



Seeing the PDB

Received for publication, April 3, 2021, and in revised form, April 26, 2021 Published, Papers in Press, May 4, 2021,
<https://doi.org/10.1016/j.jbc.2021.100742>

Jane S. Richardson^{1,*}, David C. Richardson¹, and David S. Goodsell^{2,3,*}

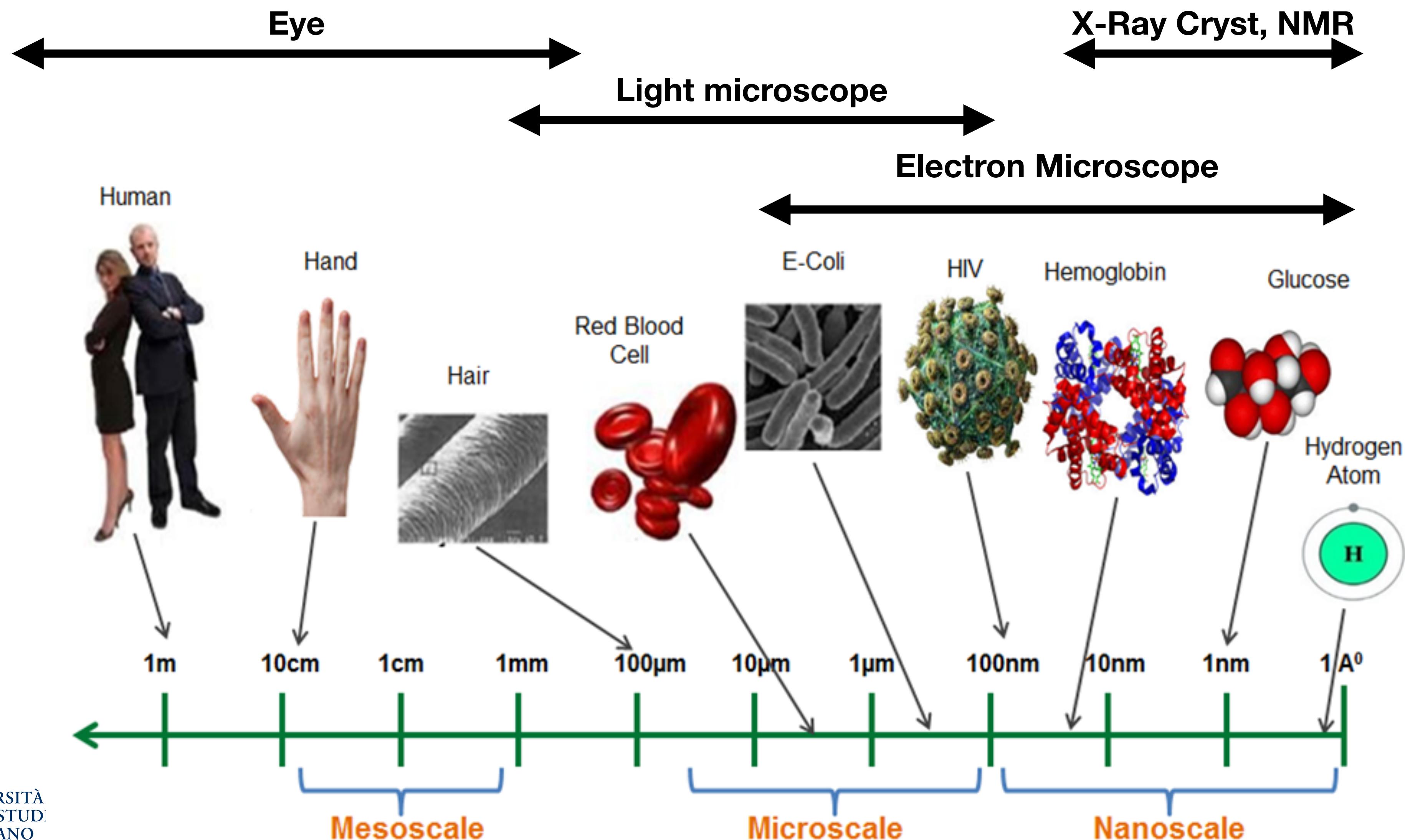
From the ¹Department of Biochemistry, Duke University, Durham, North Carolina, USA; ²Department of Integrative and Computational Biology, The Scripps Research Institute, La Jolla, California, USA; ³Research Collaboratory for Structural Bioinformatics Protein Data Bank, Rutgers, the State University of New Jersey, Piscataway, New Jersey, USA

Edited by Karin Musier-Forsyth

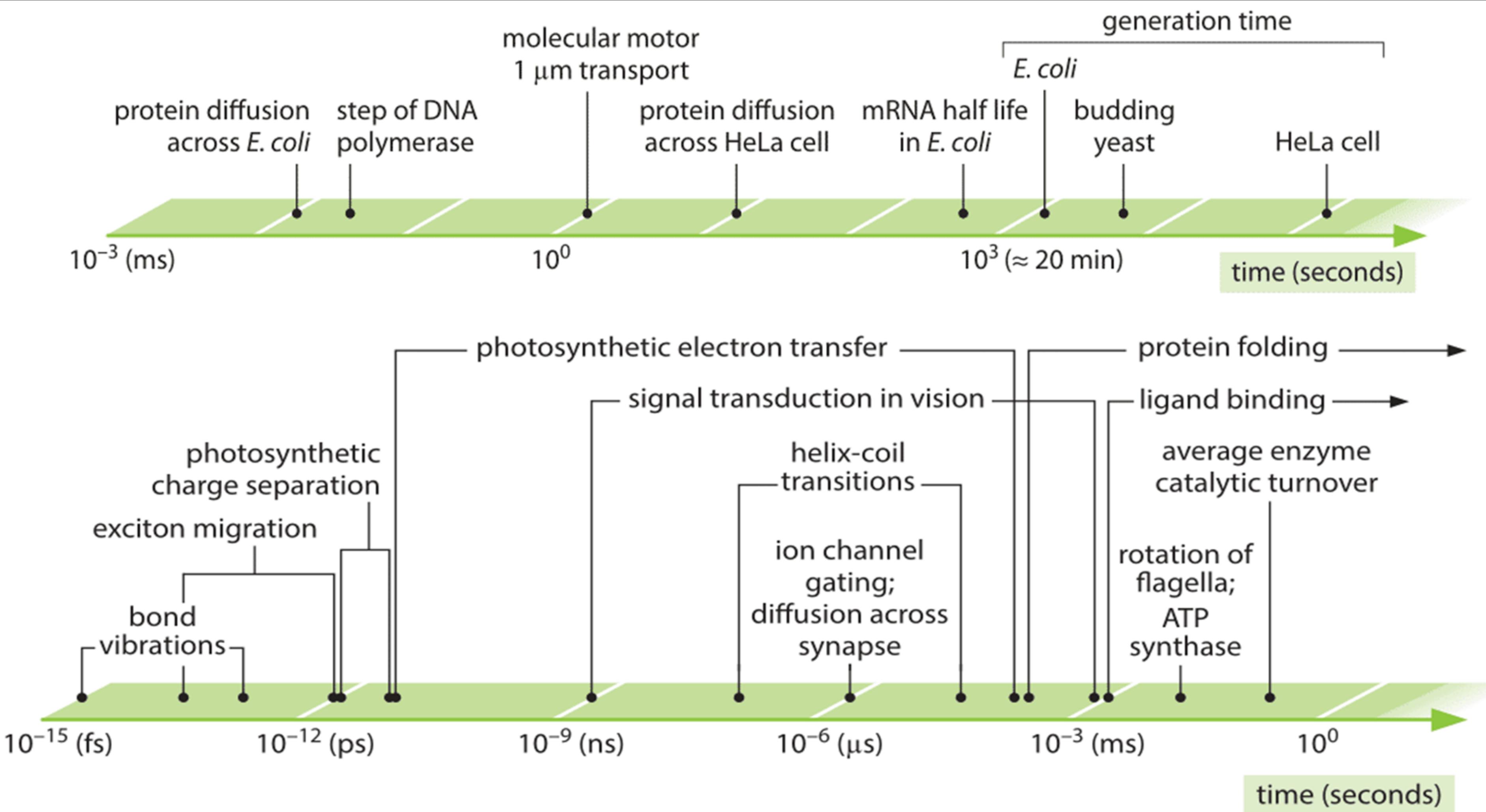
Biomolecules, Structural Biology
and Visualisation

Structural Bioinformatics

Light microscope cannot look at molecular spatial scales



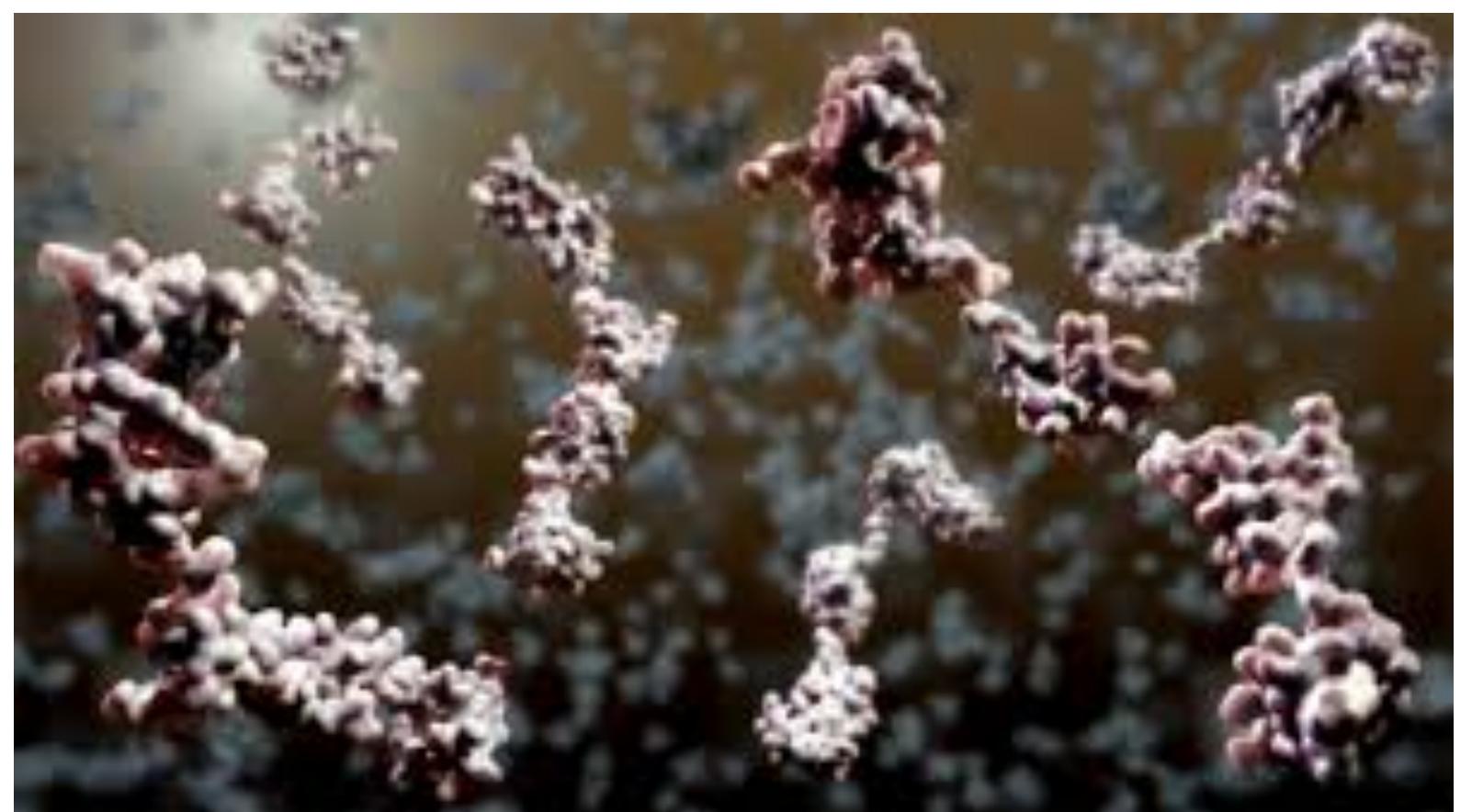
Molecules motion can span many time scales



As a consequence: observing the microscopic world is hard

We cannot use light microscopes, so we don't see directly the molecules, we need to interpret other kind of signals (shorter wavelength, electrons, magnetic fields, ...)

We do not look at a single molecule at a time but at many (~fraction of Avogadro number) with issues of rotations and translations



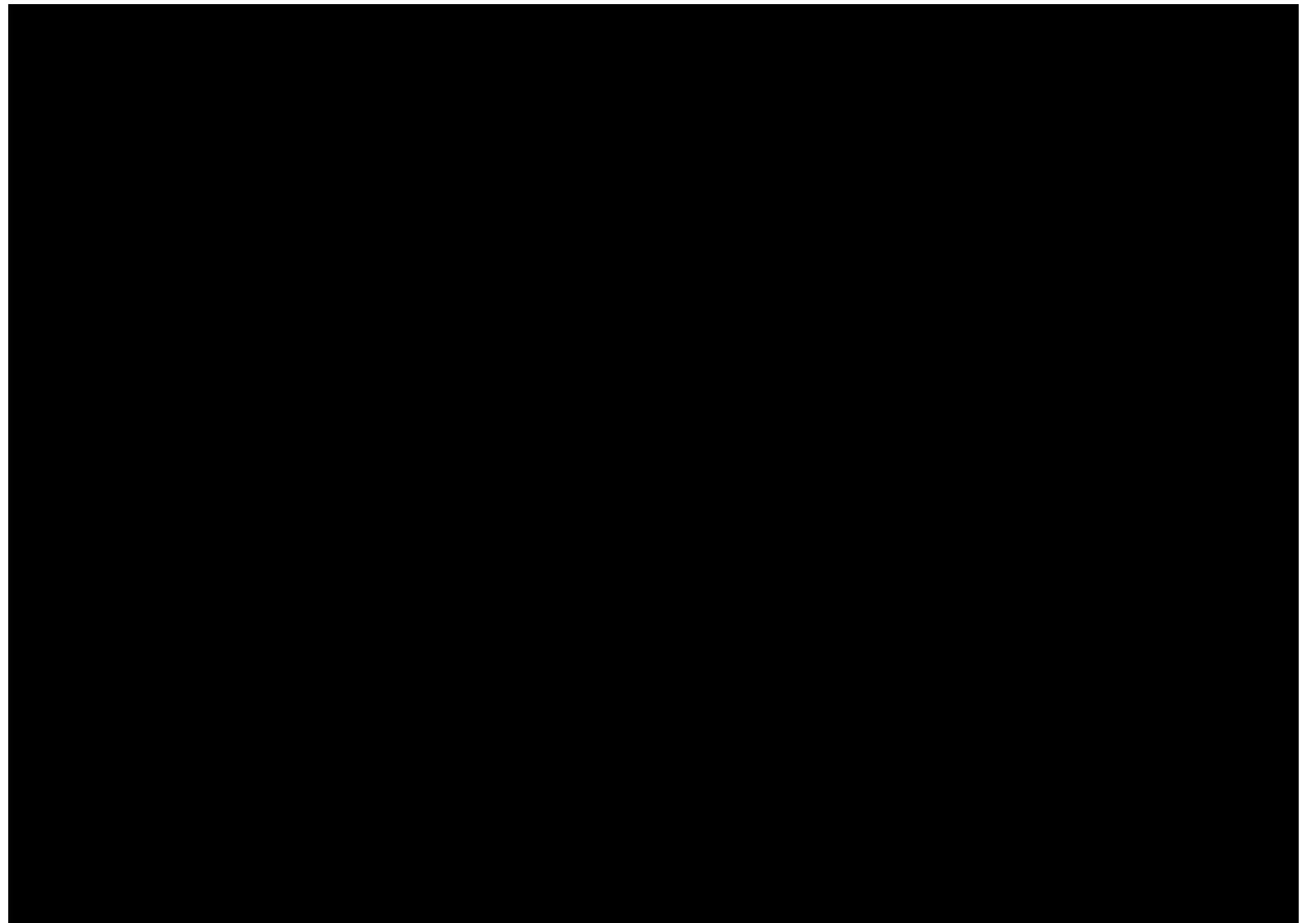
As a consequence: observing the microscopic world is hard

Molecules have different sizes, and move on multiple time scales

Long exposure: good signal vs noise ratio, but molecules that move fast may become messy

Short exposure: you could resolve the motion of all molecules, but no signal

Finding an optimal compromise is not always possible



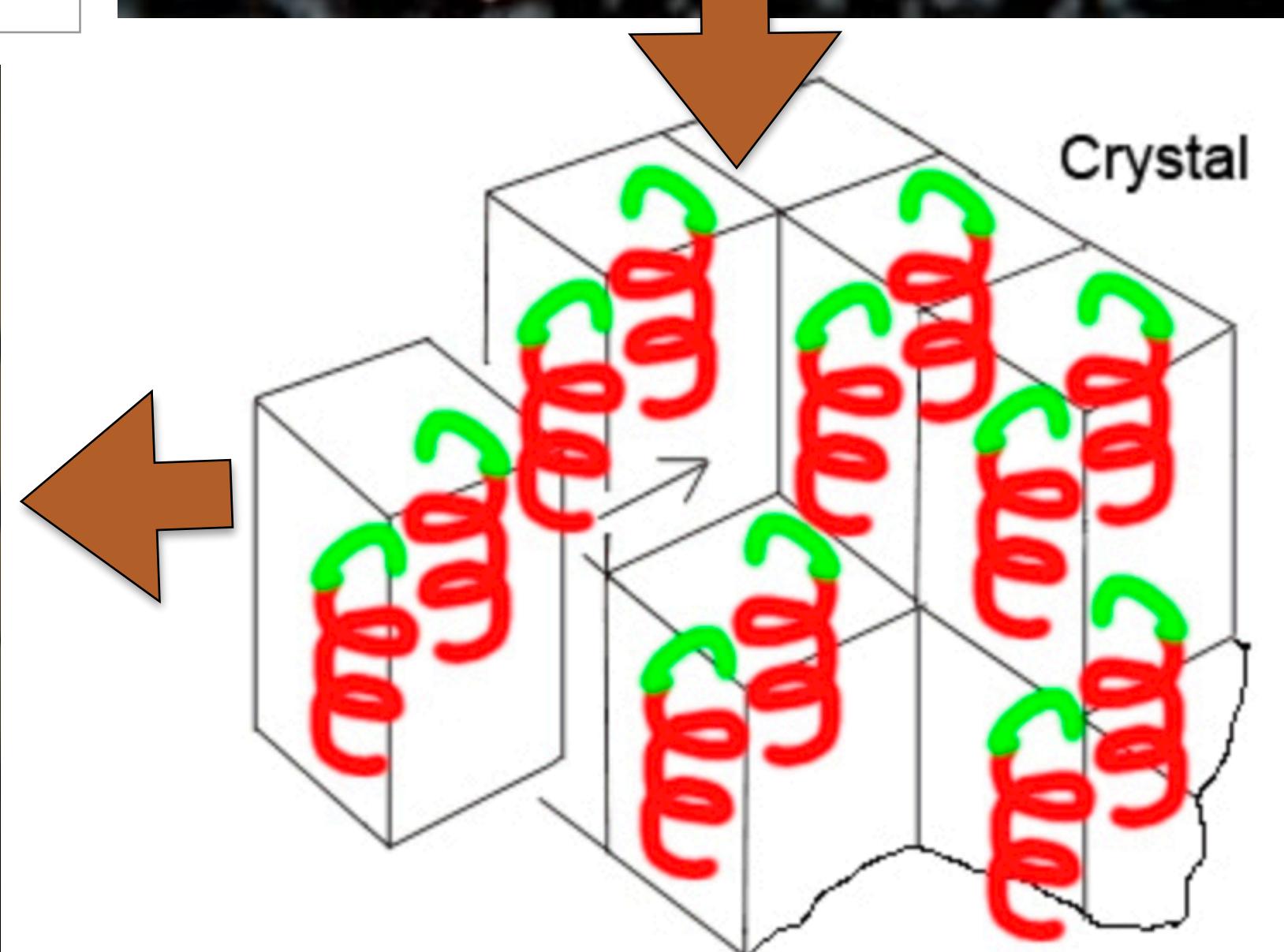
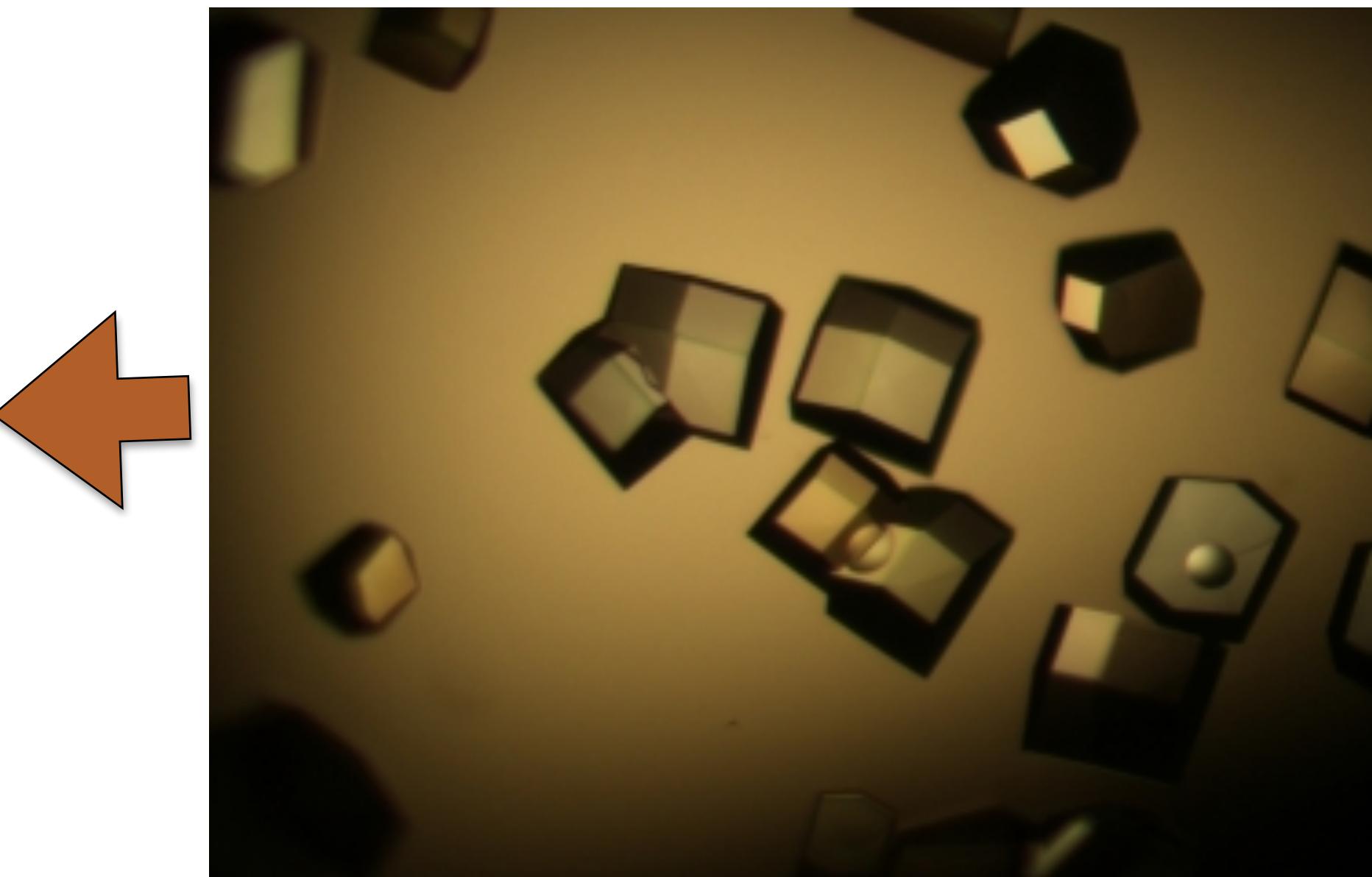
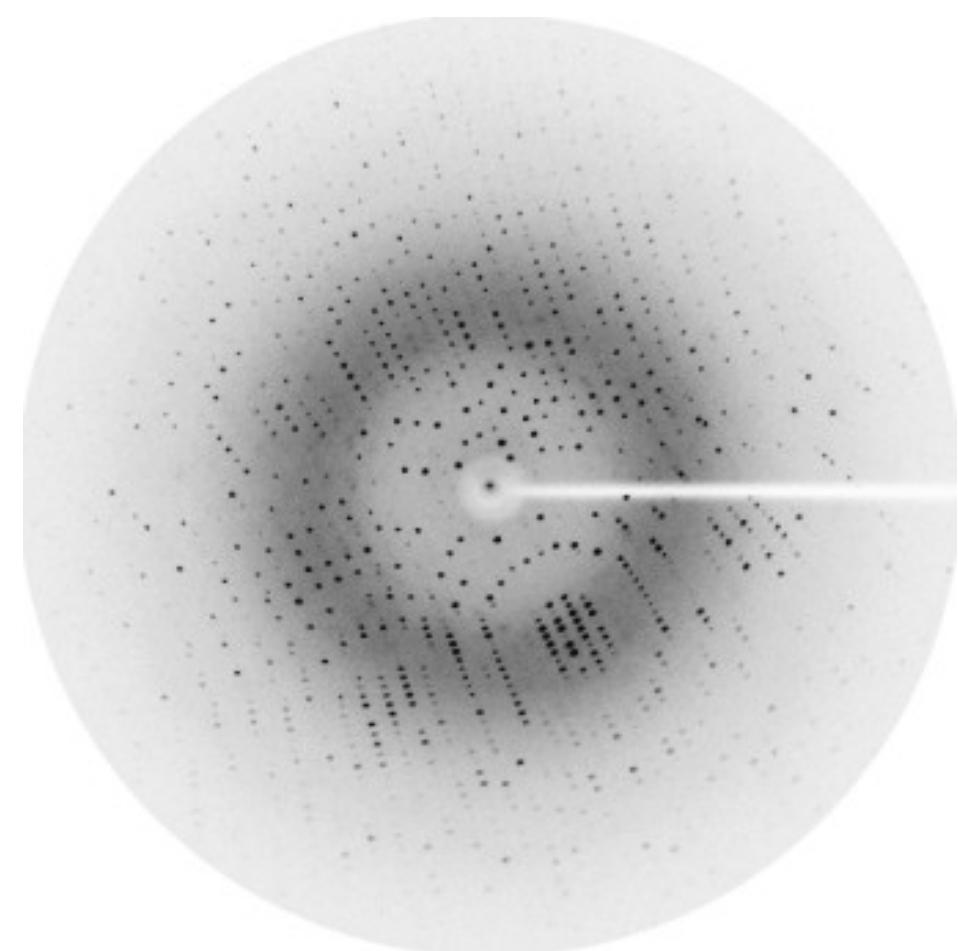
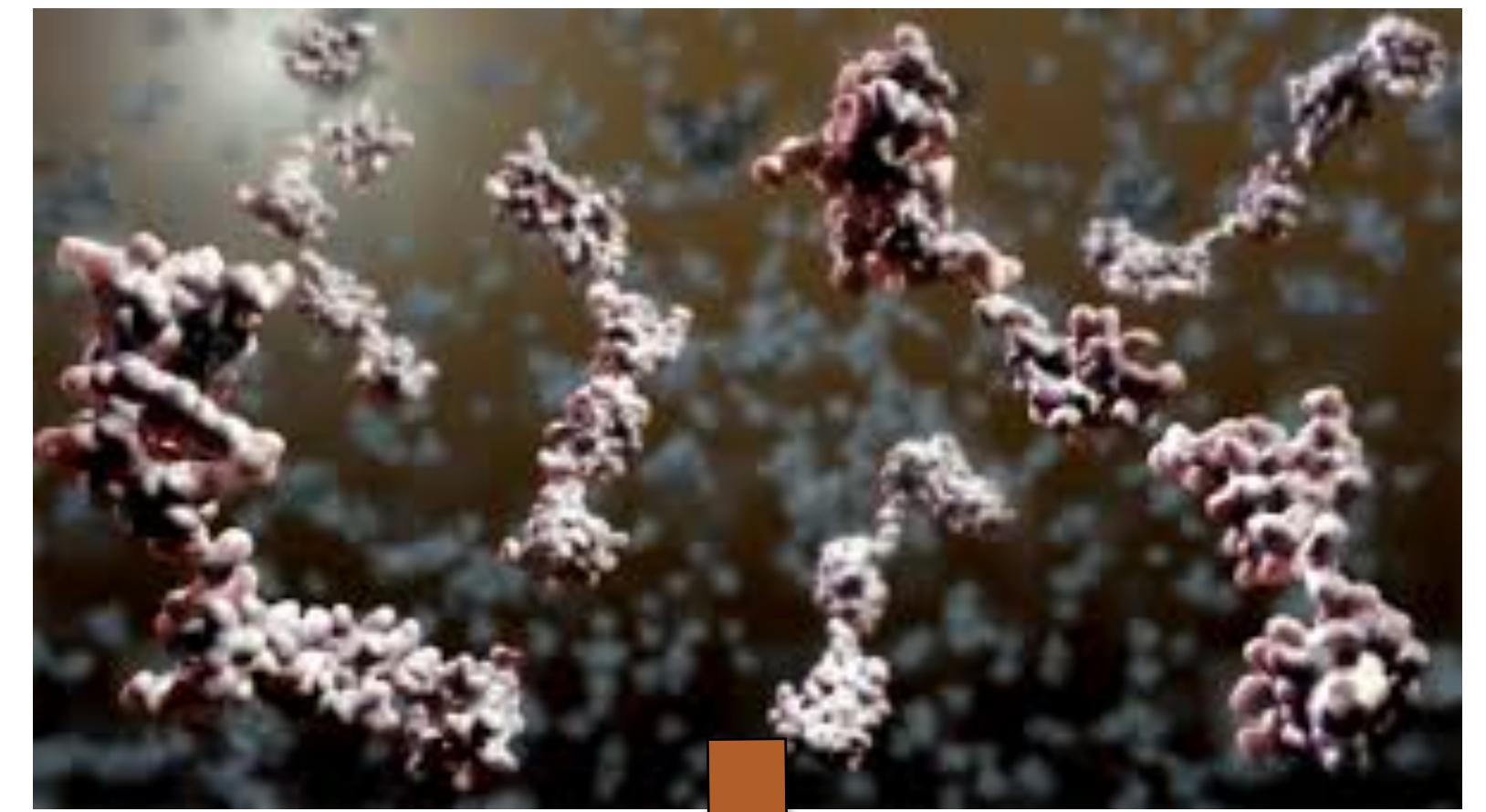
UNIVERSITÀ
DEGLI STUDI
DI MILANO



So how do we “take pictures” of molecules? 1. X-ray crystallography

Issues are signal/noise, small size, motion on many time scales

SOLUTION: We make all molecules adopt the same structure in an ordered 3D organisation: a crystal, that we observe using X-rays



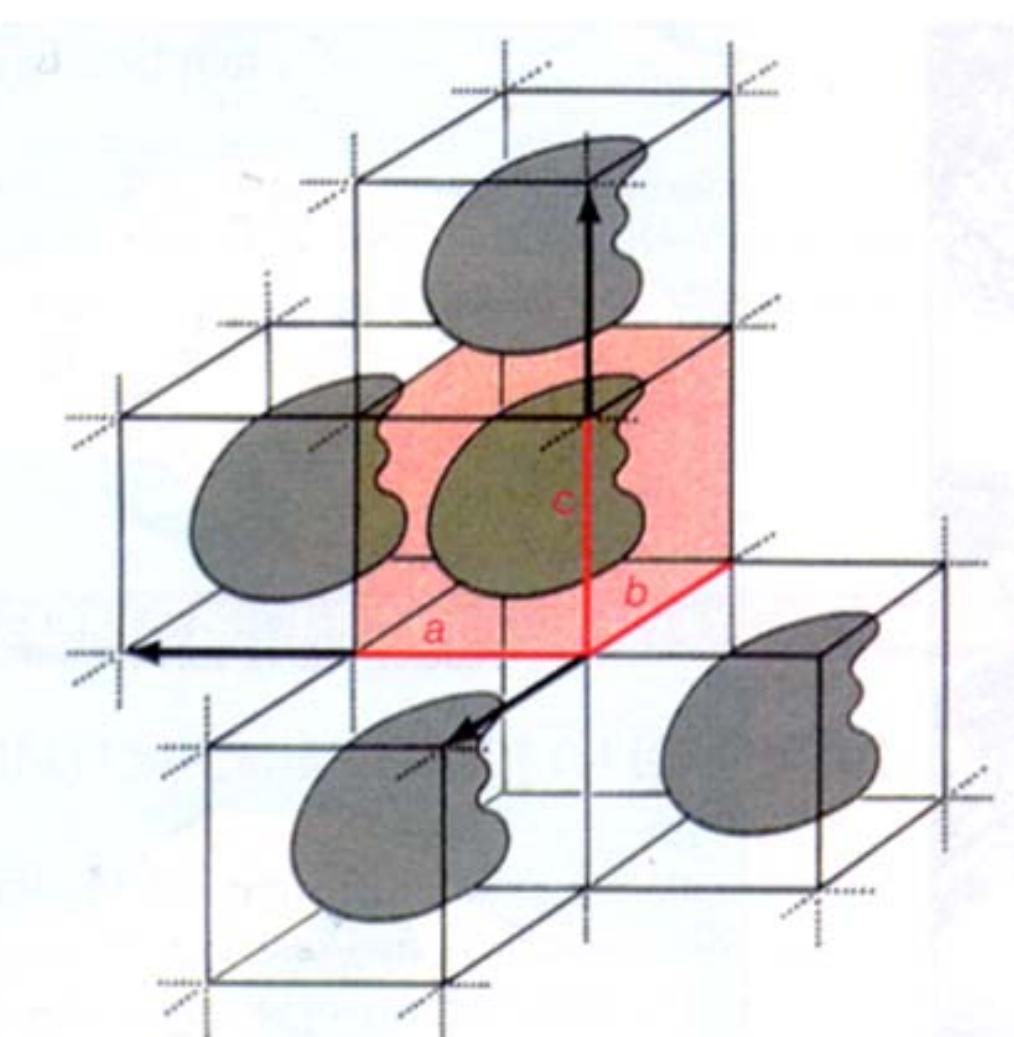
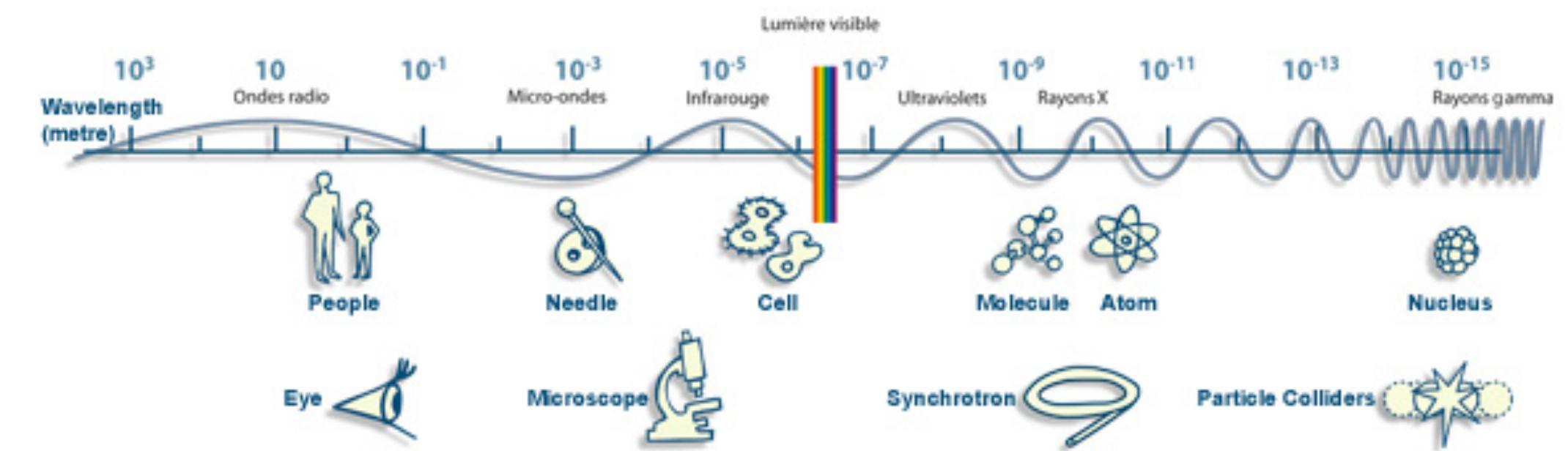
X-ray Crystallography

Why X-rays ?

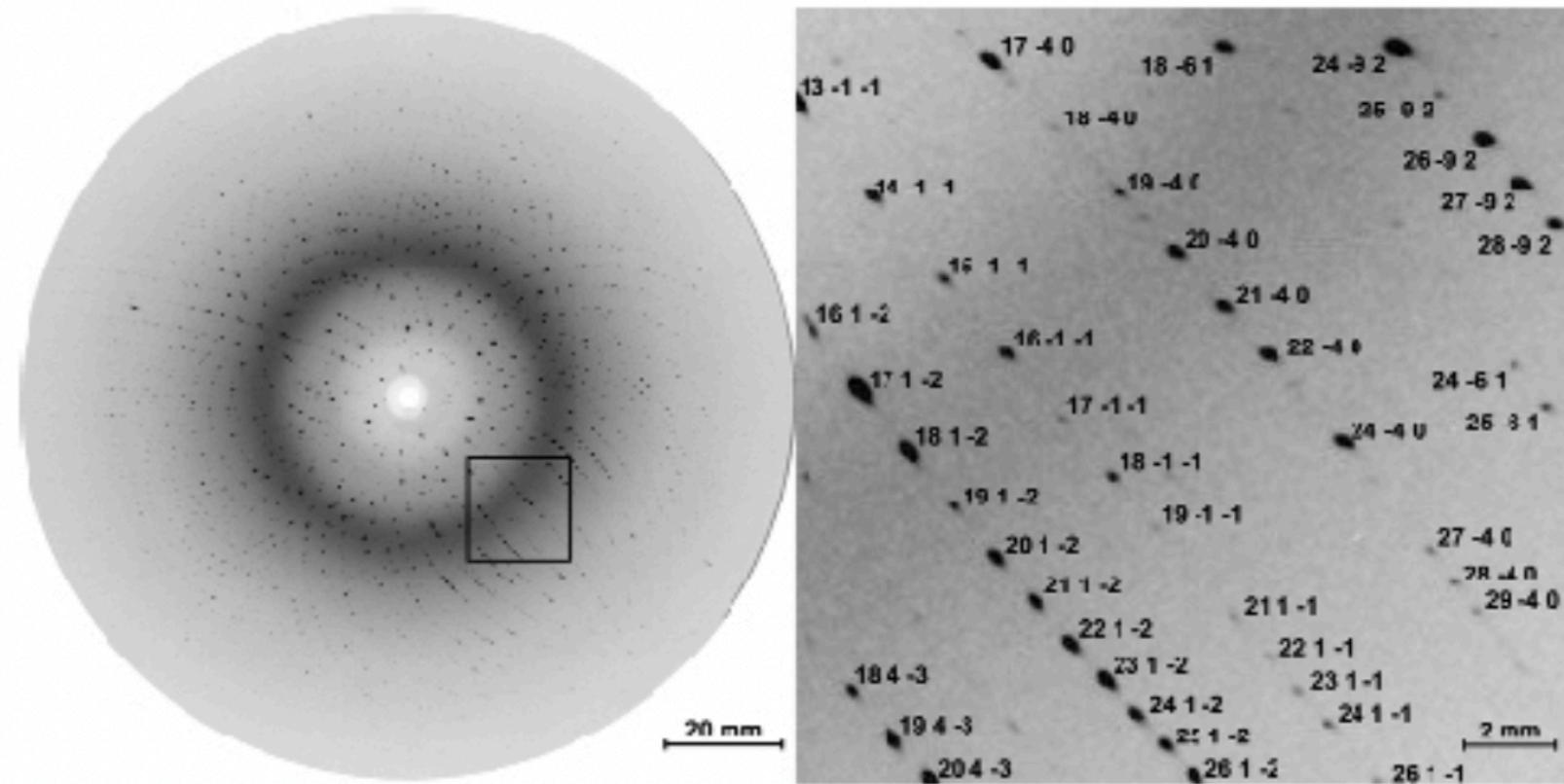
The wavelength of a X-rays is roughly **1 Å** (10^{-10} m), which is on the scale of a single atom, and it allows to have **sufficient resolution to determine the atomic positions**

Why crystals ?

X-ray crystallography requires a crystal to amplify the signal (**10^{15} - 10^{16} identical molecules**); the periodicity of the electron density is used to diffract the X-rays with manageable measurement error



X-ray Crystallography



The crystal is rotated and multiple images are collected to find out the symmetry of the crystal, that is to each black spot of each image you can assign 4 numbers: the intensity and three coordinates h, k, l (Miller indices): I_{hkl}

Each I_{hkl} is associated with a Structure Factor F_{hkl} that is associated with the electronic density $\rho(x,y,z)$ of the molecule in the crystal. With a computational procedure it is possible in some cases to obtain the electronic density that is an envelope in 3D that is then used to reconstruct the 3D coordinates of the atoms of the molecule.

$$I_{hkl} \propto I_0 \frac{V_{xtal}}{V_{Cell}^2} |F_{hkl}|^2$$

not measured

$$\rho(x,y,z) = \frac{1}{V_{hkl}} \sum |F_{hkl}| \exp(i\alpha_{hkl}) \exp[-2\pi i(hx+ky+lz)]$$



What does X-ray crystallography tell us about molecules motion?

The key idea of X-ray crystallography is that of removing protein motion. Both the motion of the different molecules, that are trapped in the crystal lattice, but also the internal motion. The structure that is obtain in crystal is generally accepted to be representative of the most populated structure found in solution.

We will see later that solution techniques like NMR and SAXS can be used to test this assumption case by case. But what do the molecules in the crystal can tell us about their motion?



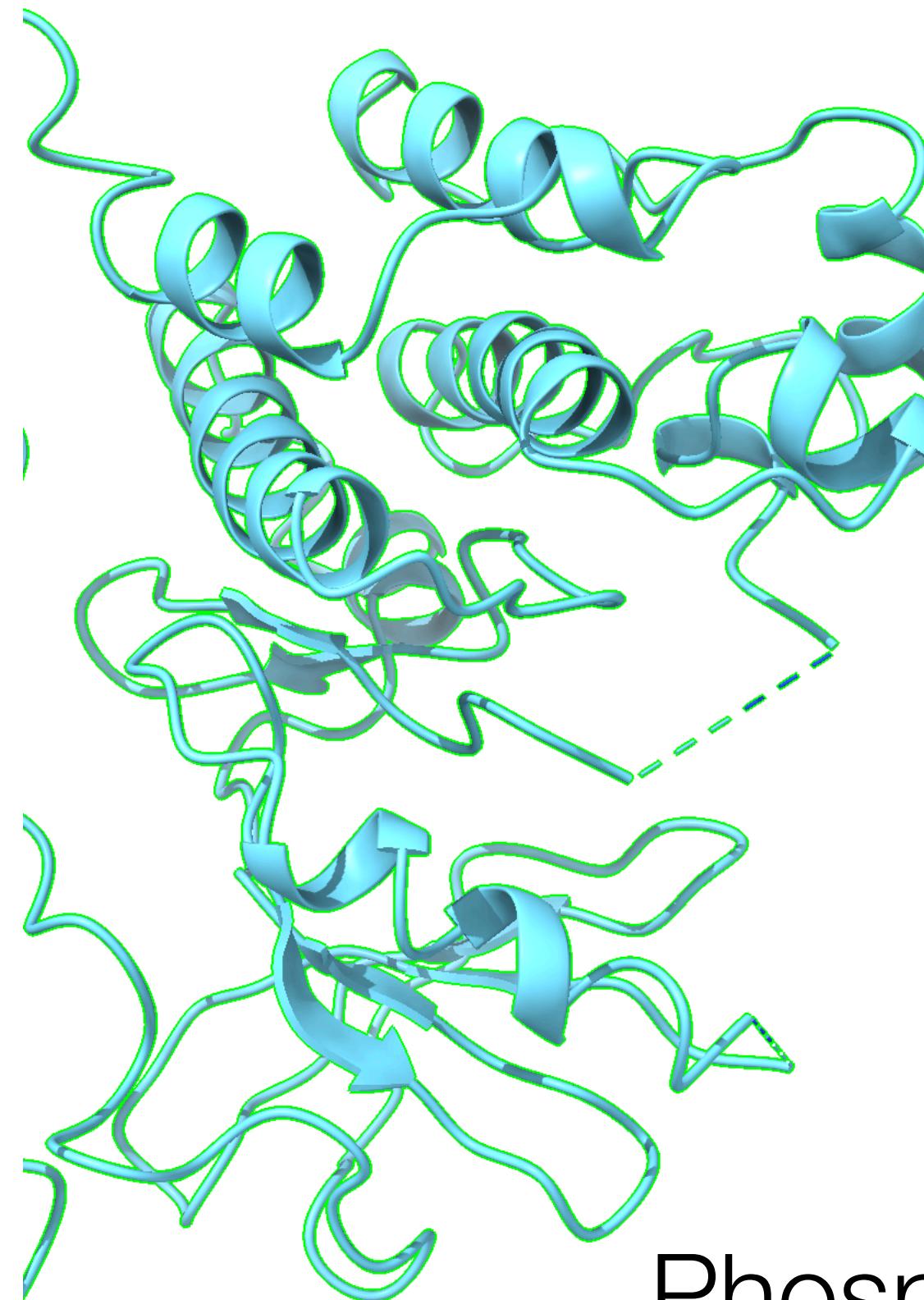
What does X-Ray crystallography tell us about molecules motion?

1. Molecules that are too flexible (intrinsically disordered proteins, proteins with many disordered loops, or with long tails) generally do not crystallise.
2. Even in proteins that do crystallise, some regions or even single atom, do not scatter and are then invisible in the 3D density (there are regions with missing density), these regions are generally deemed to be either too flexible or to populate multiple conformations in the crystal
3. Some times the same atoms scatter in different ways allowing to find multiple conformations for the same atoms in the crystal (crystal anisotropy), these motions are generally interpreted to exists also in solution



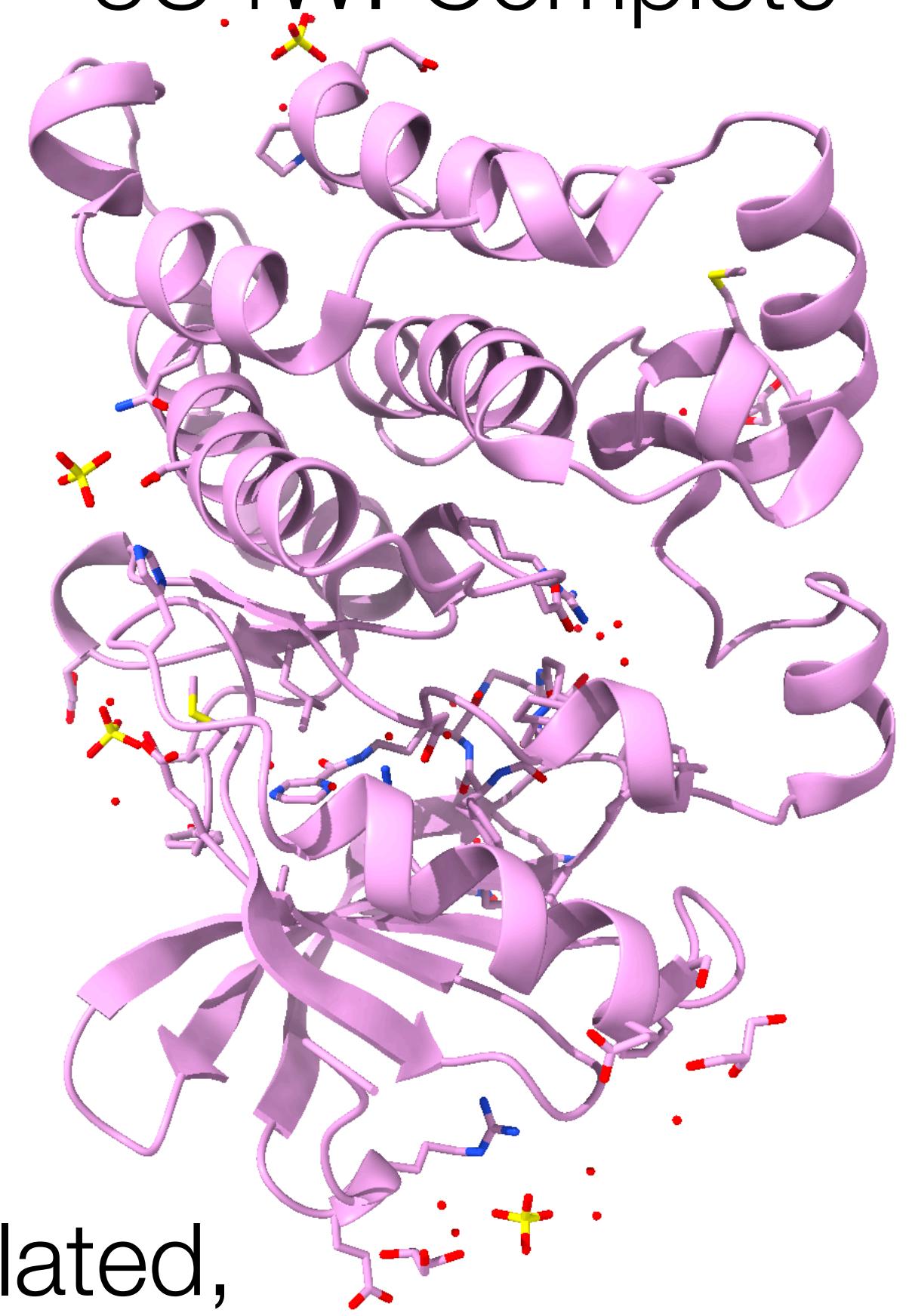
Example: the active-loop of C-src Kinases

2PTK: Missing 407-424



Phosphorylated,
inactive state

3U4W: Complete



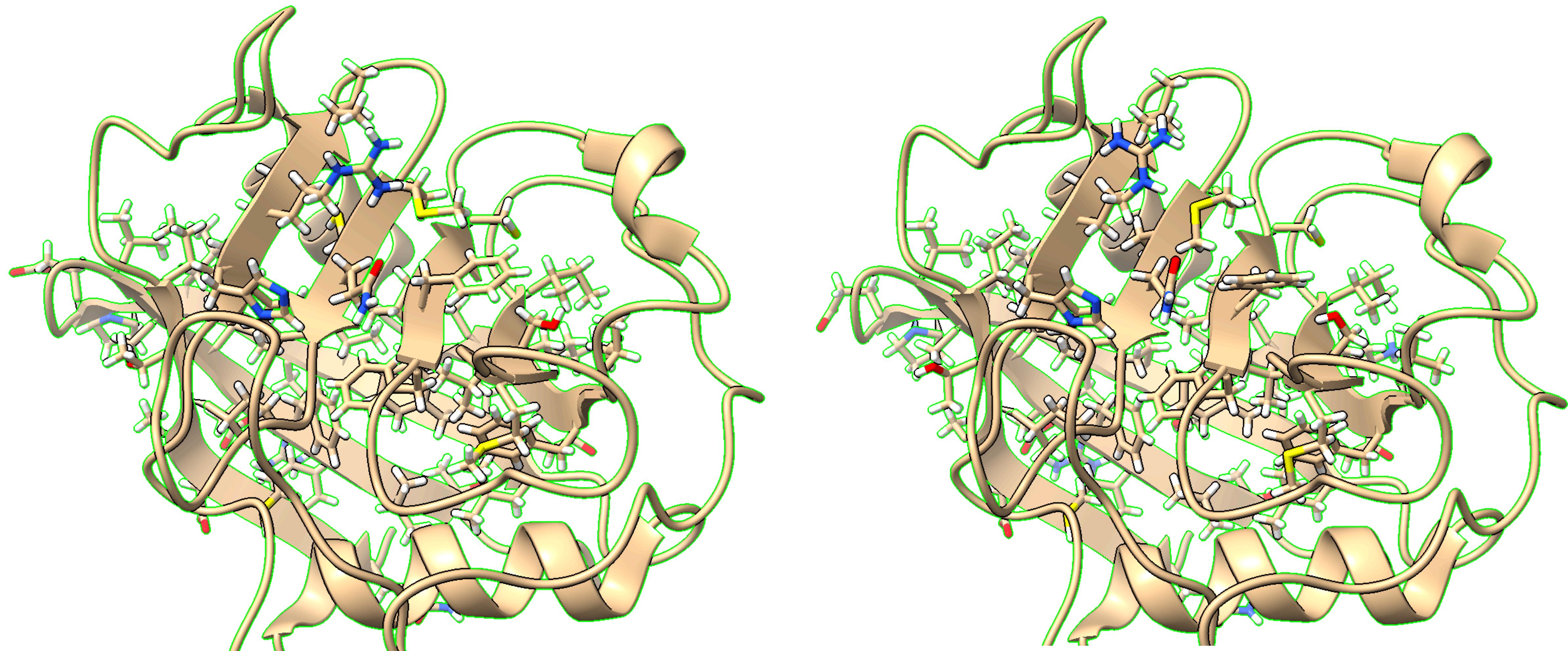
But bound to a drug

1Y57: Complete



Active state,
bound to a drug

In this Cyclophilin A structure 22 aminoacids have alternative conformations identified in the electron density (3K0N)



1. Fraser, J. S. *et al.* Hidden alternative structures of proline isomerase essential for catalysis. *Nature* **462**, 669–673 (2009).



UNIVERSITÀ
DEGLI STUDI
DI MILANO



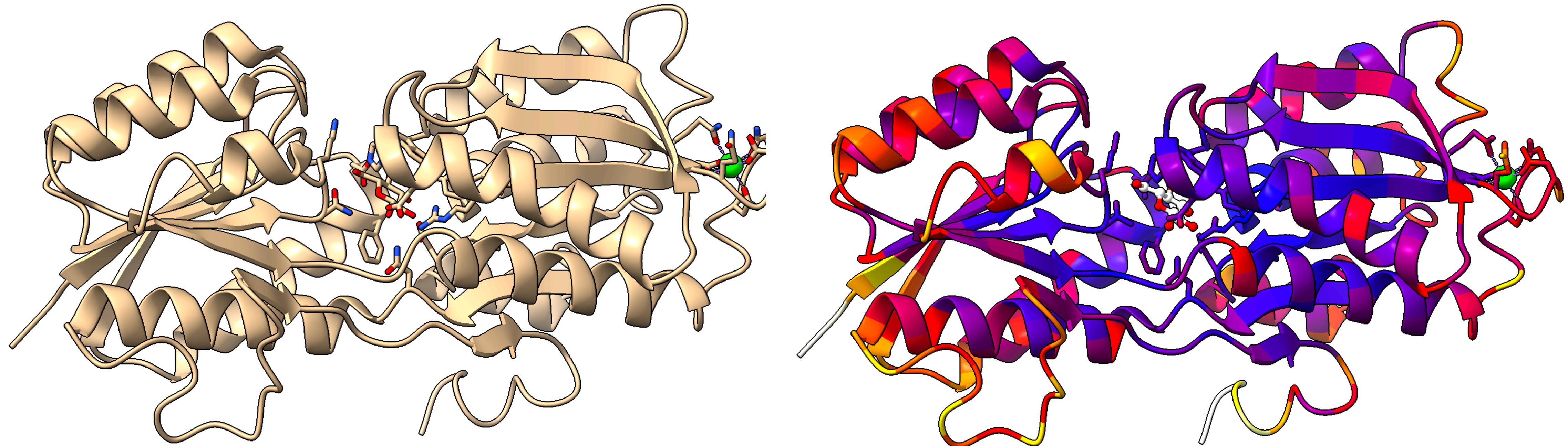
What does X-Ray crystallography tell us about molecules motion?

4. All atoms assigned with the density are associated with a B-Factor (Debye-Waller factor, Temperature factor, atomic displacement parameter). This number reports in part about the fluctuation of the atom in the crystal, but it is also the results of a number of other properties. Nonetheless, even if not to be used quantitatively, it can provide indications of different extent of protein motion in crystal, that may be associated to a similar motion in solution

Sun, Z., Liu, Q., Qu, G., Feng, Y. & Reetz, M. T. Utility of B-Factors in Protein Science: Interpreting Rigidity, Flexibility, and Internal Motion and Engineering Thermostability. *Chem. Rev.* **119**, 1626–1665 (2019).



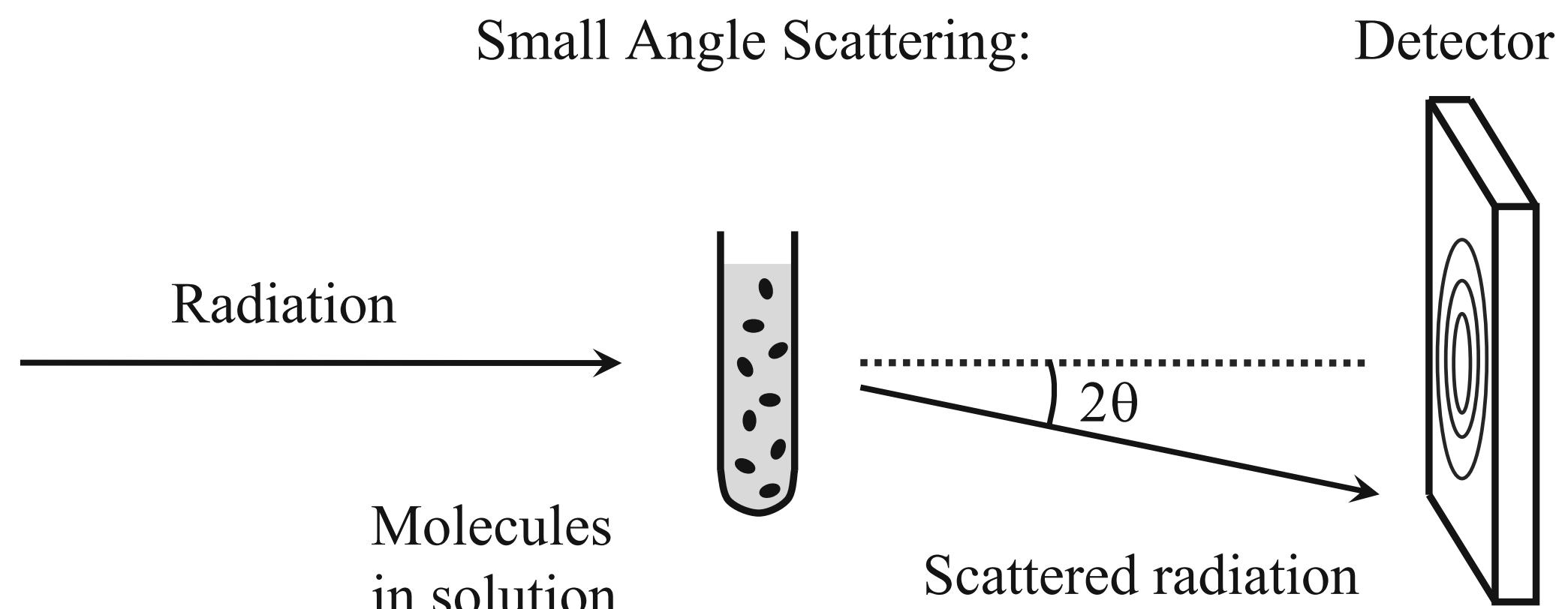
Example: B-Factors



Lighter colours highlight more noisy regions as identified by B-Factors: 2GBP

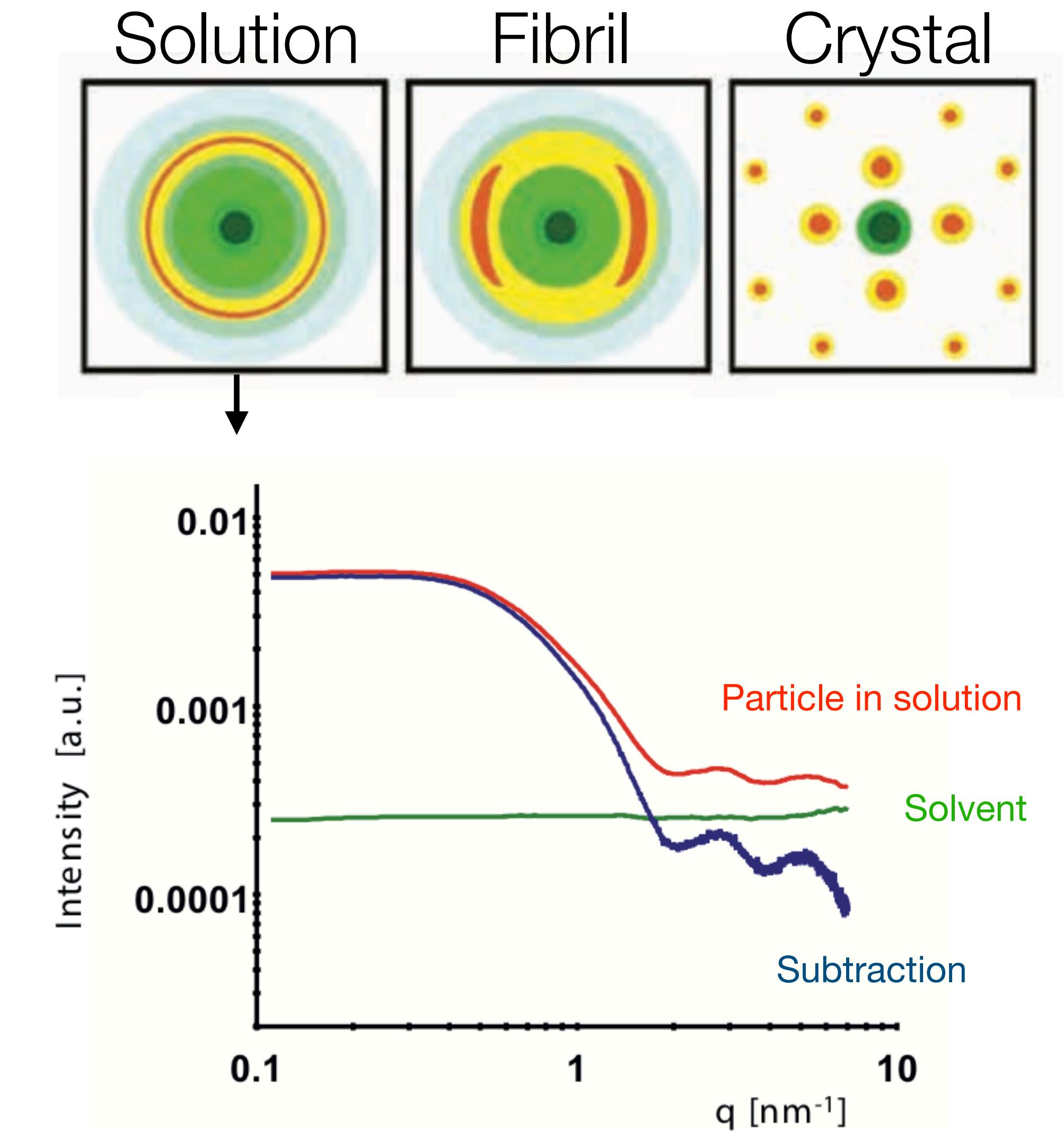


X-rays in solution: small angle X-ray scattering (SAXS)



$$I_m(q) = \sum_{j=1}^{N_A} \sum_{i=1}^{N_A} f_i(q) f_j(q) \frac{\sin(qd_{ij})}{qd_{ij}}.$$

The SAXS intensity for a dilute sample is the sum of the atomic composition $f(q)$ and their relative distance d



X-rays in solution: small angle X-ray scattering (SAXS)

$$I_m(q) = \sum_{j=1}^{N_A} \sum_{i=1}^{N_A} f_i(q) f_j(q) \frac{\sin(qd_{ij})}{qd_{ij}}.$$

Given a protein structure it is possible to calculate its theoretical intensity from the atomic positions, then it is also possible to subtract the effect of bulk water

This is useful to dock proteins together in search of quaternary structure, but also to validate the presence of high-order structures derived from crystals (e.g. is a dimeric structure observed in crystal the result of the crystallisation process or a true dimer already in solution?)

SAXS can be measured for molecules of any size, with no limitations about the extent of their dynamics.

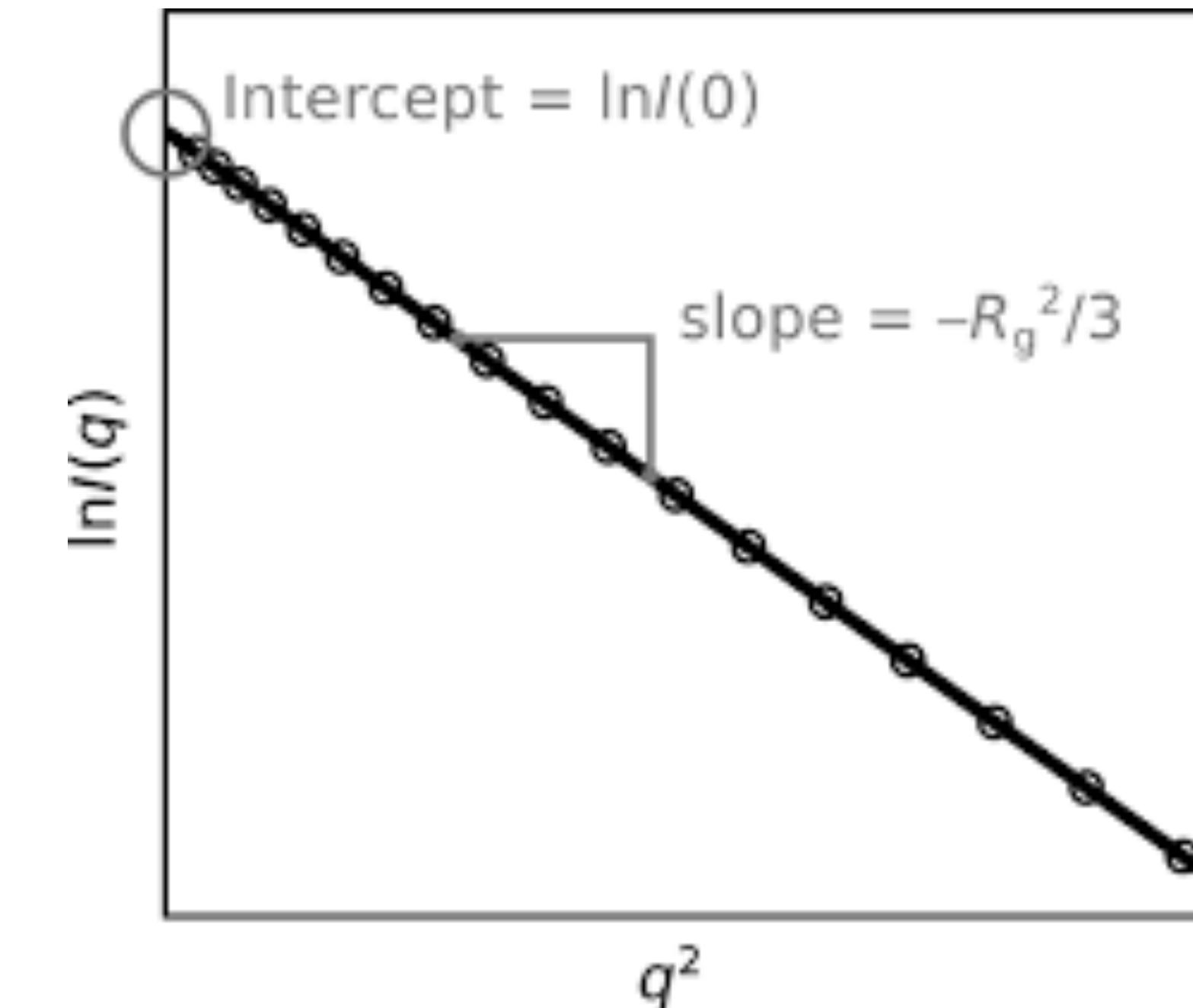


SAXS for intrinsically disordered proteins

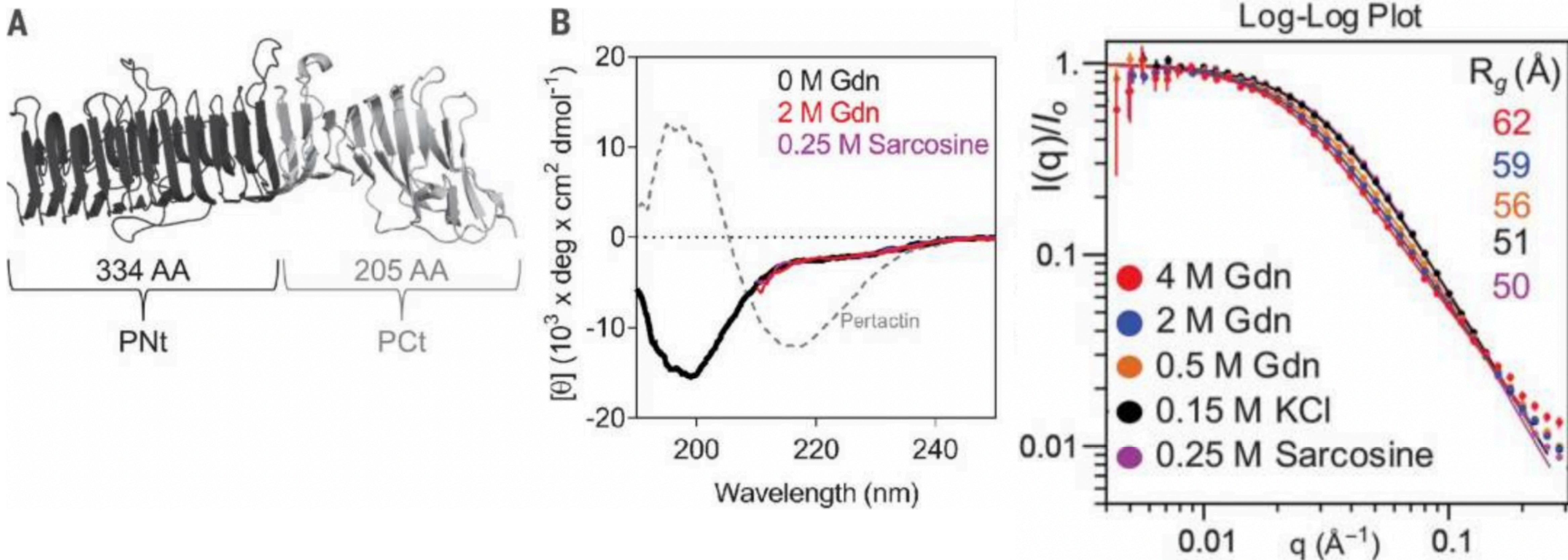
$$I(s) = I(0)e^{\frac{-s^2 R_g^2}{3}}$$

SAXS at very small angles provides information about the size of a system in solution in terms of the radius of gyration. The log of SAXS intensity is proportional to the radius (Guinier relation)

This is useful to interrogate about the oligomeric state of a folded protein but also to detect changes in the conformations explored by intrinsically disordered proteins as a consequence of the solution conditions (solvent composition, temperature, concentration, partners, ...)



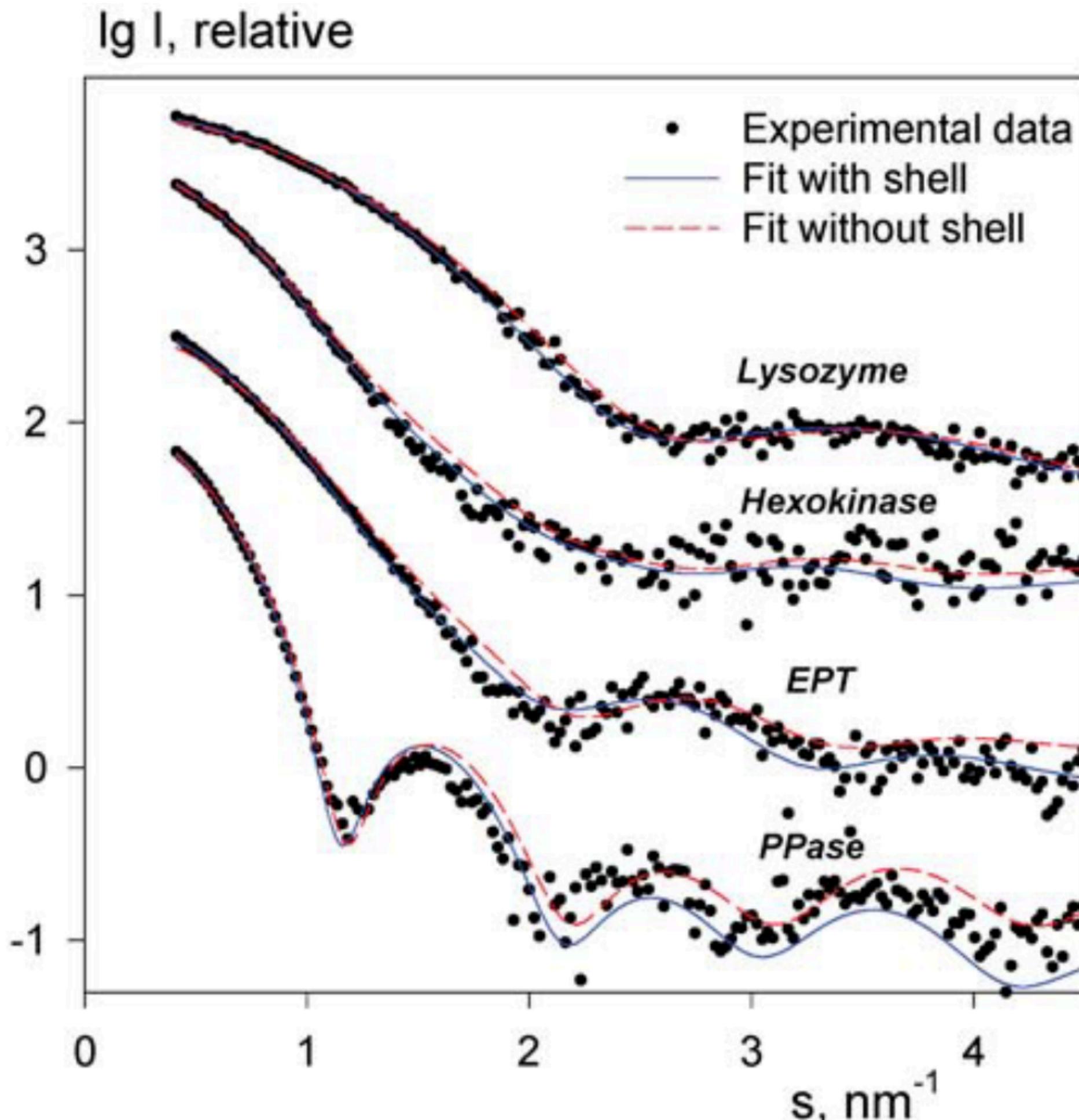
Example: effect of denaturant on the unfolded state of a protein



PNT is disordered when alone and its size depends on the solution conditions

Riback, J. A. *et al.* Innovative scattering analysis shows that hydrophobic disordered proteins are expanded in water. *Science* **358**, 238–241 (2017).

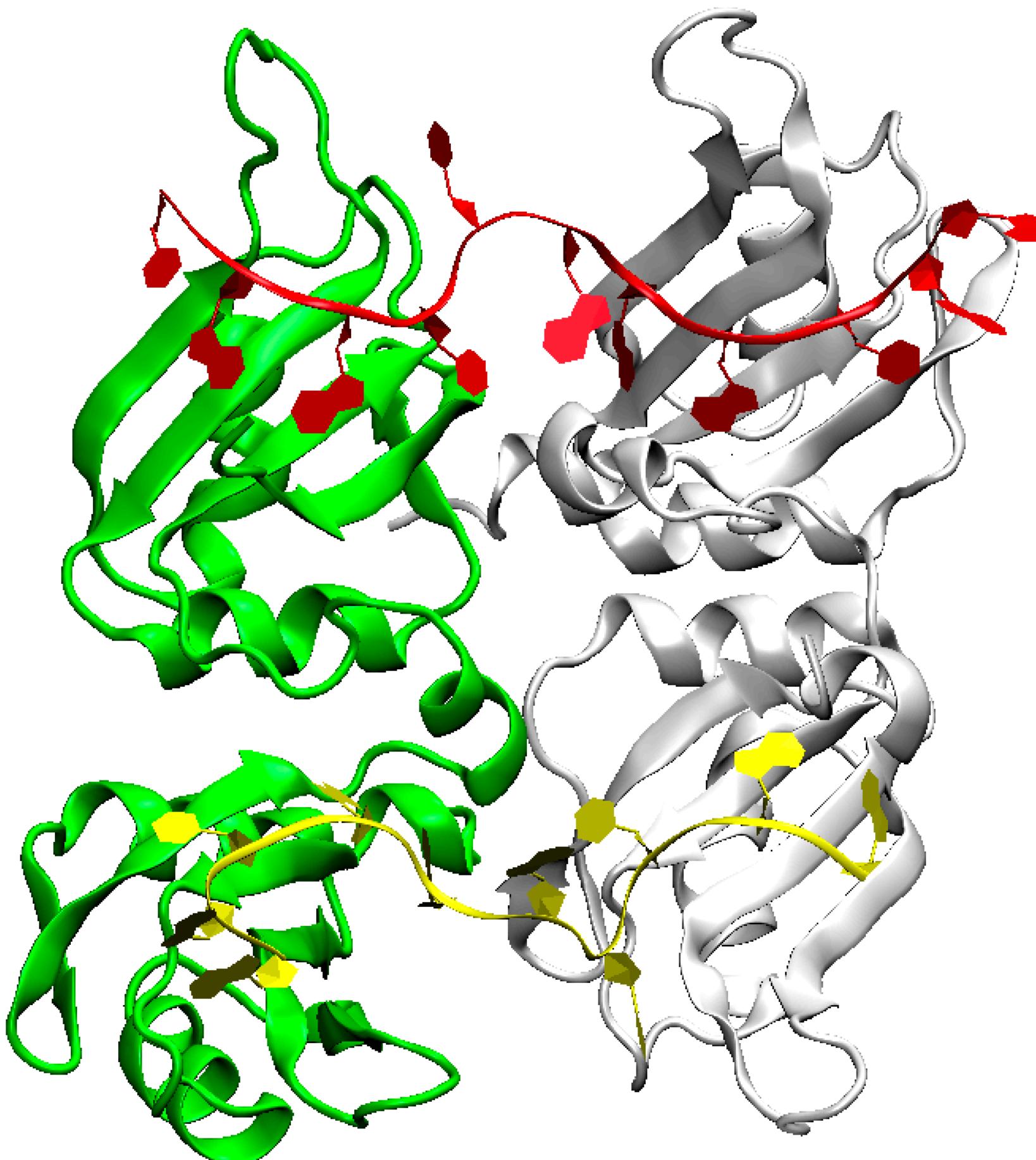
SAXS: testing the validity X-rays structures in solution



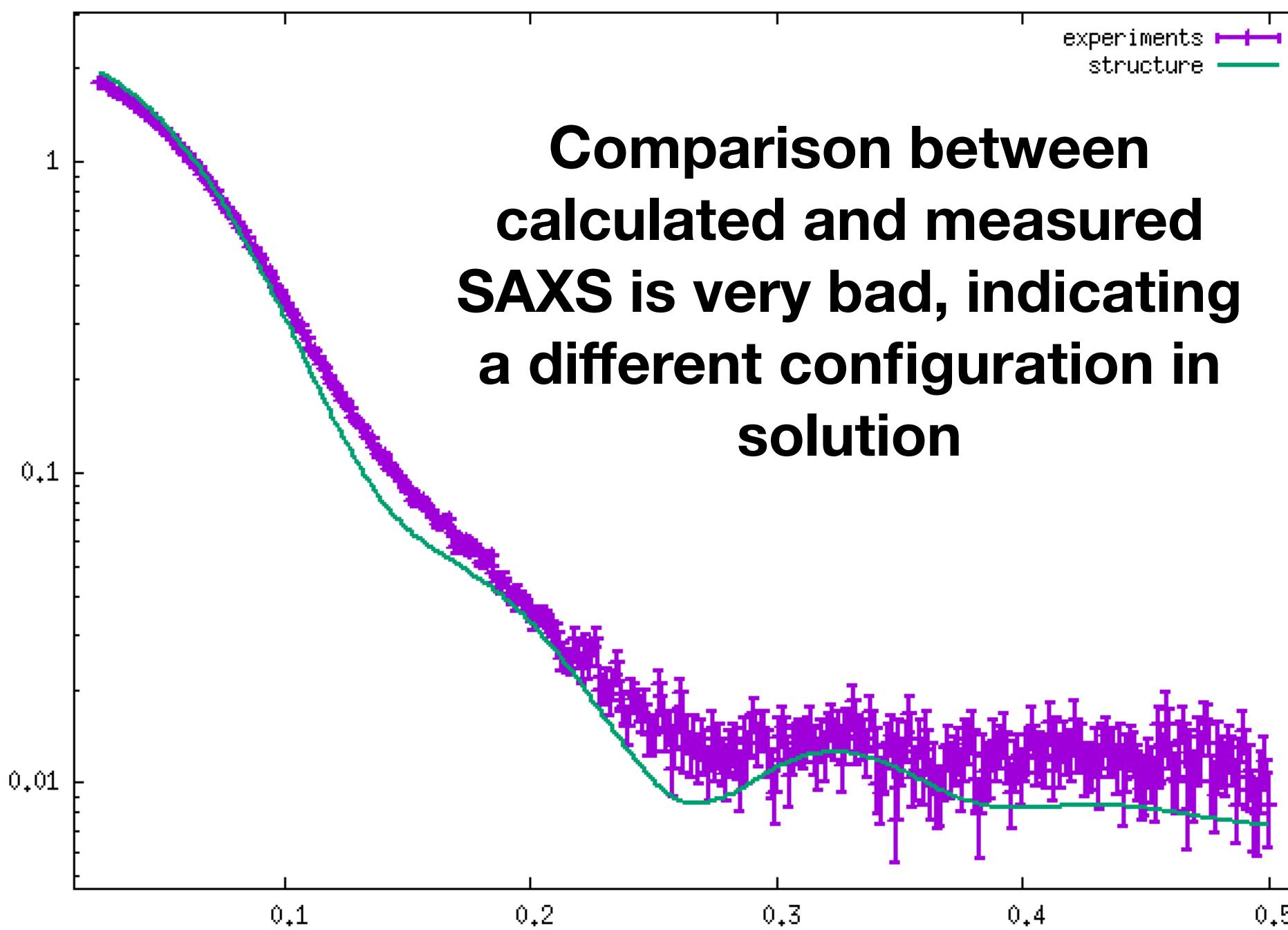
For many proteins, the comparison between the SAXS calculated from their crystal structure and the experimental data is very good, indicating that the structure is representative for the behaviour of the protein in solution.

But this is not always the case, for example for cases of multi-domain proteins with flexible linkers

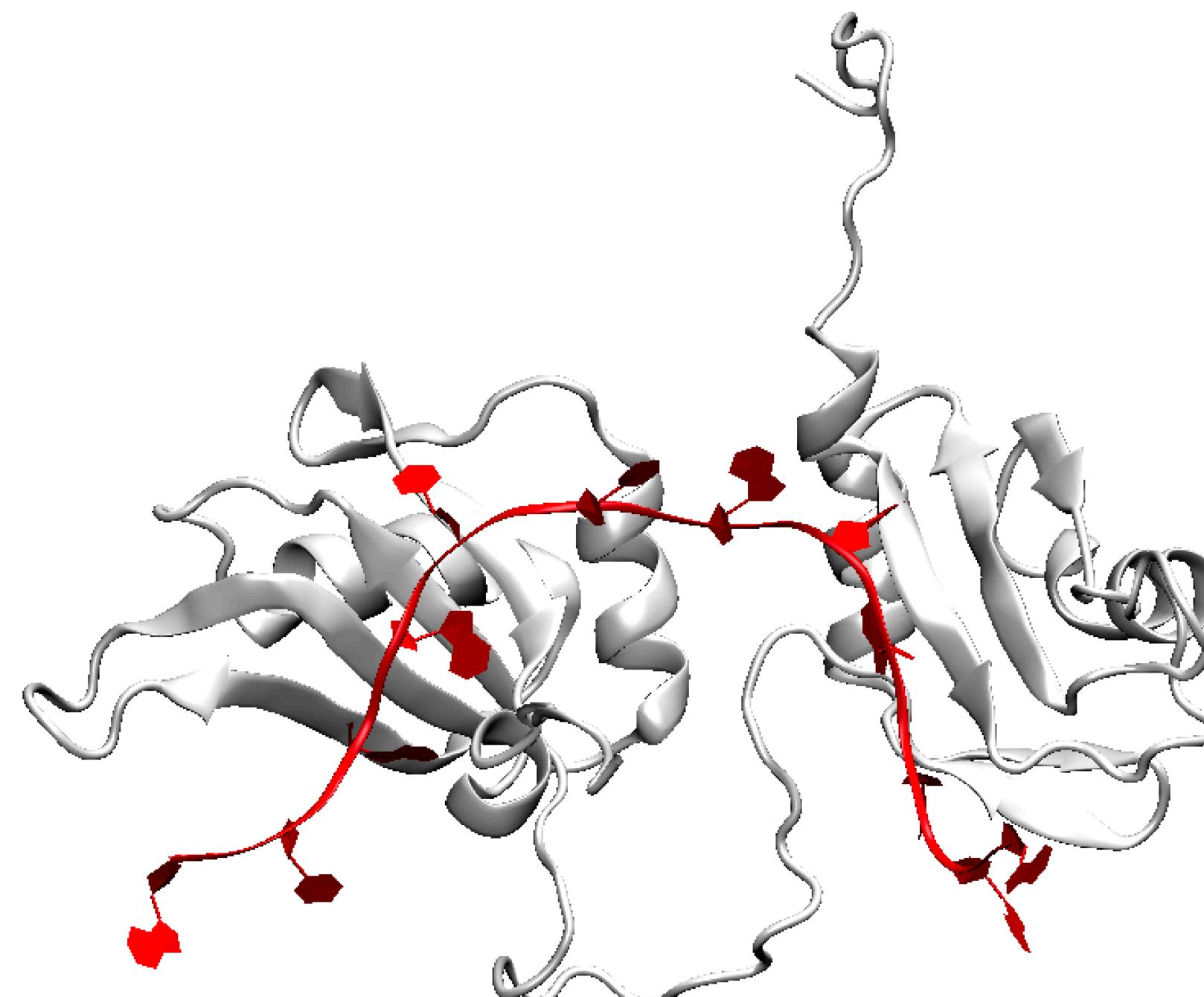
SAXS: testing the validity X-rays structures in solution



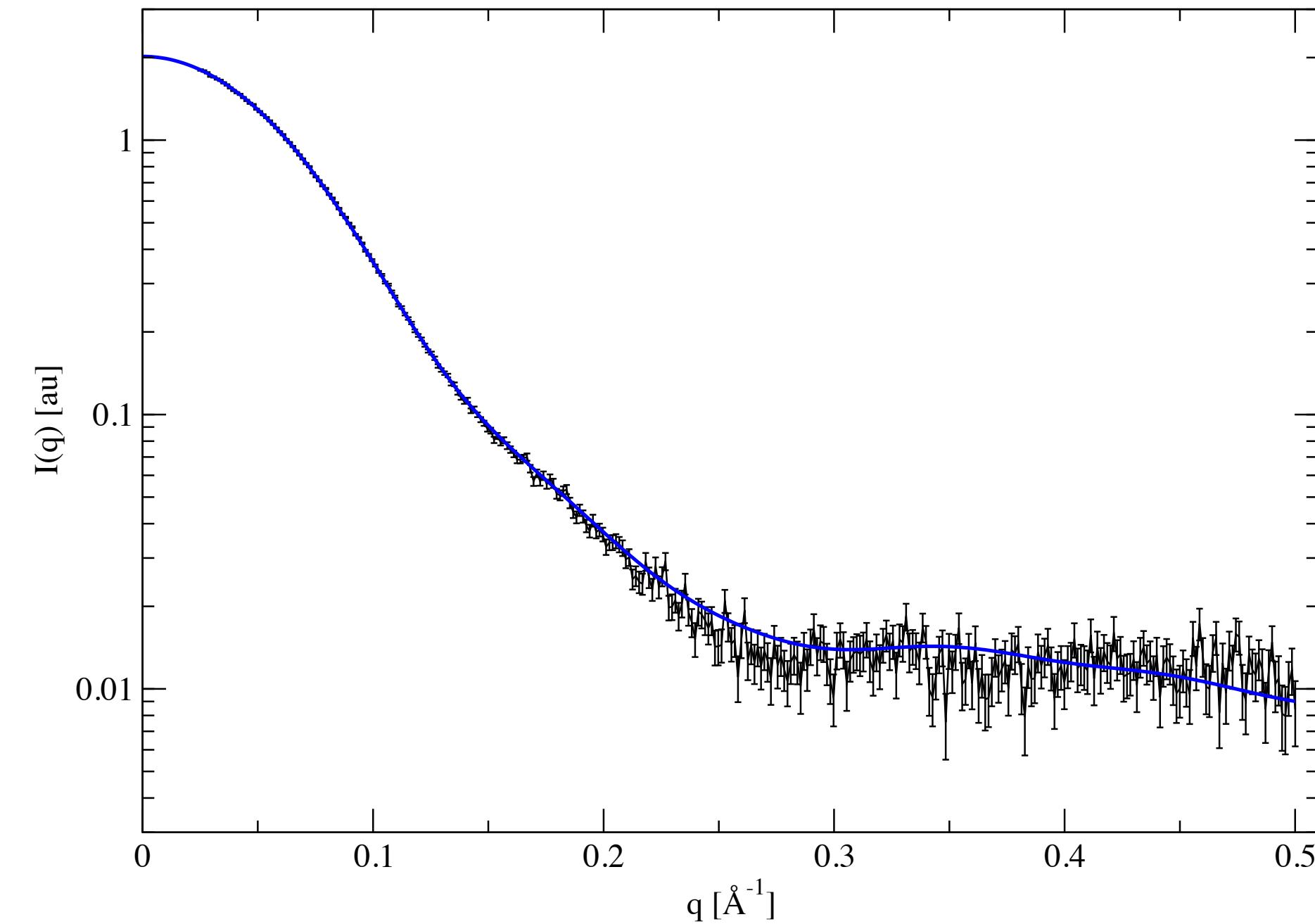
Crystal Structure of the hnRNP A1 protein (RNA binding protein) bound to a short sequence of a pri-miRNA. The structure suggest that a dimeric protein is bound transversally to two miRNA molecules.



SAXS: testing the validity X-rays structures in solution



A structure of a monomeric protein bound to a single RNA fits the SAXS data perfectly:



Kooshapur, H. *et al.* Structural basis for terminal loop recognition and stimulation of pri-miRNA-18a processing by hnRNP A1. *Nat. Commun.* **9**, 2479 (2018).



UNIVERSITÀ
DEGLI STUDI
DI MILANO



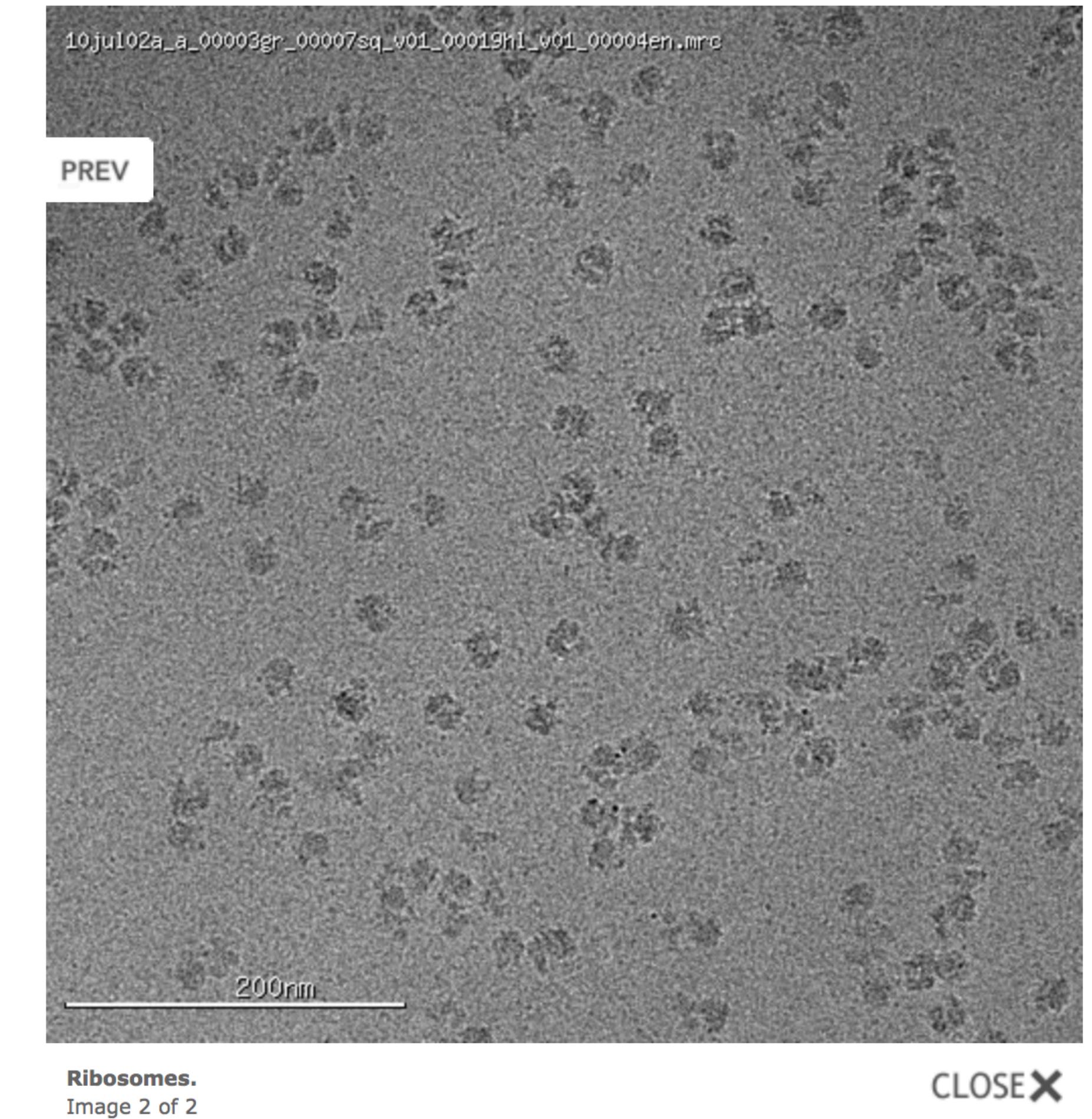
So how do we “take pictures” of molecules? 2. Cryo-Electron Microscopy

Issues are signal/noise, small size, motion on many time scales

SOLUTION: Solution is vitrified by liquid nitrogen on a plate: essentially all the molecules are on a plane, and electron (very few) are sent through the sample to avoid molecular motions so we get clear 2D shadows

See also:

<https://www.youtube.com/watch?v=BJKkC0W-6Qk>



UNIVERSITÀ
DEGLI STUDI
DI MILANO



What does cryo-EM tell us about molecules motion?

The key idea of cryo-EM is that of removing protein motion by freezing the system. Both the motion of the different molecules, that are trapped in the vitrified solution, but also the internal motion.

At odds with crystallography here different copies of the molecule can be trapped in different conformations. But as a consequence it may be difficult to accumulate enough coherent data to determine a structure.



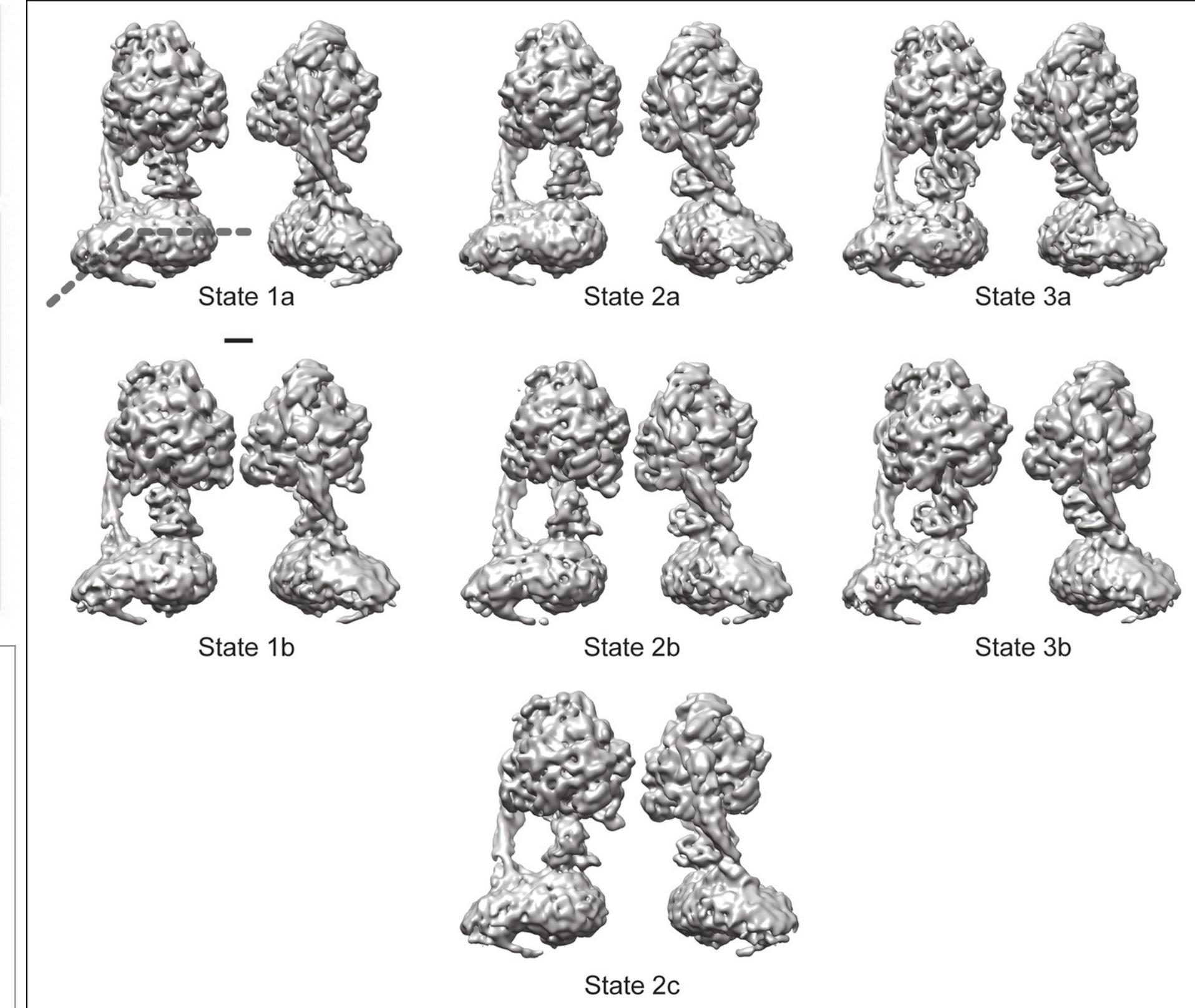
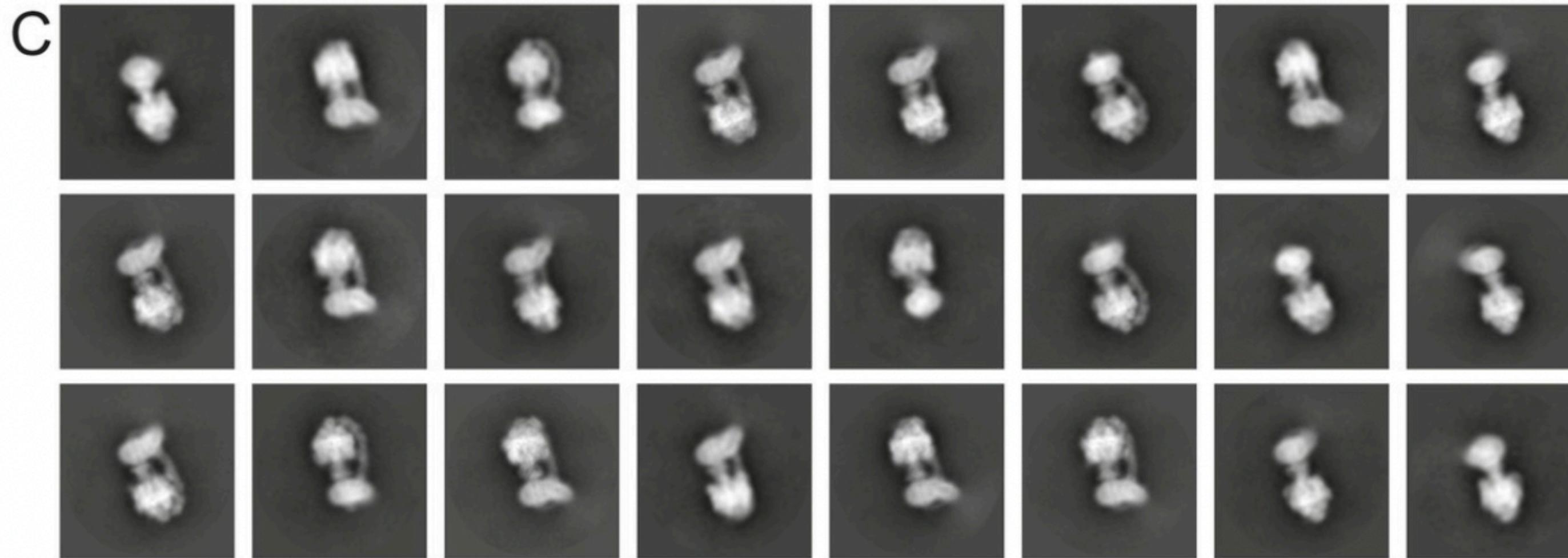
Cryo-EM can be used to identify molecules in multiple conformations

1. Shadows with the same shape (thousands) are clustered and averaged to increase the signal. Image recognition techniques are used to speed up the process.
2. Averaged 2D images are then organised in clusters: are the 2D images representing the same conformation?
3. All the images that are associated to the same conformation are then used to obtain a 3D density map in which it is then possible to build an atomic model

If the molecule populates too many conformations, or if regions of the molecule are too flexible they are invisible because it is not possible to accumulate enough 2D images to improve the signal to noise ratio. Different classes may result in different RESOLUTION both for different regions of same conformation as well as for different conformations.



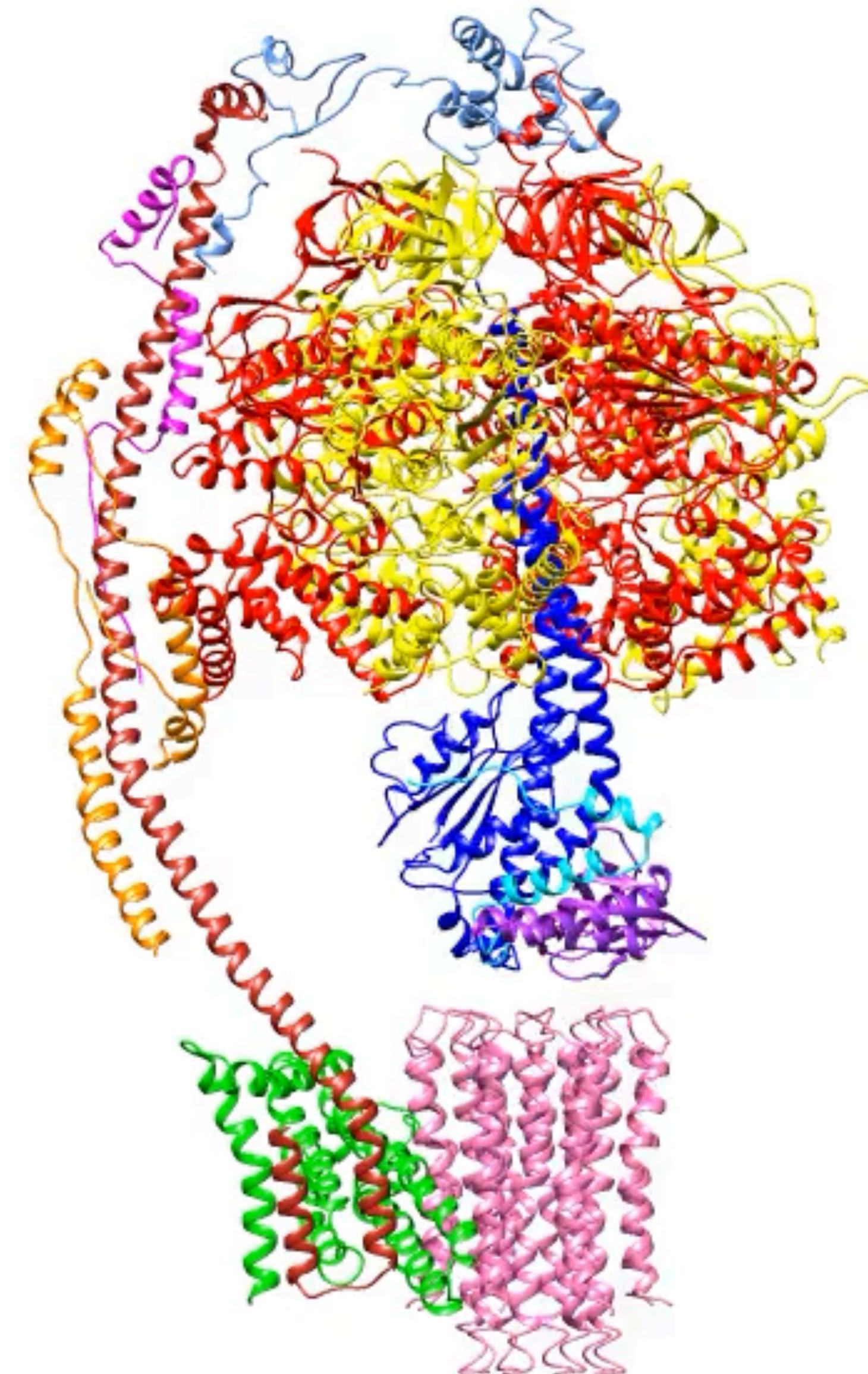
Example: Structure and conformational states of the bovine mitochondrial ATP synthase by cryo-EM



Adenosine triphosphate (ATP), the chemical energy currency of biology, is synthesized in eukaryotic cells primarily by the mitochondrial ATP synthase. ATP synthases operate by a rotary catalytic mechanism where proton translocation through the membrane-inserted FO region is coupled to ATP synthesis in the catalytic F1 region via rotation of a central rotor subcomplex.

This is an interpolation of three states to render a movie

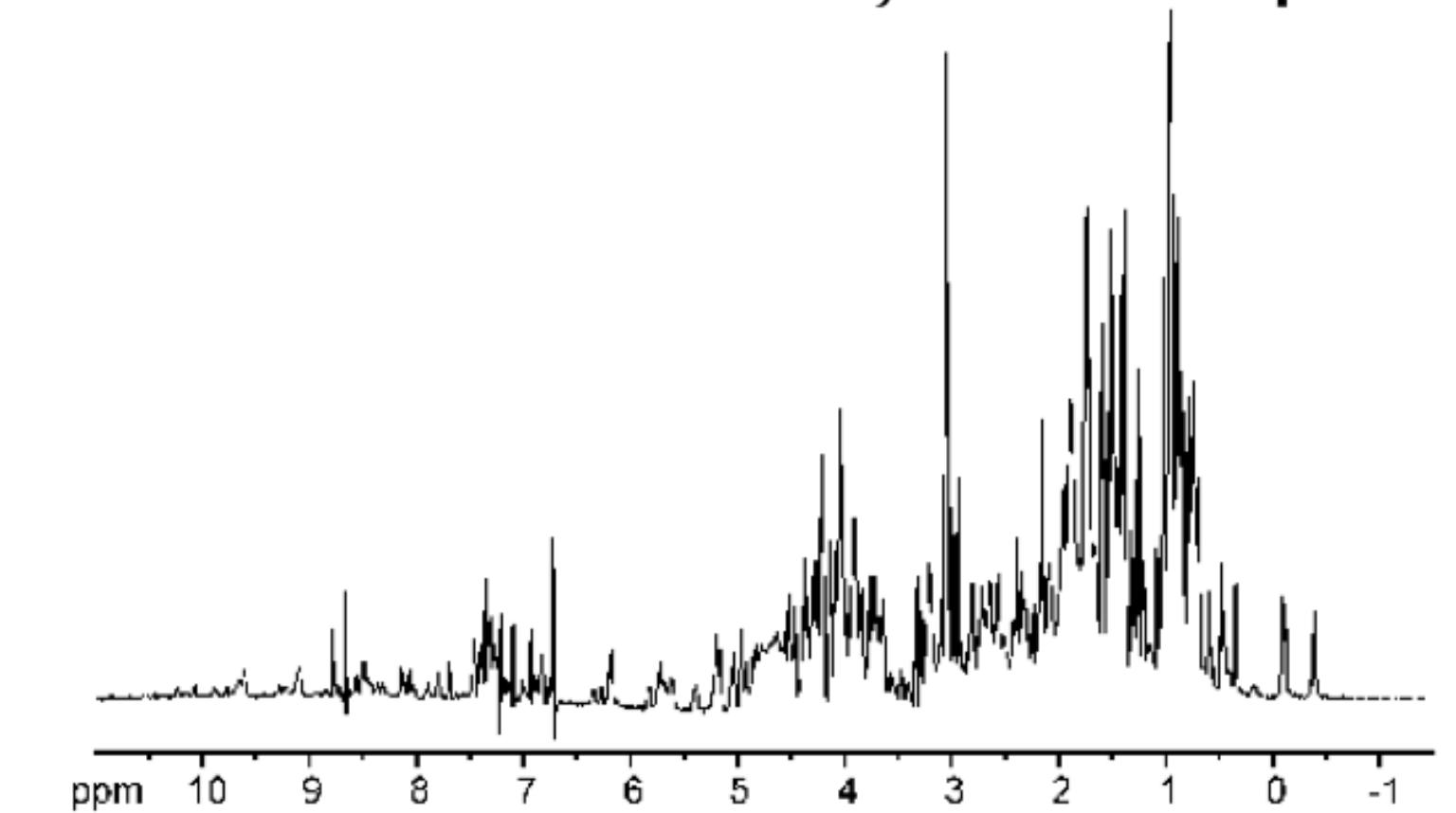
It is not necessarily how the conformational change happens



So how do we “take pictures” of molecules? 3. Nuclear Magnetic Resonance Spectroscopy

Issues are signal/noise, small size, motion on many time scales

SOLUTION: We employ the radio frequency absorption and emission properties of nuclei in presence of a magnetic field. In this way we can look at the signal of the same atom in all the molecules.



Here the signal to noise ratio is determined by the protein concentration and the size of the molecules, where the larger the molecule the weaker the signal (usually)



UNIVERSITÀ
DEGLI STUDI
DI MILANO

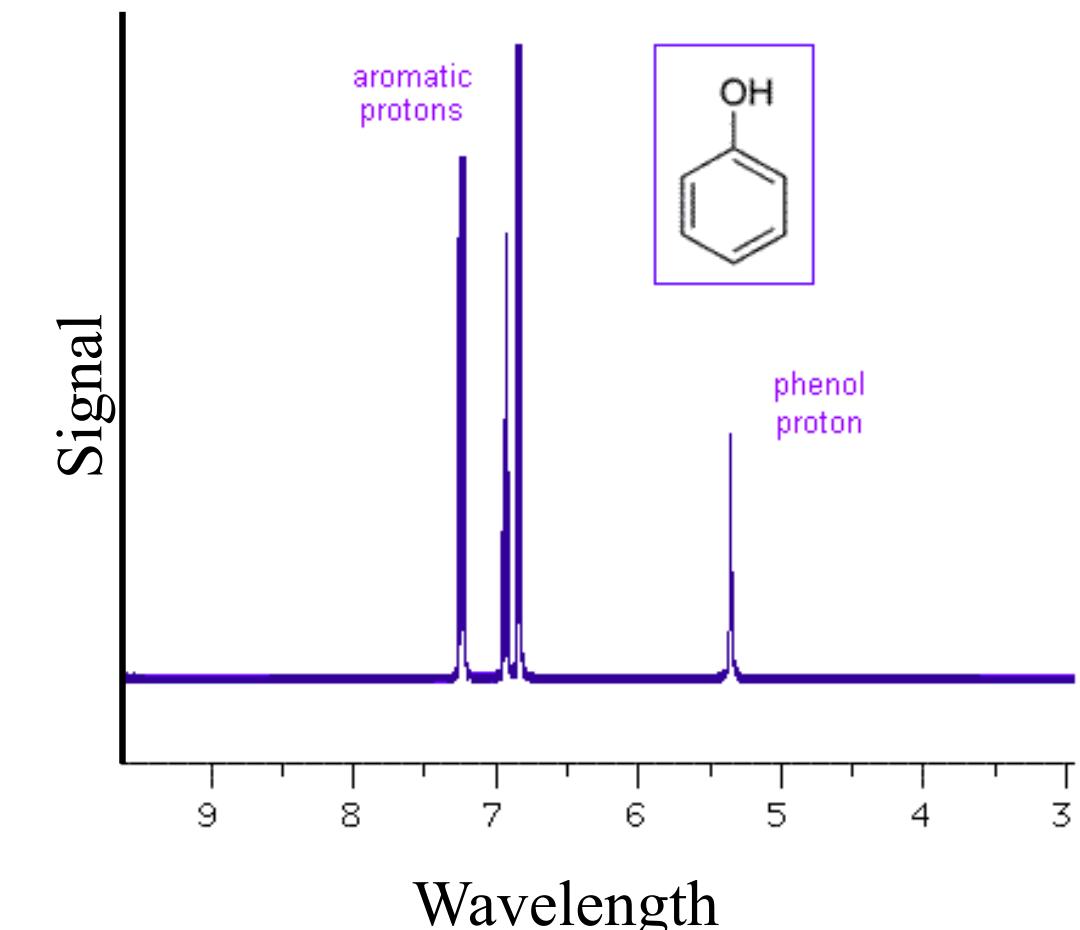


Principles of NMR spectroscopy

- **Principles:**

- A strong **external electric field** applied by the NMR machine to the sample results in **spatial alignment** of protein atoms with nuclear spin (e.g. ^1H , ^{13}C , ^{15}N)
- An **energy pulse** transiently disturbs this alignment
- The nuclei revert to their original state while **emitting radio waves** that are sensed by the NMR machine
- The waves are different for each atoms, and affected by their atom neighbors (**chemical shift**)
- By using the known chemical shifts of nuclei at different chemical environments, **the structure of the protein can be deciphered** from its NMR spectrum

- 1D NMR spectrum of protons from different locations in a phenol molecule:



Introduction to Proteins, Kessel and Ben-Tal, 2018

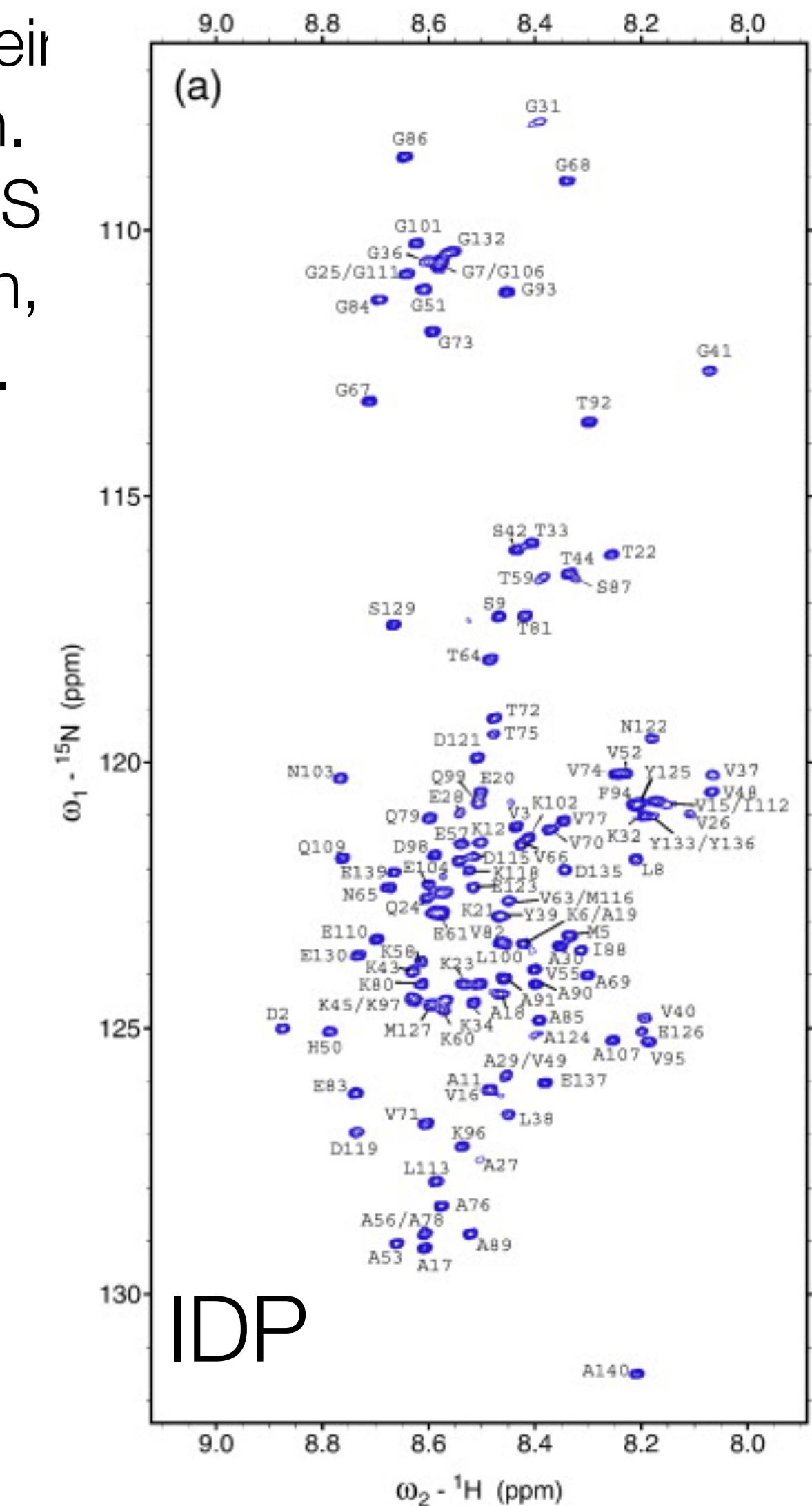
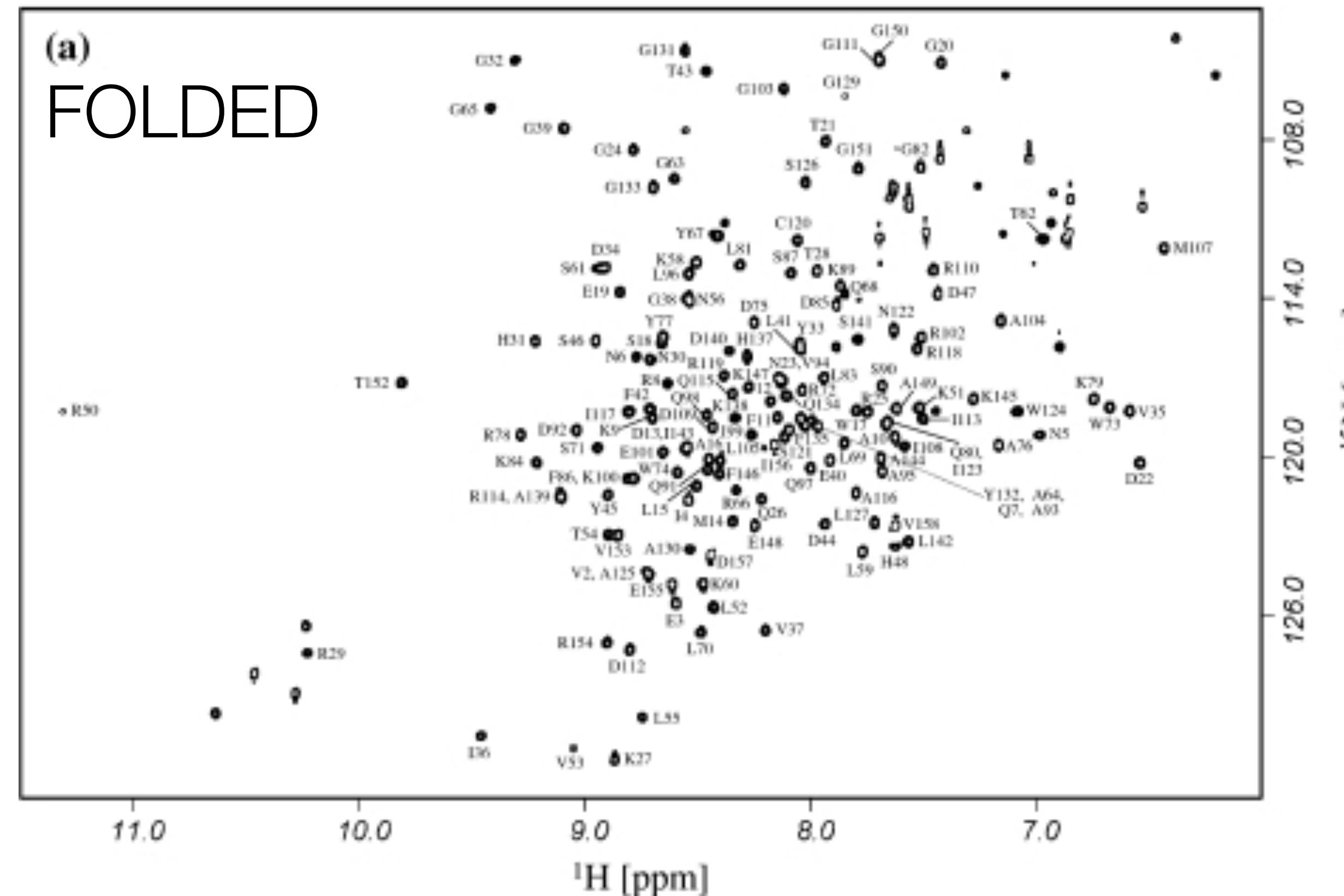


UNIVERSITÀ
DEGLI STUDI
DI MILANO



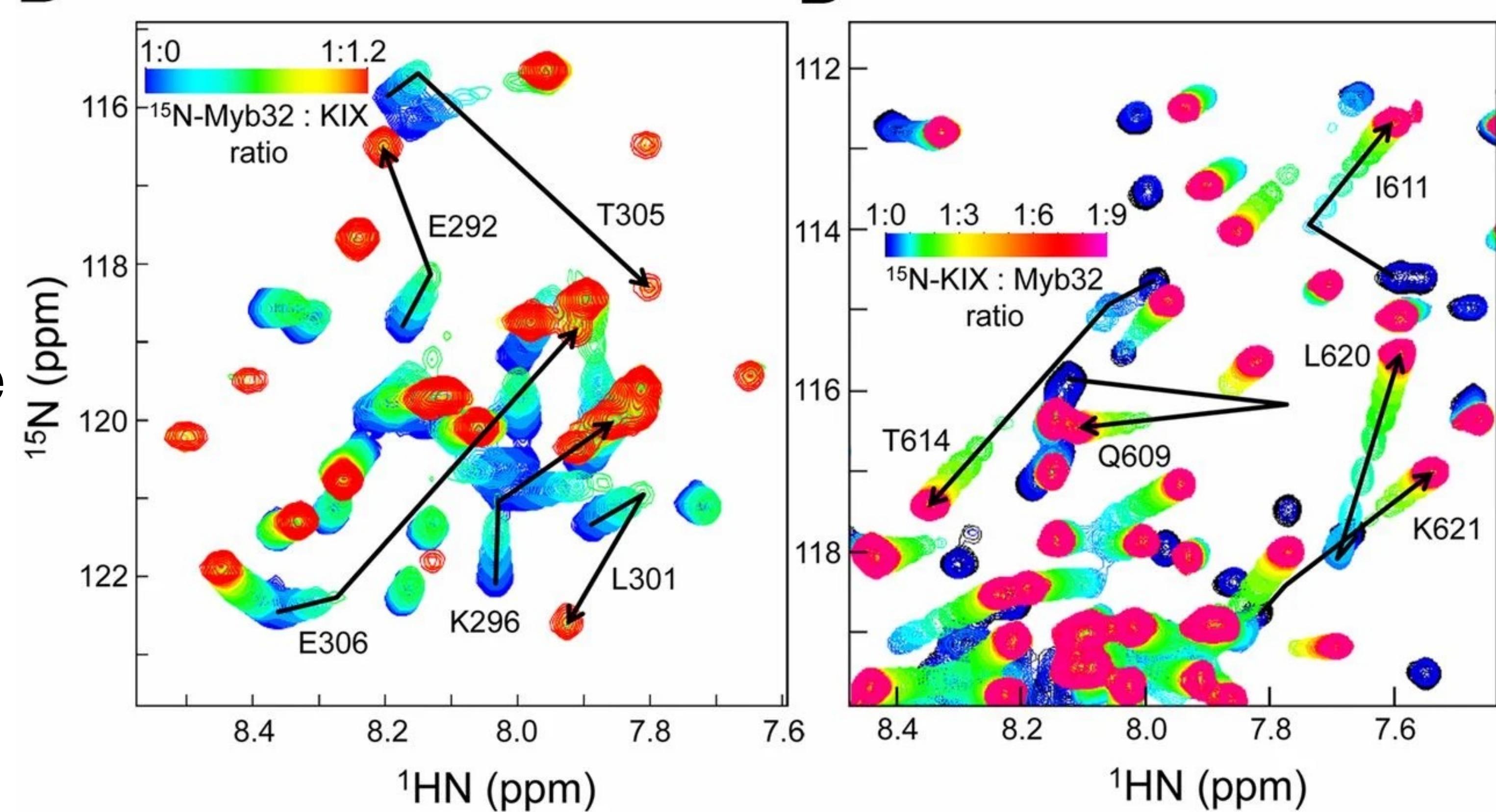
What does NMR tell us about molecules motion?

With NMR we can observe molecules in solution, without posing any constraint on their intermolecular and internal motion, apart from keeping a relatively high concentration. Indeed chemical shifts can be observed equally well for folded proteins as well as IDPs. At the same time other NMR measurements typically used for structure determination, like NOE, do work only if the protein in solution populates mostly a single conformer.



What does NMR tell us about molecules motion?

But chemical shifts, as well as other NMR experiments, can be used to follow conformational transition resulting from changes in the solution conditions as well as from interactions:



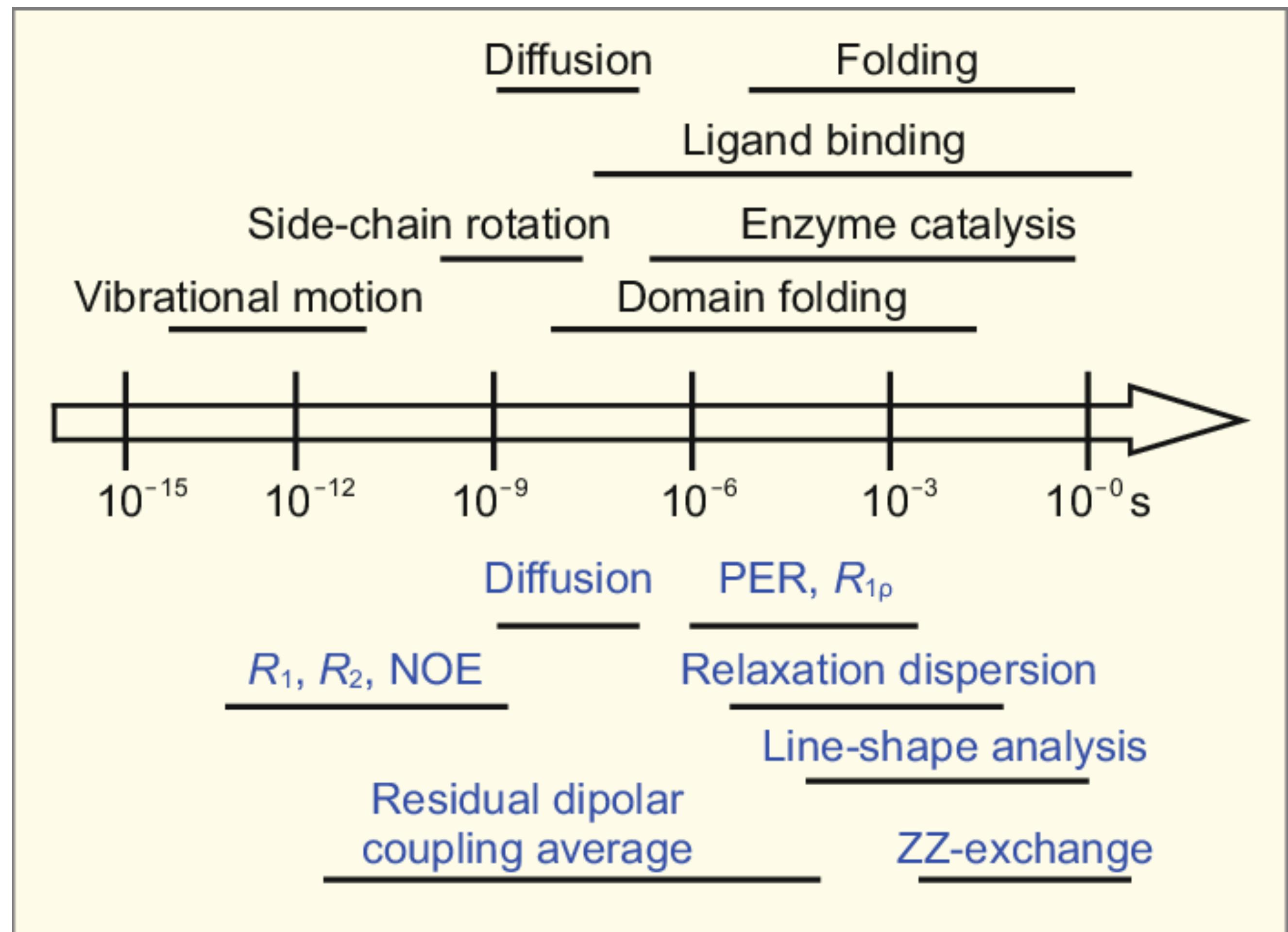
^{15}N -labeled Myb32 showing chemical shift changes upon titration with KIX.

^{15}N -labeled KIX showing chemical shift changes upon titration with Myb32.

Arai, M., Sugase, K., Dyson, H. J. & Wright, P. E. Conformational propensities of intrinsically disordered proteins influence the mechanism of binding and folding. *Proc. Natl. Acad. Sci.* **112**, 9614–9619 (2015).

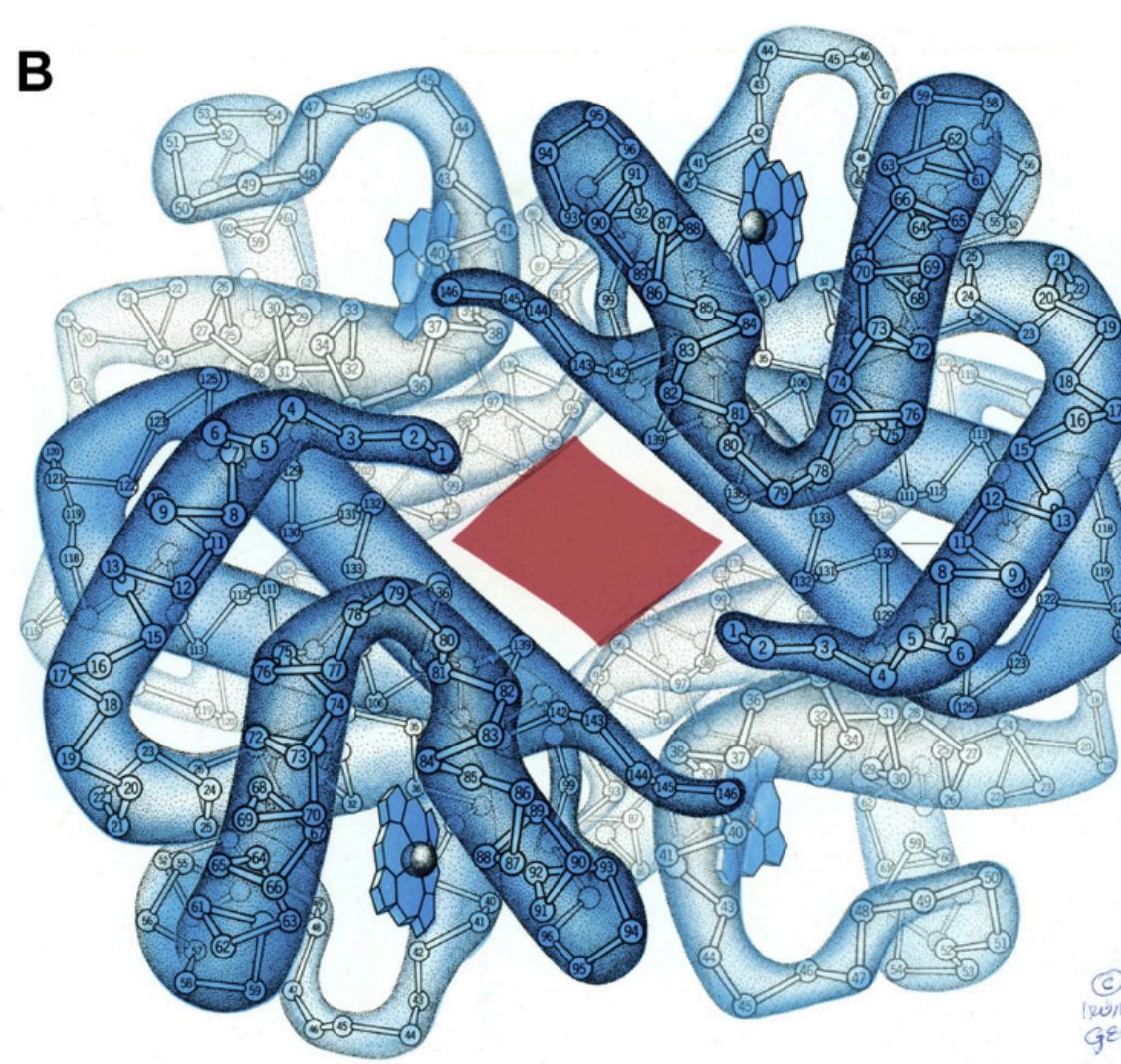
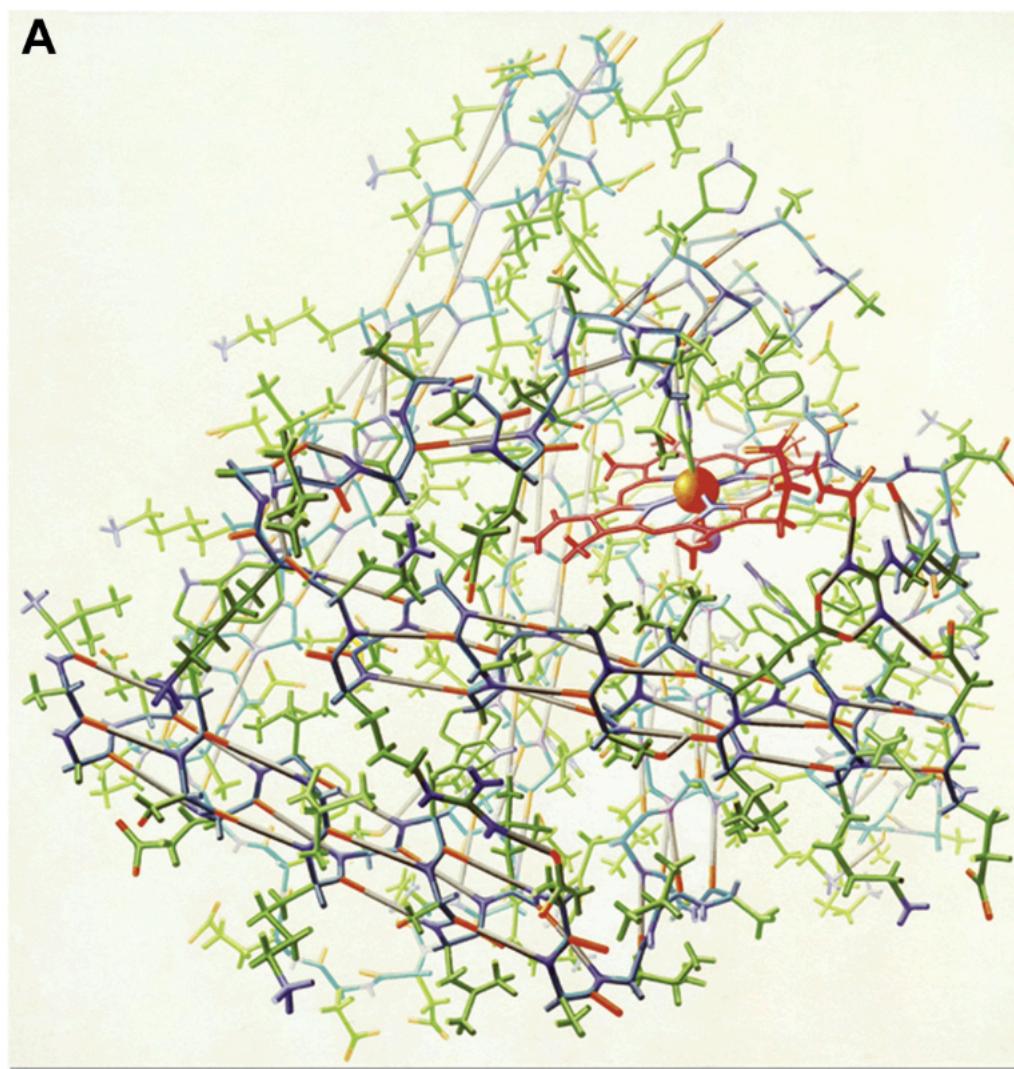
What does NMR tell us about molecules motion?

But the strength of is that of enabling a large number of different experiments that can probe intra- and inter- molecular interactions focusing on different time-scales



Biomolecular Visualisation

The result of a structure determination experiment is a structure, but in practice this is a collection of relative coordinates. How do you look at them?



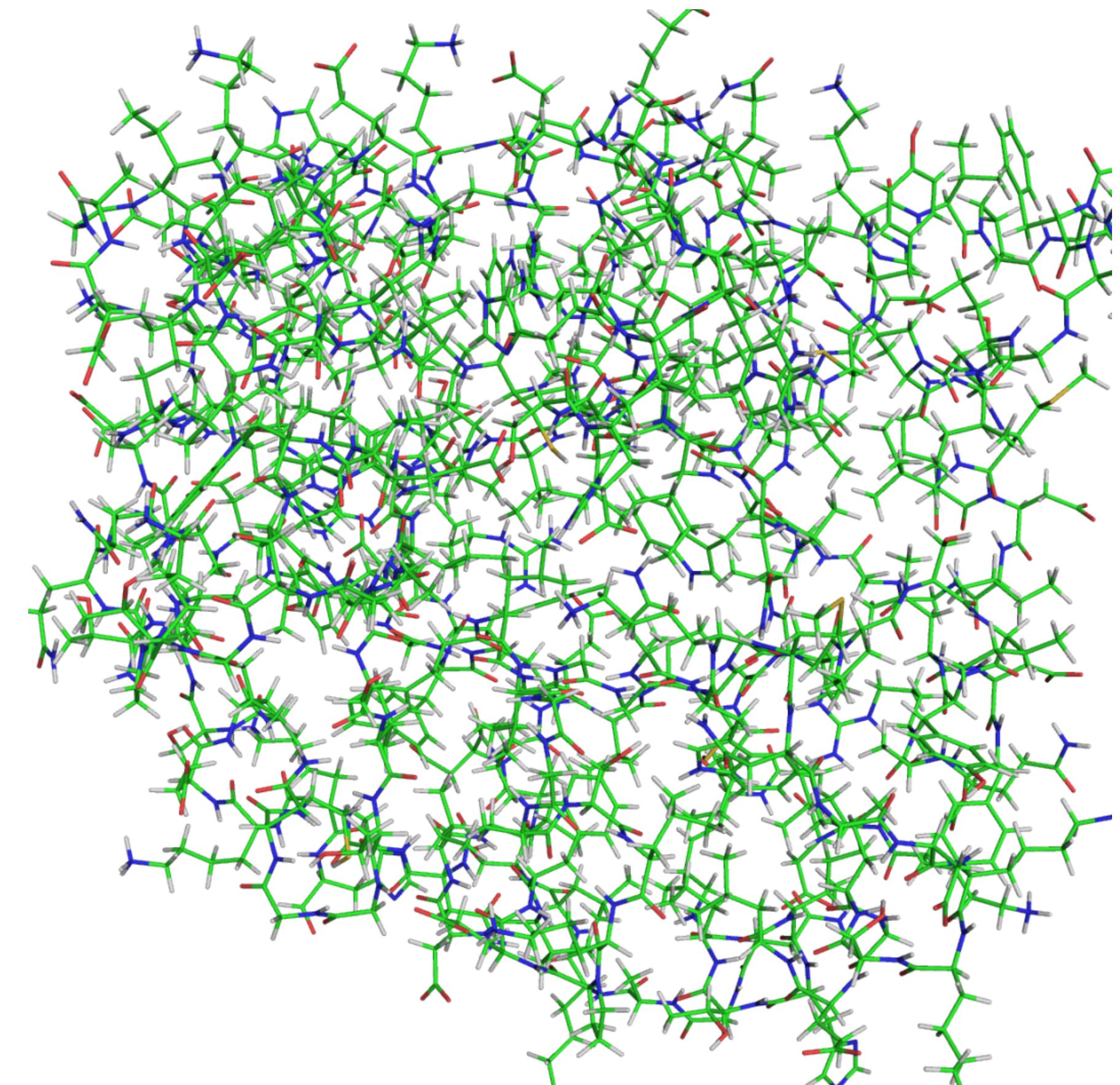
ATOM	1	N	MET	A	1	34.735	4.549	16.028	1.00	76.42	N
ATOM	2	CA	MET	A	1	35.526	3.744	17.004	1.00	75.41	C
ATOM	3	C	MET	A	1	37.035	3.970	16.806	1.00	72.52	C
ATOM	4	O	MET	A	1	37.861	3.372	17.506	1.00	73.81	O
ATOM	5	CB	MET	A	1	35.156	2.255	16.884	1.00	75.63	C
ATOM	6	CG	MET	A	1	35.743	1.350	17.962	1.00	76.82	C
ATOM	7	SD	MET	A	1	35.031	-0.302	17.938	1.00	74.07	S
ATOM	8	CE	MET	A	1	34.483	-0.461	19.683	1.00	80.55	C
ATOM	9	N	THR	A	2	37.388	4.835	15.853	1.00	67.56	N
ATOM	10	CA	THR	A	2	38.794	5.157	15.596	1.00	61.21	C
ATOM	11	C	THR	A	2	39.188	6.347	16.486	1.00	54.29	C
ATOM	12	O	THR	A	2	38.494	7.367	16.512	1.00	55.70	O
ATOM	13	CB	THR	A	2	39.040	5.515	14.099	1.00	61.40	C
ATOM	14	OG1	THR	A	2	38.131	4.776	13.263	1.00	60.26	O
ATOM	15	CG2	THR	A	2	40.466	5.160	13.699	1.00	55.32	C
ATOM	16	N	LYS	A	3	40.255	6.183	17.260	1.00	46.34	N
ATOM	17	CA	LYS	A	3	40.732	7.232	18.158	1.00	38.27	C
ATOM	18	C	LYS	A	3	41.495	8.295	17.363	1.00	35.08	C
ATOM	19	O	LYS	A	3	42.183	7.983	16.388	1.00	36.81	O
ATOM	20	CB	LYS	A	3	41.687	6.654	19.211	1.00	35.54	C
ATOM	21	CG	LYS	A	3	41.260	5.348	19.865	1.00	36.42	C
ATOM	22	CD	LYS	A	3	40.211	5.526	20.950	1.00	32.79	C
ATOM	23	CE	LYS	A	3	40.060	4.229	21.730	1.00	29.65	C
ATOM	24	NZ	LYS	A	3	39.048	4.320	22.810	1.00	32.53	N
ATOM	25	N	THR	A	4	41.427	9.535	17.829	1.00	33.00	N
ATOM	26	CA	THR	A	4	42.114	10.647	17.187	1.00	29.13	C
ATOM	27	C	THR	A	4	42.865	11.477	18.226	1.00	28.76	C
ATOM	28	O	THR	A	4	42.622	11.350	19.435	1.00	34.44	O
ATOM	29	CB	THR	A	4	41.114	11.557	16.453	1.00	31.26	C
ATOM	30	OG1	THR	A	4	40.130	12.050	17.379	1.00	31.97	O
ATOM	31	CG2	THR	A	4	40.411	10.780	15.355	1.00	29.26	C
ATOM	32	N	LEU	A	5	43.785	12.309	17.765	1.00	21.85	N
ATOM	33	CA	LEU	A	5	44.557	13.177	18.650	1.00	23.03	C
ATOM	34	C	LEU	A	5	44.296	14.618	18.200	1.00	20.84	C
ATOM	35	O	LEU	A	5	43.878	14.844	17.073	1.00	28.21	O
ATOM	36	CB	LEU	A	5	46.054	12.838	18.558	1.00	18.36	C
ATOM	37	CG	LEU	A	5	46.504	11.499	19.167	1.00	20.69	C
ATOM	38	CD1	LEU	A	5	47.884	11.128	18.668	1.00	18.37	C
ATOM	39	CD2	LEU	A	5	46.497	11.561	20.697	1.00	14.10	C

Biomolecular Visualisation

The advent of computers has allowed to dig into the structures and consequently has pushed to define standards to help the rationalisation of the structural features

- **Wireframe (bonds connectivity)**

The covalent bonds in the protein are shown as wires, coloured by the type of atoms they connect. The atoms are not shown.



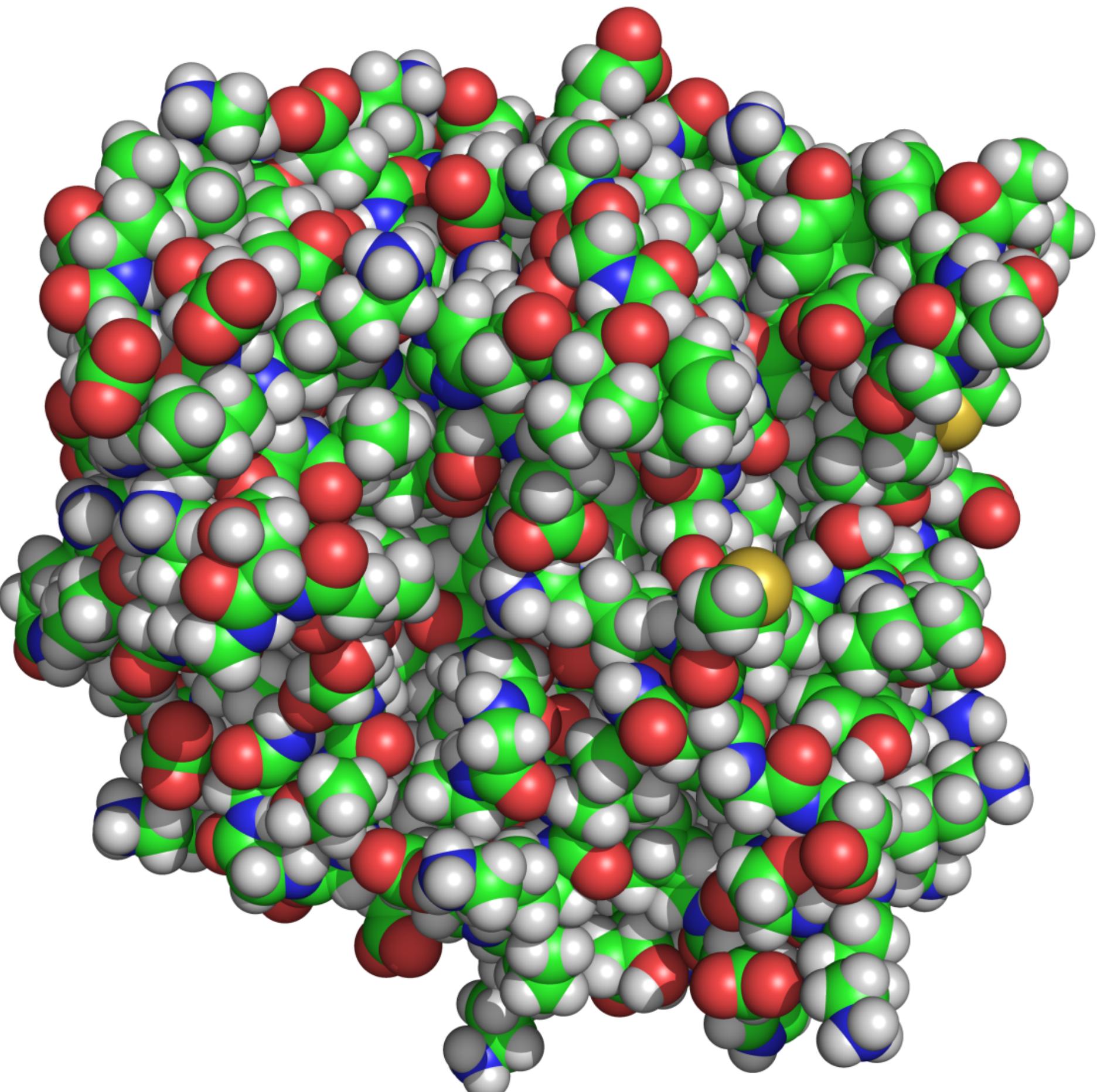
Different graphical representations reveal different properties

The atoms are represented as spheres with a van der Waals radius, and coloured according to the CPK convention

In chemistry, the **CPK coloring** (for **Corey–Pauling–Koltun**) is a popular color convention for distinguishing **atoms** of different **chemical elements** in molecular models.

- White for **hydrogen**
- Black for **carbon**
- Blue for **nitrogen**
- Red for **oxygen**
- Deep yellow for **sulfur**
- Purple for **phosphorus**
- Light, medium, medium dark, and dark green for the **halogens** (**F, Cl, Br, I**)
- Silver for metals (**Co, Fe, Ni, Cu**)

- **Space-fill (general shape, size)**



Different graphical representations reveal different properties

- **Ribbon (topology, secondary structures)**

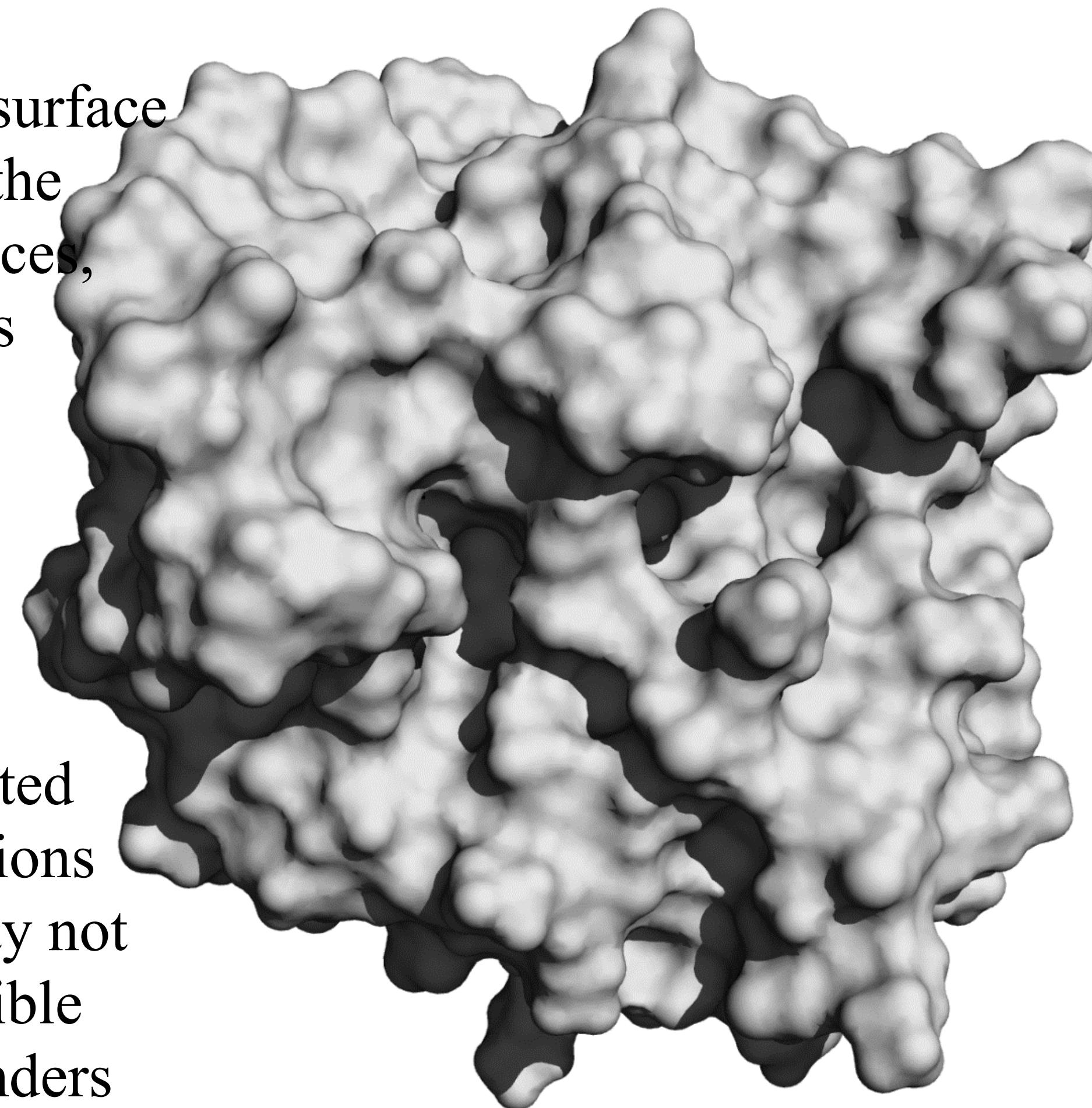
The atoms and bonds of the protein are not shown. Instead, the backbone is represented as a ribbon with different colors, according to the secondary structure of the chain in that region: α -helices are colored in red, β -strands in yellow, and loops in green, as well as the disordered parts of the chain. The shape of the ribbon was calculated as the line going through all the C_α atoms of the protein.



Different graphical representations reveal different properties

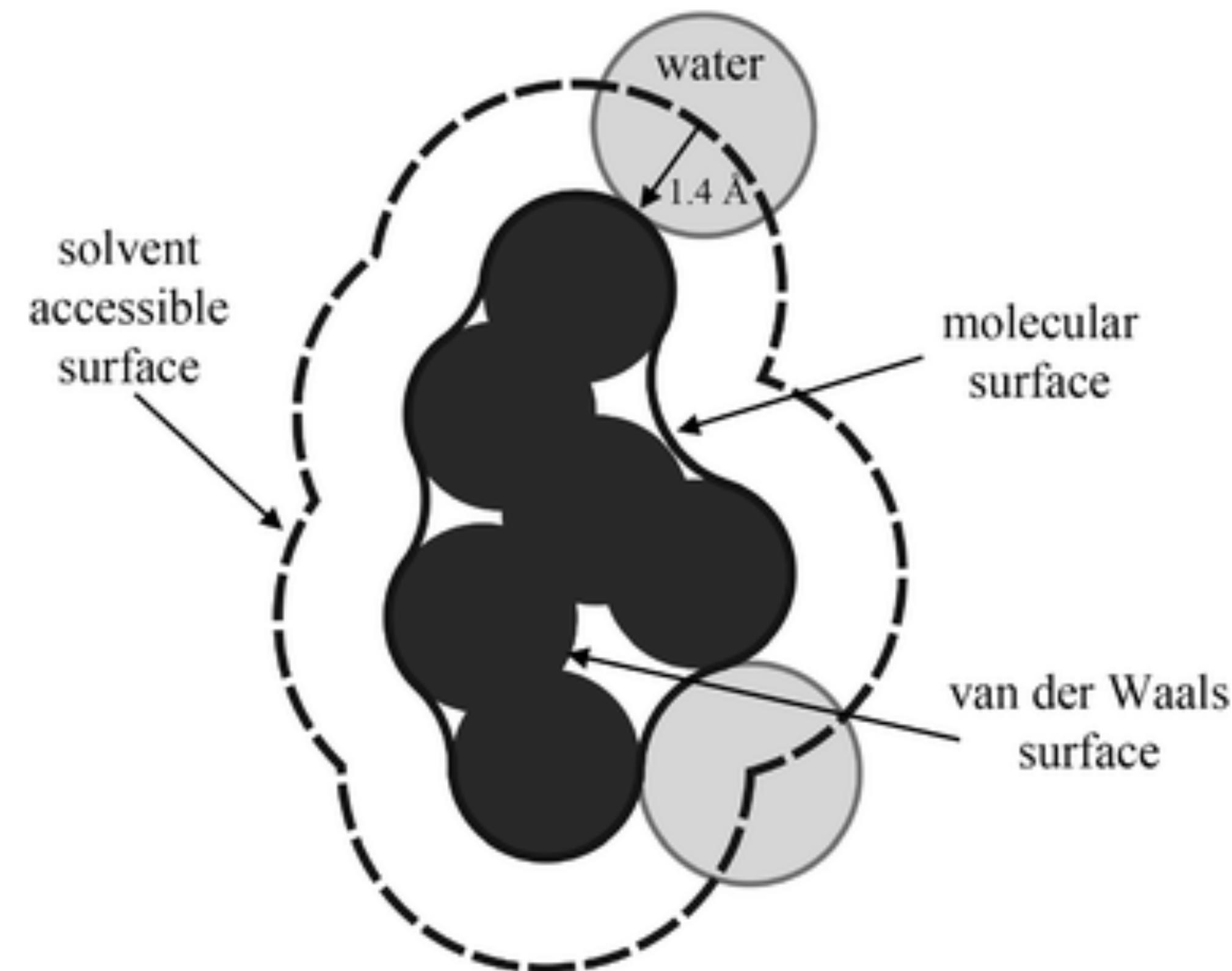
- **Surface (potential binding sites)**

The water-accessible surface is shown, illustrating the indentations and crevices, which may function as binding sites.



Surface can be calculated in different ways. Regions accessible to water may not represent those accessible to larger or smaller binders

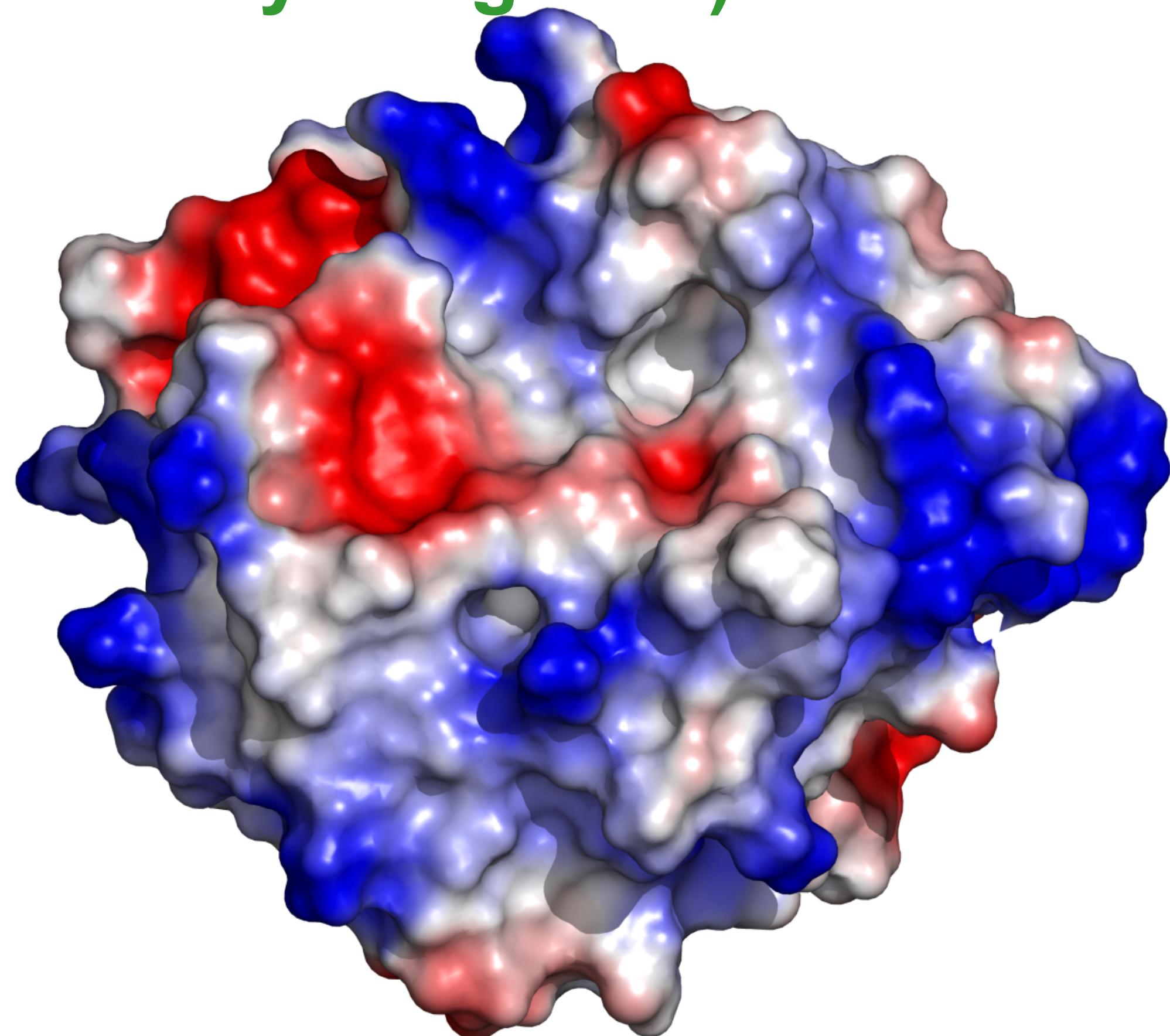
Types of surfaces



Different graphical representations reveal different properties

The figure was produced using PyMol. Negative potentials ($0k_B T/e > F > -60k_B T/e$) are red, positive potentials ($0k_B T/e < F < 60k_B T/e$) are blue, and neutral potentials are white. The electrostatic potential was calculated using APBS.

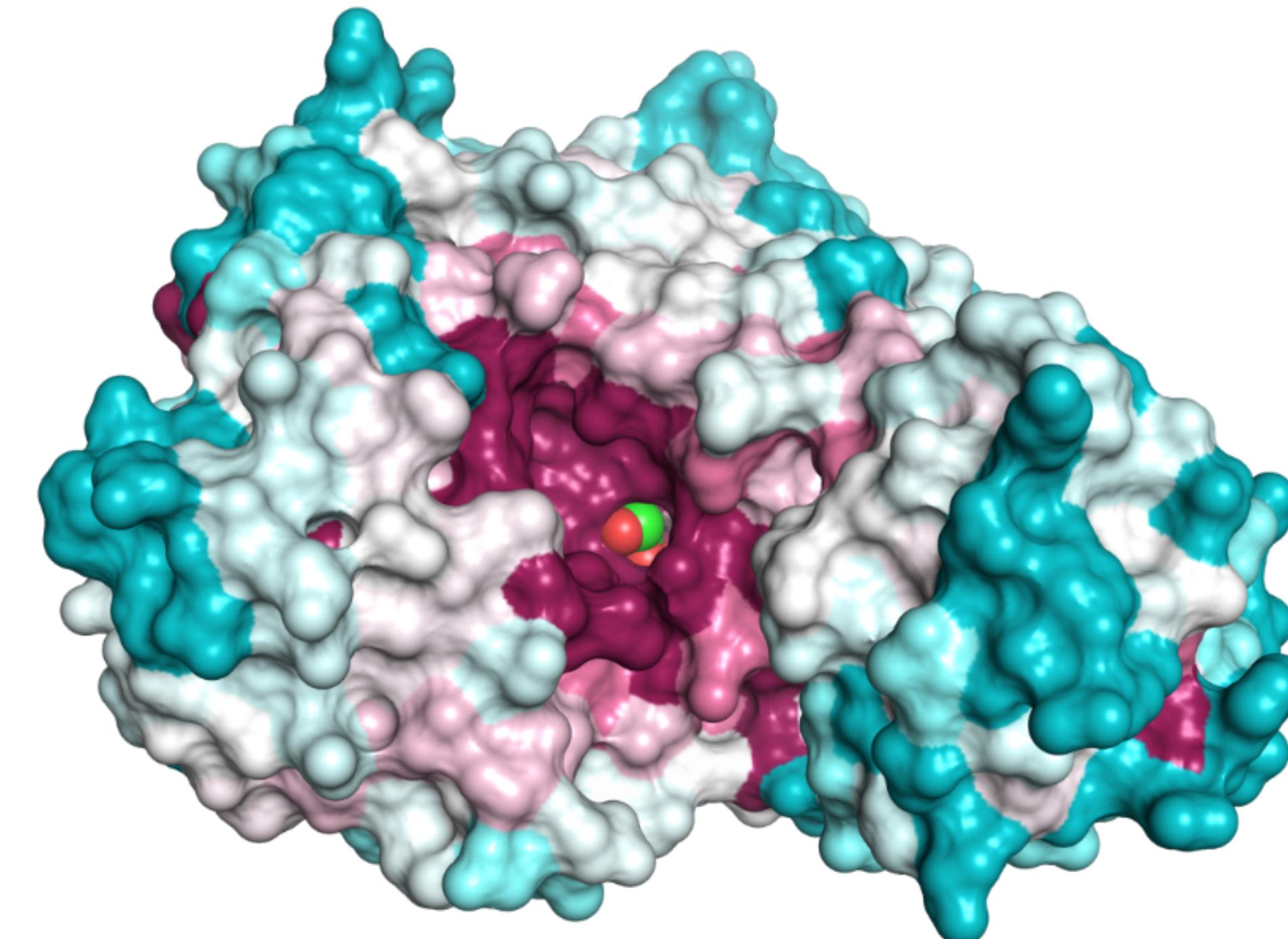
- Surface, colored by **electrostatic potential (Φ) (complementarity to ligands)**



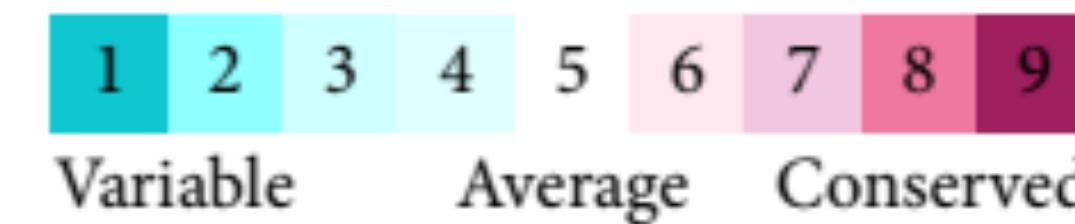
Different graphical representations reveal different properties

- **Space-fill, colored by evolutionary conservation (biologically important regions)**

The protein in the figure (triose phosphate isomerase, PDB entry 1amk) is shown in spacefill representation, with each residue colored by evolutionary conservation (turquoise - lowest, maroon - highest. See color-code in figure). The most conserved region is in the middle, where the natural substrate of the enzyme is bound.



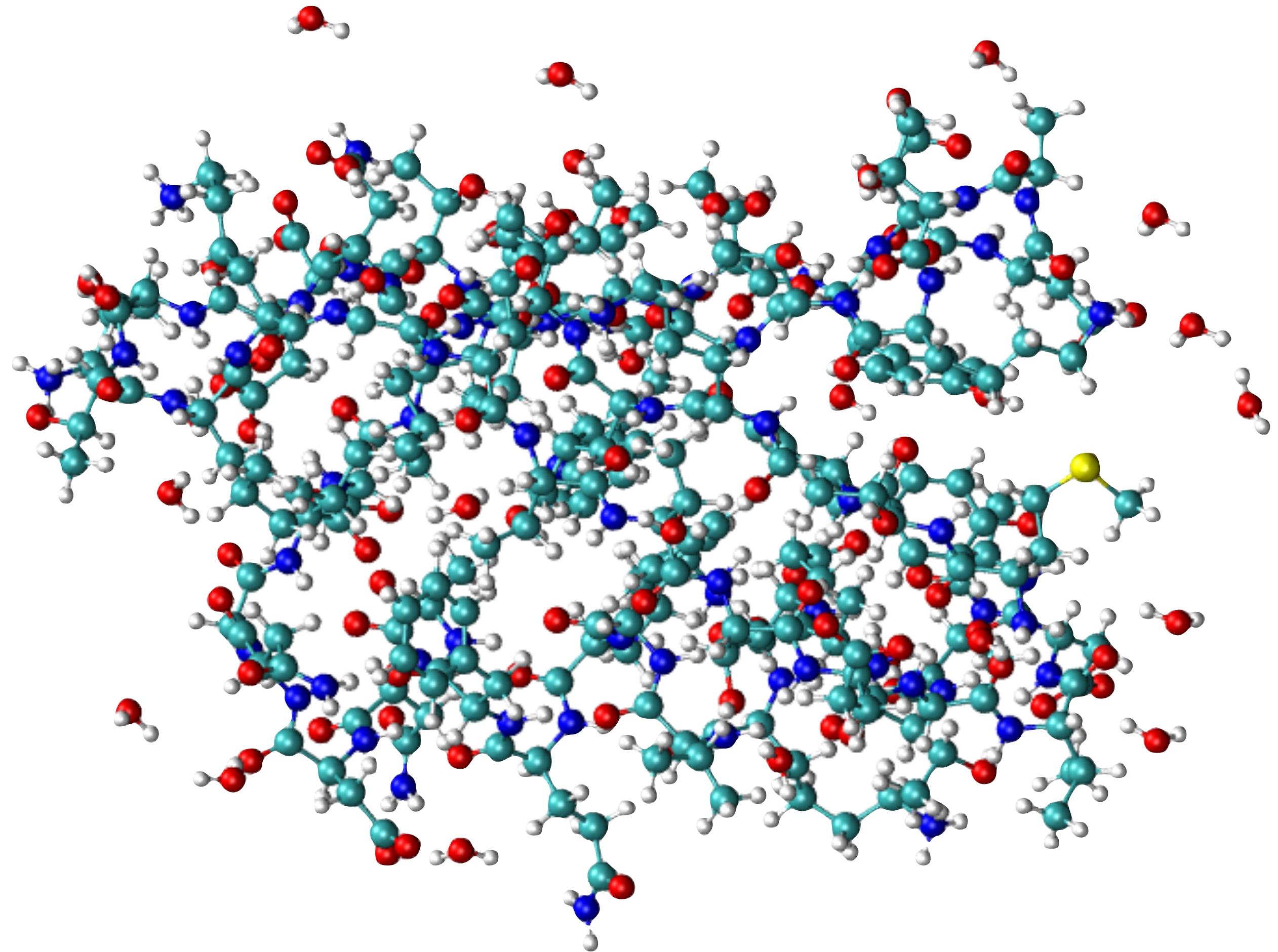
(<http://consurf.tau.ac.il>)



UNIVERSITÀ
DEGLI STUDI
DI MILANO



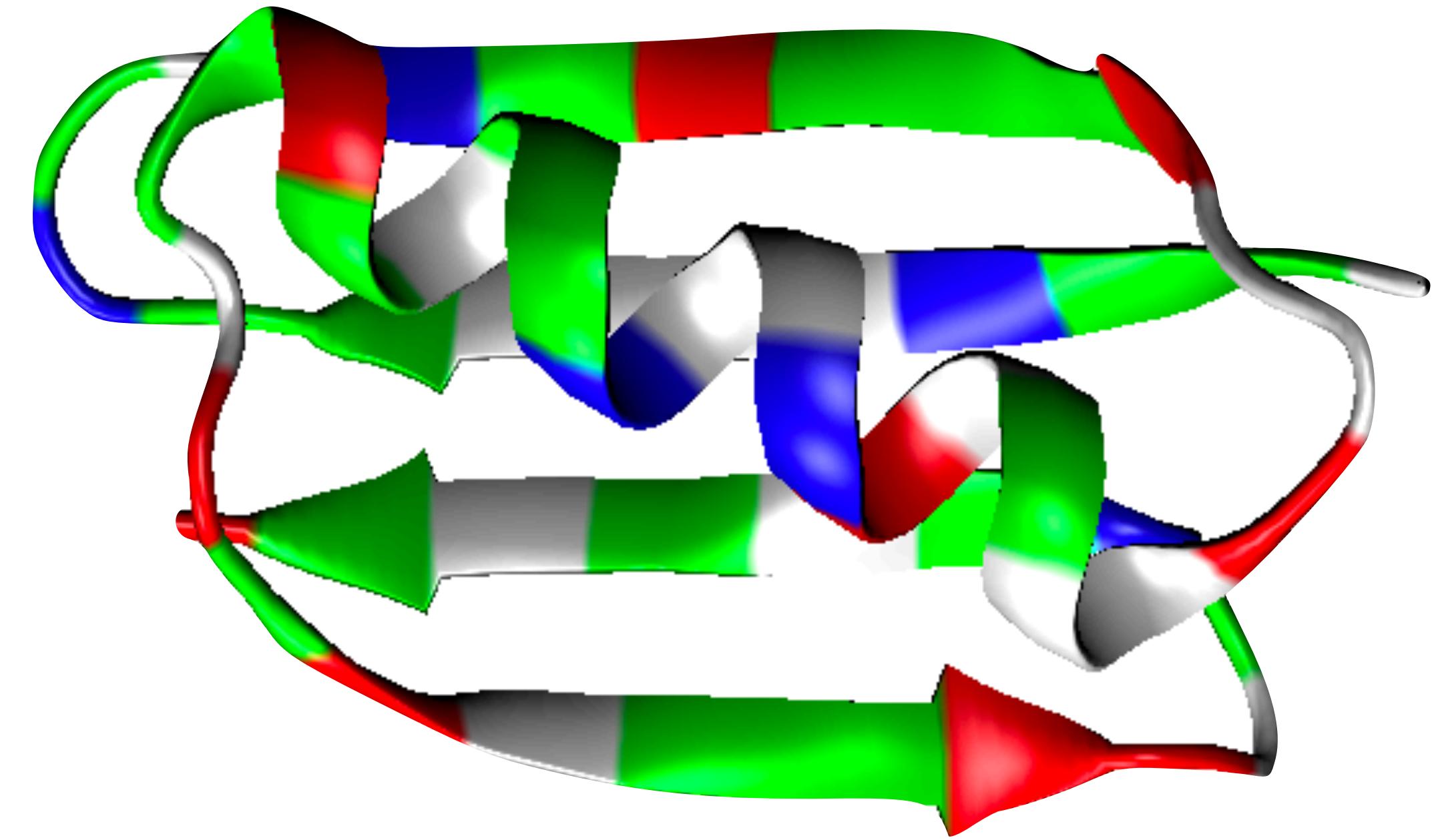
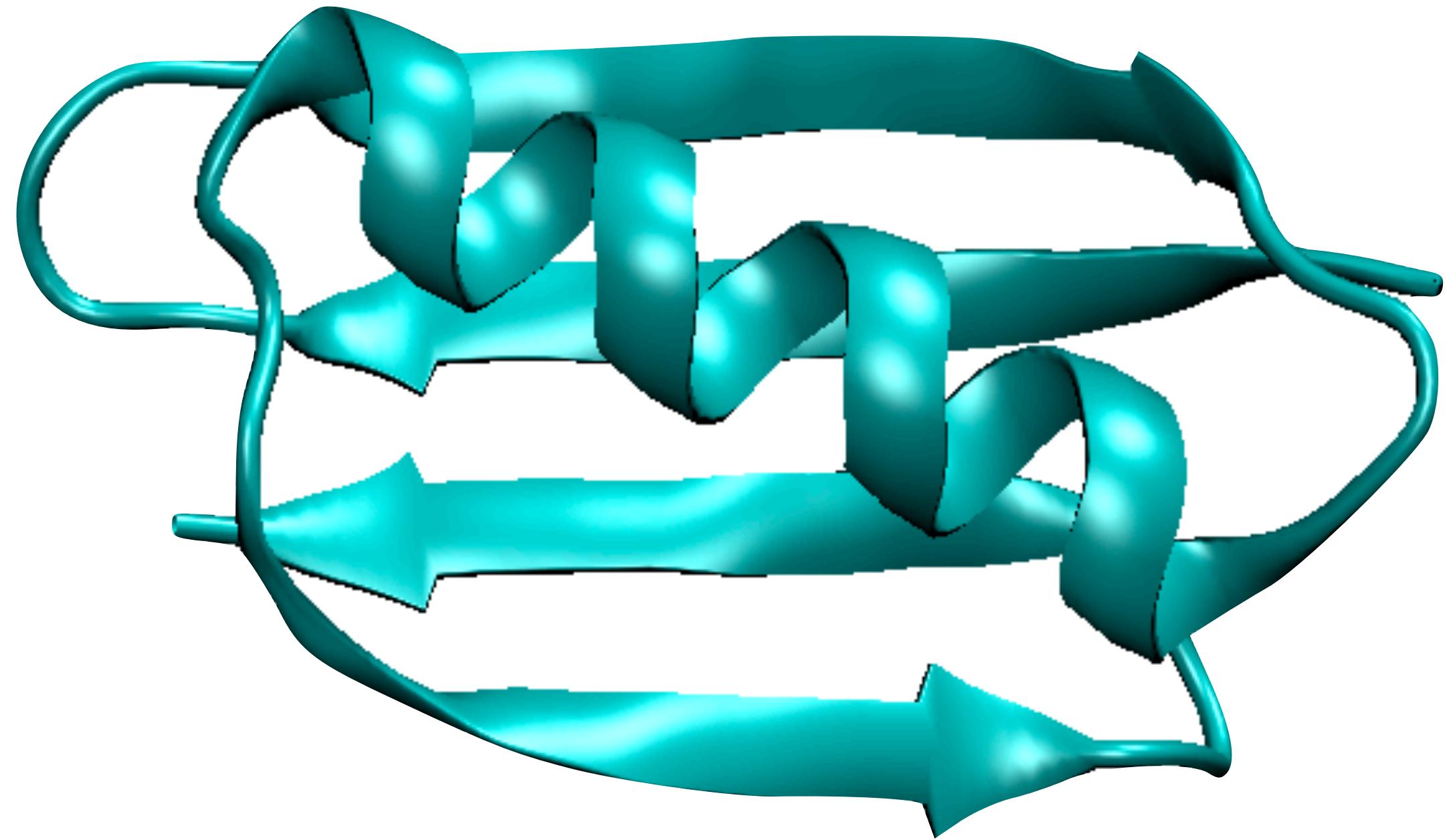
1PGB in Ball-and-Stick: showing all atoms determined



This is a 56 residues protein, determined by X-ray crystallography. In addition to all the protein atoms, also some trapped water are present.

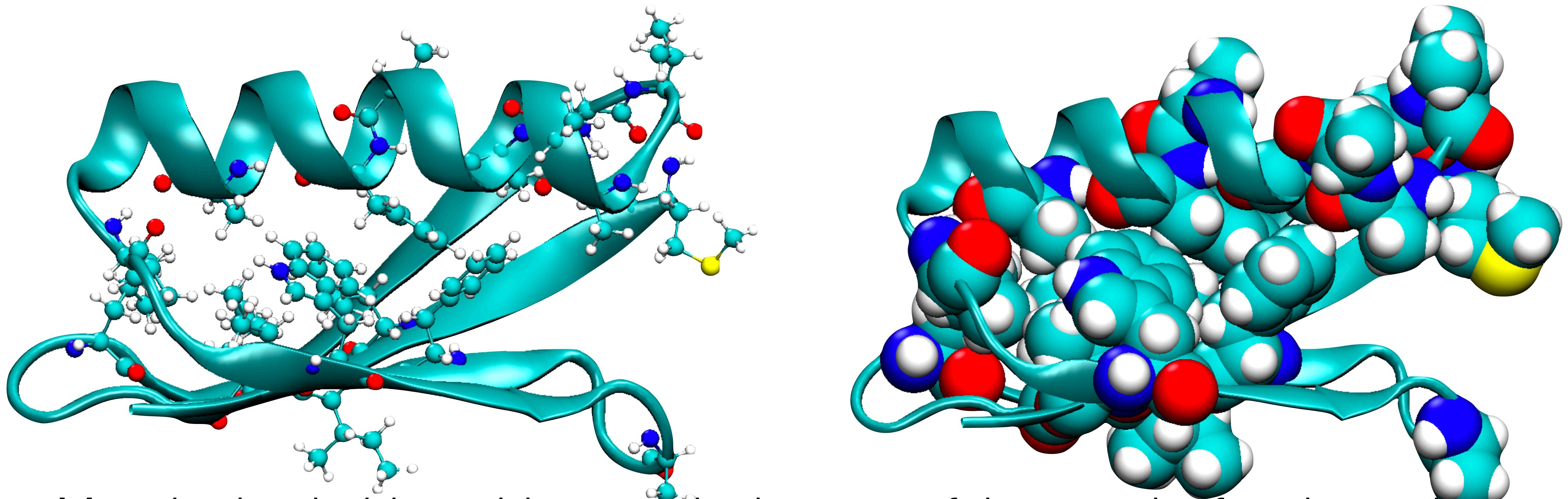


1PGB in cartoon: focusing on secondary structures



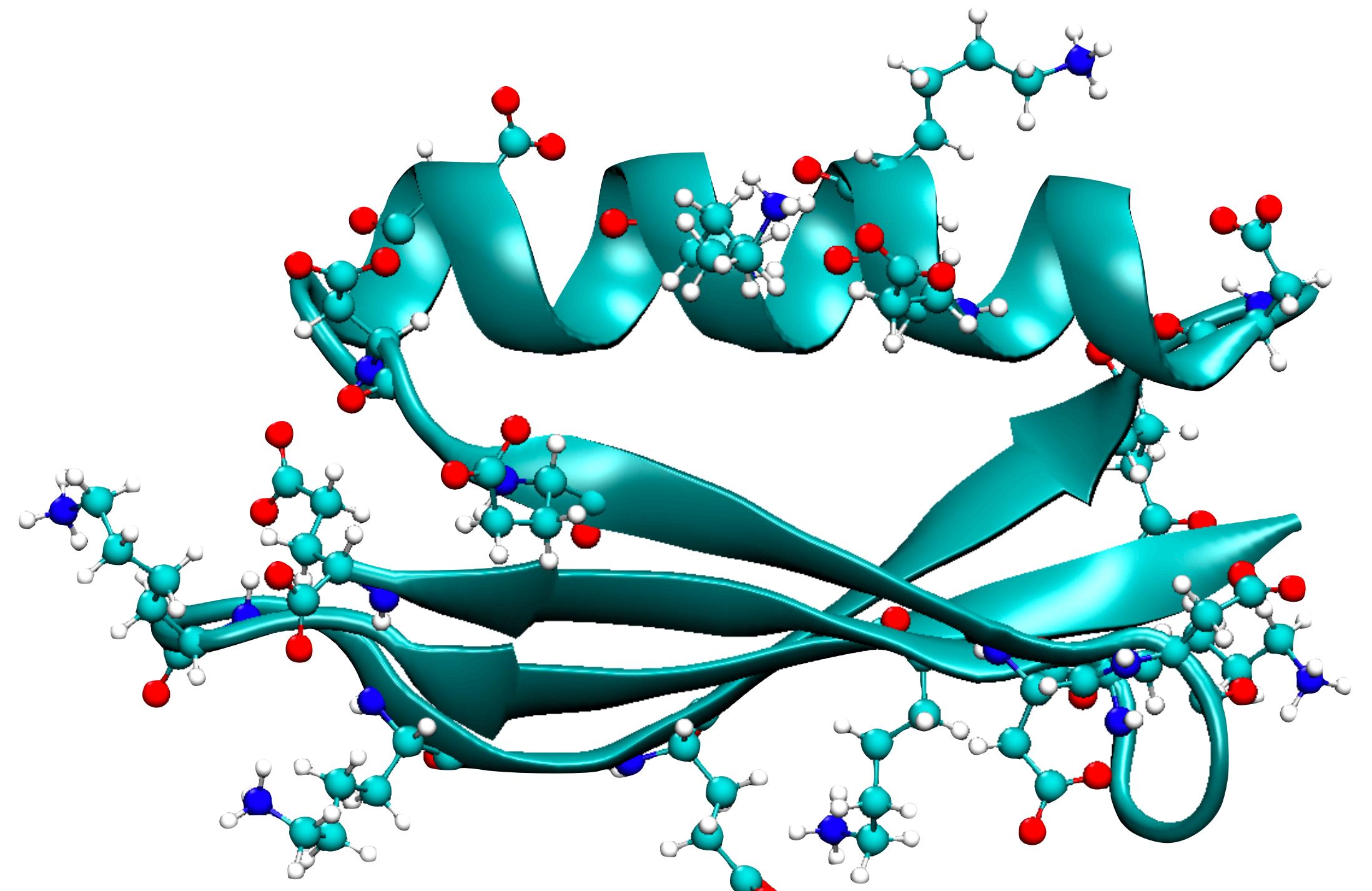
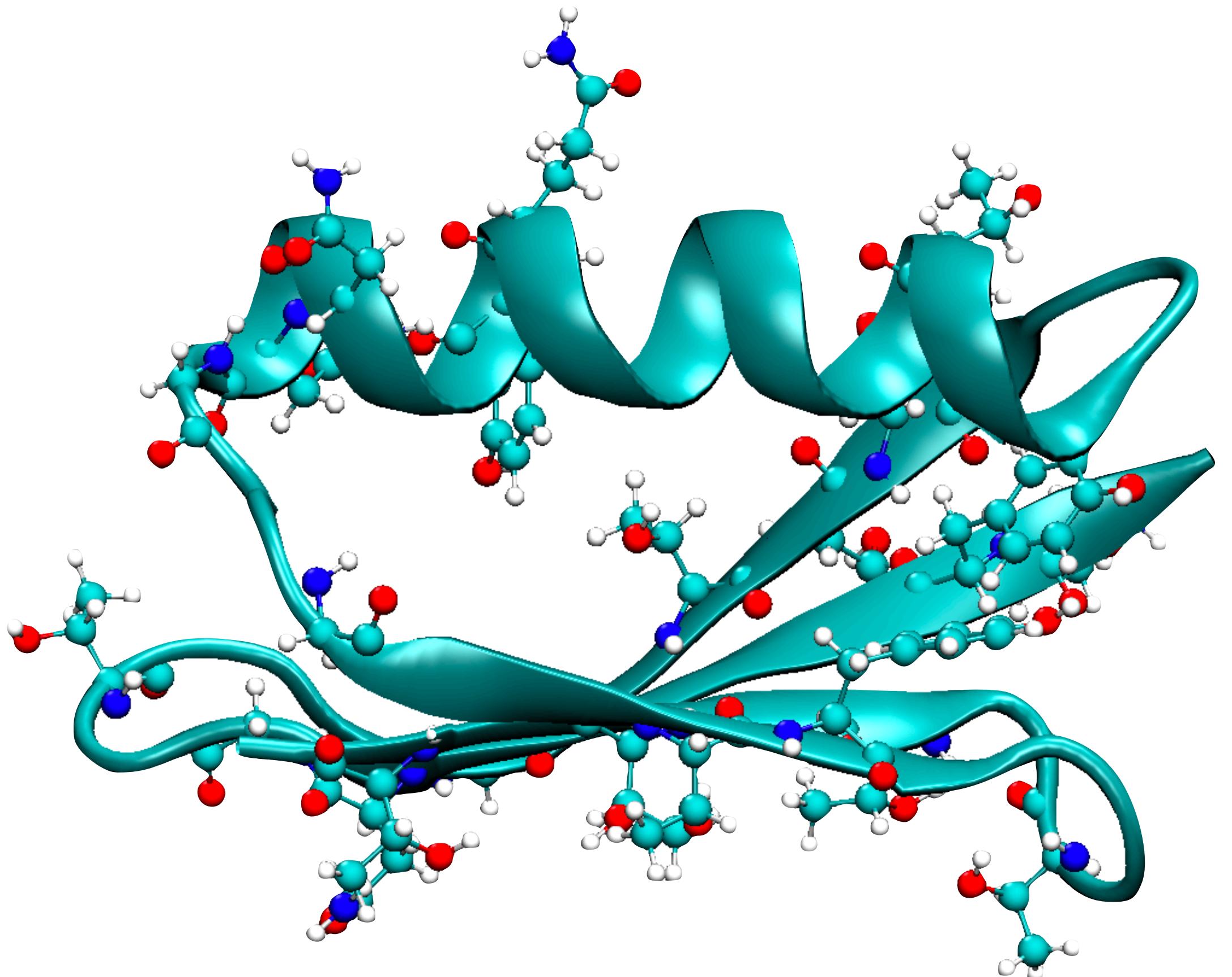
Coloring by residue type (polar: green/non-polar: white/
positive: blue/negative: red) highlights typical pattern of
secondary structure amino acidic composition

1PGB adding hydrophobic residues



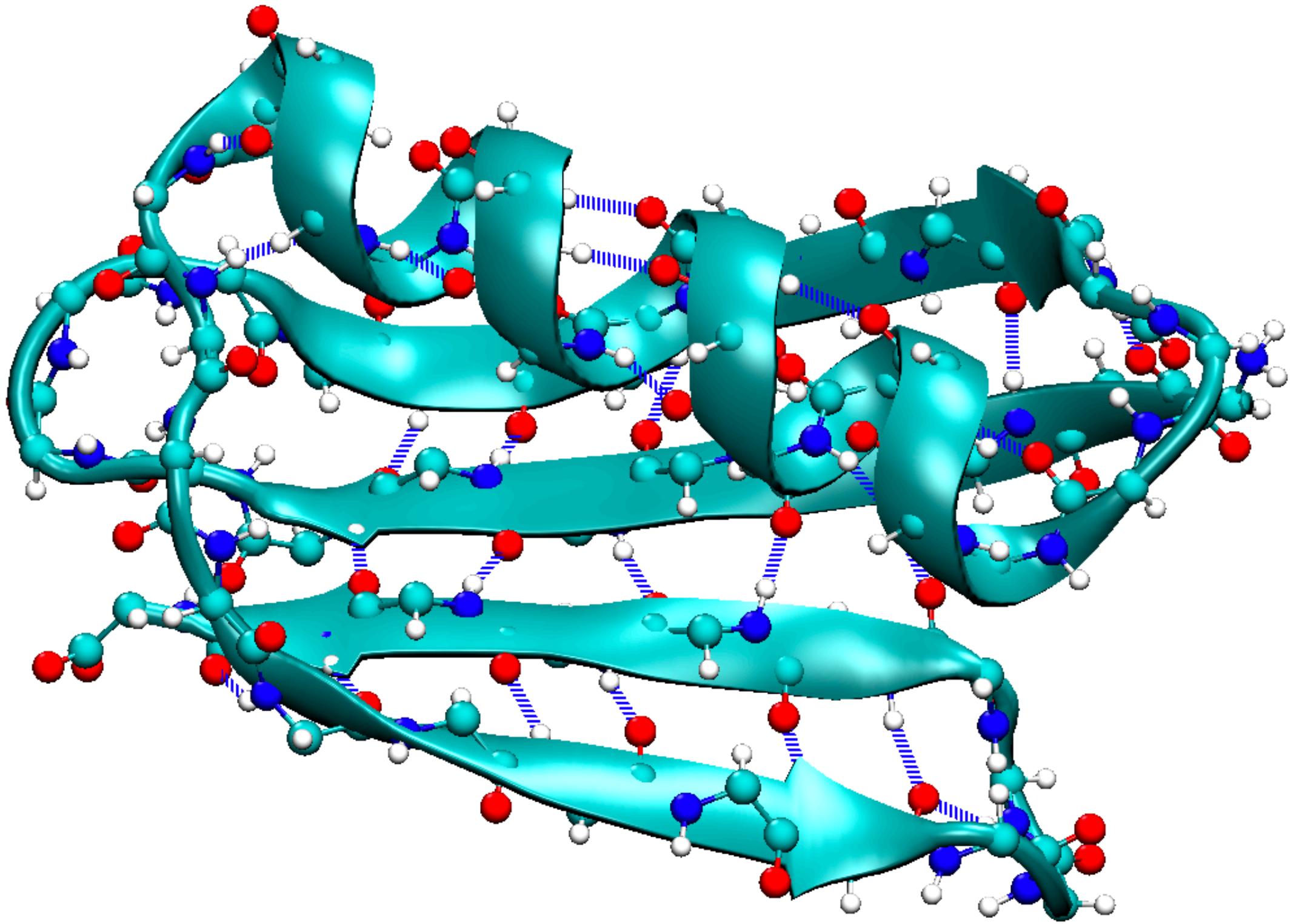
Most hydrophobic residues are in the core of the protein, forming a dense packing (looking using correct atomic radii), but some are also on the surface, possible hint for protein-protein interaction sites?

1PGB showing only Polar and Charged residues

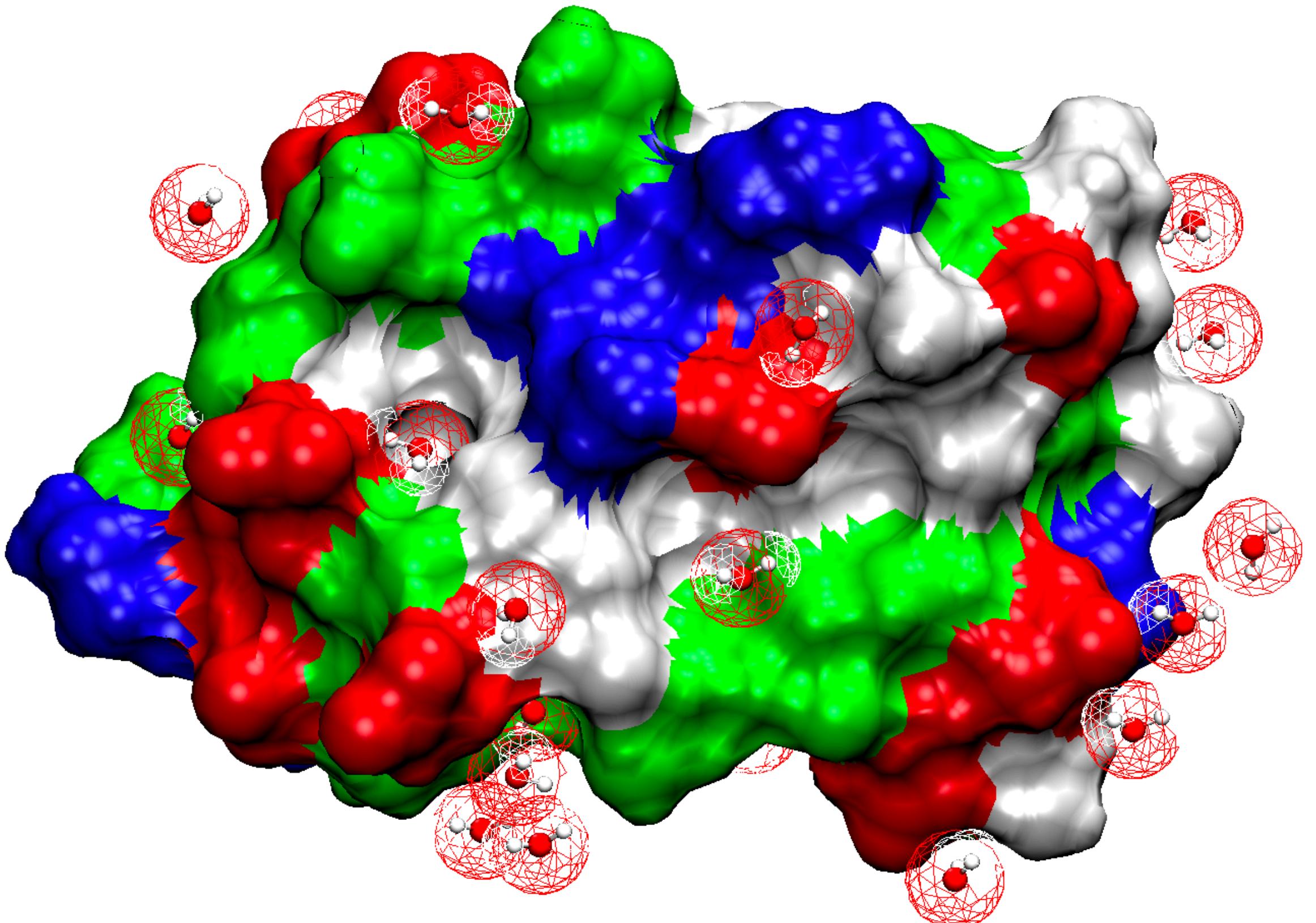


Polar and charged aminoacid are mostly on the surface and in the loops

1PGB: looking at hydrogen bonds and solvent accessible surface area



Hydrogen bonds are usually defined geometrically
distance between donor and acceptor < 3.5
Angle between donor, acceptor and hydrogen < 30 deg



SASA is coloured by residue type
highlighting polar and hydrophobic regions

Visualisation: Software

VMD: Visual Molecular Dynamics, most used for molecular dynamics simulations

PyMol: high-quality graphics, scriptable in python, very good for protein superposition

ChimeraX: high-quality graphics, very good for cryo-EM

Coot: focused on structure determination for X-Ray and Cryo-EM

UnityMol: based on gaming Unity Engine

NGLView: to be programmed in python



Practicals

Install VMD if you will
use your own laptop

We are going to use
COLAB:

colab.research.google.com

<https://www.ks.uiuc.edu/Research/vmd/>

To use them (essentially to
save the results of your work)

You need a google Drive
account, please do one

