Proteins: theory and preparation for simulation

PHRMSCI 275. Drug Discovery Computing Techniques

The University of California, Irvine

03/01/2022

Mary Pitman, PhD, contact: mpitman@uci.edu

Outline

1. Background: why the starting configuration can make or break your simulation
   - Theoretical considerations: statistical mechanics, protein physics
   - Limitations of atomic resolution simulations
2. Structure files
   - Protein Data Bank: crystallography, NMR, cryo-EM, AlphaFold
3. Tools and considerations.

Scan to access pptx and follow along, get links

Statistical Mechanics Postulate

Ergodic Hypothesis: The (long) time average of any mechanical property in a real macroscopic system is equal to the average value of that property over all microscopic states of the system. Each state is weighted by the probability of occurrence, provided that the microscopic sates replicate the thermodynamic state and environment of the actual system.

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t \mathrm{d}t' \left\langle A(q(\boldsymbol{r}^N); \boldsymbol{r}^N(0), \boldsymbol{p}^N(0), t) \right\rangle_{NVE} = \langle A(q) \rangle_{NVE}$$

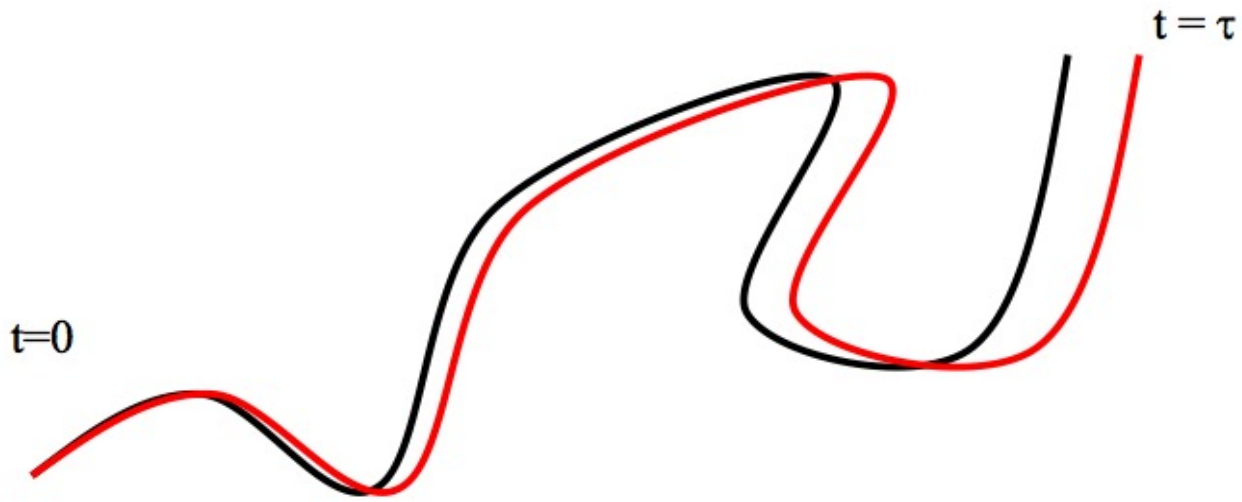
If the ergodic hypothesis is true, then time averages equal ensemble averages (Ehrenfests 1912).

Further reading: (1) *Compendium of the Foundations of Classical Statistical Physics*, Jos Uffink
(2) *An Introduction to Applied Statistical Thermodynamics*, Sandler, Section 1.4

Starting conditions can have unpredictable trajectory effects

The dynamics of a well behaved, classical, many body system is chaotic.

∴ Trajectories that differ in their initial conditions diverge exponentially ("Lyapunov stability")

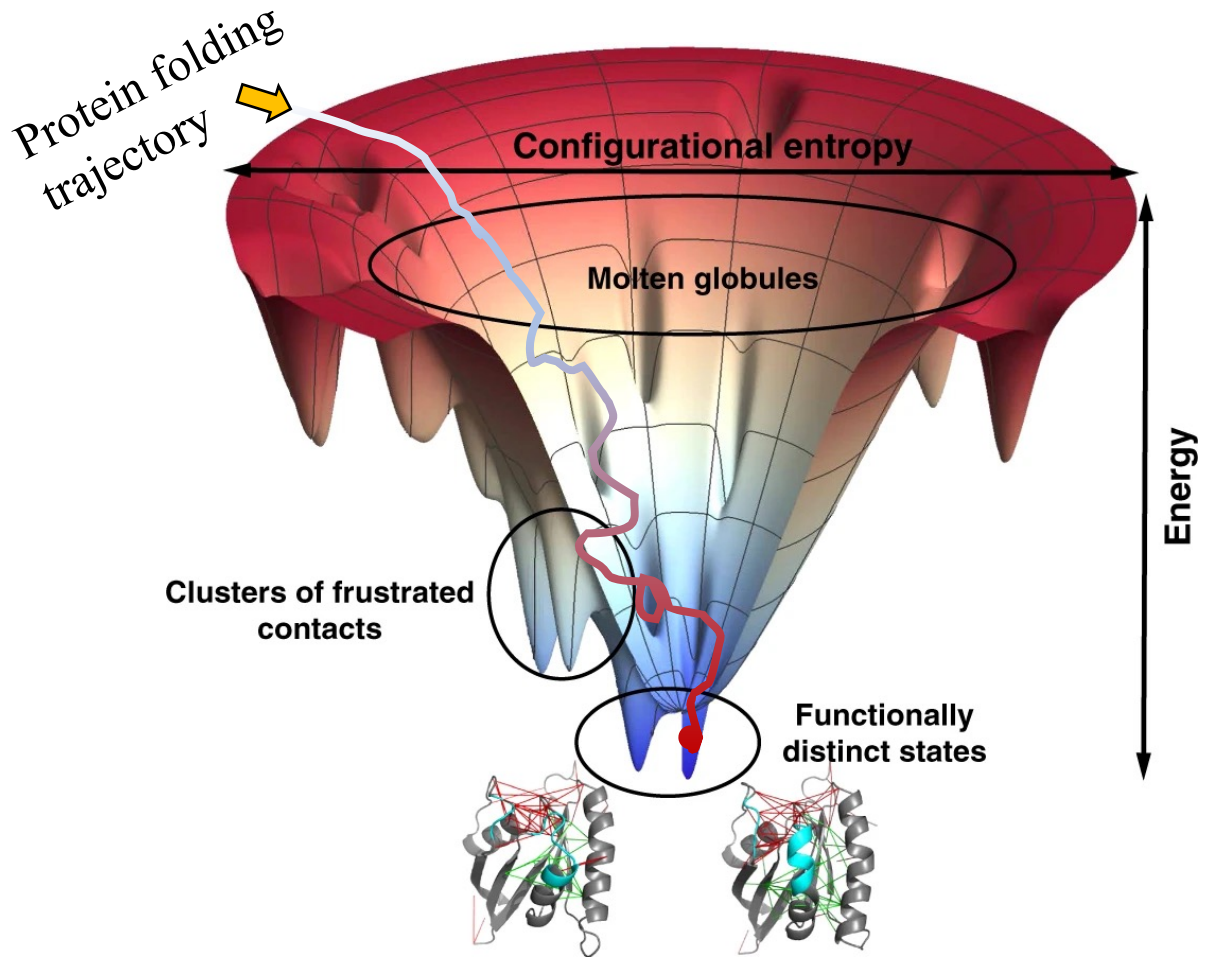


… but we can obtain meaningful thermodynamic quantities from MD due to the "shadow therom"

Further reading: *Statistical Mechanics and Molecular Dynamics*, Luca M. Ghiringhelli, link

Starting conditions can have unpredictable trajectory effects

The physics of proteins is ‘glassy’ and so overall the assumption of ergodicity does not hold.

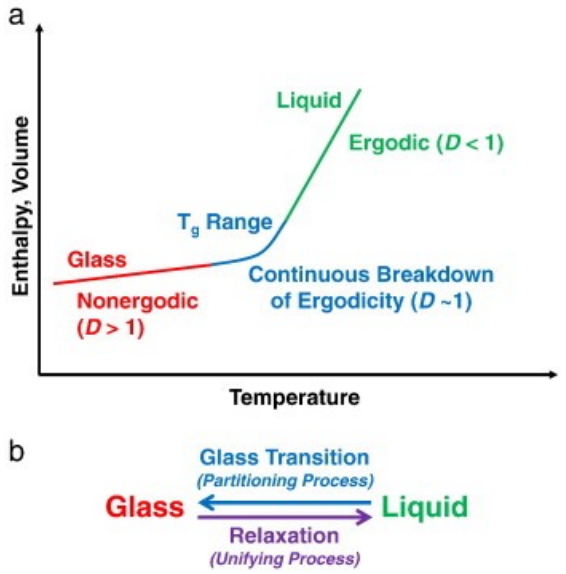


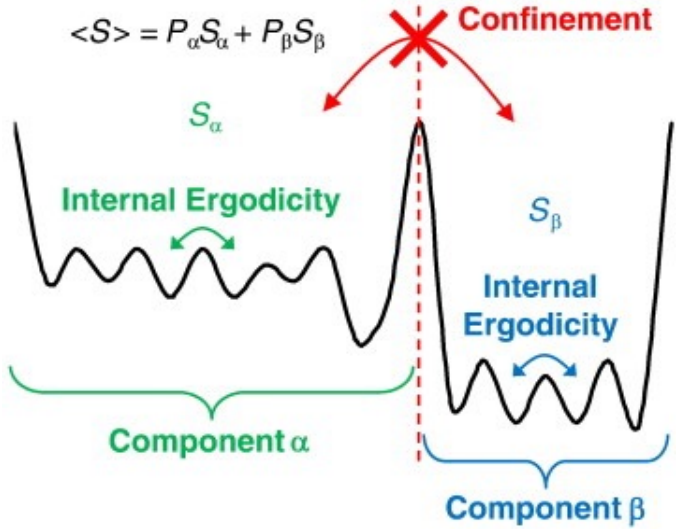
(1) *Coevolutionary information…and the thermodynamics of natural selection,* Morcos et al.

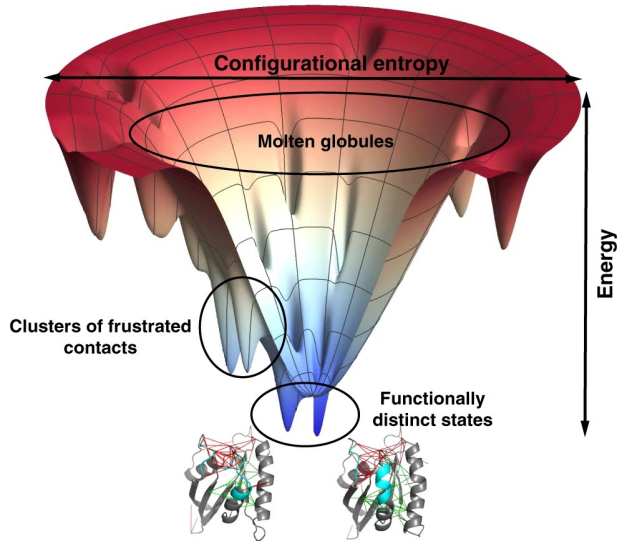
(2) *Funnels, Pathways, and the Energy Landscape of Protein Folding: A Synthesis,* Bryngelson et al.

Starting conditions can have unpredictable trajectory effects

- Proteins undergo a glass transition as they fold
- The free energy landscape is rugged, or frustrated. As the protein folds it becomes trapped in local minima.
- Within these minima, there is *pseudo-ergodicity (*Ehrenfests)



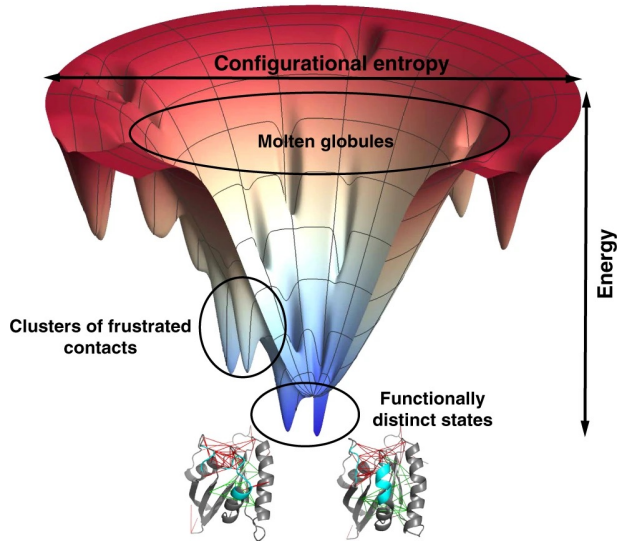




(1) *Statistical mechanics of glass,* Mauroa & Smedskjaer [paper link]

A minima does not mean “native”. The crystal structure may not be ”native”.

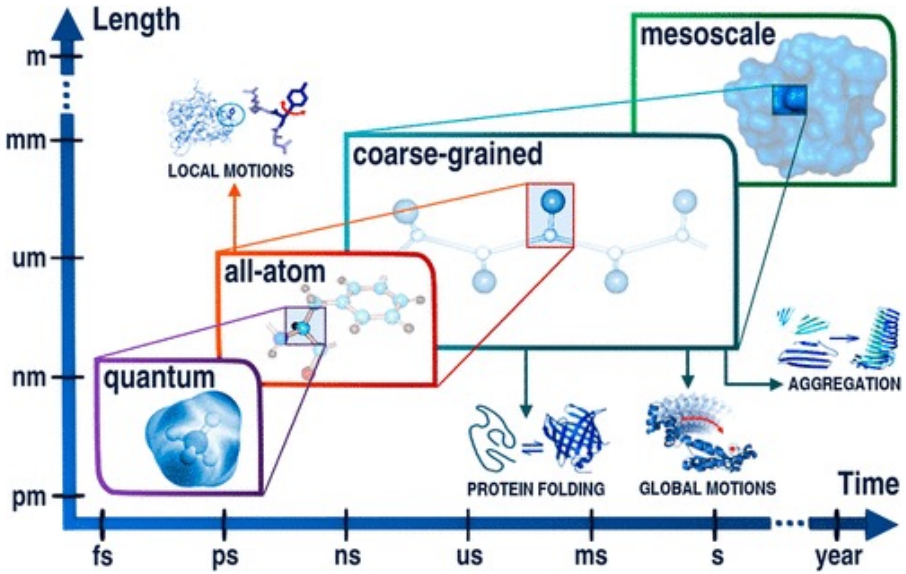
- Proteins undergo a glass transition as they fold
- The free energy landscape is rugged, or frustrated. As the protein folds it becomes trapped in local minima.
- Within these minima, there is *pseudo-ergodicity (*Ehrenfests)

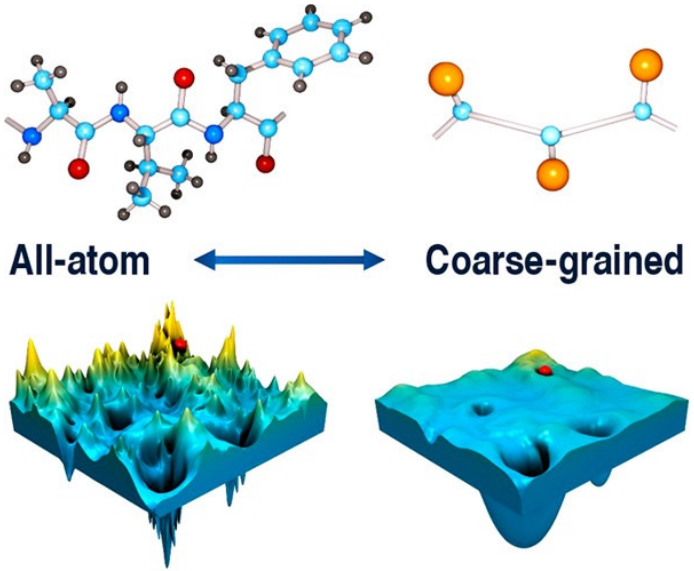


(1) *Statistical mechanics of glass,* Mauroa, Smedskjaer paper link

In all-atom MD, the phase space sampled will be limited.

- Mistakes in the starting configuration are unlikely to be resolved.



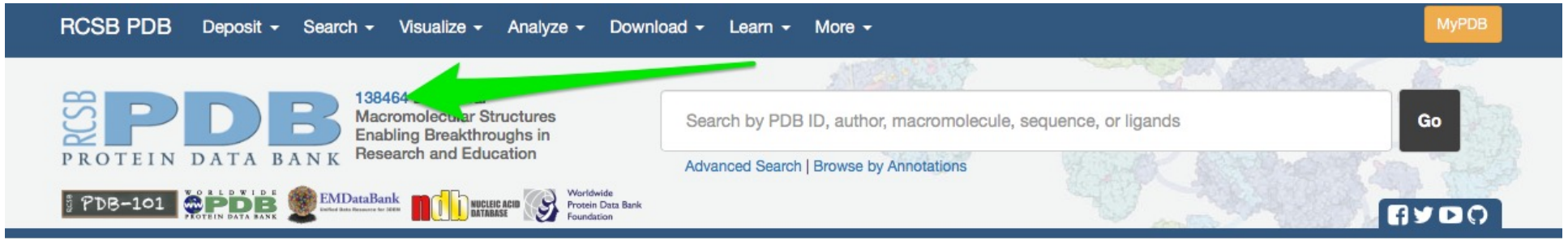


(1) *Coarse-Grained Protein Models and Their Applications,* Kmiecik *et al.*

Outline

1. Background: why the starting configuration can make or break your simulation
   - Theoretical considerations: statistical mechanics, protein physics
   - Limitations of atomic resolution simulations
2. Structure files
   - Protein Data Bank: crystallography, NMR, cryo-EM, , AlphaFold
3. Tools and considerations

The Protein Data Bank (PDB) is a repository which stores structures of macromolecules



www.rcsb.org

Most structures are obtained via crystallography

Other Statistics

PDB Data Distribution by Experimental Method and Molecular Type

Copy | CSV

Experimental Method	Proteins	Nucleic Acids	Protein/NA Complex	Other	Total
X-Ray	116115	1916	5922	10	123963
NMR	10660	1236	249	8	12153
Electron Microscopy	1459	31	506	0	1996
Other	210	4	6	13	233
Multi Method	112	4	2	1	119
Total	128556	3191	6685	32	138464

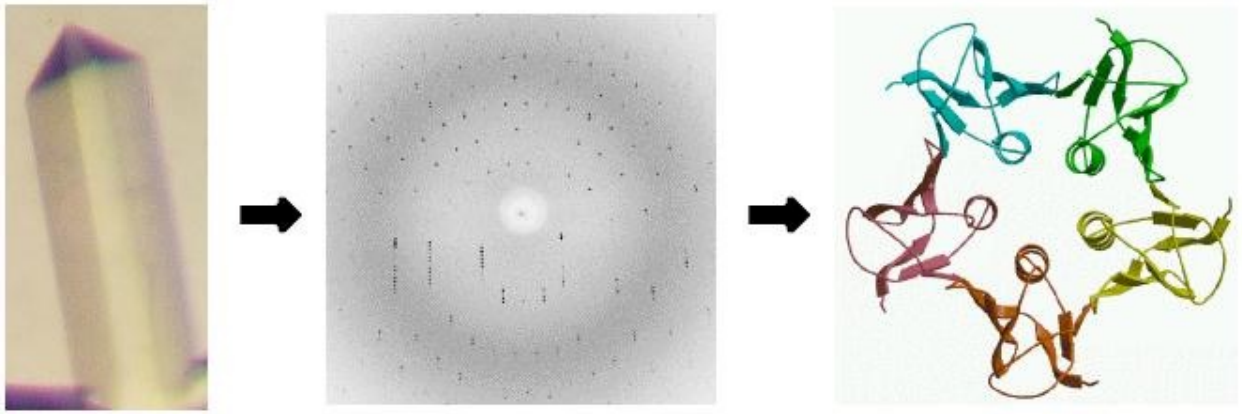
113777 structures in the PDB have a structure factor file.
9490 structures in the PDB have an NMR restraint file.
3242 structures in the PDB have a chemical shifts file.
2022 structures in the PDB have a 3DEM map file.

(RCSB PDB)

What's in a PDB file?

- Coordinates:
  - Atom list
  - 3D position in space
  - "occupancy"
  - Missing information!
- Experimental details
  - What the system is, how the structure was solved, secondary structure information, protein sequence
- Other info:
  - i.e. citations

Crystallography tries to turn diffraction data into structures

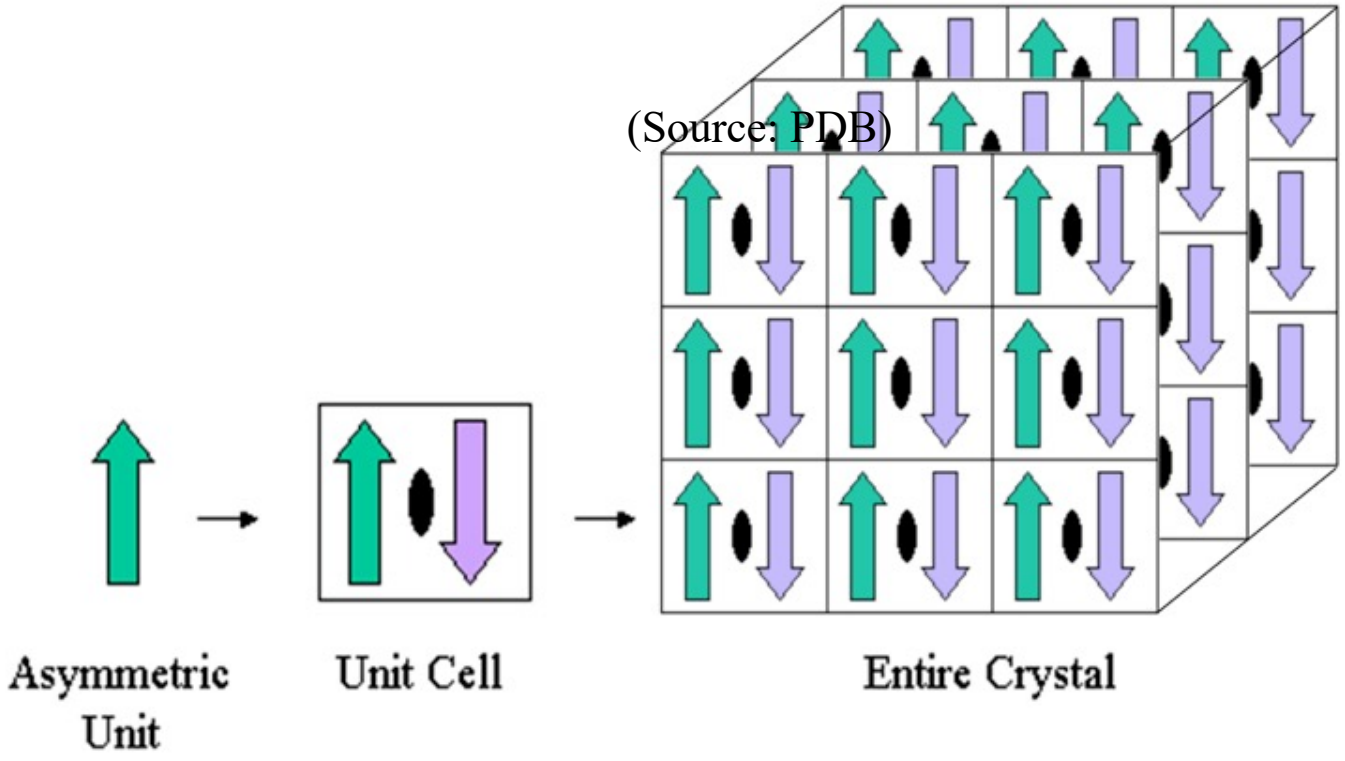


Really, there is not quite enough data -- each atom has about 3-4 parameters, and we have about one observation per parameter

This means we fit a model to the data

http://www-structmed.cimr.cam.ac.uk/Course/Overview/Overview.html

PDB files for crystals contain just a portion of the whole crystal structure -- the “asymmetric unit”



(Source: PDB)

(Source: PDB)

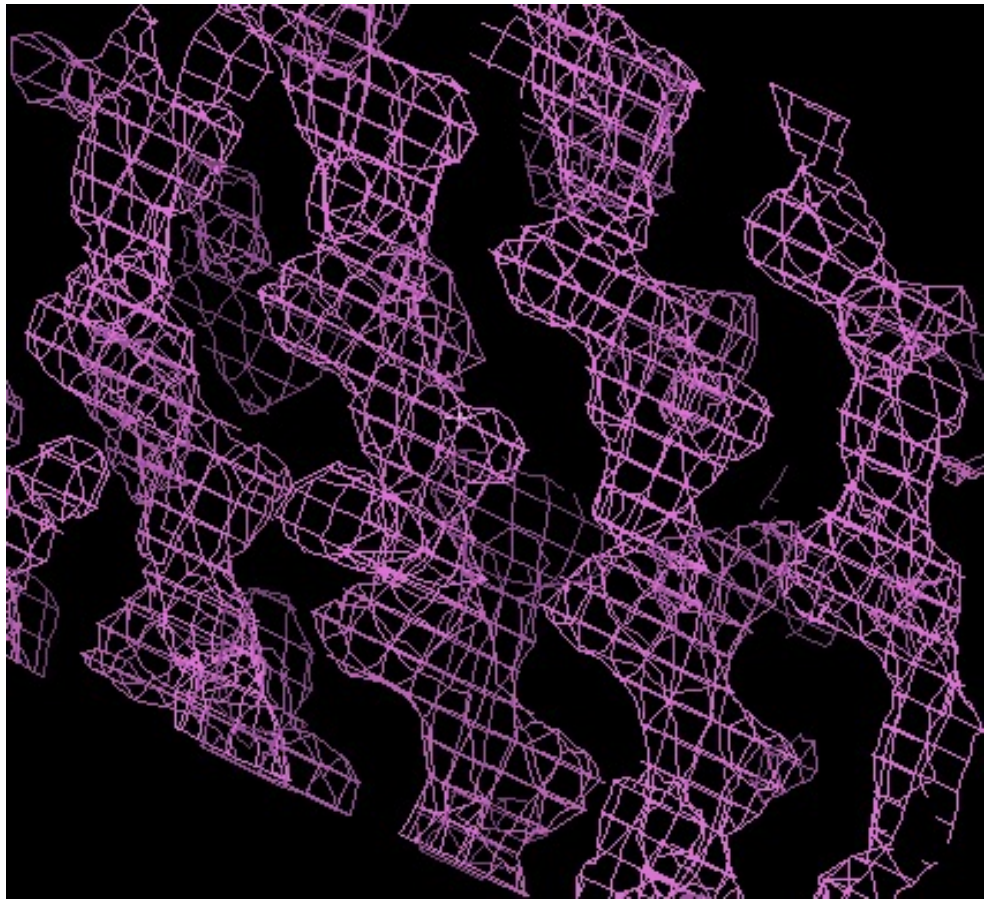
The biological assembly may or may not be the same as the asymmetric unit

- Asymmetric unit can be:
  - A single biological assembly
  - Multiple biological assemblies
  - Part of a biological assembly

(Source: PDB)

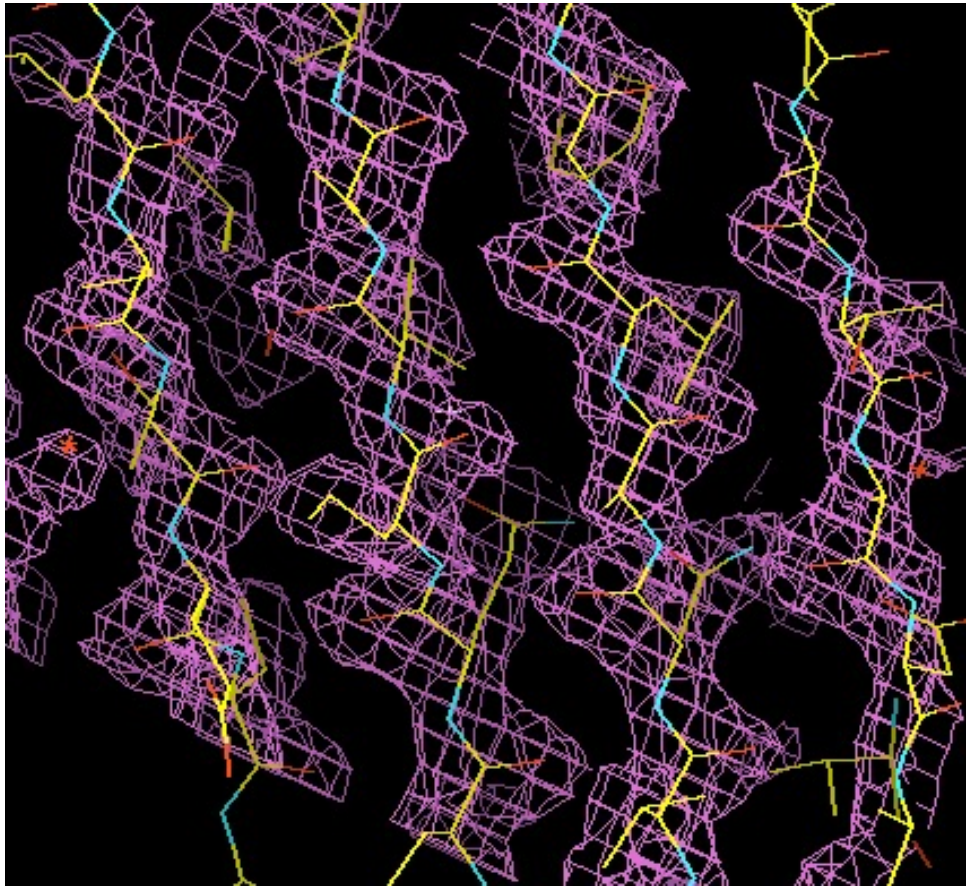
Asymmetric unit with one biological assembly	Asymmetric unit with a portion of a biological assembly	Asymmetric unit with multiple biological assemblies
		
Entry **2hhb** contains **one** hemoglobin molecule (**4 chains**) in the asymmetric unit.	Entry **1hho** contains **half** a hemoglobin molecule (**2 chains**) in the asymmetric unit. A crystallographic two-fold axis generates the other 2 chains of the hemoglobin molecule.	Entry **1hv4** contains **two** hemoglobin molecules (**8 chains**) in the asymmetric unit.

But at some point, we have to fit a model to the "data" -- the electron density

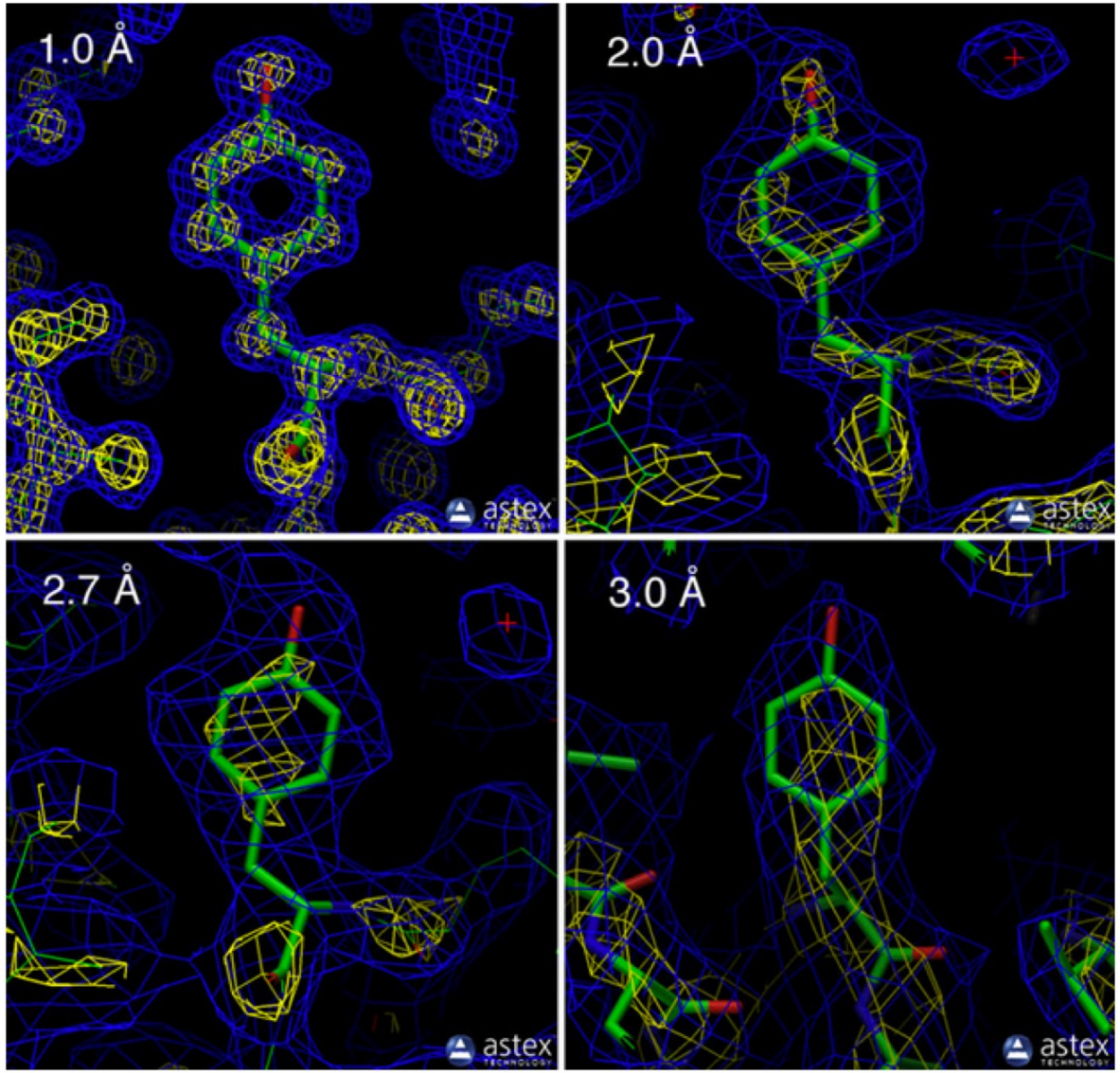


It is nonobvious where the atoms should be -- and this is good data

The actual model IS consistent with this data, but the model is not the data. So we should visualize the maps to verify consistency



Resolution is an important aspect when looking at structures

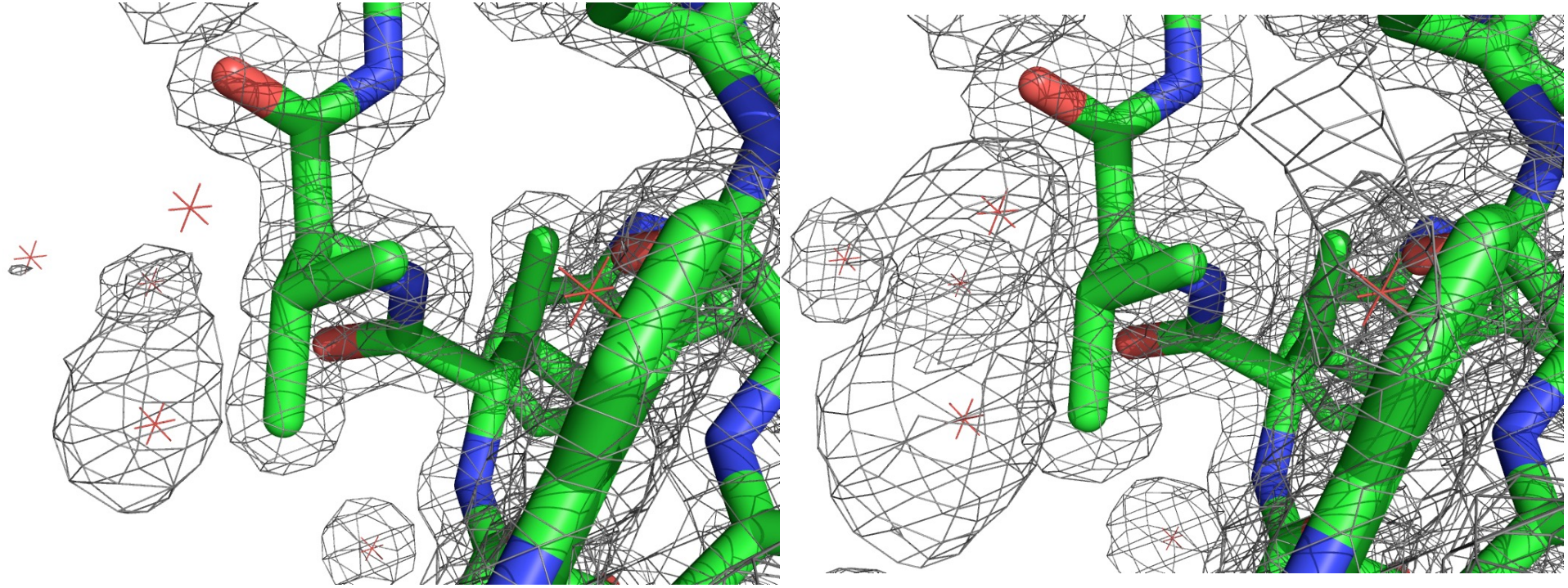


(Source: PDB)

There are two common ways of visualizing the electron density

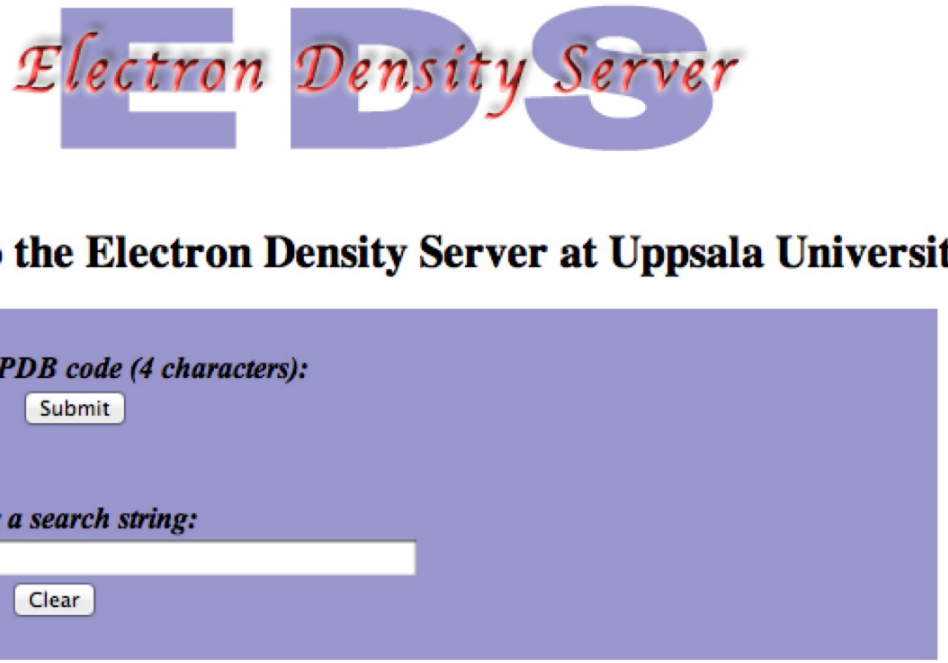
- "Difference map" $\mathbf{F}_o - \mathbf{F}_c$
  - Shows difference between calculated and observed structure factors
- The map itself, usually as $2\mathbf{F}_o - \mathbf{F}_c = 2\left|\mathbf{F}_o\right| - \left|\mathbf{F_c}\right|$
  - Shows the actual observations, with the difference from the calculated values also appearing to a limited extent

But, there is a value for the electron density everywhere, so we have to select contour levels



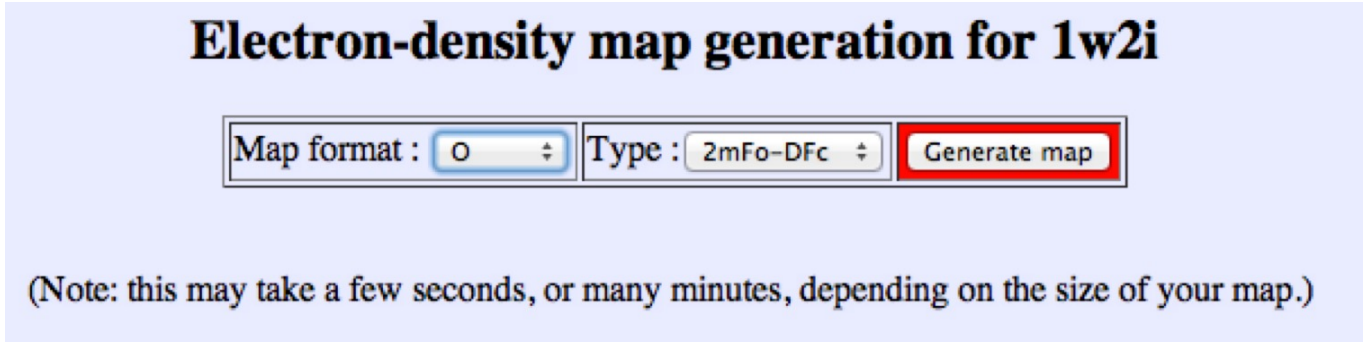
PDB files don't actually contain electron density, though modern ones contain calculated structure factors

- Calculated structure factors yield the electron density
- Usually, see the Electron Density Server to get electron densities

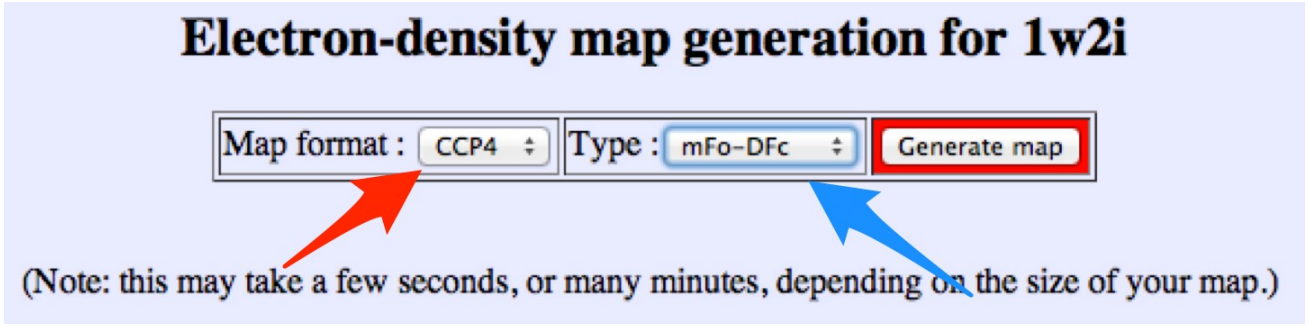


http://eds.bmc.uu.se/eds/

From EDS, you choose a type of map and a format



Let's do a difference map for visualization in PyMol



Next, I download the map to my computer, extract it, and get ready to load into PyMol

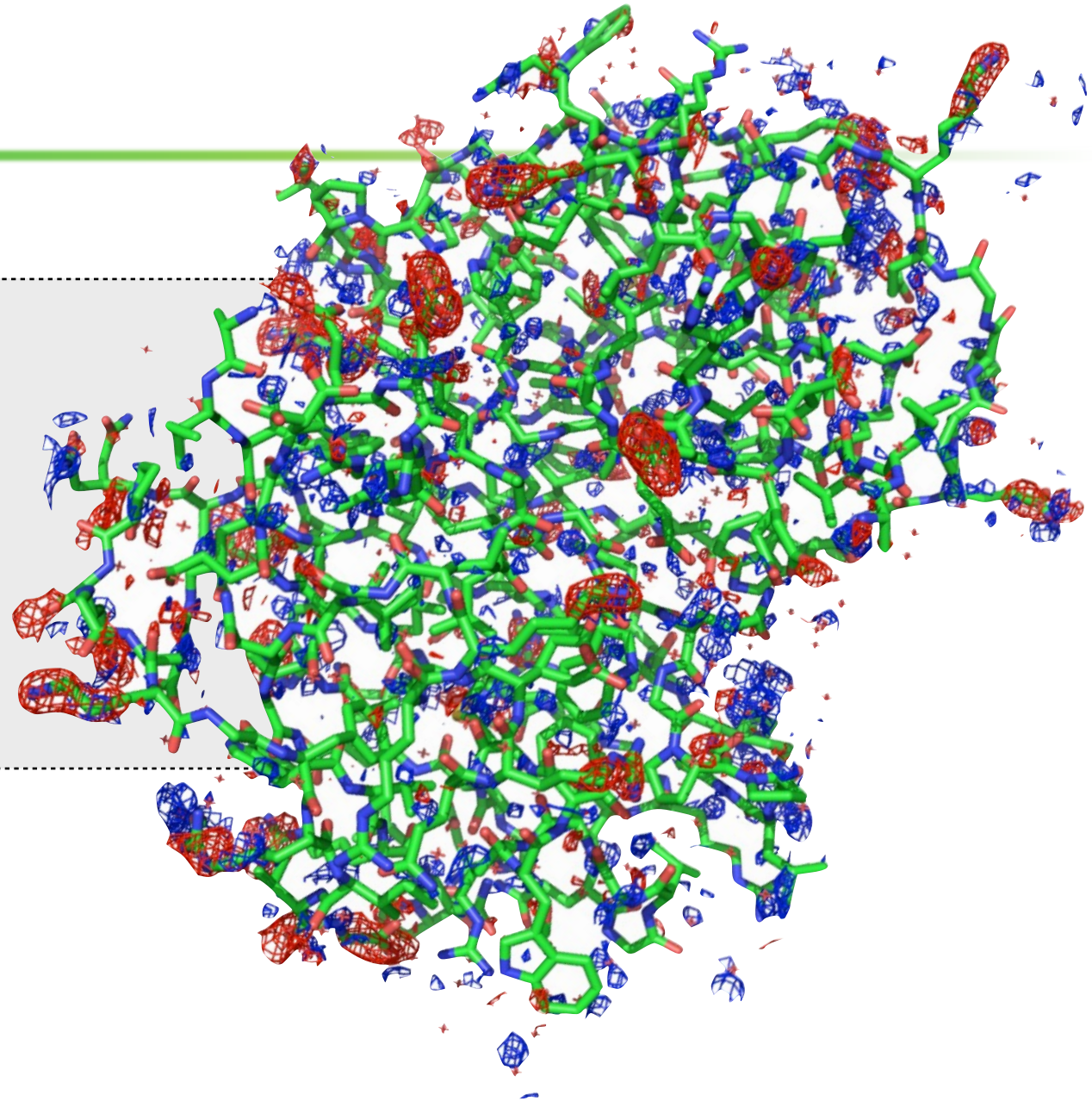
Here is your gzipped map : 1w2i_diff.ccp4.gz

- Check out this PyMol tutorial:
  - http://pymolwiki.org/index.php/Display_CCP4_Maps

- Note: Difference maps usually contoured at +/- 2.5 sigma

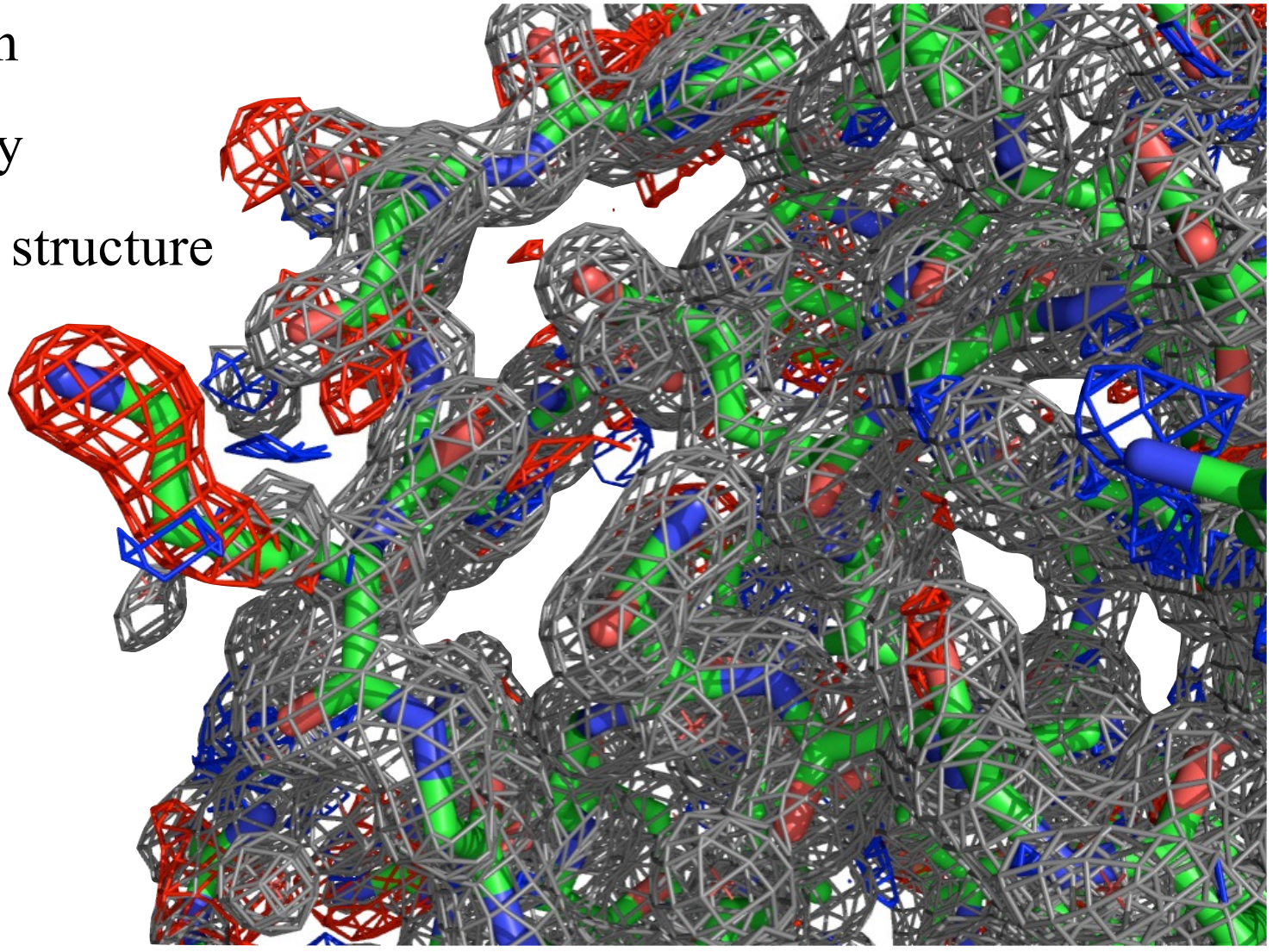
In PyMol:

```
load 1W2I.pdb
load 1w2i_diff.ccp4, 1W2I_map
load 1w2i.ccp4, 1W2I_2fo
isomesh map_pos, 1W2I_map, 2.5, 1W2I, carve = 1.6
isomesh map_neg, 1W2I_map, -2.5, 1W2I, carve = 1.6
isomesh map, 1W2I_2fo, 1.0, 1W2I, carve = 1.6
color gray, map
color blue, map_pos
color red, map_neg
show sticks
```



Let's focus on an area with large differences and look at the actual density there as well

- The data: The diffraction pattern
- The results: The electron density
- The author's interpretation: The structure

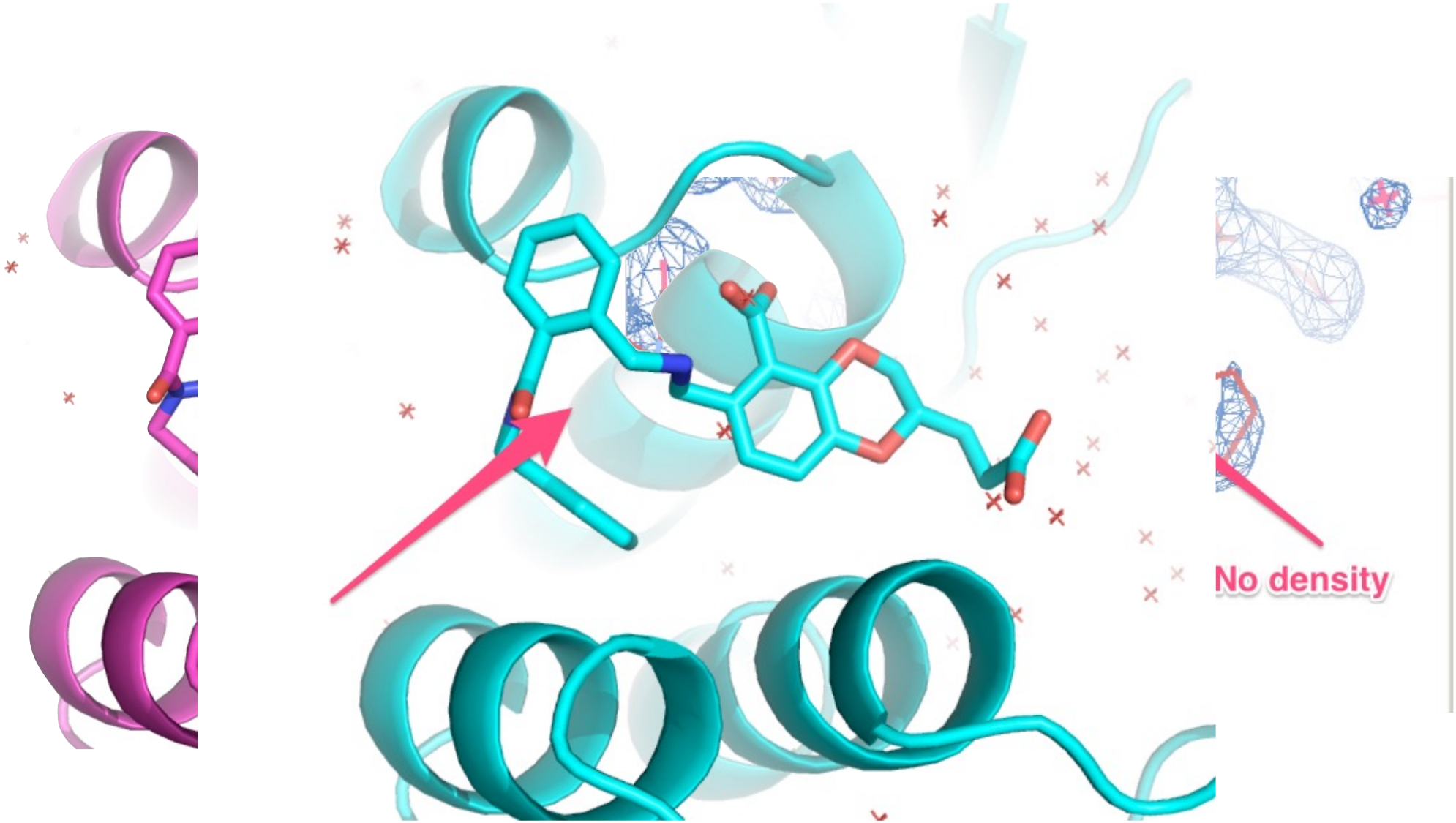


R-value and R-free are important factors in reliability

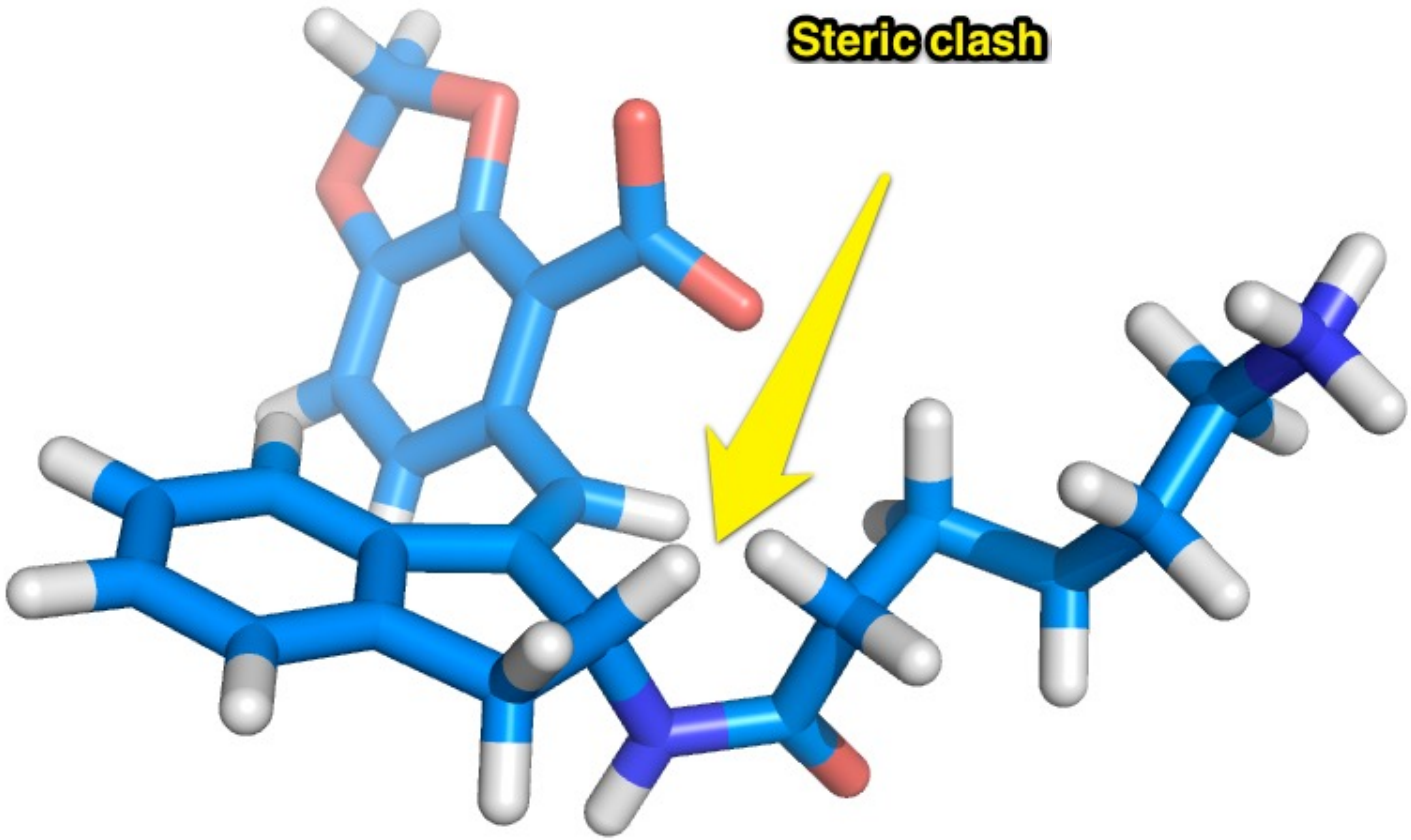
- R value: Measures how well simulated diffraction pattern matches experiment
  - Ideal: R=0
  - Typical: R~0.2
  - Random: R about 0.63
- But R value is used in building model, which introduces bias
- Better: Use R-free
  - R value on 10% of data which is held back and not used in building model. Typically slightly higher, around 0.26
  - Big concern if it is ever a lot higher than R -- suggests overfitting

(Source: PDB)

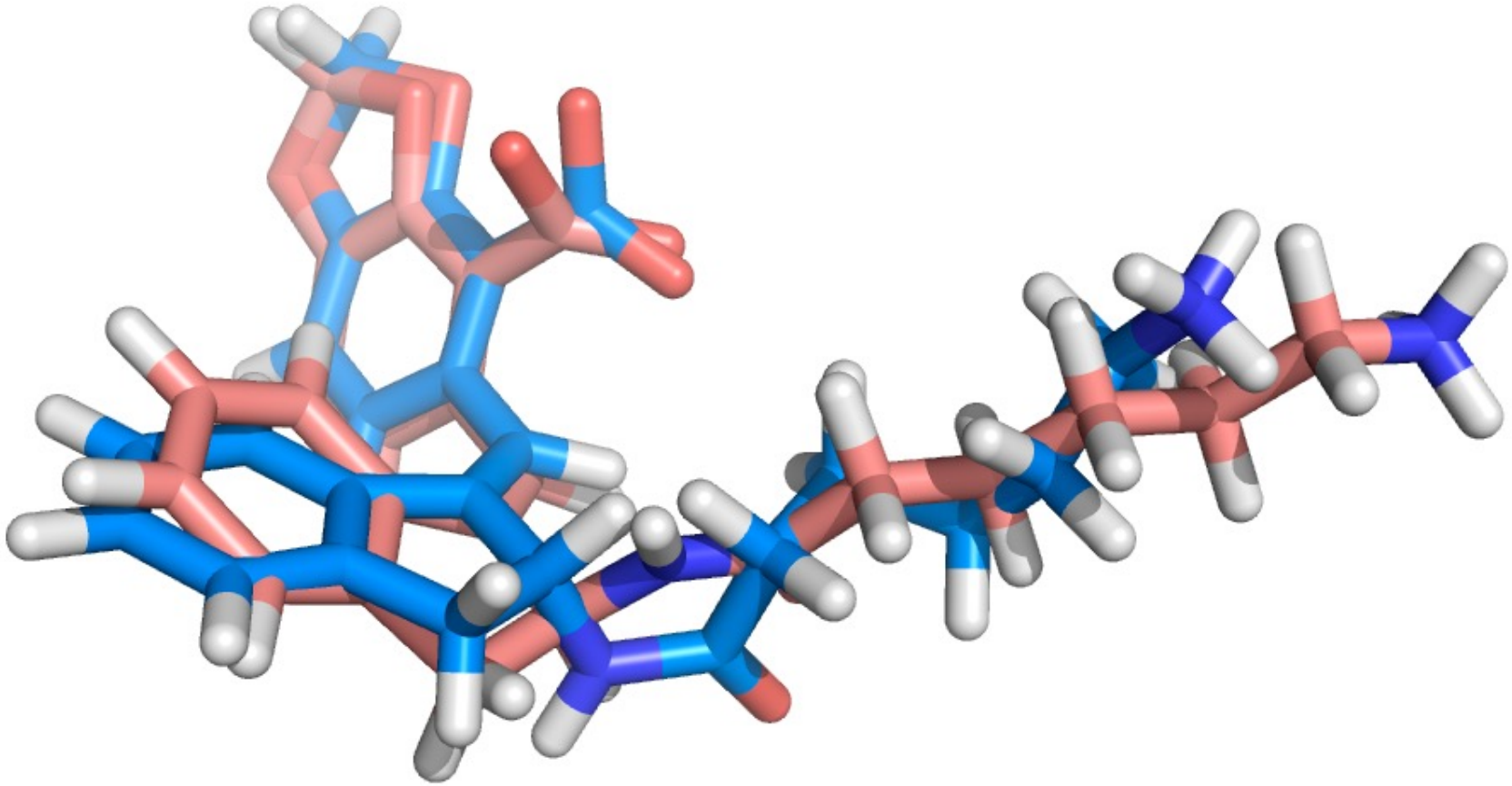
A lot of other things can get in the way: How sure are you of what you're looking at?



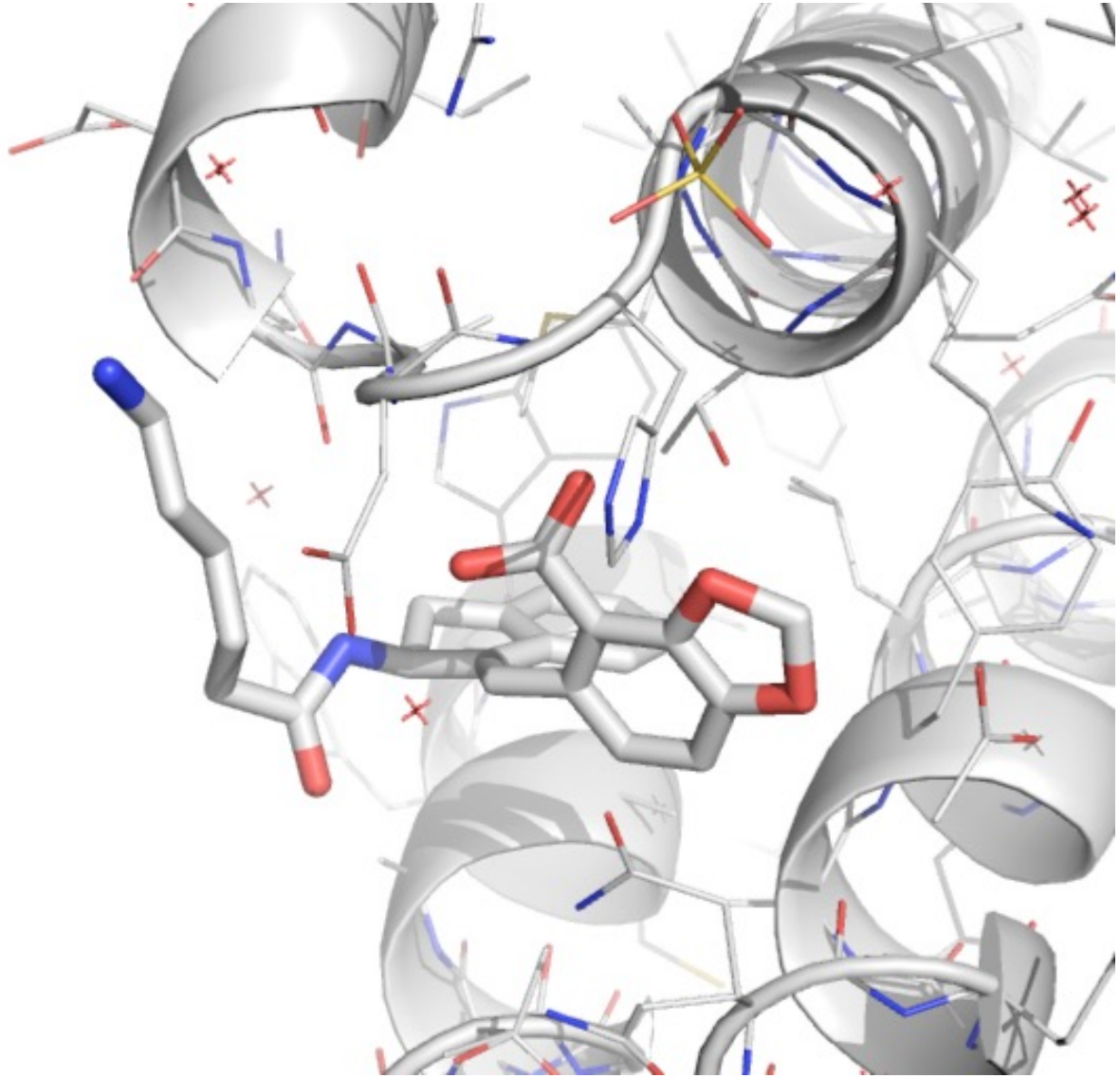
Molecular modeling is used in refinement. Do you have the right balance of forces and density?



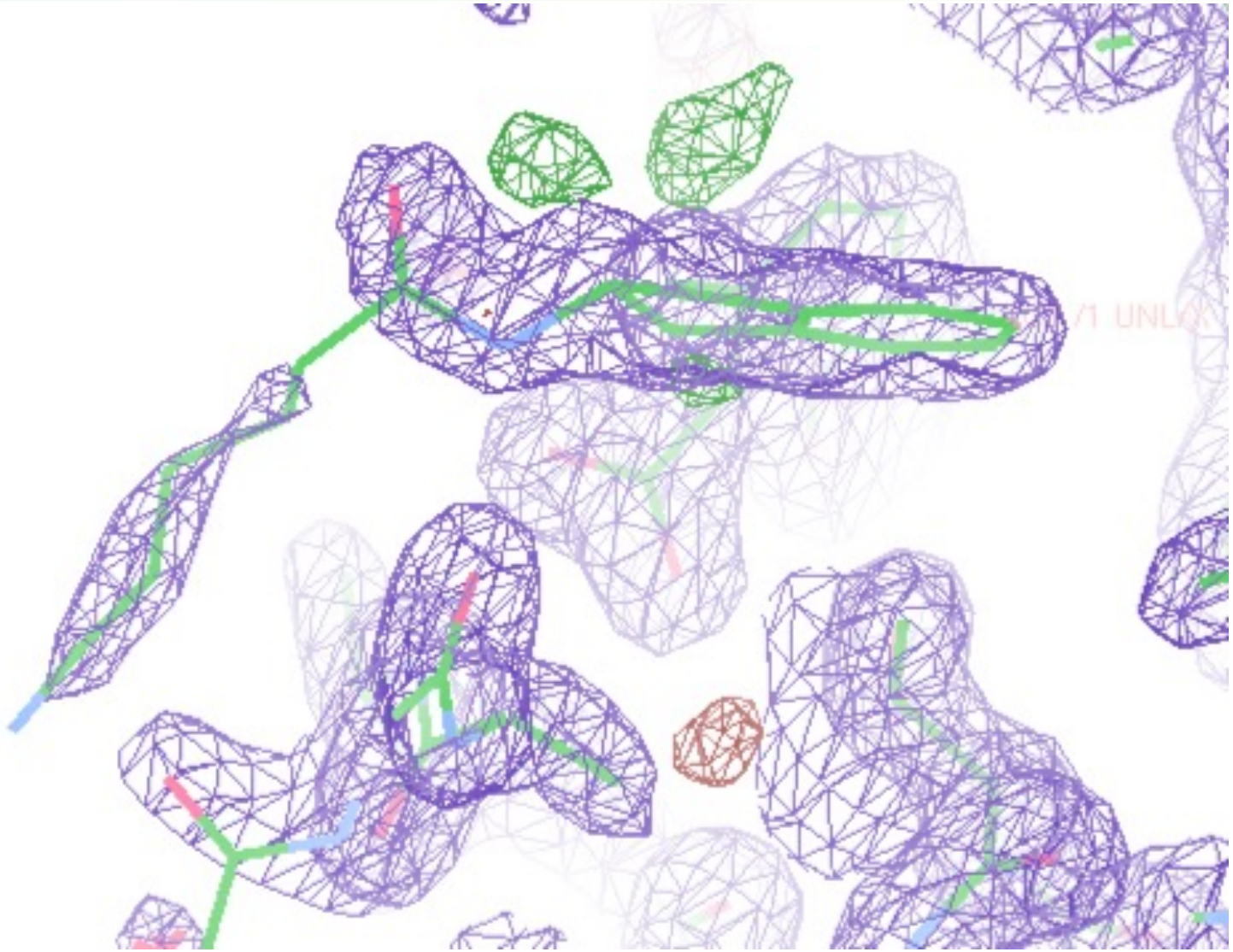
We proposed the opposite stereoisomer, which would eliminate the steric clash



The newly refined model eliminates the steric clash



And it fits the density as well as the original model

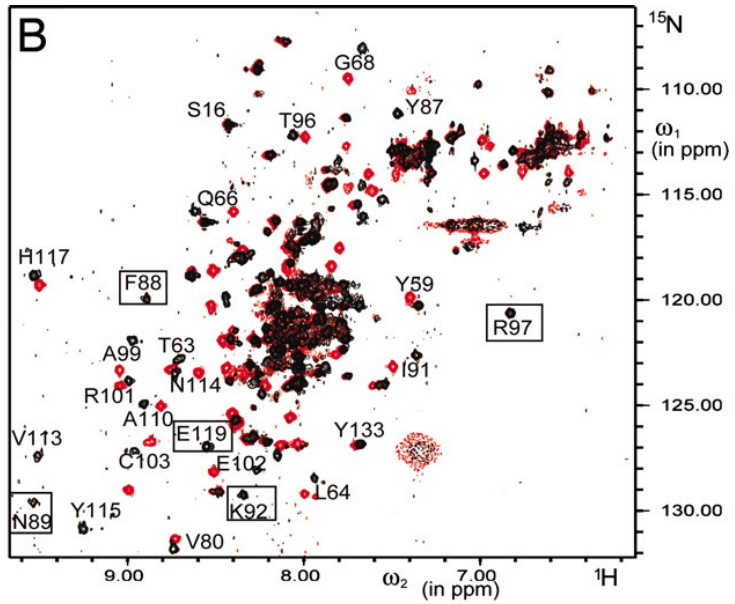


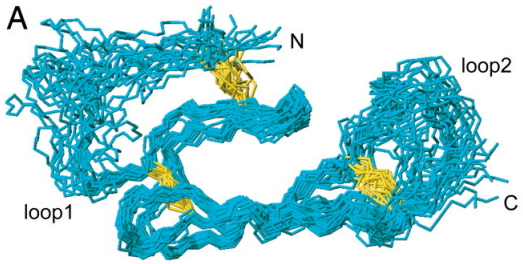
Crystal structures are limiting in some respects

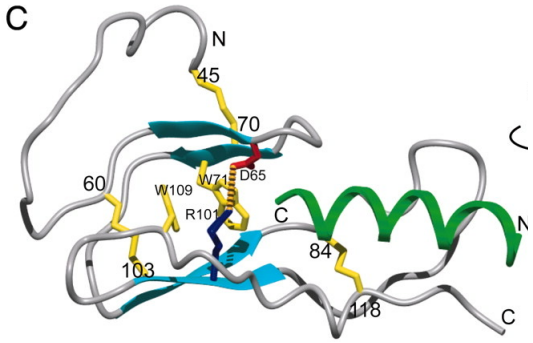
- Proteins are not really static
- They are a *model* fit to the data (the data is the diffraction pattern/electron density)
  - It's easy to confuse them with the data itself
  - Bad modeling yields a bad model!
  - Hetero atoms are particularly problematic (i.e. ligands)
- Crystal conditions, cosolvents, temperature can be important

NMR structures are also deposited on the PDB

- Proteins are not static and these structures are resolved in solution. Uses isotope labels.
- So, the model includes multiple conformations.
- NMR data can illuminate more than just structures, but it is size limited for what can be resolved.







(1) Grace *et al.* "Structure of the N-terminal domain …a peptide ligand" paper link

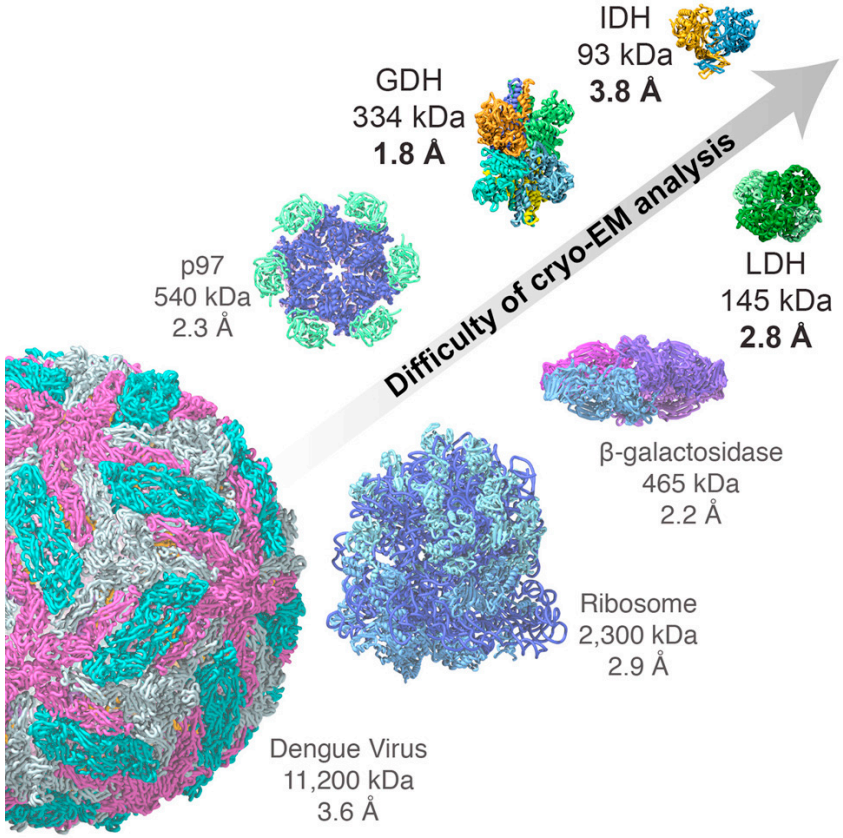
The next experimental revolution? Cryo-Electron Microscopy (cryo-EM)

Pros, for our purposes:

- Rapid freezing and actual fixation in vitreous ice preserves the hydrated state. The structural information may better reflect the state of the sample before it was frozen.
- Samples can be inhomogeneous allowing for observation of conformational ensembles.
- The chemical environment can be controlled

Cons:

- Potentially low resolution (but labs are working on applying ML to improving the signal to noise ratio detection)



Source: Merk et al.

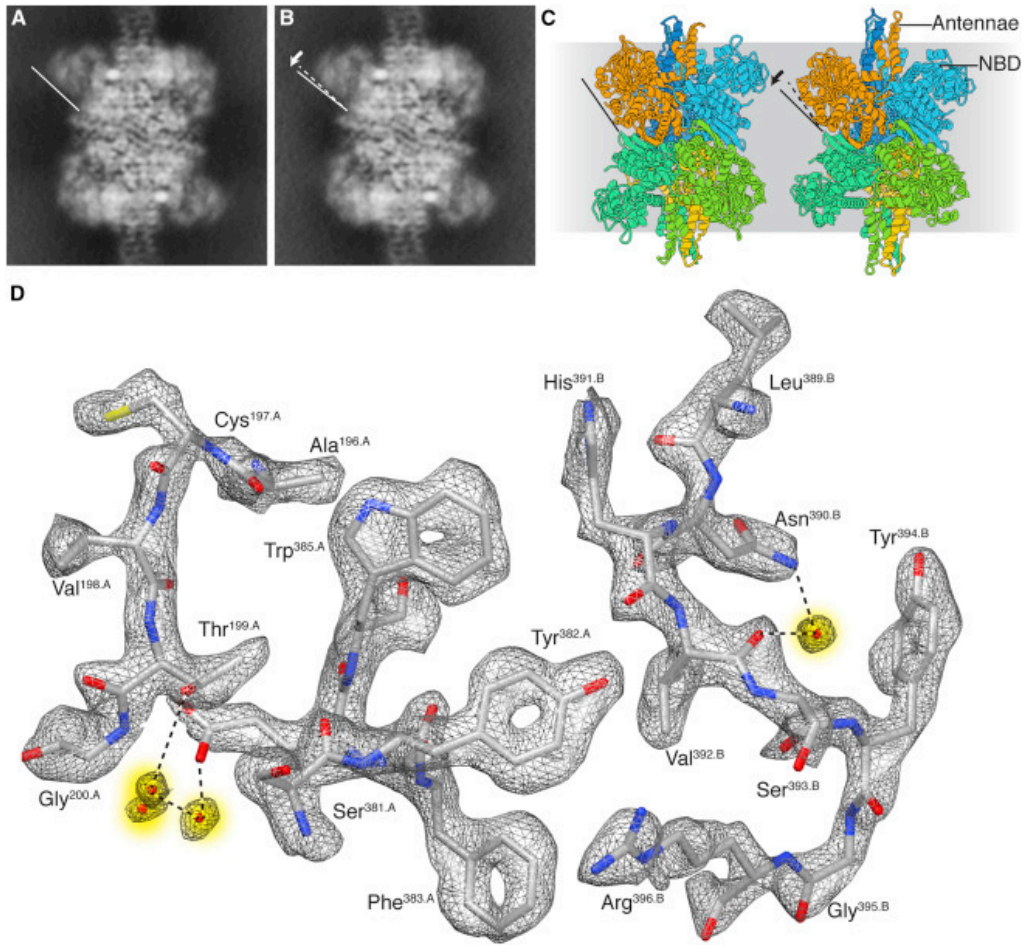
Further reading: Revolutionary cryo-EM is taking over structural biology, *Nature* **578**, 201 (2020)

The next experimental revolution? Cryo-EM

Cryo-EM can now cross the 2 Å resolution barrier and resolve structures of proteins with sizes < 100 kDa.

- This technique and applications of it will likely become rapidly more prevalent

- Increasing applications in pharma for sm drug discovery and biologics



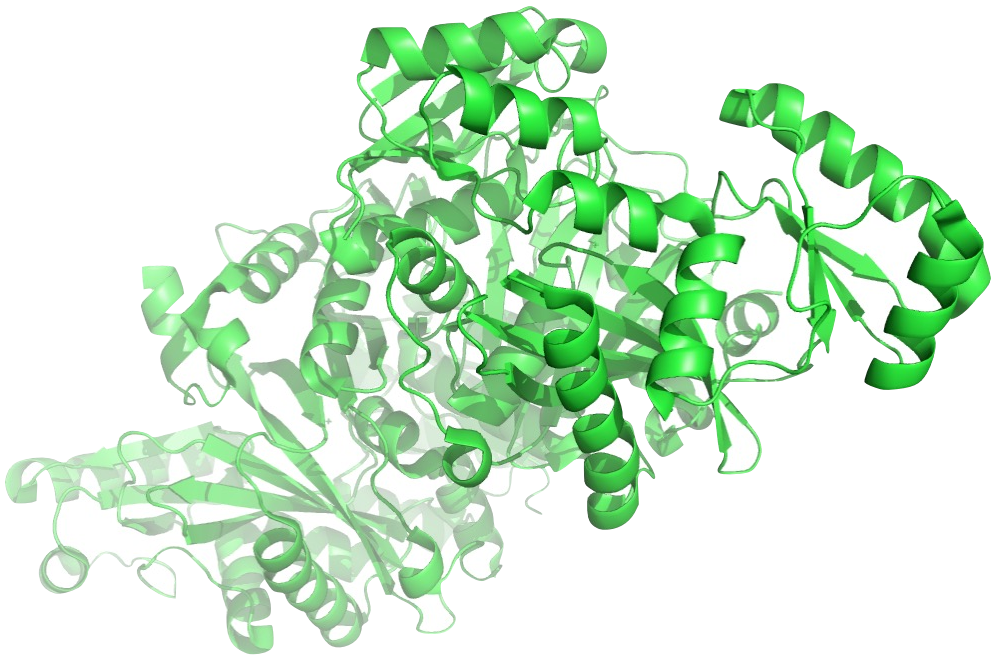
(1) Merk *et al.* "Breaking Cryo-EM …to Facilitate Drug Discovery" https://doi.org/10.1016/j.cell.2016.05.040

Homology modeling seeks to solve a common protein structure problem

- We have a protein sequence of interest
  - ex. MET ALA ILE VAL ARG ALA HIS LEU LYS ILE TYR GLY ARG VAL GLN GLY VAL GLY PHE ARG TRP SER MET GLN ARG GLU ...
  - that is, MAIVRAHLKIYGRVQGVGFAWSMNRE...
- We want to predict its structure
  - There are structures of related proteins available
- How do we use the structures of these related proteins to help with our problem?

Homology or comparative models predict new structures based on knowns

- Example: Could we predict the structure of BC in *S. Aureus* based on its structure in *E. Coli?*

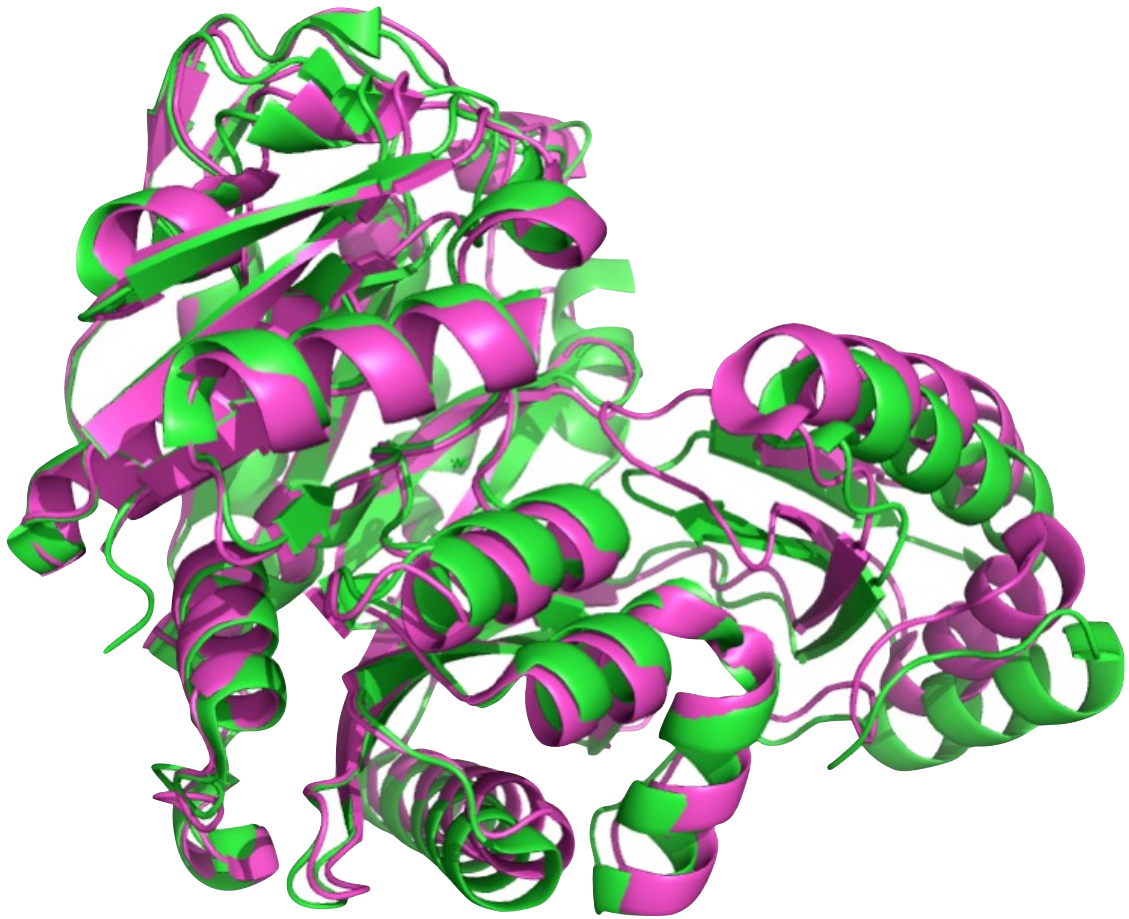


MLDKIVIANRGEIALRILRACKELGIKTVAVHSSADRDLKHVLLADETVCIGPAPSVKSY
LNIPAIISAAEITGAVAIHPGYGFLSENANFAEQVERSGFIFIGPKAETIRLMGDKVSAI
AAMKKAGVPCVPGSDGPLGDDMDKNRAIAKRIGYPVIIKASGGGGGRGMRVVRGDAELAQ
SISMTRAEAKAAFSNDMVYMEKYLENPRHVEIQVLADGQGNAIYLAERDCSMQRRHQKVV
EEAPAPGITPELRRYIGERCAKACVDIGYRGAGTFEFLFENGEFYFIEMNTRIQVEHPVT
EMITGVDLIKEQLRIAAGQPLSIKQEEVHVRGHAVECRINAEDPNTFLPSPGKITRFHAP
GGFGVRWESHIYAGYTVPPYYDSMIGKLICYGENRDVAIARMKNALQELIIDGIKTNVDL
QIRIMNDENFQHGGTNIHYLEKKLGLQEK

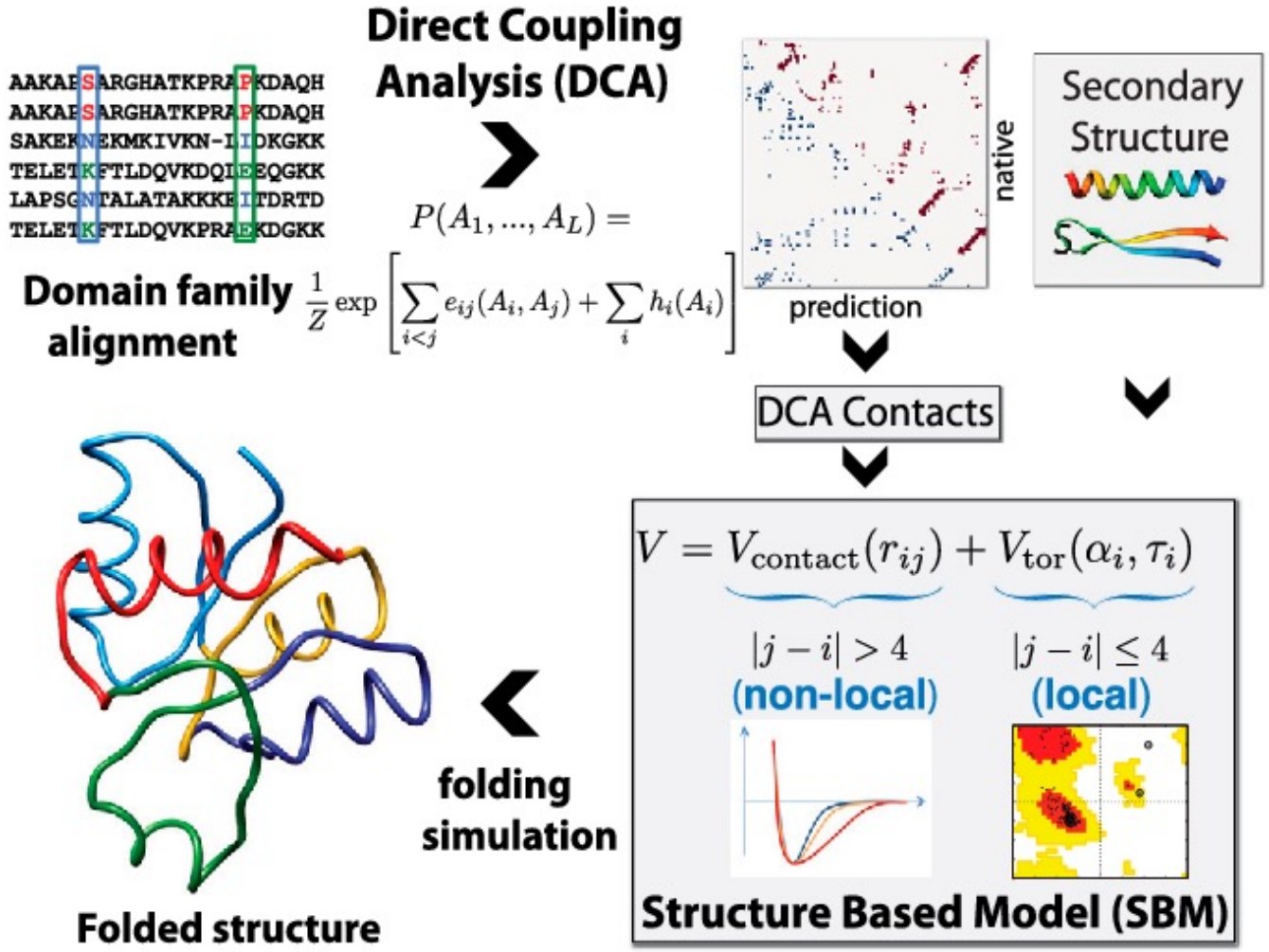
MKKVLIANRGEIAVRIIRACRDLGIQTVAIYSEGDKDALHTQIADEAYCVGPTLSKDSYL
NIPNILSIATSTGCDGVHPGYGFLAENADFAELCEACQLKFIGPSYQSIQKMGIKDVAKA
EMIKANVPVVPGSDGLMKDVSEAKKIAKKIGYPVIIKATAGGGGKGIRVARDEKELETGF
RMTEQEAQTAFGNGGLYMEKFIENFRHIEIQIVGDSYGNVIHLGERDCTIQRRMQKLVEE
APSPILDDETRREMGNAAVRAAKAVNYENAGTIEFIYDLNDNKFYFMEMNTRIQVEHPVT
EMVTGIDLVKLQLQVAMGDVLPYKQEDIKLTGHAIEFRINAENPYKNFMPSPGKIEQYLA
PGGYGVRIESACYTNYTIPPYYDSMVAKLIIHEPTRDEAIMAGIRALSEFVVLGIDTTIP
FHIKLLNNDIFRSGKFNTNFLEQNSIMNDEG

It turns out the structure is already available, and almost identical despite sequence differences

```
sp|P24182|ACCC_ECOLI        MLDKIVIANRGEIALRILRACKELGIKTVAVHSSADRDLKHVLLADETVC 50
tr|Q99TW7|Q99TW7_STAAM      -MKKVLIANRGEIAVRIIRACRDLGIQTVAIYSEGDKDALHTQIADEAYC 49
                             :.*::*********:**:****::***:***::*..*:*  *. :****: *

sp|P24182|ACCC_ECOLI        IGPAPSVKSYLNIPAIISAAEITGAVAIHPGYGFLSENANFAEQVERSGF 100
tr|Q99TW7|Q99TW7_STAAM      VGPTLSKDSYLNIPNILSIATSTGCDGVHPGYGFLAENADFAELCEACQL 99
                             :***: * .****** *:* *  **. .:********:***:***  * . :

sp|P24182|ACCC_ECOLI        IFIGPKAETIRLMGDKVSAIAAMKKAGVPCVPGSDGPLGDDMDKNRAIAK 150
tr|Q99TW7|Q99TW7_STAAM      KFIGPSYQSIQKMGIKDVAKAEMIKANVPVVPGSDG-LMKDVSEAKKIAK 148
                             ****. ::*: ** *  * * * **.** ****** * .*:.: : ***

sp|P24182|ACCC_ECOLI        RIGYPVIIKASGGGGGRGMRVVRGDAELAQSISMTRAEAKAAFSNDMVYM 200
tr|Q99TW7|Q99TW7_STAAM      KIGYPVIIKATAGGGGKGIRVARDEKELETGFRMTEQEAQTAFGNGGLYM 198
                             :**********:.****:*:**.*.: **  .: **. **::**.*. :**

sp|P24182|ACCC_ECOLI        EKYLENPRHVEIQVLADGQGNAIYLAERDCSMQRRHQKVVEEAPAPGITP 250
tr|Q99TW7|Q99TW7_STAAM      EKFIENFRHIEIQIVGDSYGNVIHLGERDCTIQRRMQKLVEEAPSPILDD 248
                             **:** **:****::.*. **.*:*.****::*** **:******:* :

sp|P24182|ACCC_ECOLI        ELRRYIGERCAKACVDIGYRGAGTFEFLFENGE--FYFIEMNTRIQVEHP 298
tr|Q99TW7|Q99TW7_STAAM      ETRREMGNAAVRAAKAVNYENAGTIEFIYDLNDNKFYFMEMNTRIQVEHP 298
                             * ** :*: ..:*.  :.*..***:**::: .:  ***:***********

sp|P24182|ACCC_ECOLI        VTEMITGVDLIKEQLRIAAGQPLSIKQEEVHVRGHAVECRINAEDP-NTF 347
tr|Q99TW7|Q99TW7_STAAM      VTEMVTGIDLVKLQLQVAMGDVLPYKQEDIKLTGHAIEFRINAENPYKNF 348
                             ****:***:**:* **::* *: *. ***:::: ***:* ******:* :.*

sp|P24182|ACCC_ECOLI        LPSPGKITRFHAPGGFGVRWESHIYAGYTVPPYYDSMIGKLICYGENRDV 397
tr|Q99TW7|Q99TW7_STAAM      MPSPGKIEQYLAPGGYGVRIESACYTNYTIPPYYDSMVAKLIIHEPTRDE 398
                             :****** :: ****:*** **  *:.**:********:.*** :  .**

sp|P24182|ACCC_ECOLI        AIARMKNALQELIIDGIKTNVDLQIRIMNDENFQHGGTNIHYLEKKLGLQ 447
tr|Q99TW7|Q99TW7_STAAM      AIMAGIRALSEFVVLGIDTTIPFHIKLLNNDIFRSGKFNTNFLEQNSIMN 448
                             **    .**.*::: **.*.: ::*:::*:: *: *  * ::***::  ::

sp|P24182|ACCC_ECOLI        EK- 449
tr|Q99TW7|Q99TW7_STAAM      DEG 451
                             ::
```



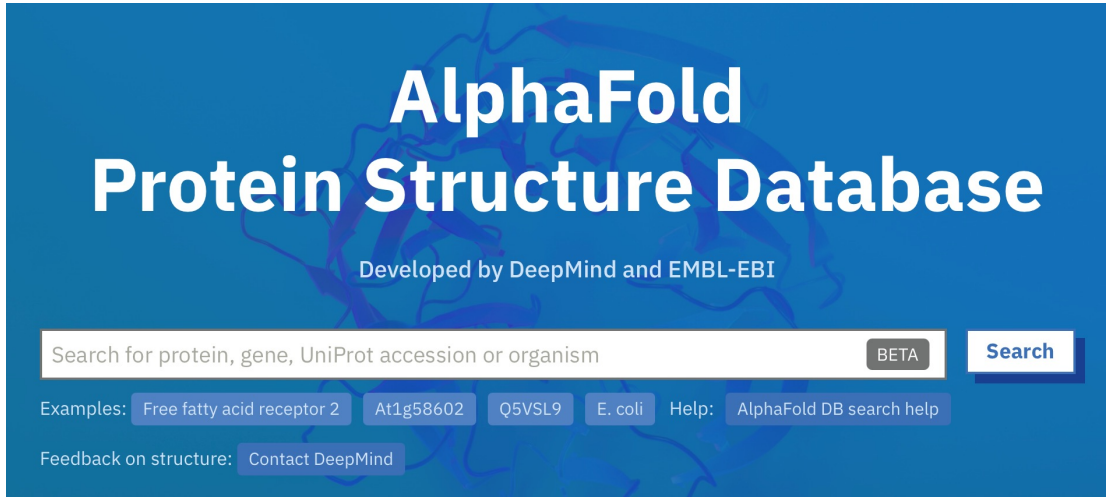
Leveraging evolution to learn protein structures



(1) Sulkowska *et al.* "Genomics-aided structure prediction" DOI:10.1073/pnas.1207864109

AlphaFold: The future for protein folding? Sometimes

AlphaFold is an AI system developed by DeepMind that predicts a protein's 3D structure from its amino acid sequence. It regularly achieves accuracy competitive with experiment.



Do not use the resource blindly. Apply towards research projects when you have a strong foundation of what a reasonable result should be. *Remember: you must justify your choices in review!*

(1) https://alphafold.ebi.ac.uk/search/uniprotDescription/Free%20fatty%20acid%20receptor%202

Outline

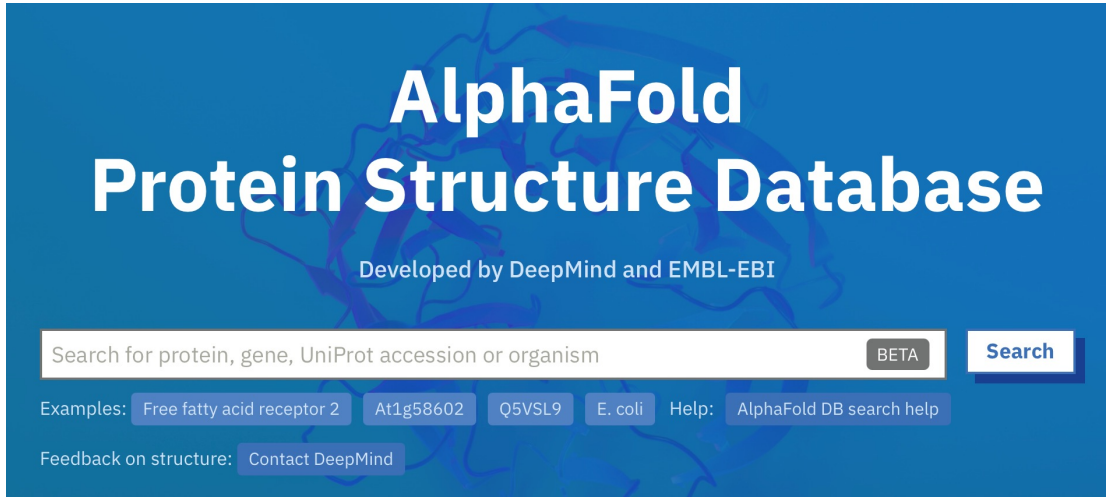
1. Background: why the starting configuration can make or break your simulation
   - Theoretical considerations: statistical mechanics, protein physics
   - Limitations of atomic resolution simulations
2. Structure files
   - Protein Data Bank: crystallography, NMR, cryo-EM, alphaFold
3. Tools and considerations

AlphaFold: The future for protein folding? Sometimes

(1) Go to link

(2) Search “free fatty acid receptor 2”

Q: How would you assess which parts of this model are most trustworthy?

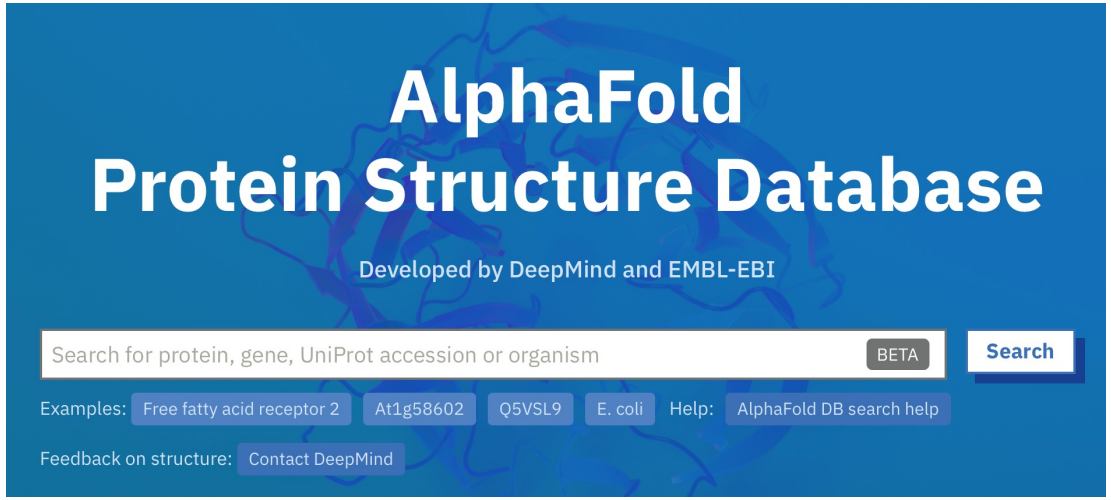


(1) https://alphafold.ebi.ac.uk/search/uniprotDescription/Free%20fatty%20acid%20receptor%202

AlphaFold: The future for protein folding? Sometimes

(1) Go to link

(2) Search “Histone domain containing protein, uniprot: A0A077ZLJ1”



Qs:

(1) What part of the protein did AlphaFold struggle in prediction?

(2) Is this a bug or a feature?

(3) Can it be improved?

(1) https://alphafold.ebi.ac.uk/search/uniprotDescription/Free%20fatty%20acid%20receptor%202

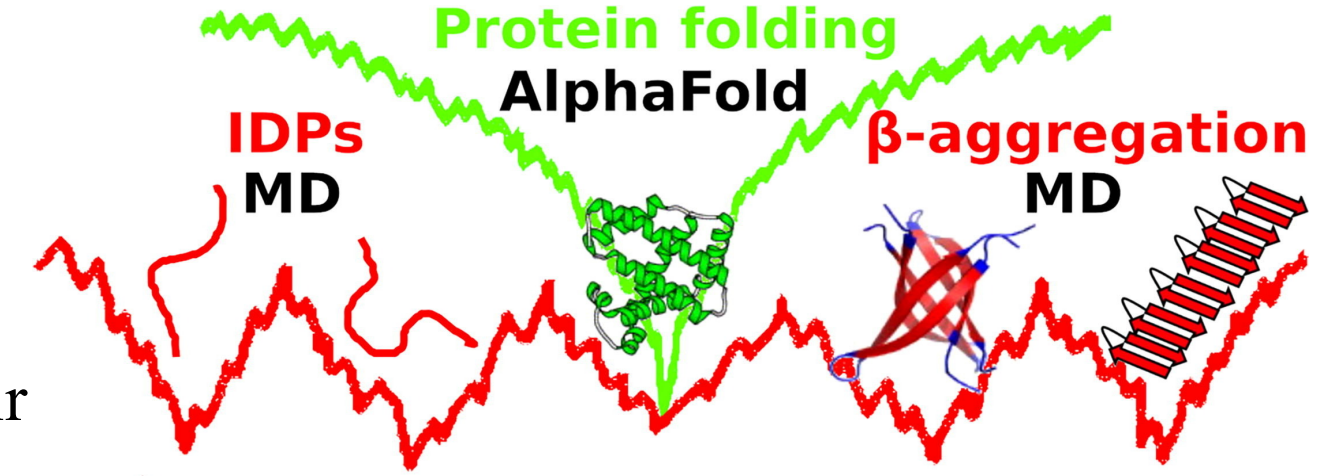
This motivates homology or comparative modeling: Often, proteins have high structural similarity even at low sequence identity

- Globular, homologous proteins have sequence similarity:
  - Above 50% sequence identity, models are generally quite good, with errors mostly in sidechain positioning
  - In 30-50% range, errors can be more severe, but often still tolerable (generally considered acceptable)
  - Below 30%, all bets are off

- Sequence identity: Fraction of amino acids that are exactly the same
- Sequence similarity: Fraction that have similar properties (i.e. polar, positively charge, negatively charged, hydrophobic)

Sequence identity and sequence similarity are related but different concepts

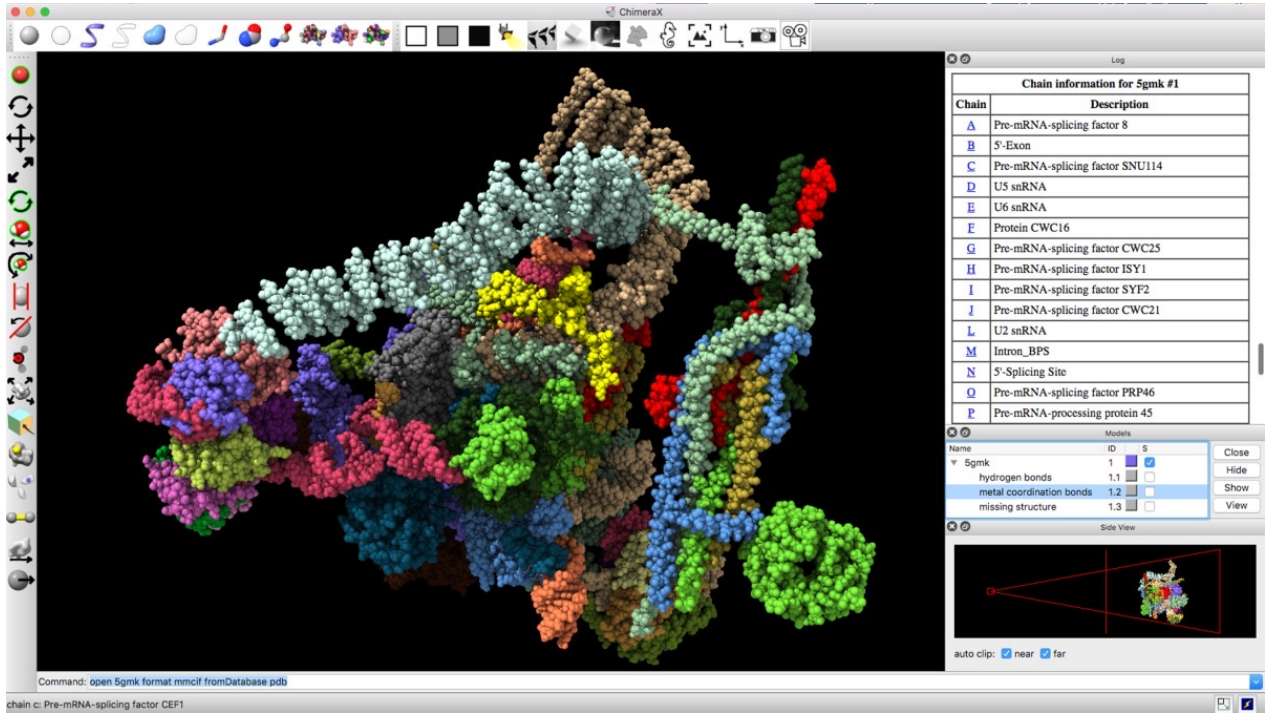
- But Intrinsically Disordered Proteins (IDPs) and Intrinsically Disordered Regions (IDRs: loops, tails) evolve faster.
- There is less constraint on their evolution so there is **less signal** in their homology data
- **AND** their free energy landscapes are broad and shallow.



(1) Birgit Strodel, "Energy Landscapes … Disordered Proteins" https://doi.org/10.1016/j.jmb.2021.167182

AlphaFold, docking, visualization tool: UCSF ChimeraX

- Uses MODELLER: you can edit structures for simulation with GUI visualization.
- It can search for solved AlphaFold structures, or you can run yourself:
  - https://www.rbvi.ucsf.edu/chimerax/data/alphafold-nov2021/af_sbgrid.html
- Tools for docking and ligand visualization:
  - https://www.youtube.com/watch?v=iZPDRLH9W2U

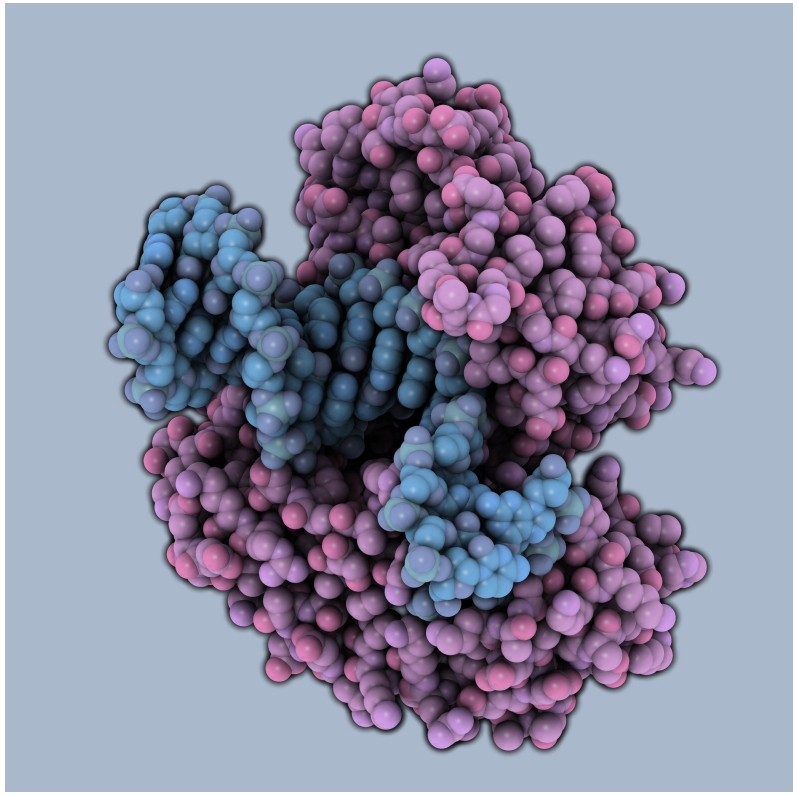


In a limited pool of GUIs, which should you use?

- Visualization:

difficulty

- QuteMol: Impressive images, zero effort.
- Pymol: great for a lot of things, sometimes lacking for MD. Python interface, most people start here.
- Chimera: high degree of customization, some learning curve. But, added features that are hard to replace in a GUI.
- VMD: Useful for trajectories, but unpolished publication quality images. Scripting for Tk console can be tedious.



QuteMol image: attribution unknown

(1) Wiki: https://en.wikipedia.org/wiki/QuteMol

Pause before you proceed, what to do before MD

- Checklist, open your structure files and preferred GUI:
  - Is the structure I got the right one?? Is it the correct species? No substitutions? Highest resolution?
  - Are loops or bonds unphysically crossing each other?
  - Are there large clashes detected? (VMD, or Chimera 'find clashes' command (https://www.cgl.ucsf.edu/chimera/docs/ContributedSoftware/findclash/findclash.html)
  - Are there heavy or exotic residues left as an artifact of experimental resolution?
  - Are any atoms missing?
  - Are protonation states defined? Calculated? If not, why not? To calculate: H++ Server.
  - Should any heavy atoms, such as metals, be included in specific locations?

Remember: any and every decision you make at the start in MD can be questioned in review.
Keep notes of what you did and why you did it. Reproducibility matters.

Prepping structures the automated way

- Checklist, open your structure files and preferred GUI:
  - Orion
  - OESpruce, OEBio (https://docs.eyesopen.com/toolkits/python/oechemtk/index.html#oebio-theory)
  - scripting with MODELLER
  - Scripting with tleap and PARMED.

- These tools are good once you have a dedicated workflow. But, don't just hit play on a downloaded script.

Remember: any and every decision you make at the start in MD can be questioned in review.
Keep notes of what you did and why you did it. Reproducibility matters.