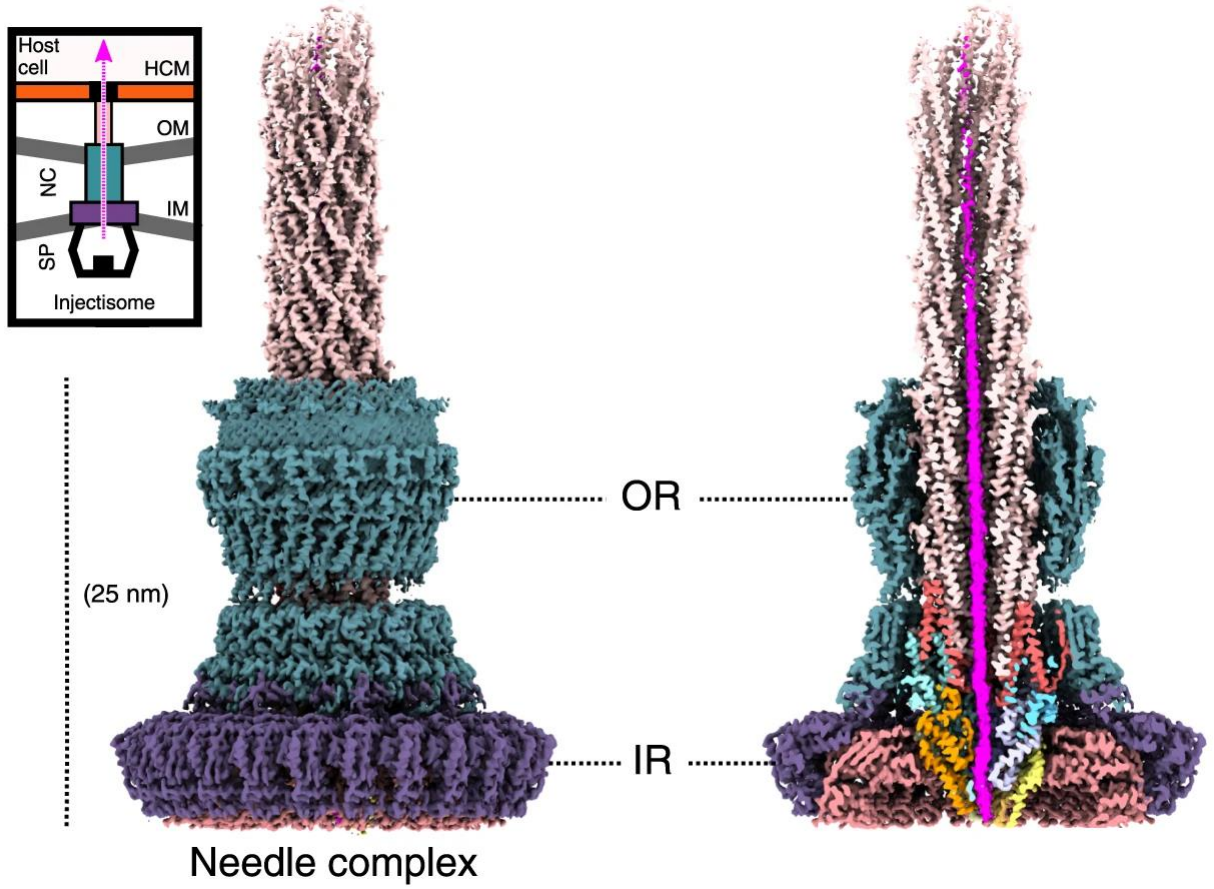
S. McGuffee et al., ***PLoS Comp. Biol.****, 2010*

Protein quaternary structure: assembly



ATP syntase

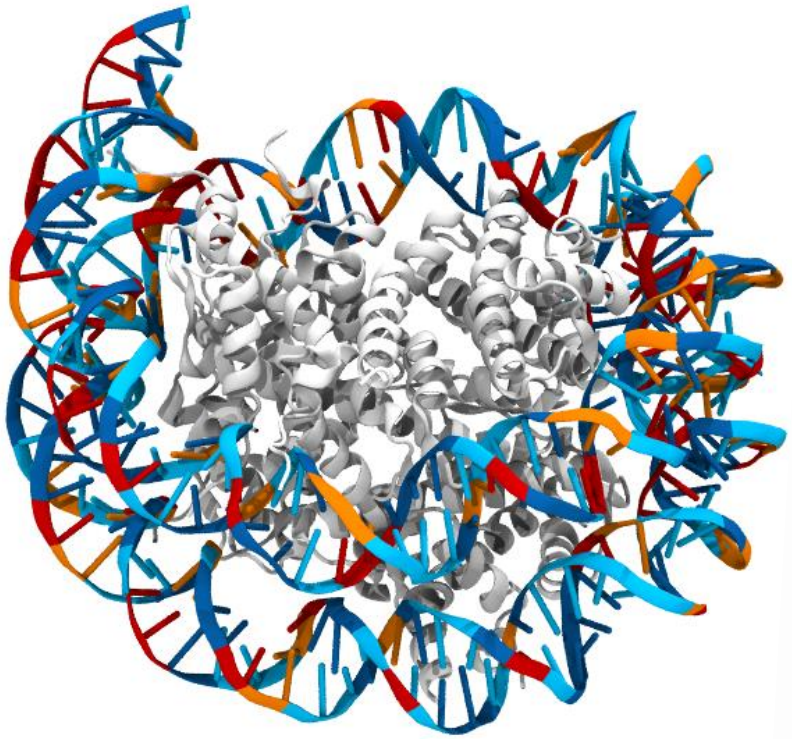
L. Zhou and L. A. Sazanov, *Structure and conformational plasticity of the intact Thermus thermophilus V/A-type ATPase,* ***Science****, 2019*



Injectisome

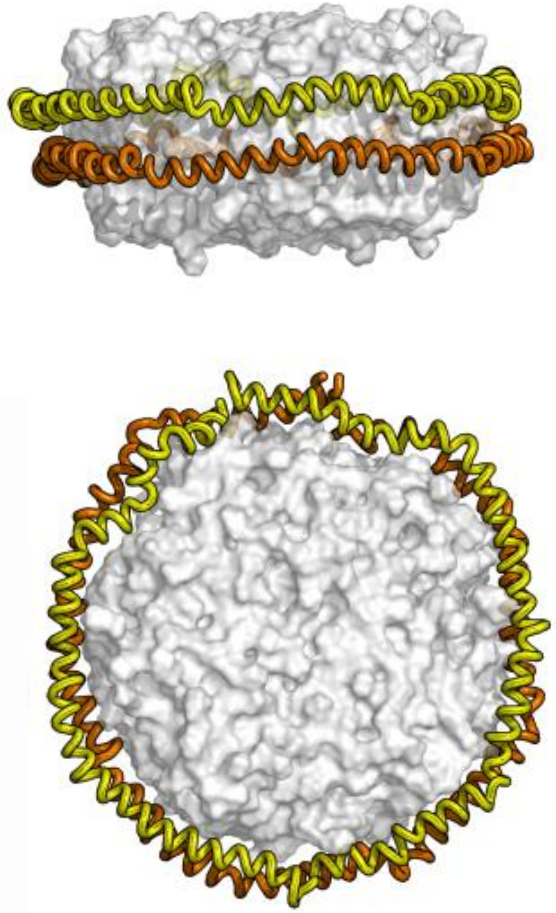
S. Miletic *et* al., *Substrate-engaged type III secretion system structures reveal gating mechanism for unfolded protein translocation****, Nature Comms***, 2021

The structure determines the function



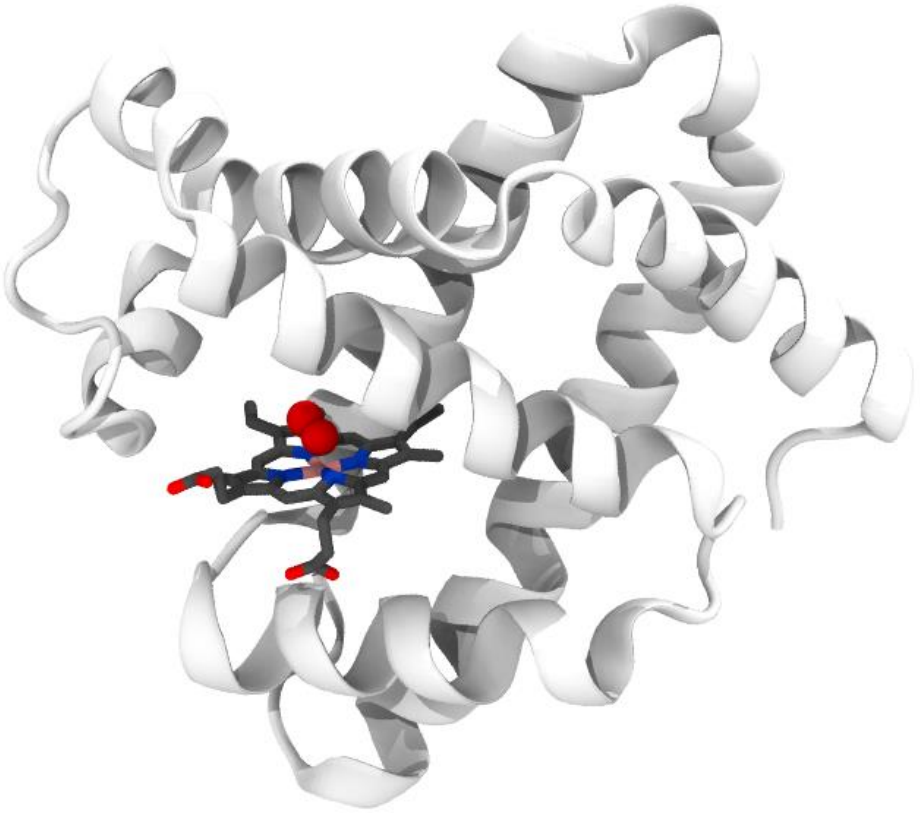
Nucleosome (PDB: 5CPI)

K. Luger et al., *Crystal Structure of the nucleosome core particle at 2.8 A 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

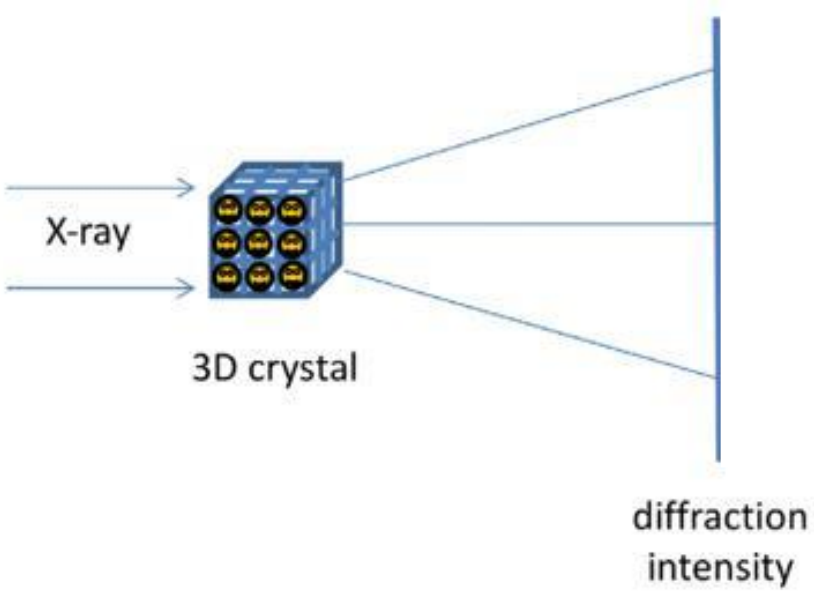
- **Proteins in diseases** (e.g., COVID-19, salmonella, flu, ...)
  - pathogen's own metabolism/structure
  - pathogen's weapon

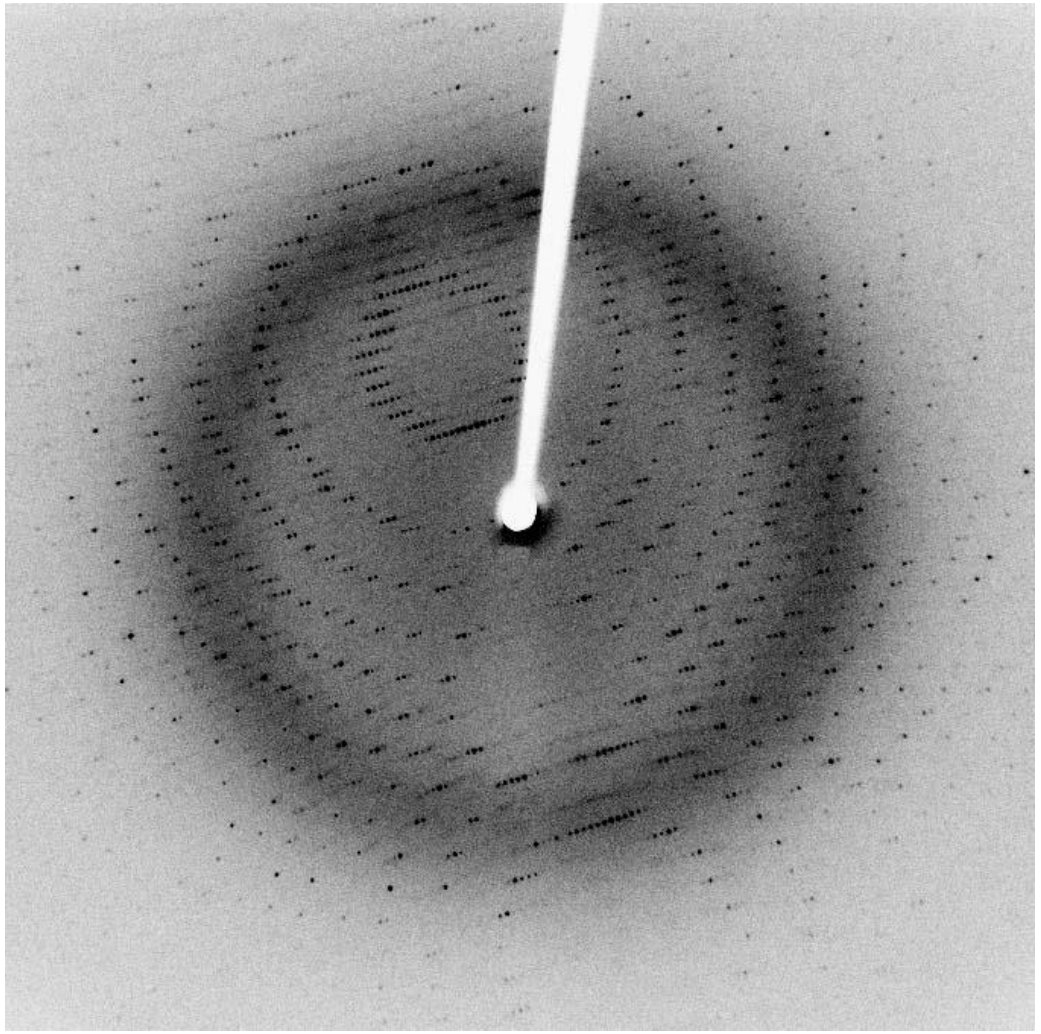
- **Proteins in disorders** (e.g., Cancer, Alzheimer's, ...)
  - own protein misfolds
  - own protein folds, but has different dynamics

- **Drug**
  - Small molecule designed to specifically bind to a protein, so as to affect its function

Part 2: structure determination

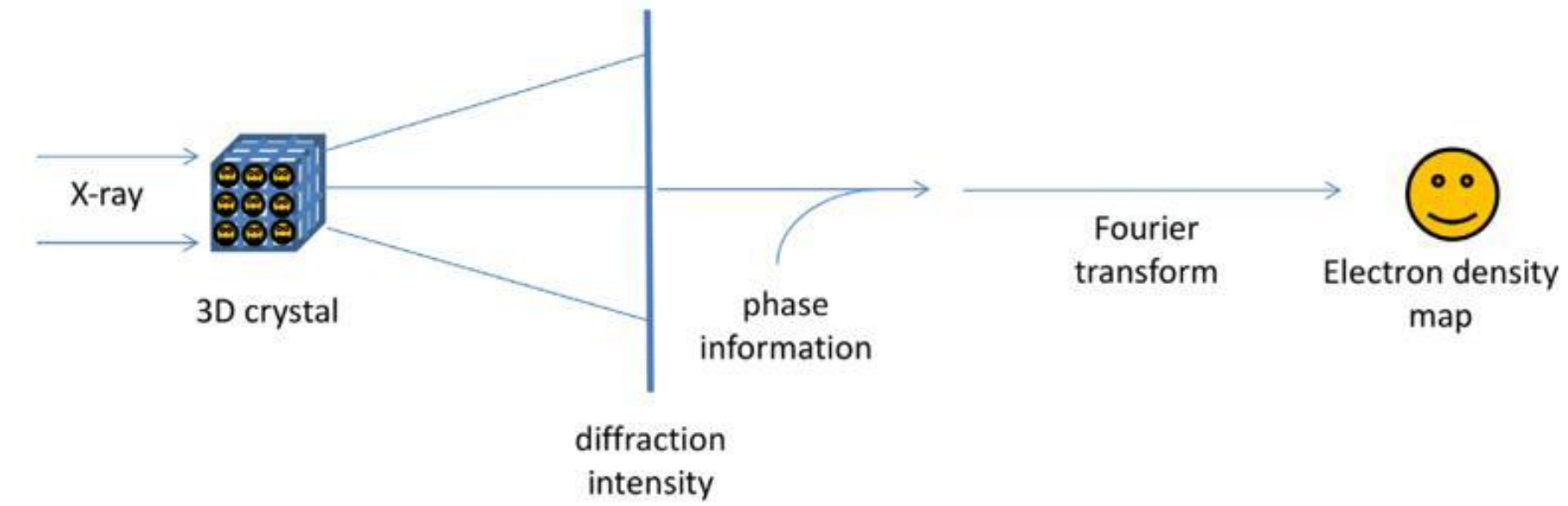
Structure determination: X-ray crystallography

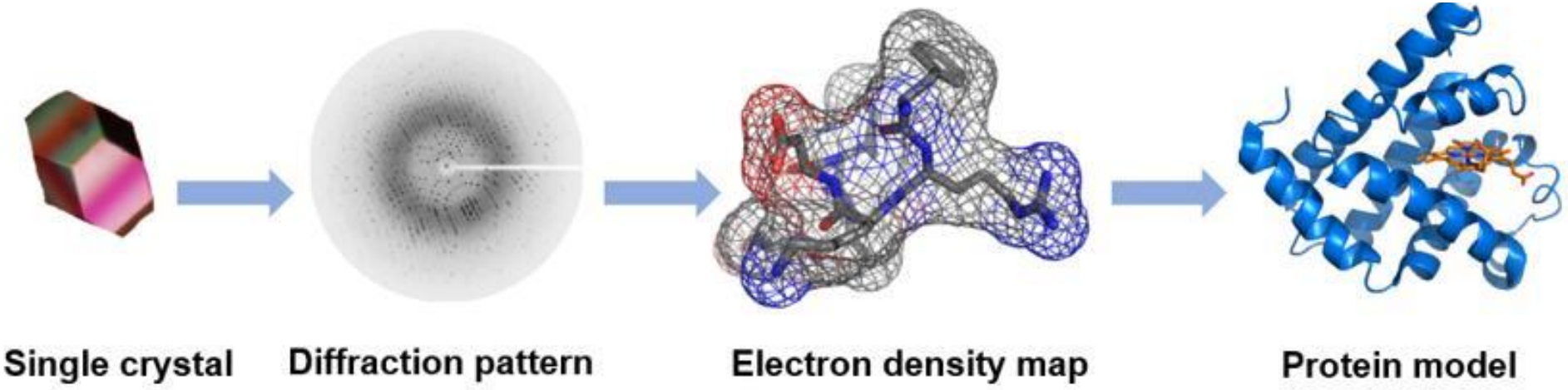




www.creative-biostructure.com/comparison-of-crystallography-nmr-and-em_6.htm

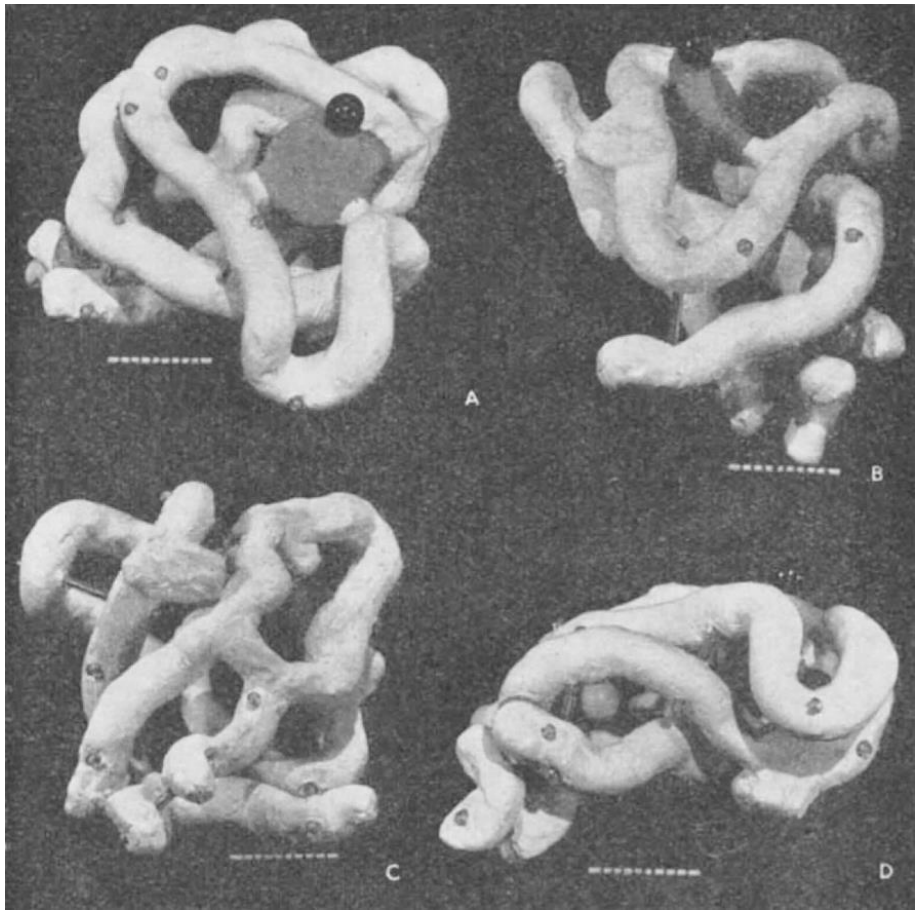
Structure determination: X-ray crystallography

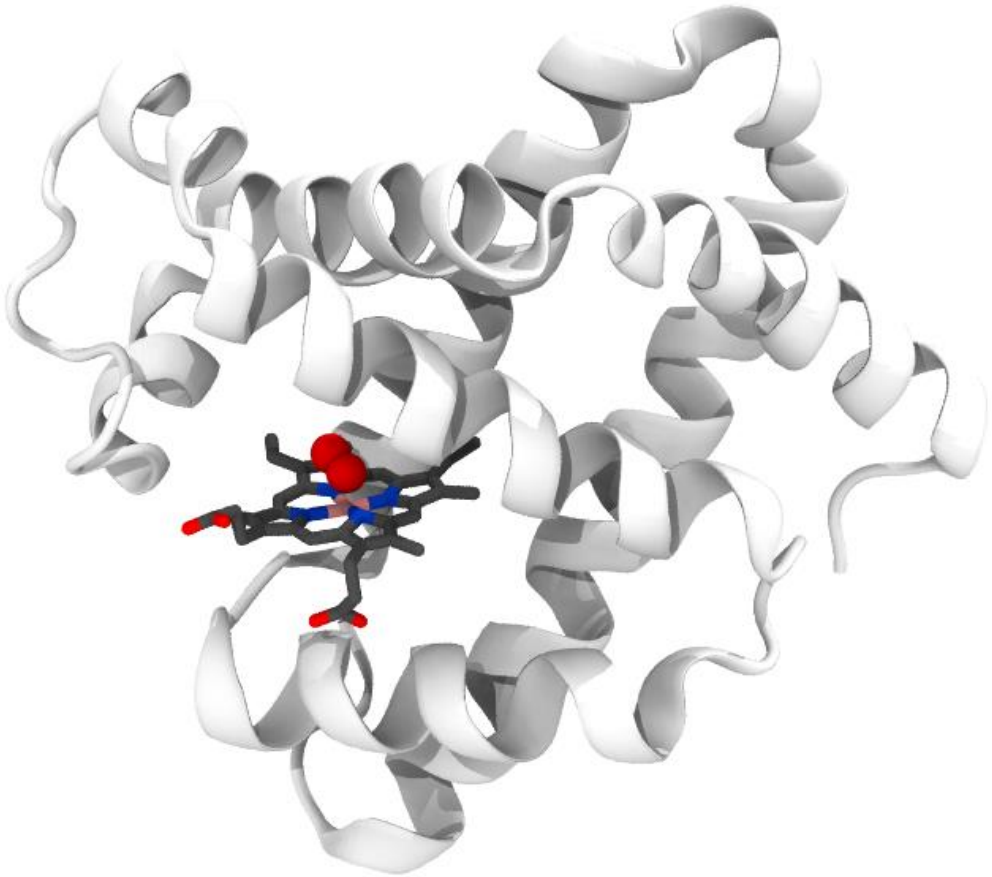




www.creative-biostructure.com/comparison-of-crystallography-nmr-and-em_6.htm

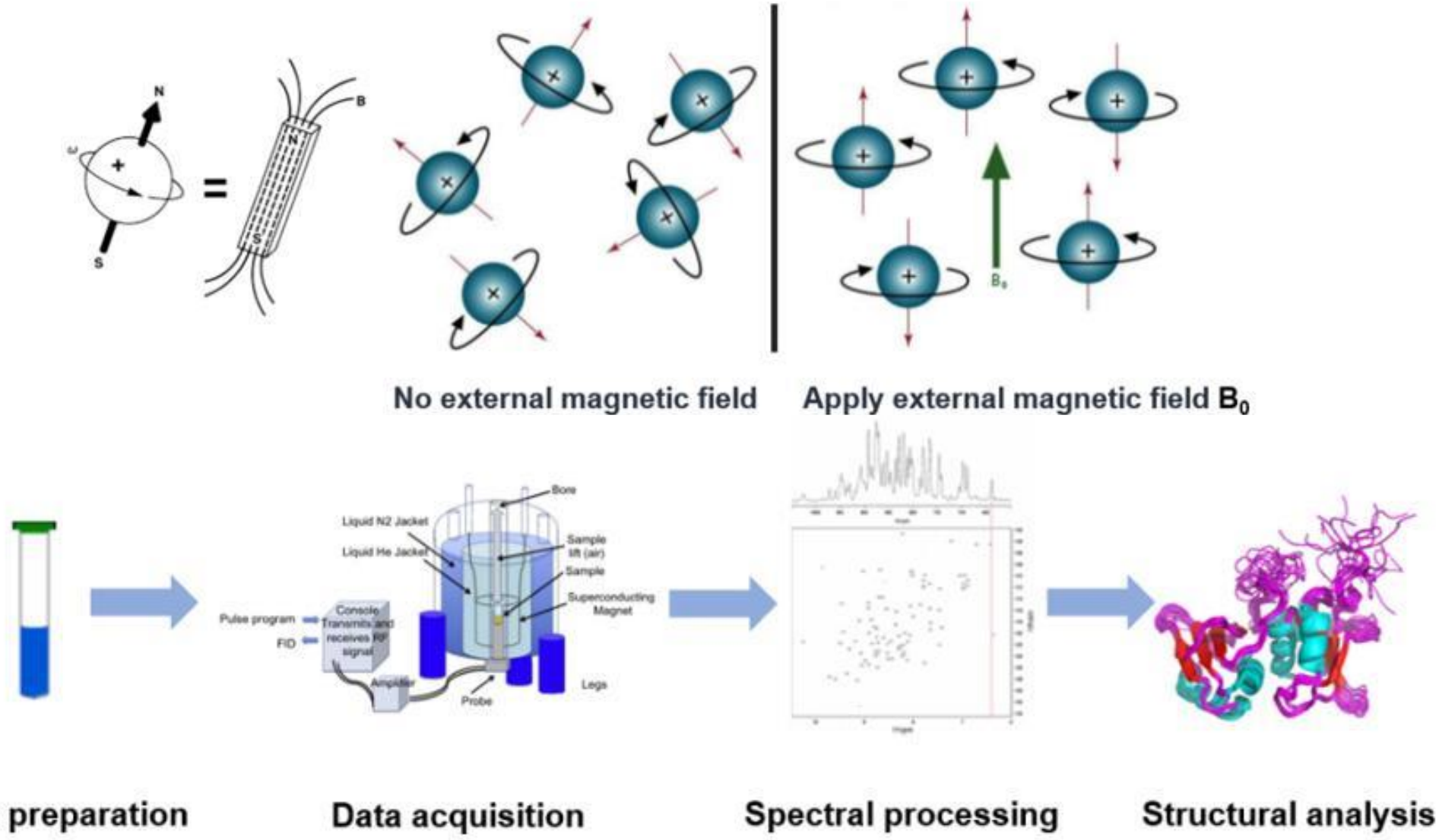
Structure determination: X-ray crystallography





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

Structure determination: Nuclear Magnetic Resonance (NMR)



www.creative-biostructure.com/comparison-of-crystallography-nmr-and-em_6.htm

[Extra] Structure determination: Electron Microscopy (EM)

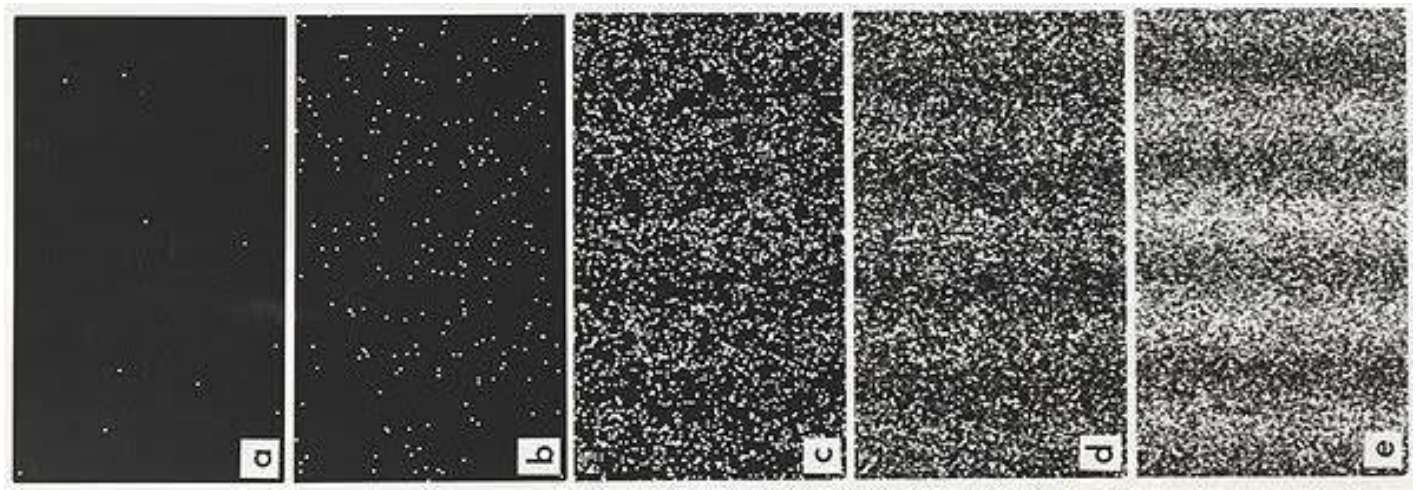
- Particles as «waves that transfers energy and momentum»

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

λ : wavelength
p : momentum
h : Planck constant

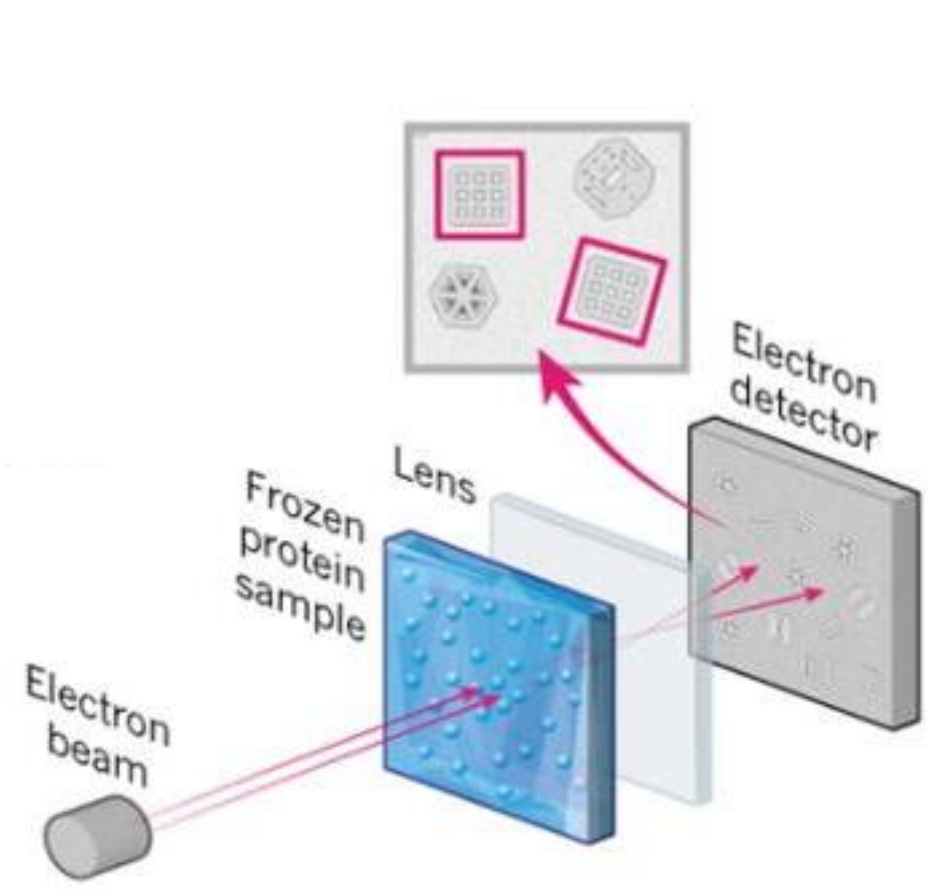
Louis De Broglie
1892-1987

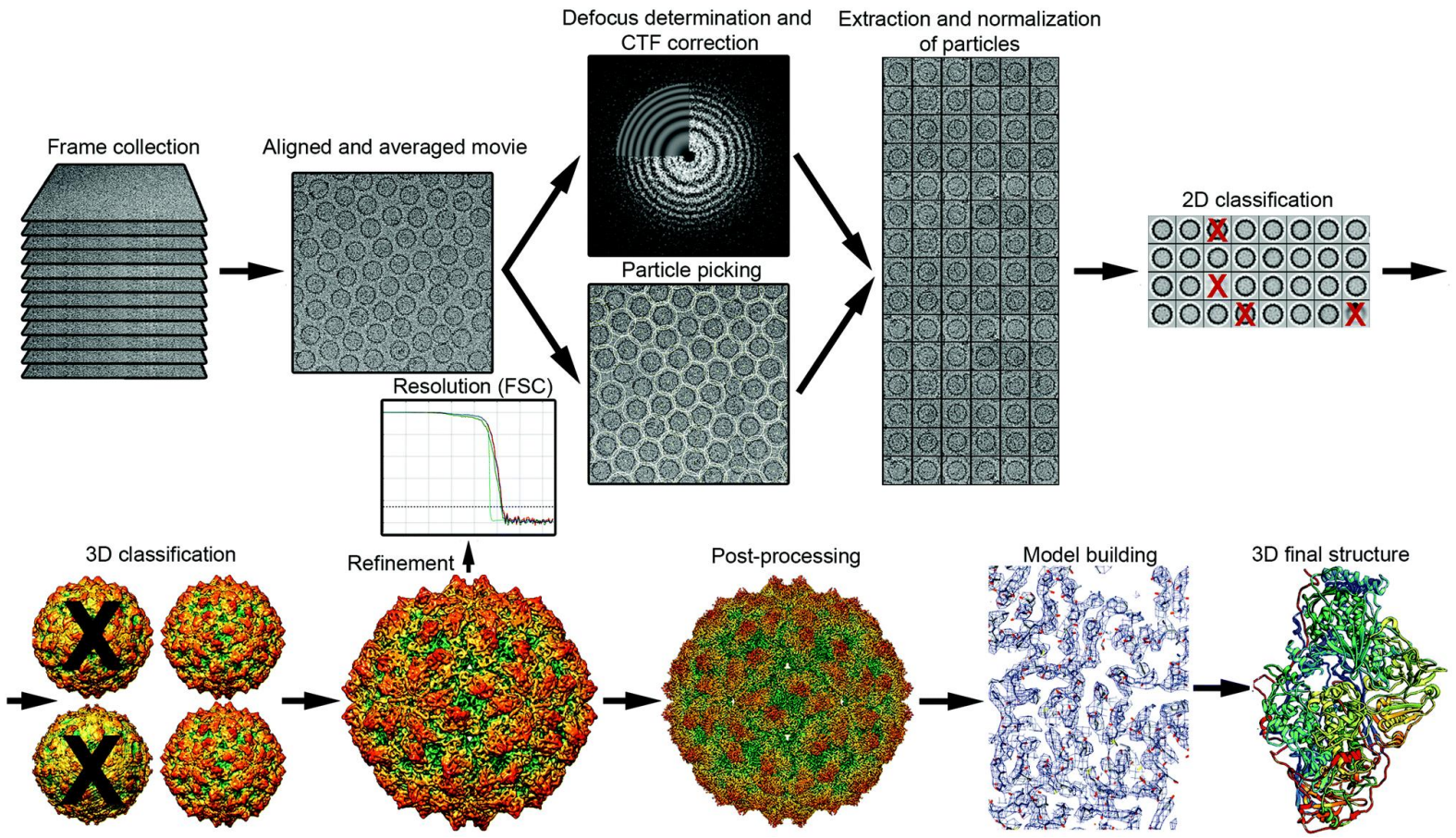
- Davisson–Germer experiment: electrons diffract too!



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

Structure determination: Electron Microscopy (EM)

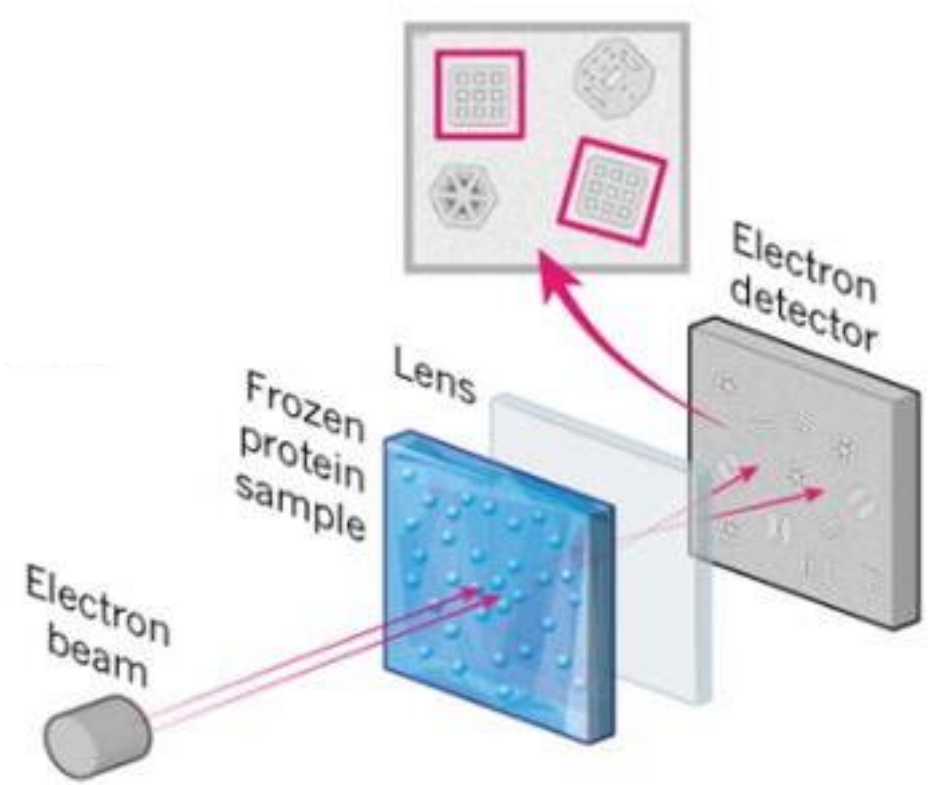


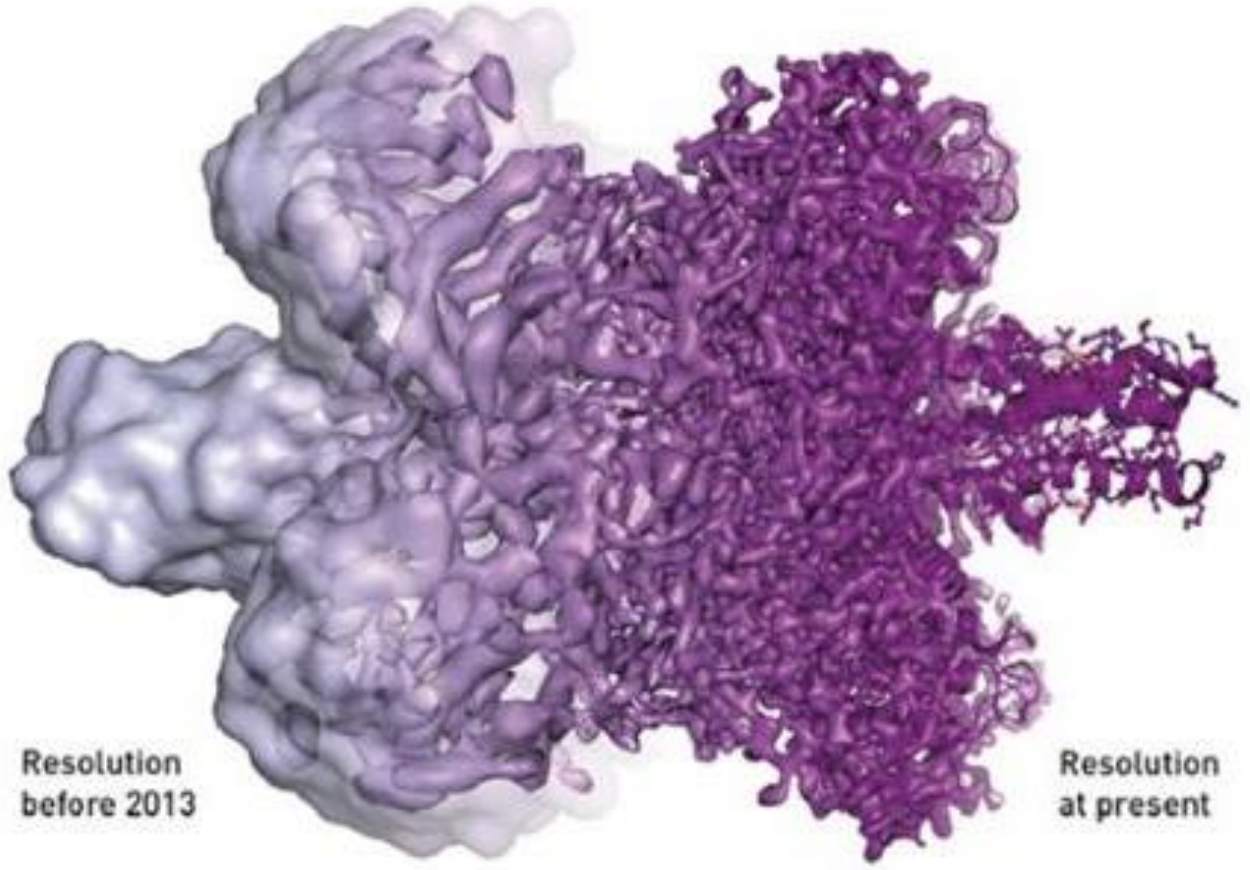


M. V. Ruiter *et al.*, *Structural nanotechnology: three-dimensional cryo-EM and its use in the development of nanoplatforms for in vitro catalysis,* ***Nanoscale***, 2019

Structure determination: Electron Microscopy (EM)

Resolution Revolution

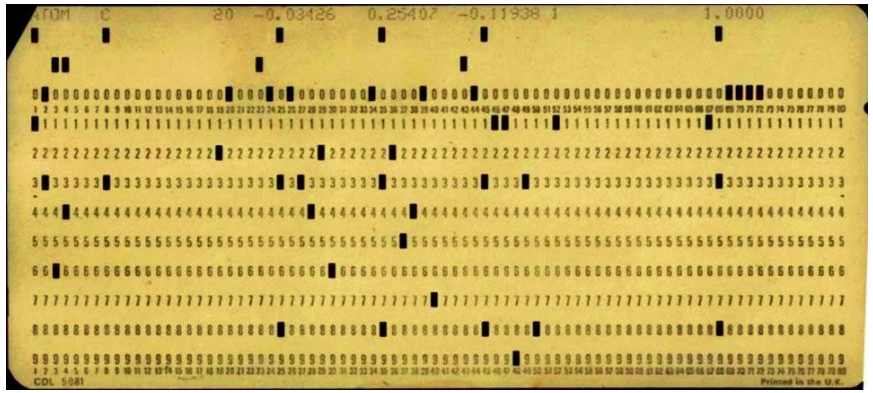




www.creative-biostructure.com/comparison-of-crystallography-nmr-and-em_6.htm

The Protein Data Bank (PDB)

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



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

RCSB PDB PROTEIN DATA BANK — www.rcsb.org

PDBe Protein Data Bank in Europe — www.ebi.ac.uk/pdbe

PDBj Protein Data Bank Japan — www.pdbj.org

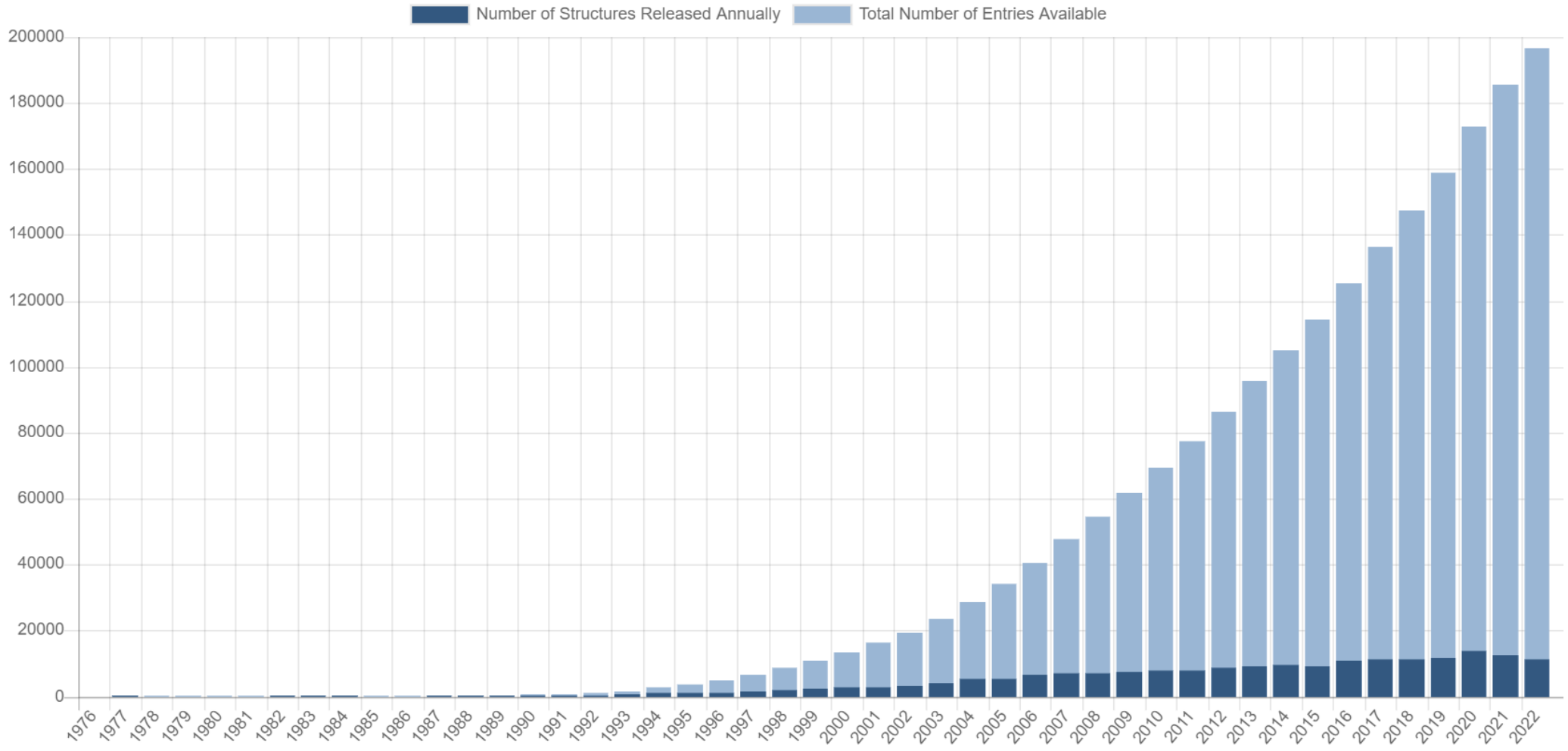
BMRB — www.bmrb.io

- Each molecule assigned a unique **4-characters code** (e.g., 1MBO, 1AV1, …)

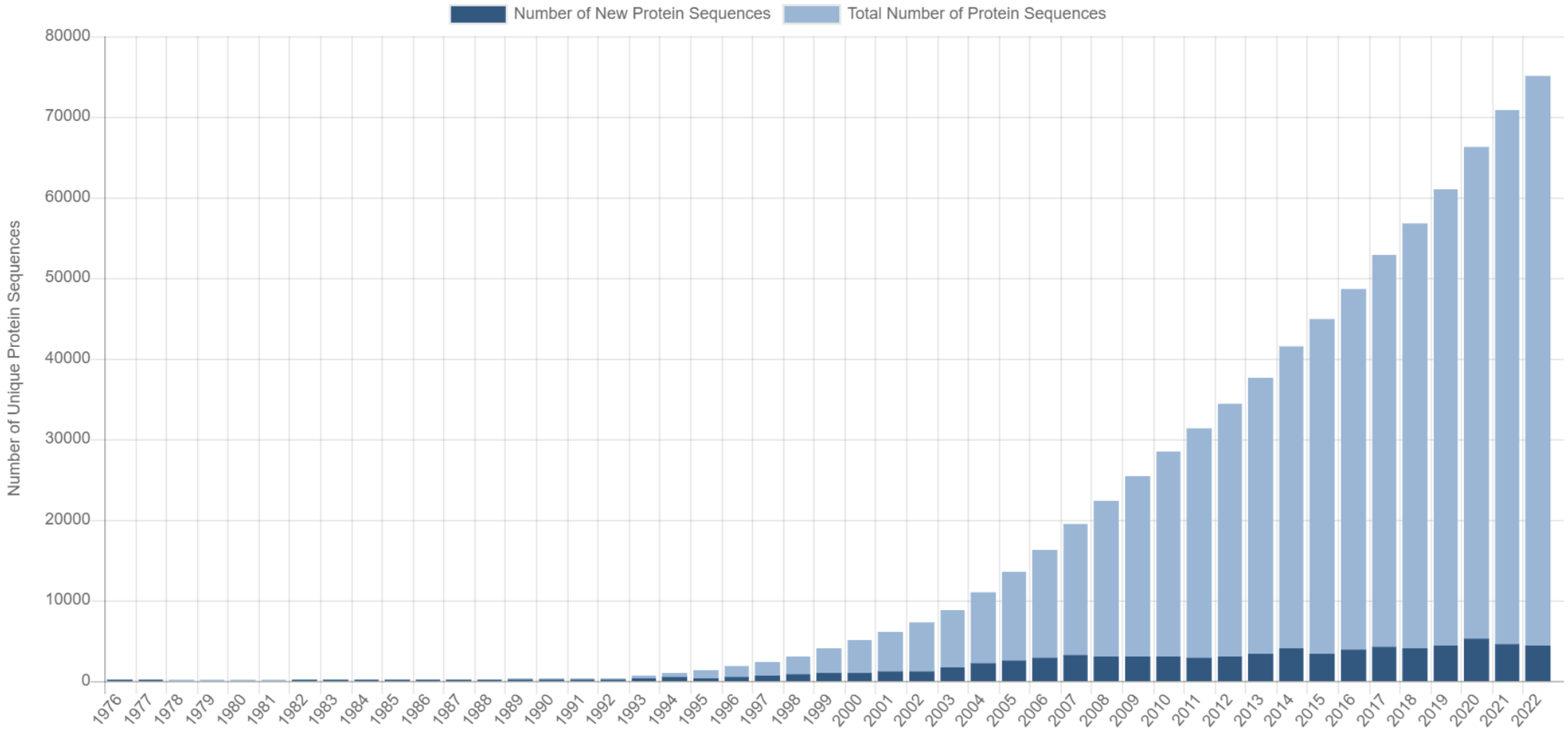
H.Berman, K.Henrick and H. Nakamura, *Announcing the worldwide Protein Data Bank,* ***Nature Structural & Molecular Biology***, 2003

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

The Protein Data Bank (PDB)

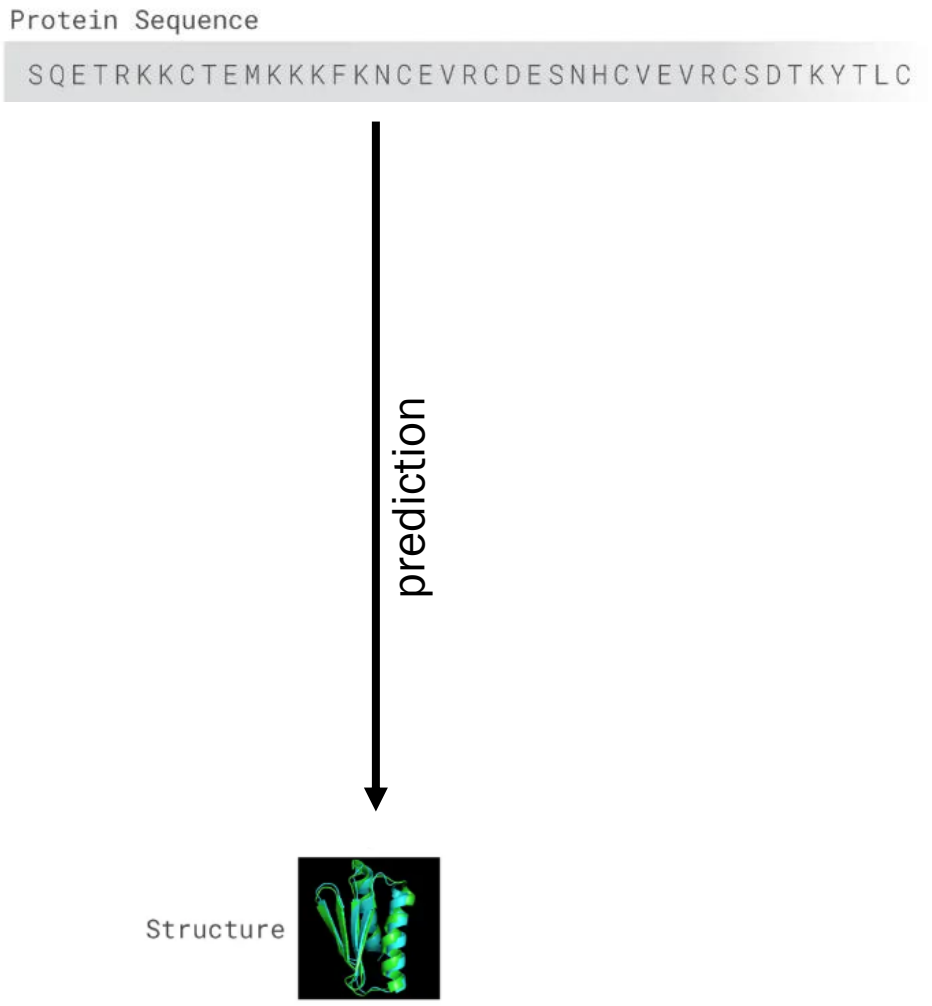


The Protein Data Bank (PDB)



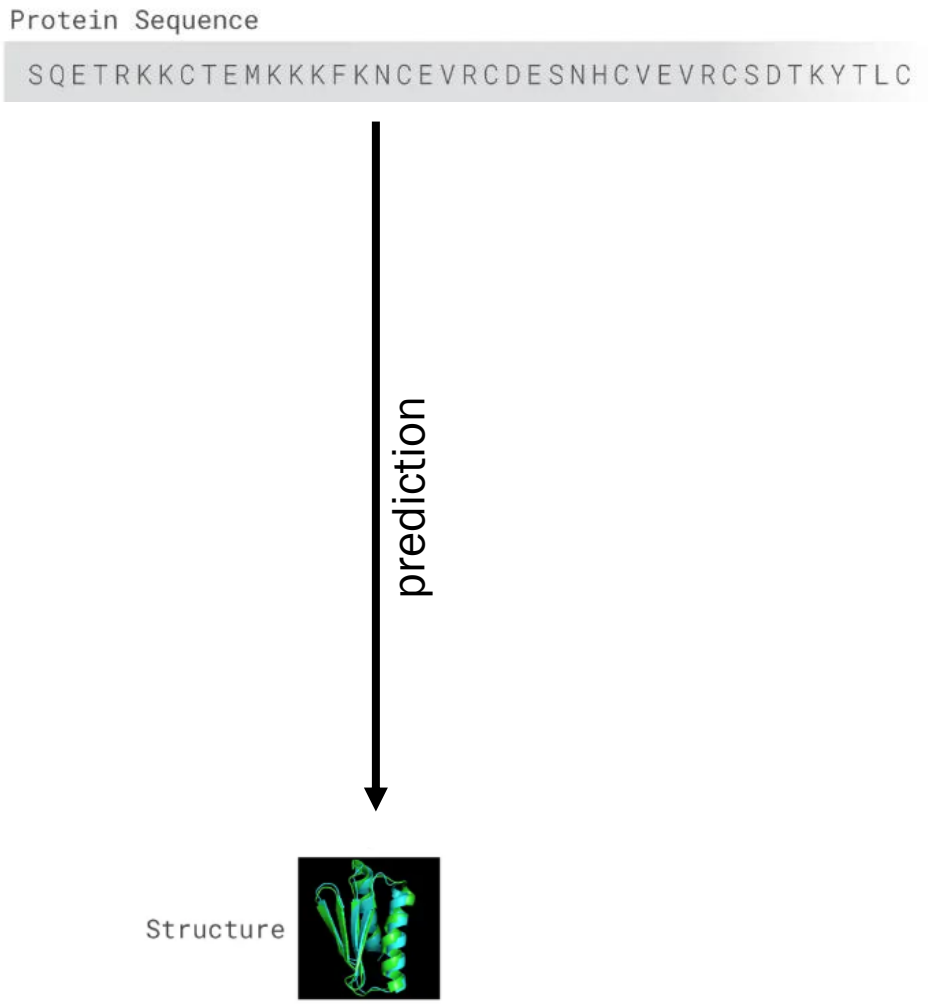
known protein **structures**: ~80’000 (*PDB, 90% identity*)
known protein **sequences**: ~190’000’000 (*UNIPROT*)

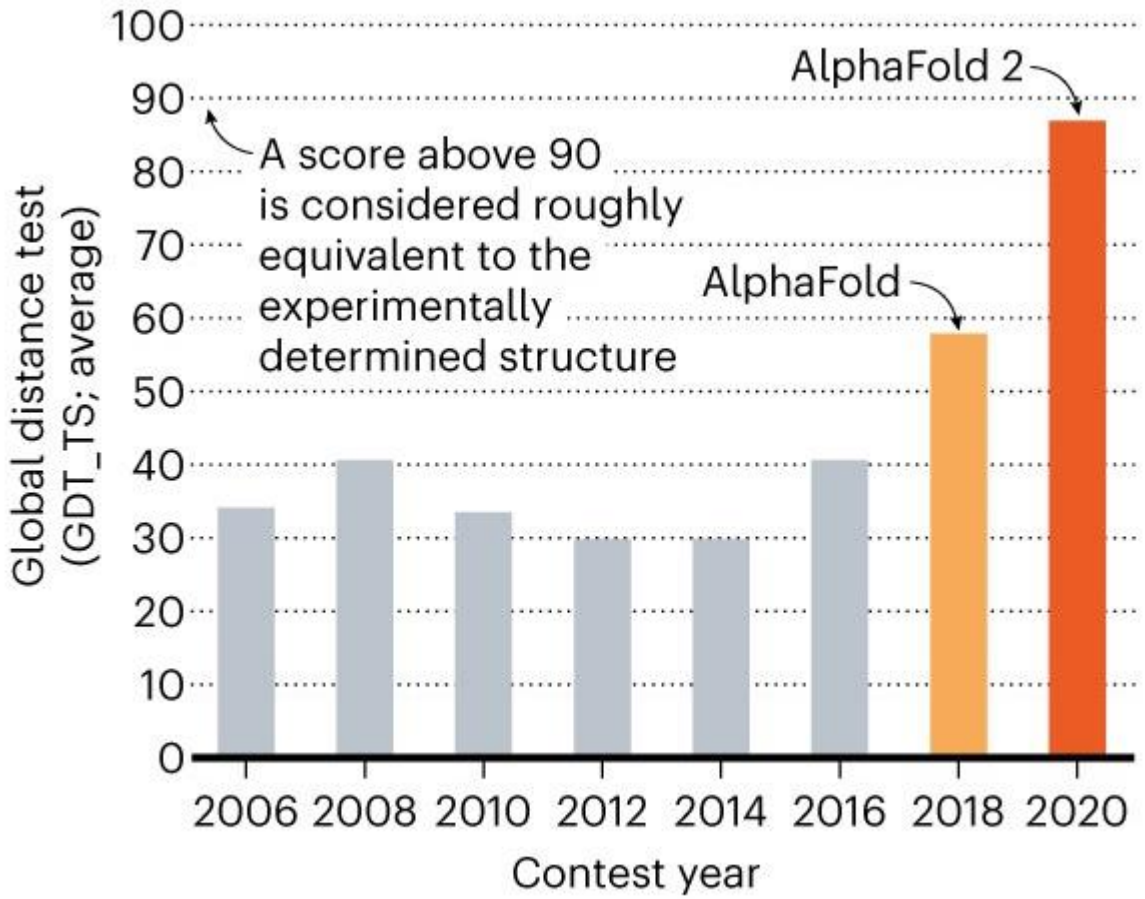
Protein fold prediction



CASP, since 1994 biennial competition on protein fold prediction: predictioncenter.org

Protein fold prediction



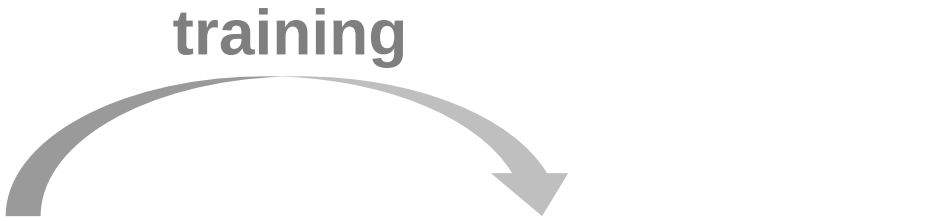


A.W. Senior et al., *Improved protein structure prediction using potentials from deep learning*, ***Nature***, 2020

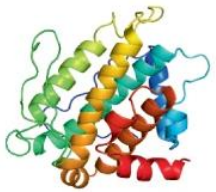
J. Jumper et al., *Highly accurate structure prediction with AlphaFold*, ***Nature***, 2021

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

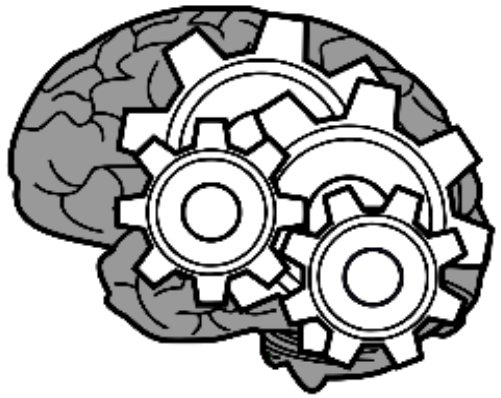
Protein fold prediction: AlphaFold

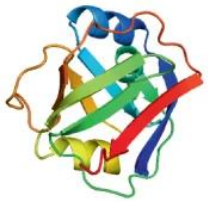


NITPLAKRTYNYRAVL…



HTIWKLSRLWSLQ…





LIYRICMPGILCYEND…

ALRIKAIVVPRILGPQ…

Protein fold prediction: AlphaFold

prediction

MAPVYEGMAS…

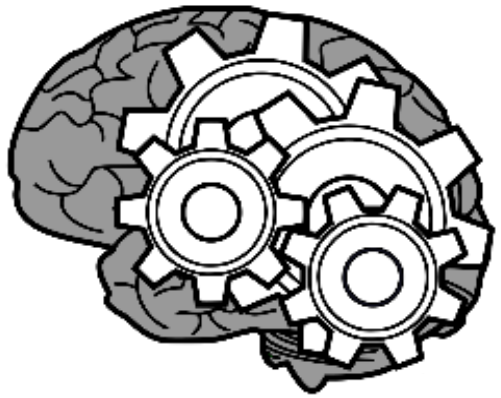
HVQVFSPHTL…

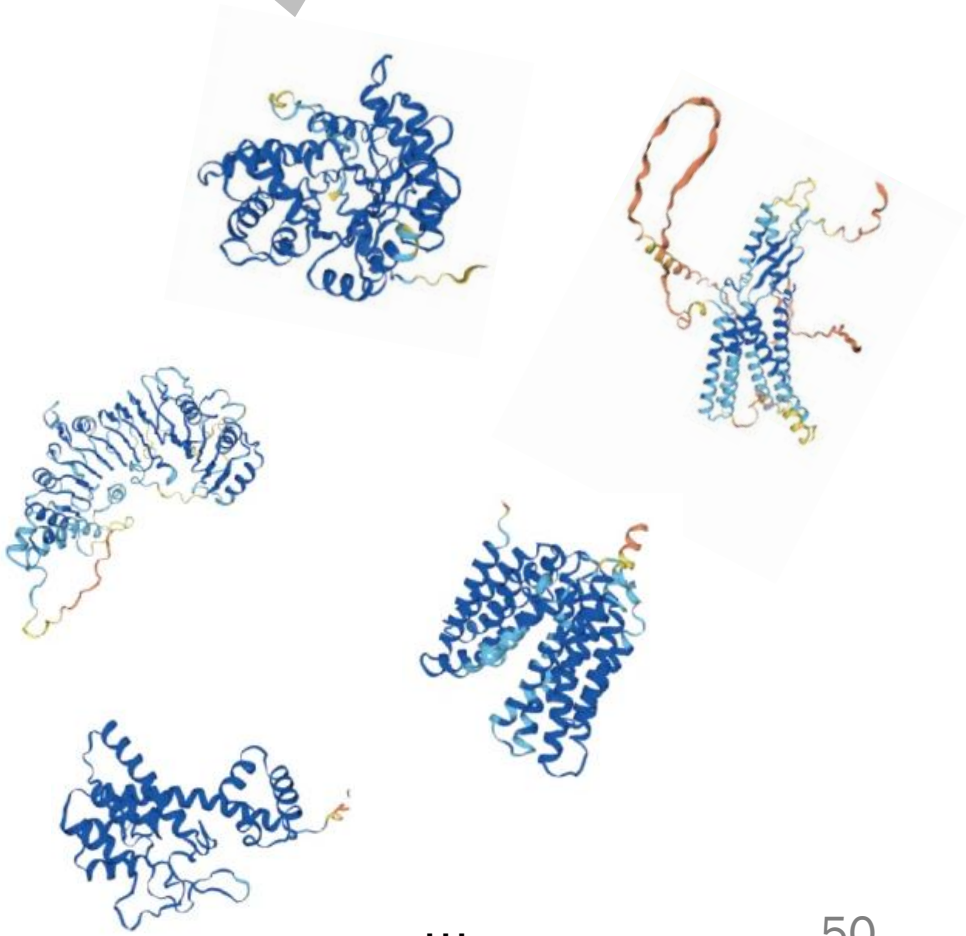
QSSAFCSVKK…

LKIEPSSNWD…

MTGYGSHSKV…

…



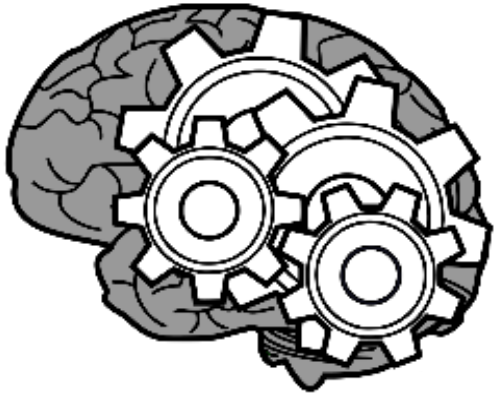


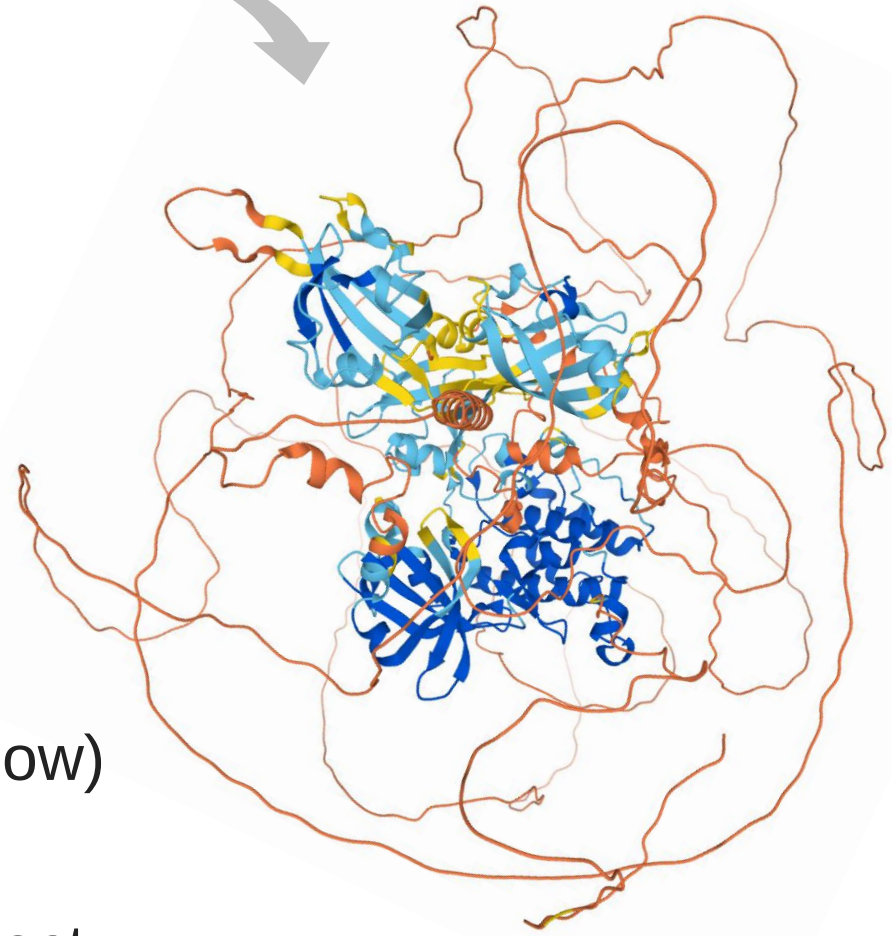
…

Protein fold prediction: AlphaFold

prediction

YSQSKNIPLS…

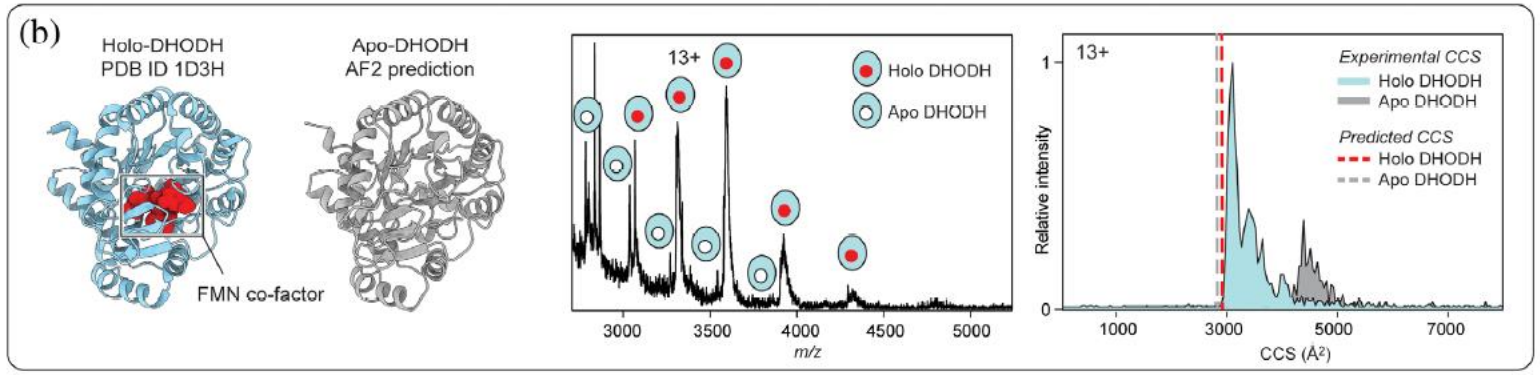


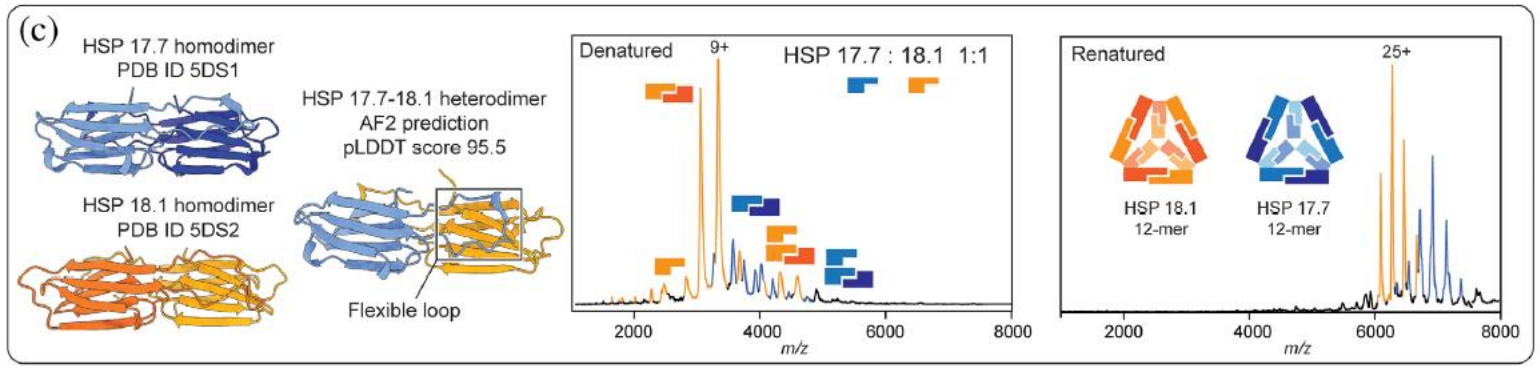


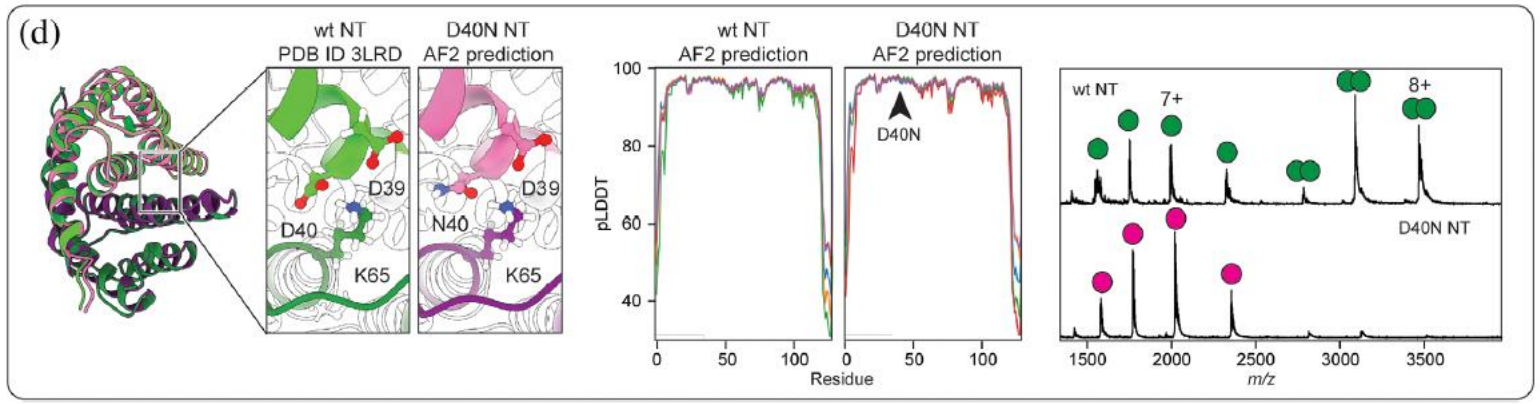
PLDDT score: confidence score (**100** = high, **<50** = low)

Warning: even high-confidence models can be incorrect.

Protein fold prediction: warning!







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

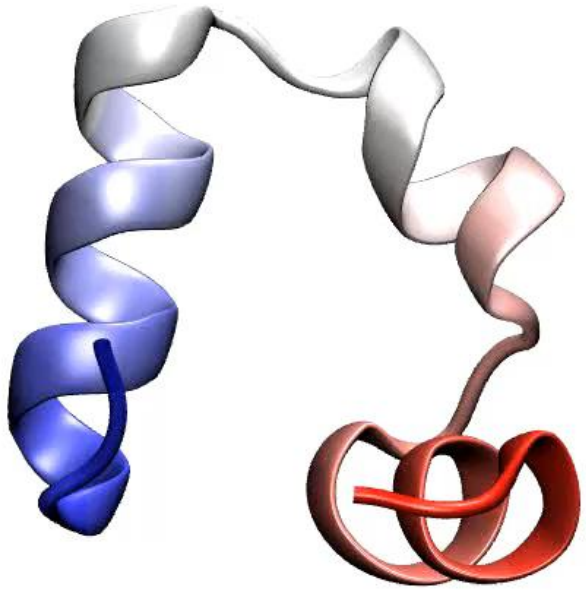
High-confidence hetero multimer predicted, but proteins do not co-assemble

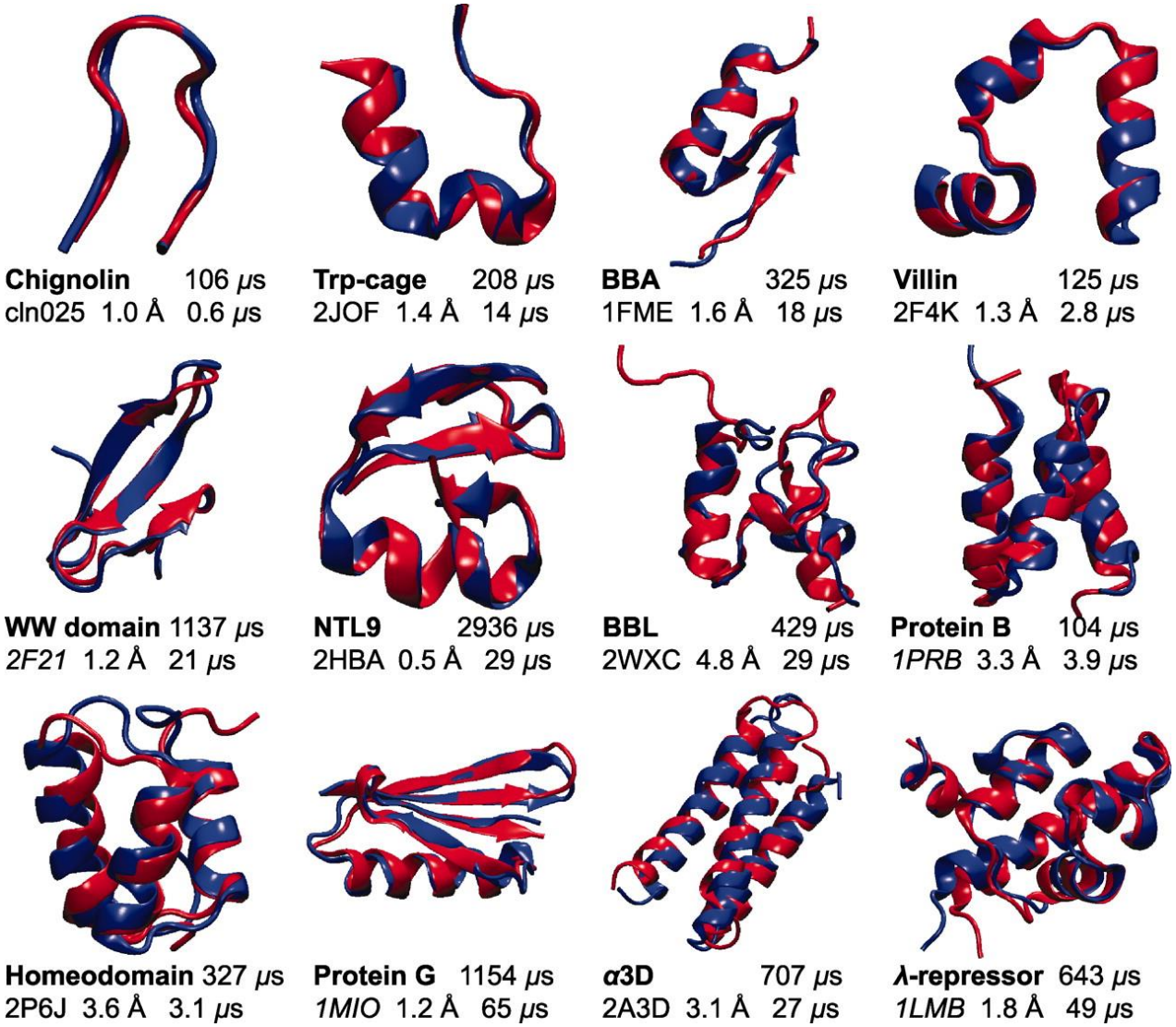
High-confidence homodimer of mutated protein predicted, but mutation abolishes complex formation

T.M.Allison et al., ***Protein Science***, 2022

Watching proteins fold: simulation

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



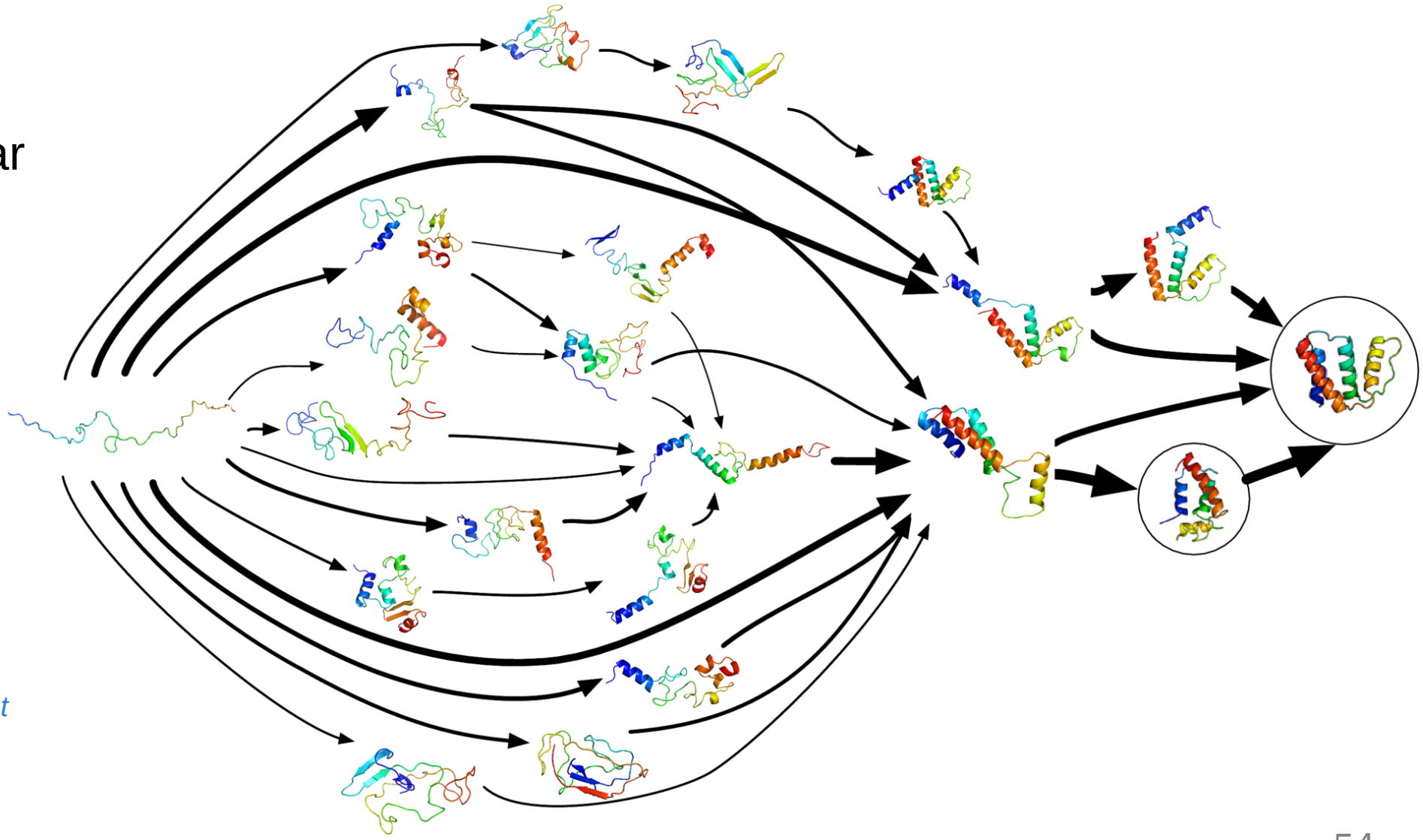


P. Freddolino and K. Schulten. *Common structural transitions in explicit-solvent simulations of villin headpiece folding,* ***Biophysical J.***, 2009

K. Lindorff-Larsen et al. *How fast-folding proteins fold,* ***Science***, 2011

Folding@Home

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

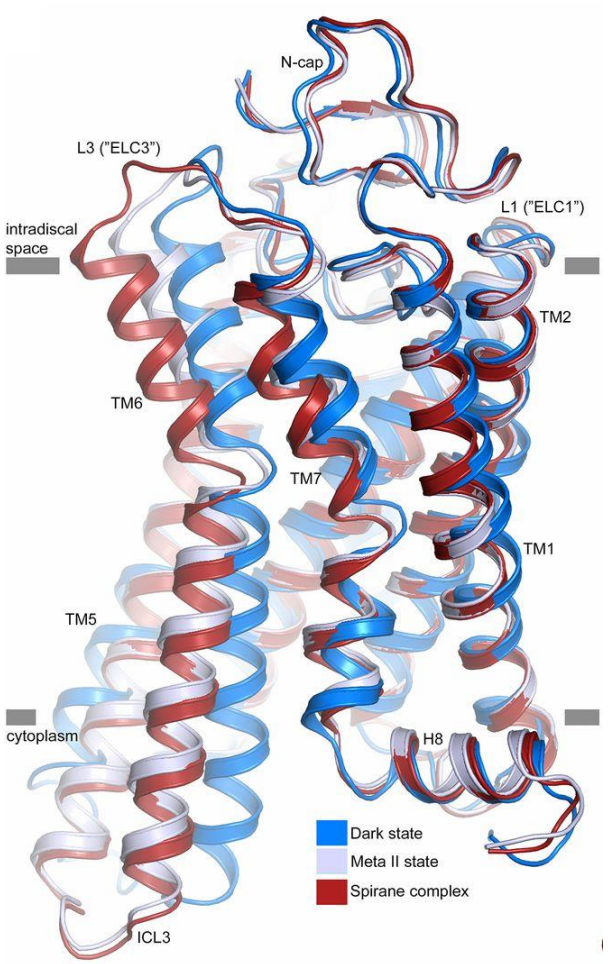


www.foldingathome.org

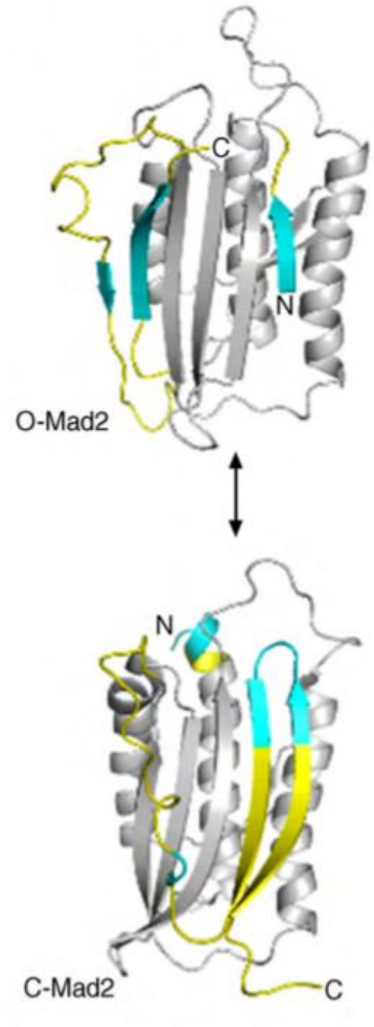
V. S. Pande, K. Beauchamp, G. R. Bowman, *Everything you wanted to know about Markov State Models but were afraid to ask*. ***Methods***, 2010

Protein dynamics

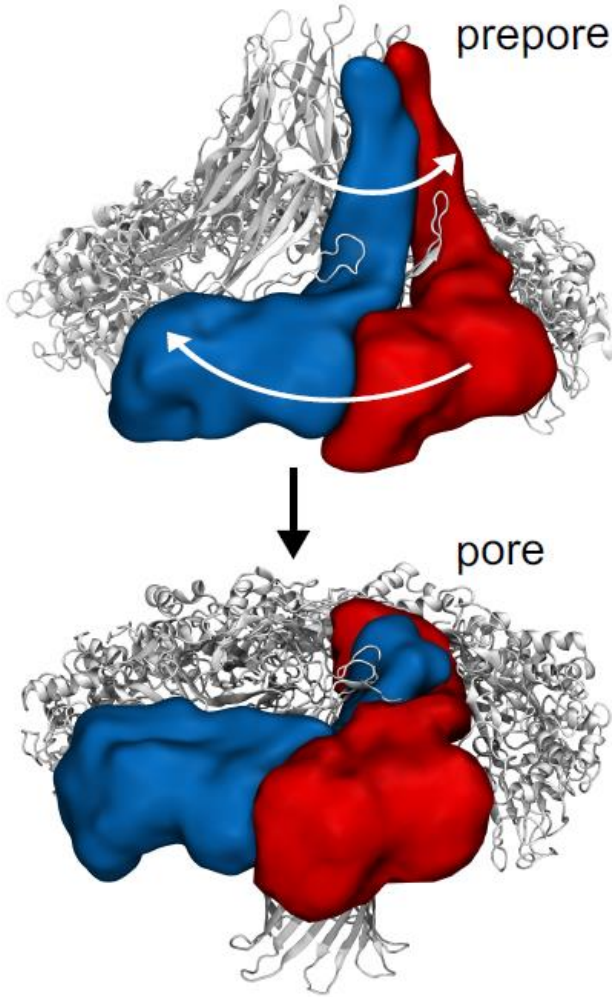
sequence + interaction with environment = conformational space



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

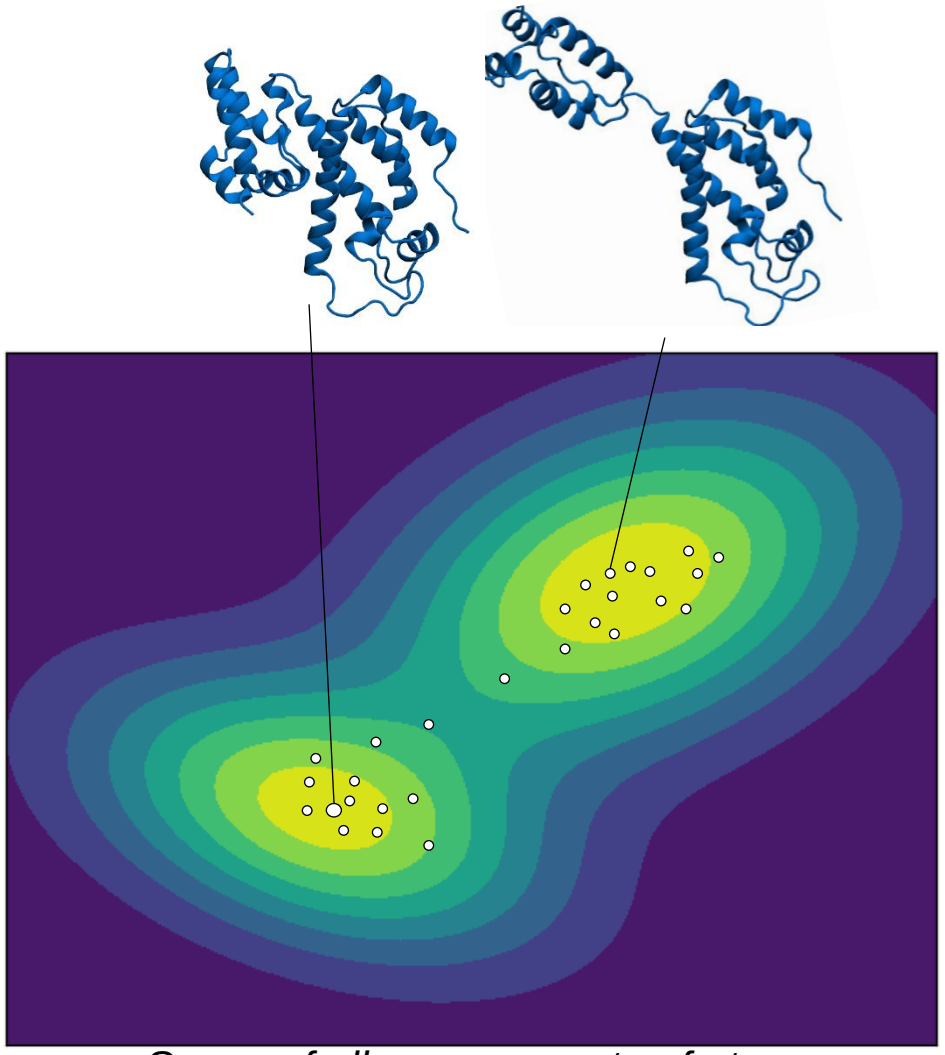


M.T. Degiacomi et al., *Molecular assembly of the aerolysin pore reveals a swirling membrane-insertion mechanism*, ***Nat. Chem. Biol.***, 2013

Sampling the conformational space

Low-energy conformations are explored more often than high-energy ones

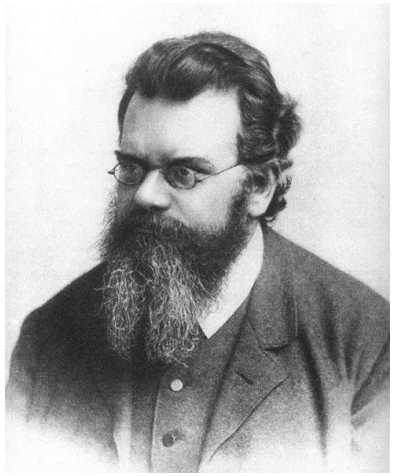
The higher the energy barrier between two states, the longer the time needed to observe a transition



Space of all arrangements of atoms

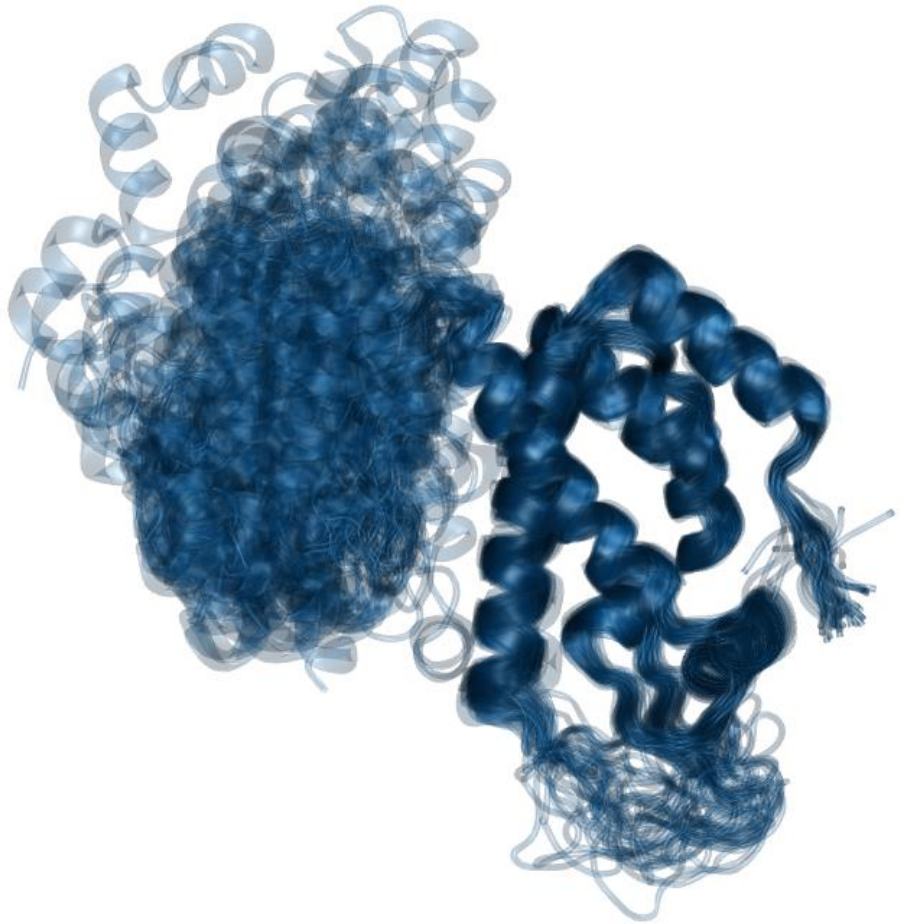
[Extra] sampling the conformational space

The Boltzmann distribution



$$p_i = \frac{1}{Q} e^{-\varepsilon_i / kT}$$

$$Q = \sum_{i=1}^{M} e^{-\varepsilon_i / k_B T}$$

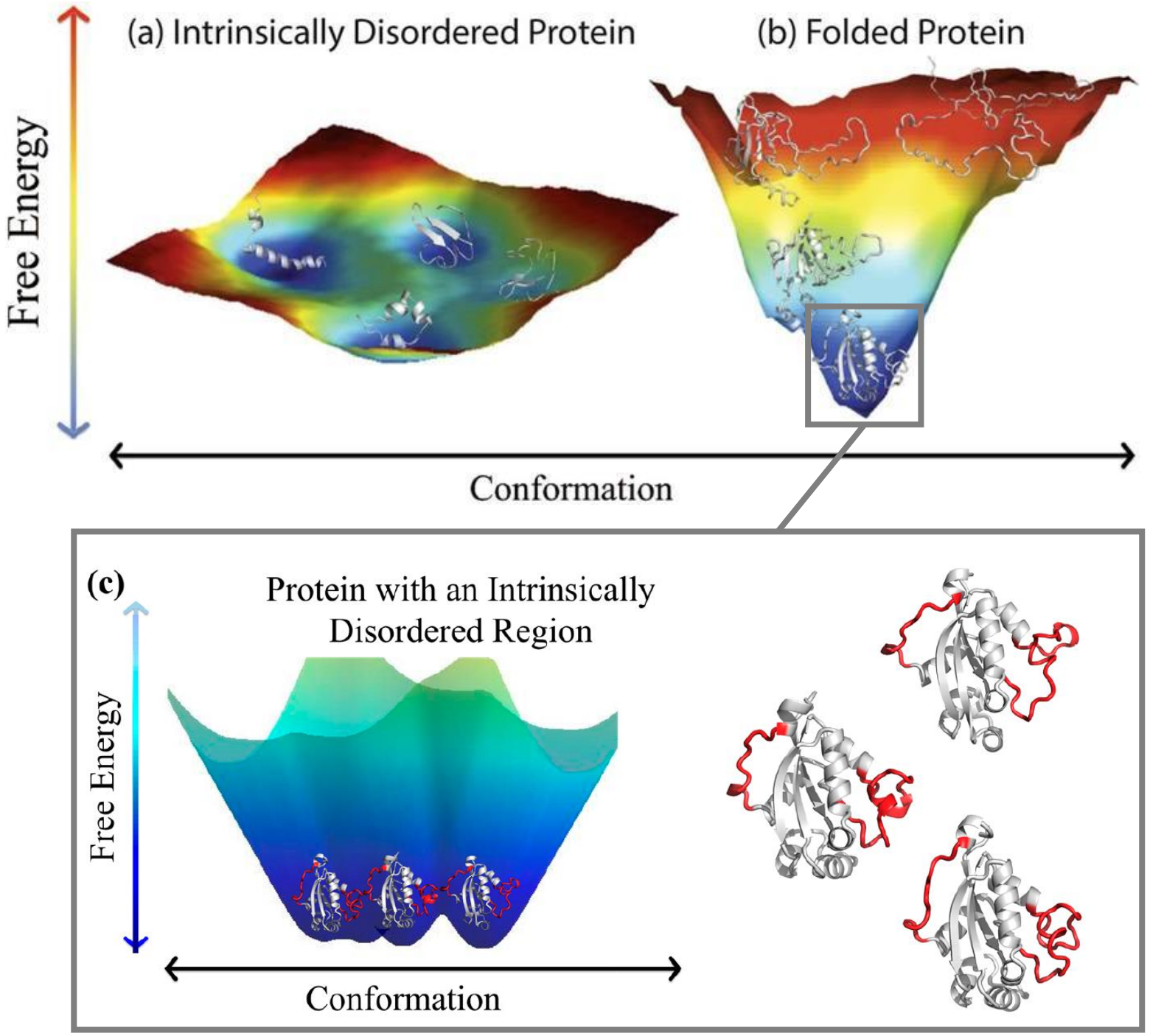


Not all proteins fold

Intrinsically Disordered Proteins (IDP):

- have a shallow energy landscape
- do not occupy a well-defined ensemble of states

Proteins may contain Intrinsically Disordered Regions (IDR)



V.M. Burger et al., *Intrinsically Disordered Proteins: Where Computation Meets Experiment*, ***Polymers***, 2014

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