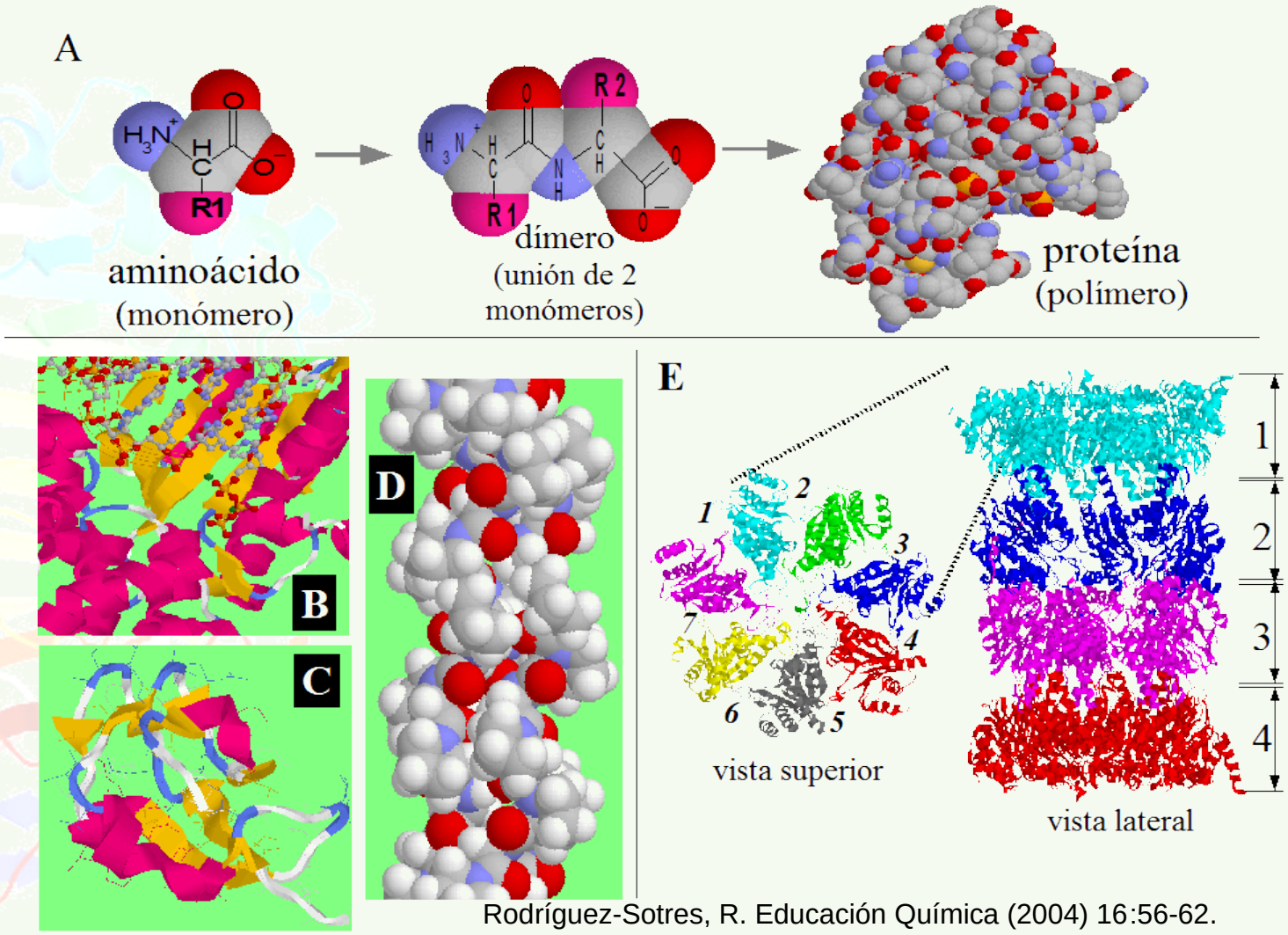
Three-dimensional protein structure

Prediction and assessment at the dawn of the *Artificial Intelligence* era

Rogelio Rodríguez Sotres

Facultad de Química,
UNAM

Las proteínas son hetero-polímeros con gran diversidad estructural y funcional



Rodríguez-Sotres, R. Educación Química (2004) 16:56-62.

The problem of complexity

There are 20^{10} peptides of 10 amino acids

that is 10.2×10^{12} possible sequences

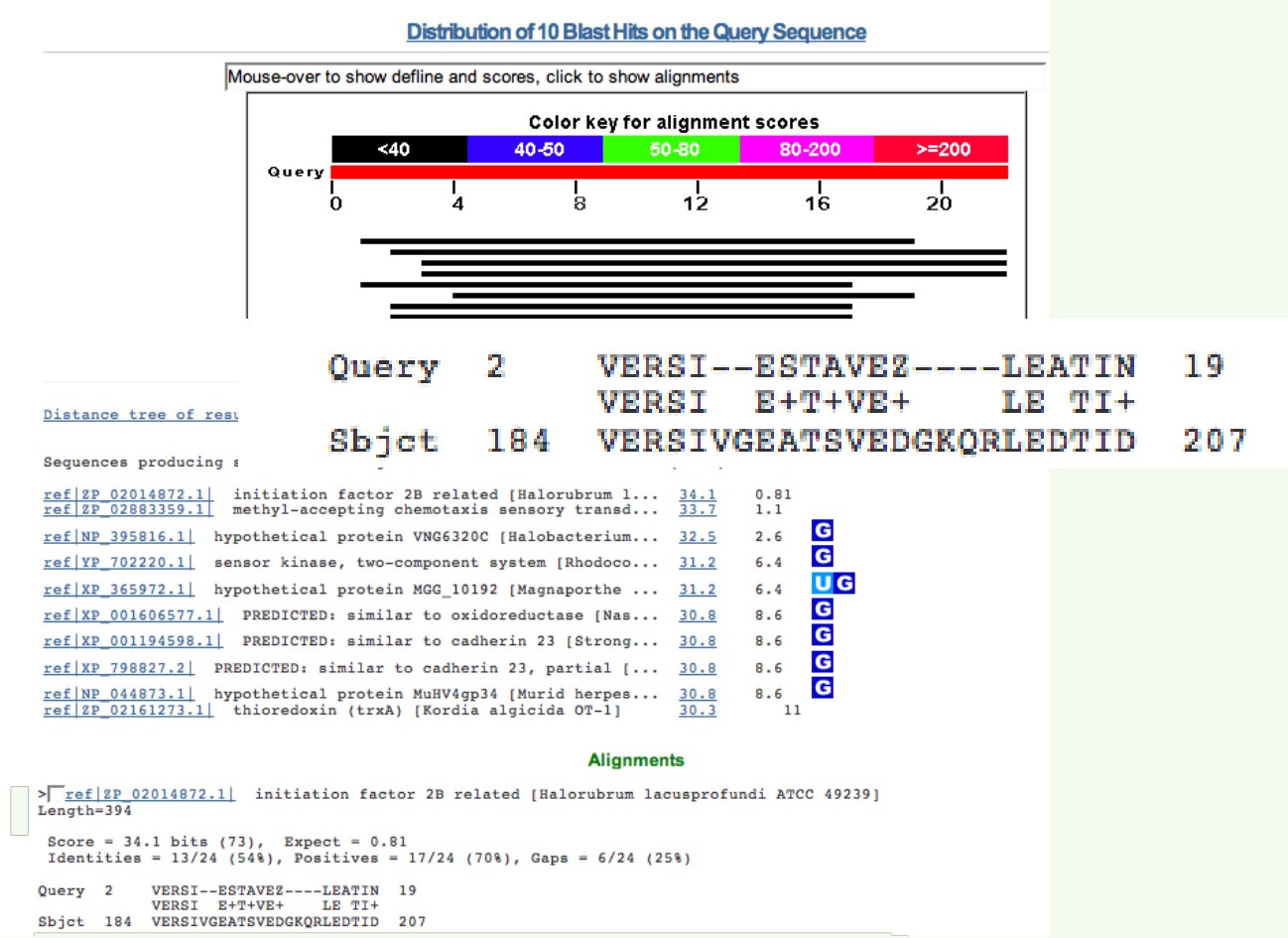
- ***The NCBI reference data base (RefSeq, protein) comprises 289'333'423 protein sequences* for a total length of nearly 10 billion letters.***

There is less than 9.7 in 10'000 chances to find in RefSeq a random sequence of 10 amino acids.

*Useful fact: An average protein has 345 aa for a MW of 38 KDa.

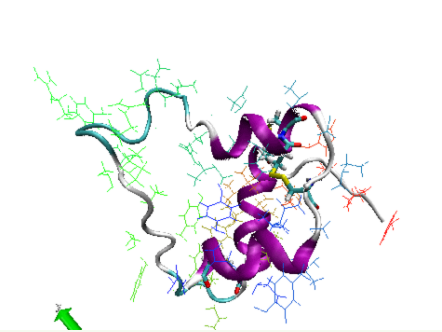
Example: BLAST some weird sequence:

A
VER
SI
ESTA
VEZ
LE
ATINARE



Conformation complexity

- Each residue can adopt many conformations
- *Levinthal*[*‡] estimated in millions of times the age of the universe the time required for a protein to explore all of its possible conformations.



- But proteins do fold in a fraction of a second
- for 58 years predicting the fold of a protein was not possible[†]

*This is known as the *Levinthal*'s paradox

‡Zwanzig et al. PNAS USA (1992) 89:20-2, DOI: 10.1073/pnas.89.1.20

†Dill et al. Sci (2012) 338:1042-6, DOI: 10.1126/science.1219021

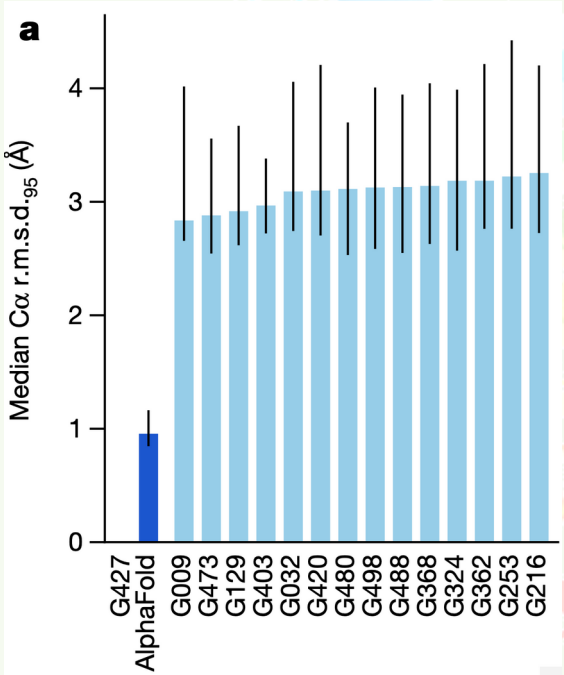
Highly accurate protein structure prediction with AlphaFold

Jumper J , Evans R , Pritzel A , Green T , Figurnov M , Ronneberger O , Tunyasuvunakool K , Bates R , Žídek A , Potapenko A , Bridgland A , Meyer C , Kohl SAA , Ballard AJ , Cowie A , Romera-Paredes B , Nikolov S , Jain R , Adler J , Back T , Petersen S , Reiman D , Clancy E , Zielinski M , Steinegger M , Pacholska M , Berghammer T , Bodenstein S , Silver D , Vinyals O , Senior AW , Kavukcuoglu K , Kohli P , & Hassabis D (2021) Nature 596:583—589. DOI: 10.1038/s41586-021-03819-2

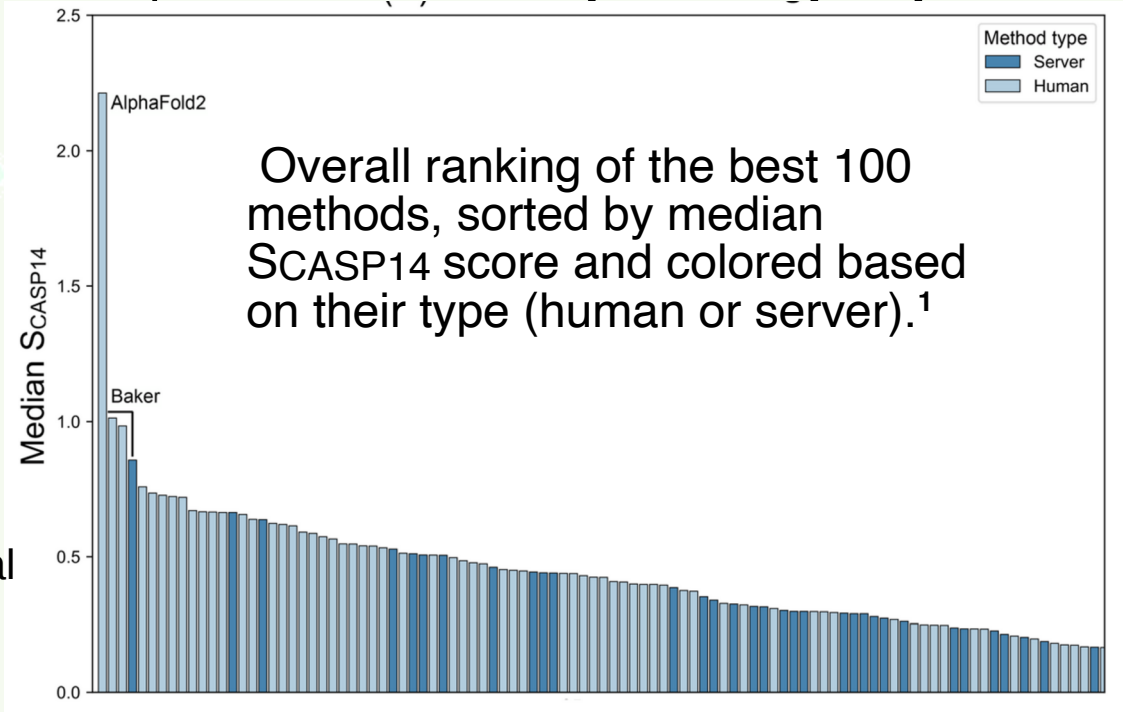
IN 2021 SOMETHING REMARKABLE HAPPENED:

An Artificial Intelligence was able to predict the protein folding using the evolutionary information of a protein (*i.e*. a sequence alignment)

In the CASP competition teams try to predict the folding of target proteins, known only to a jury.



(a) The performance of AlphaFold on CASP14 (n = 87 protein domains) relative to the top-15 entries, group numbers correspond CASP group id. Data are median ± 95% confidence interval (10,000 bootstrap samples).[2]

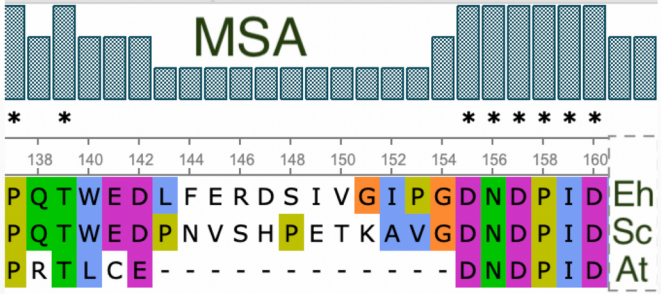


Overall ranking of the best 100 methods, sorted by median SCASP14 score and colored based on their type (human or server).[1]

[1]Pereira, J. et al. Proteins (2021) , DOI 10.1002/prot.26171
[2]Jumper, J. *et al*. Nature (2021) 596: 583—589 , DOI 10.1038/s41586-021-03819-2

How evolution helps us to understand the folding of proteins?

Selective pressure acts on both:
close and distant contacts



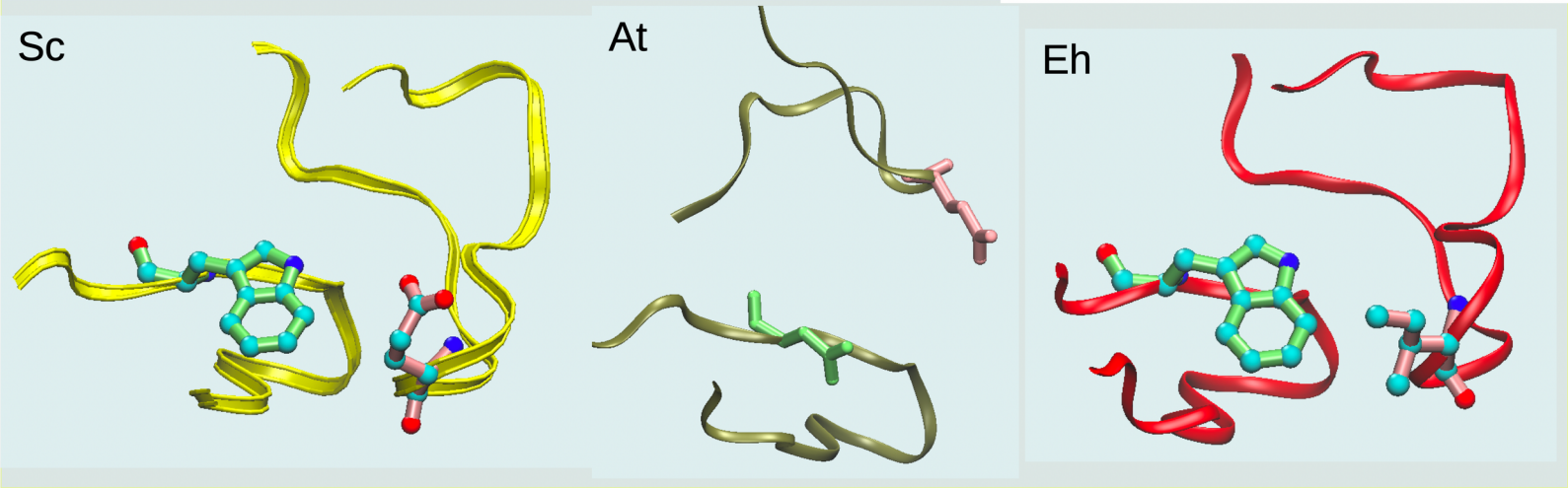


Figure 1. Example of change in contact constrains for inorganic pyrophosphatases from *Enthamoeba histolytica, Saccharomyces cereviciae and Arabidopsis thaliana.*

Deep Mind promise‡

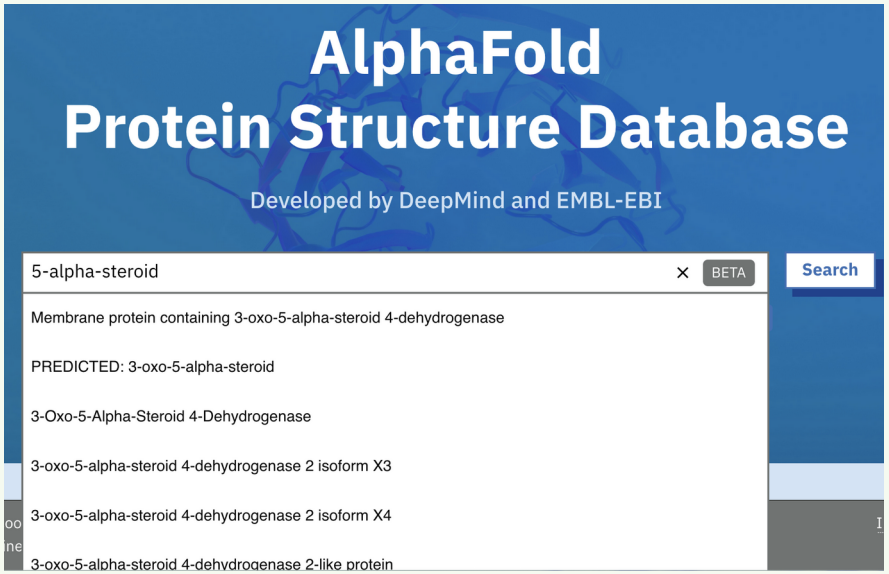
- Depp Mind promised to predict the nearly complete proteomes of all organisms with sequenced genomes at:
- UNIREF genomic data base
- As for now, prediction for the entire UniprotKB database are already available:

https://alphafold.ebi.ac.uk

‡A notorious contribution to science coming from a private company.

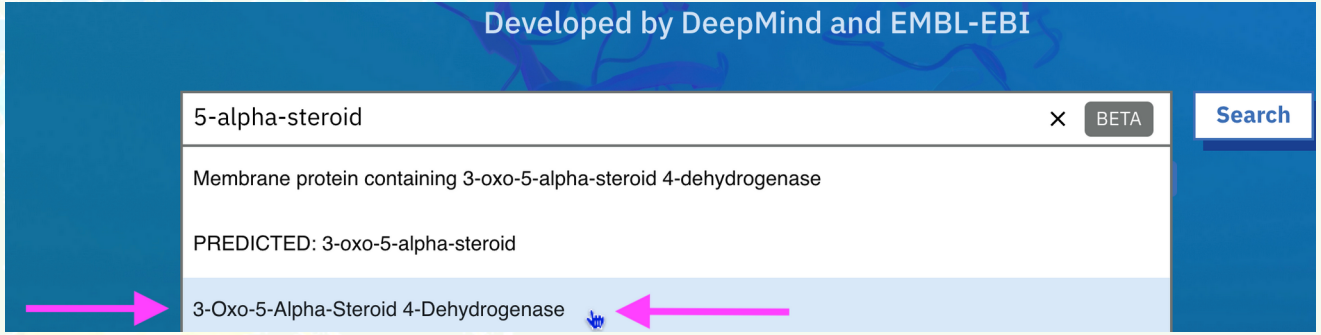
Let us test this

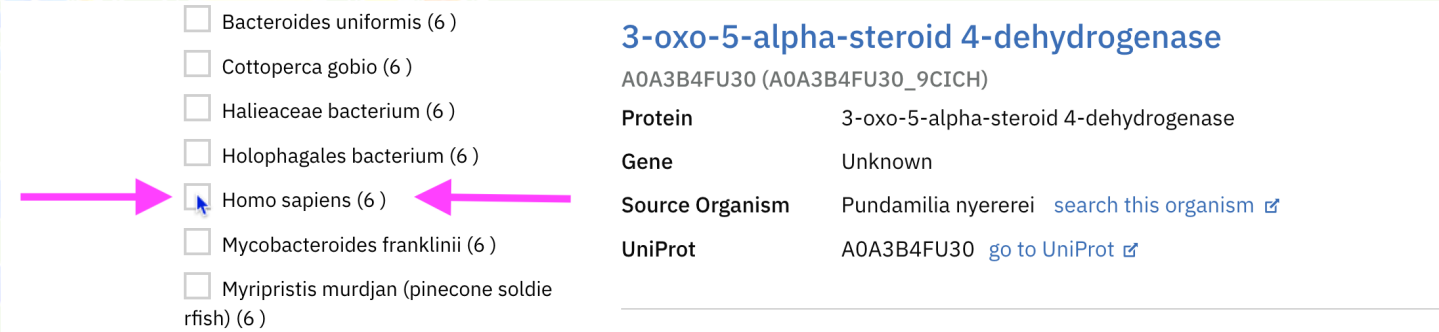
- Proteins data at the AlphaFold resource are organized by UniprotKB accession codes



- And you may search by key word or protein name:
- type “5-alpha-steroid” and choose the link:
  - 3-Oxo-5-Alpha-Steroid-4-Dehydrogenase

Check on “Homo sapiens” box





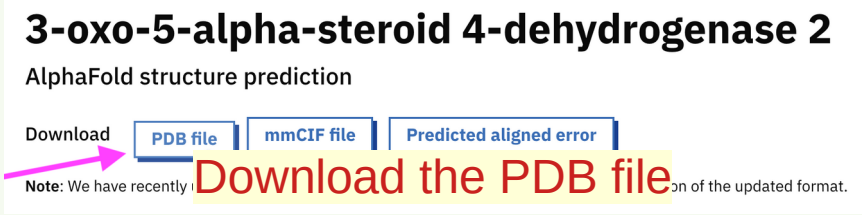
Chose the isoform Nº 2

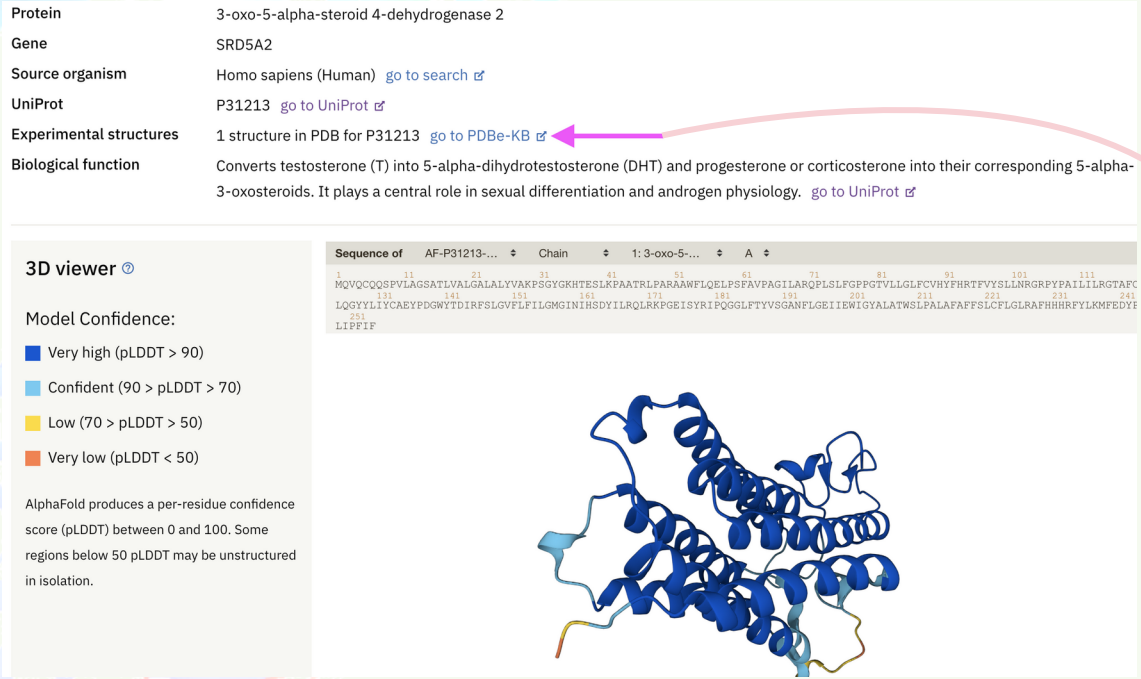
3-oxo-5-alpha-steroid 4-dehydrogenase 2

P31213 (S5A2_HUMAN)

Protein	3-oxo-5-alpha-steroid 4-dehydrogenase 2
Gene	SRD5A2
Source Organism	Homo sapiens search this organism
UniProt	P31213 go to UniProt
PDBe-KB	1 PDB structure for P31213 go to PDBe-KB

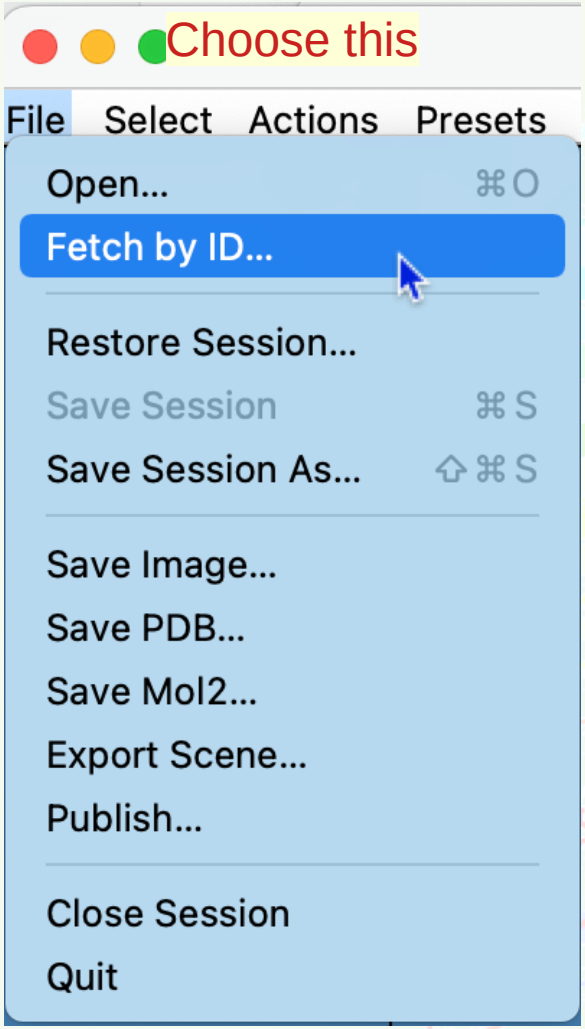
Now we get the live 3D-picture of the prediction

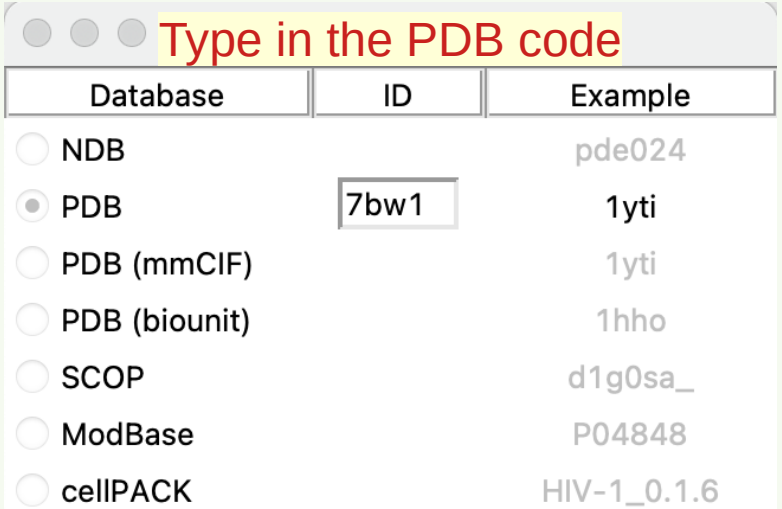




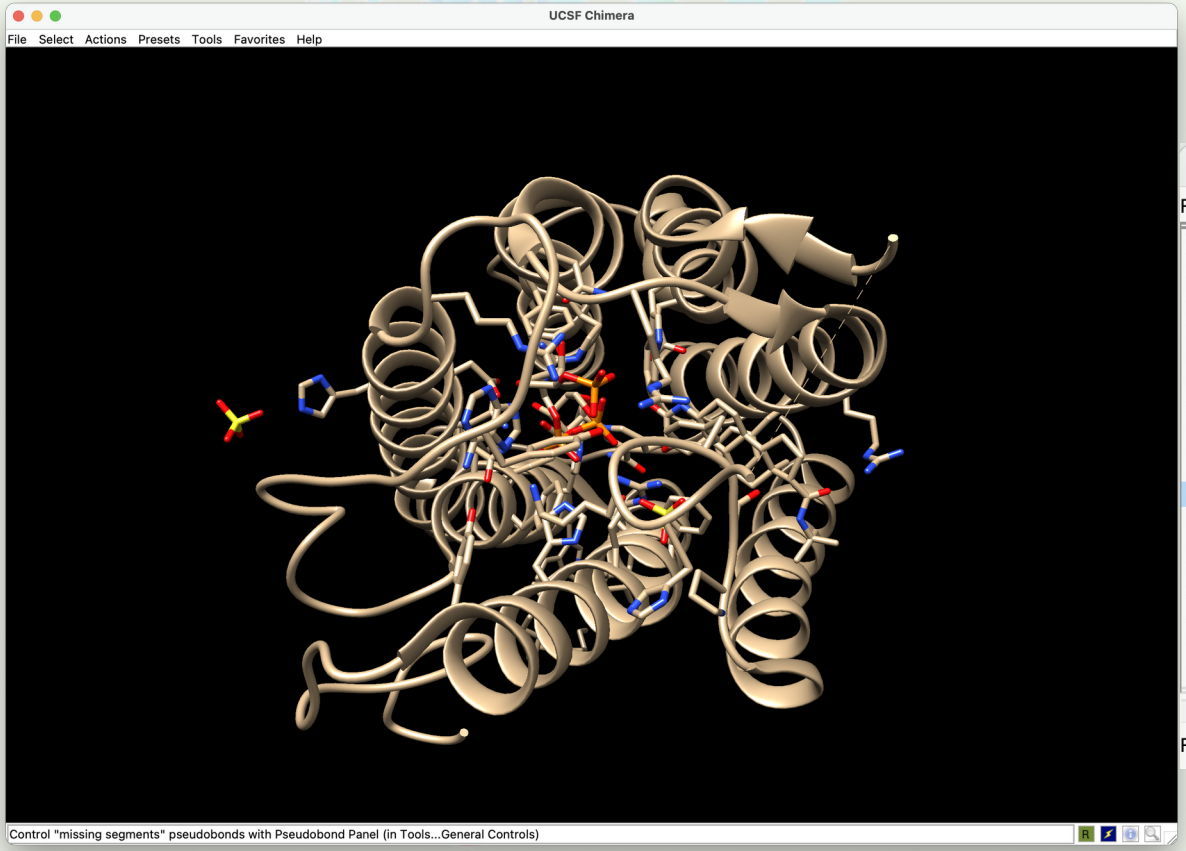
and there is an experimental structure of this protein, linked here

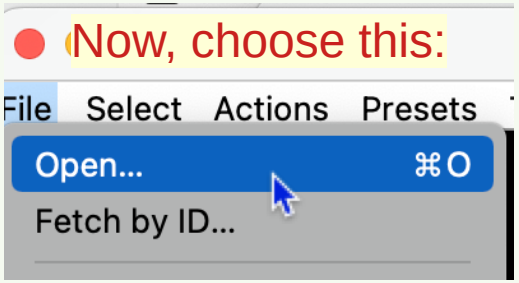
Open Chimera

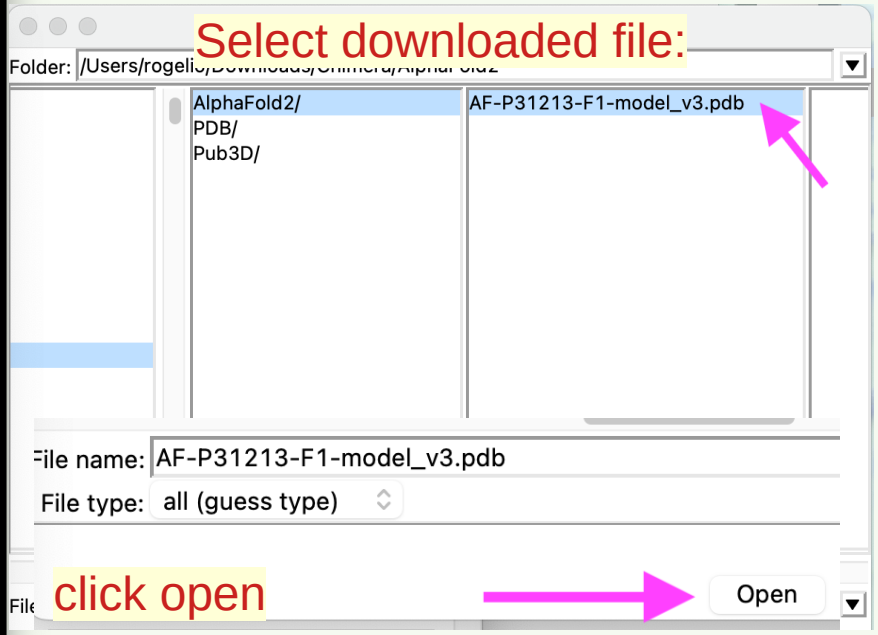




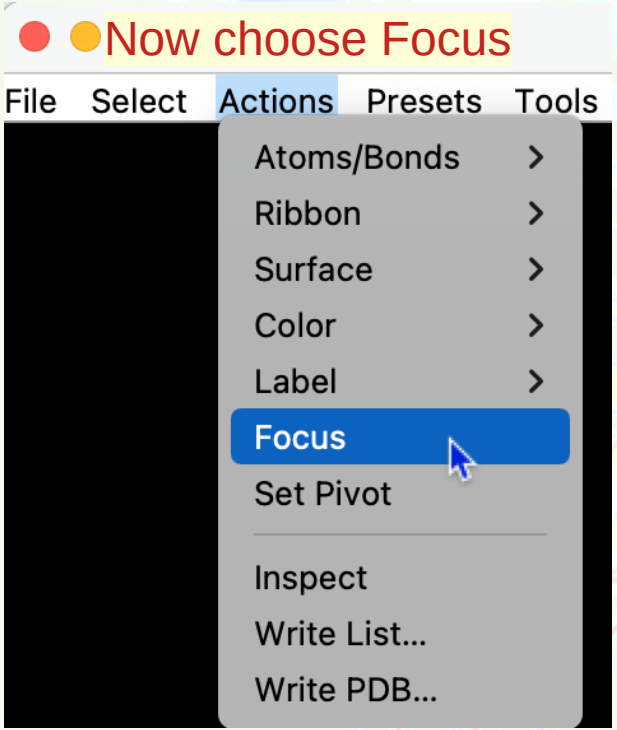
We get this image



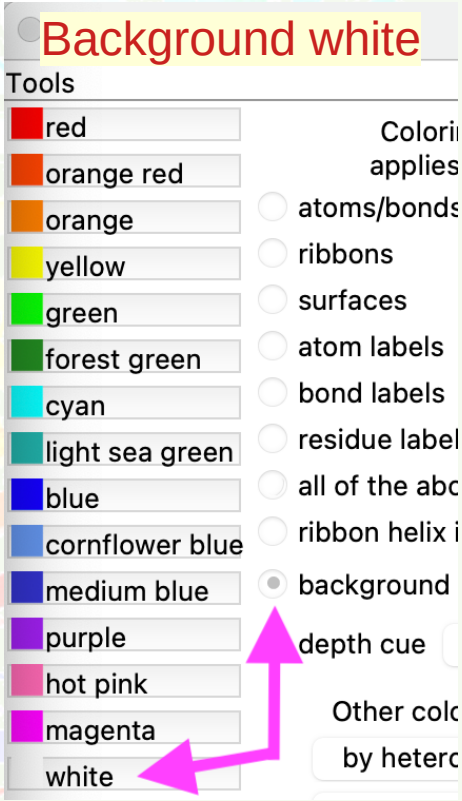




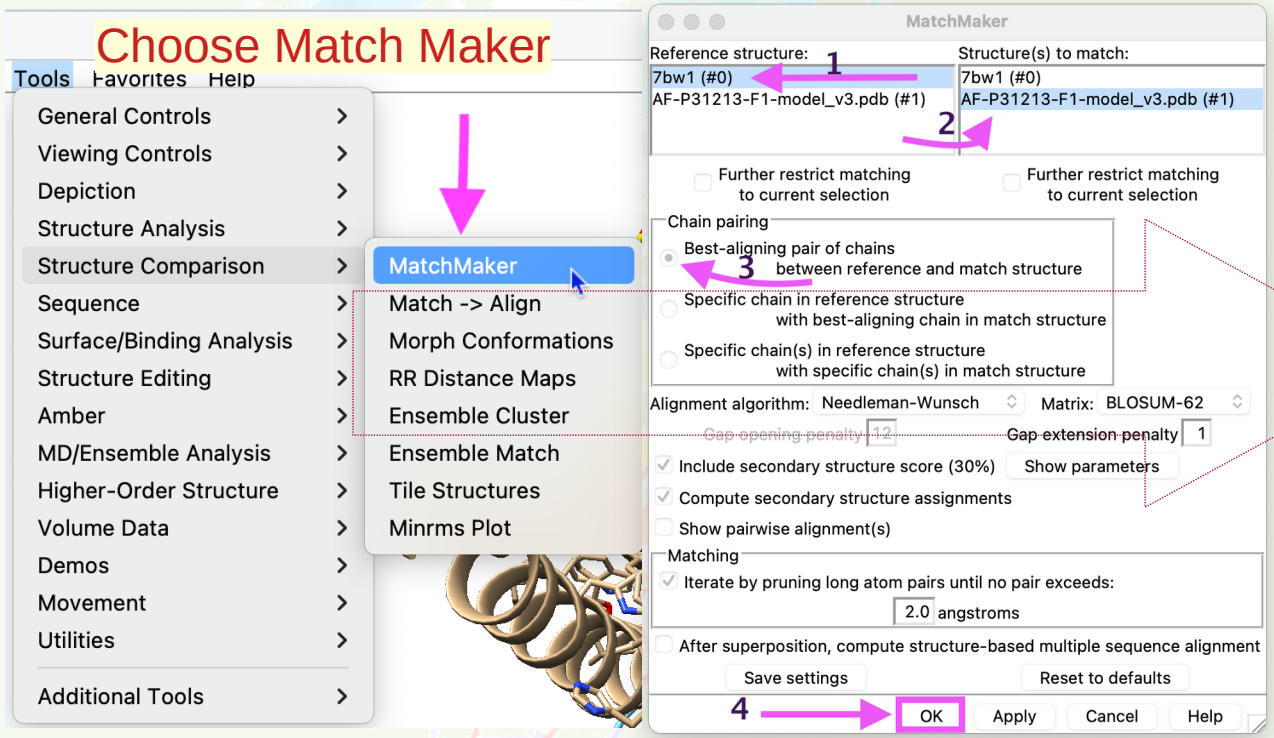
Get loaded the model also

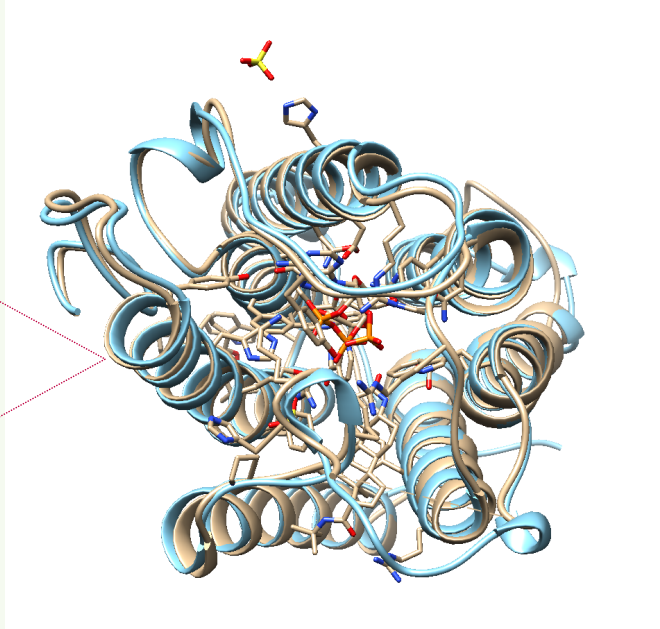


Make some decorations

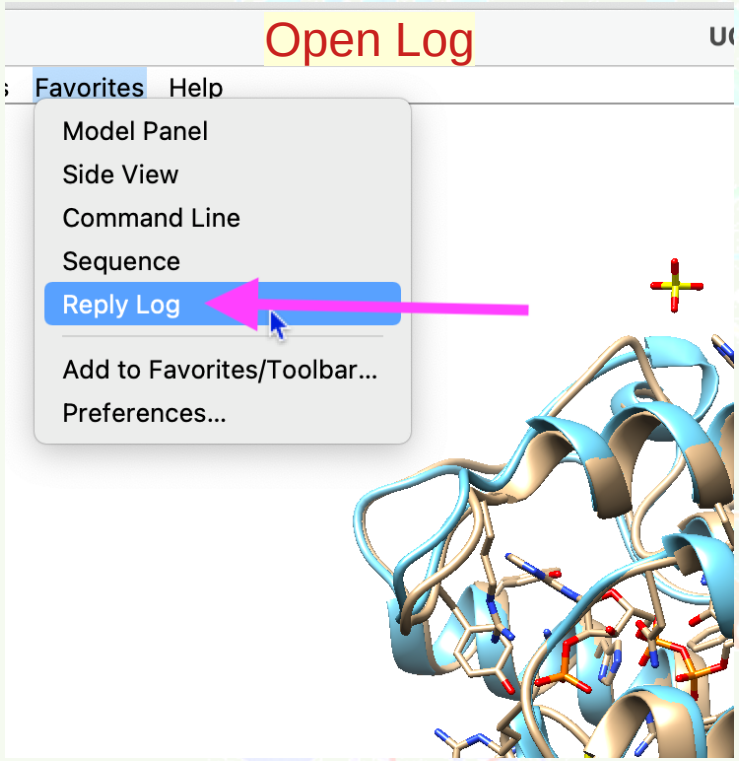


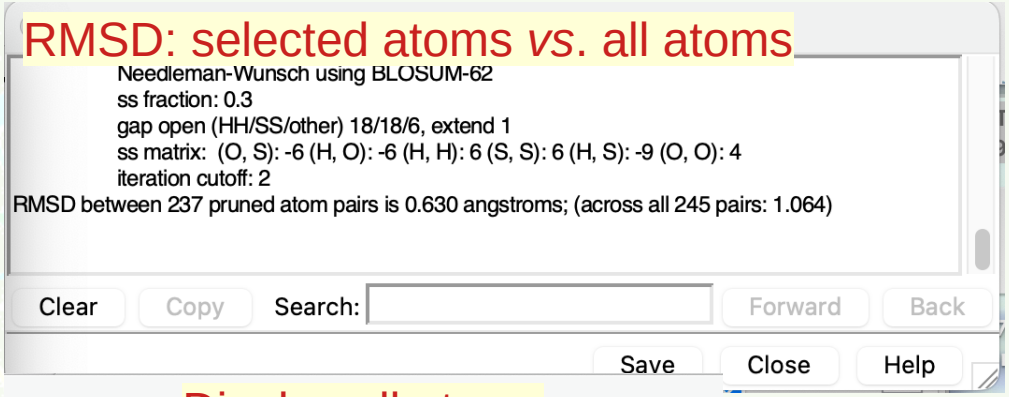
Superimpose the structures

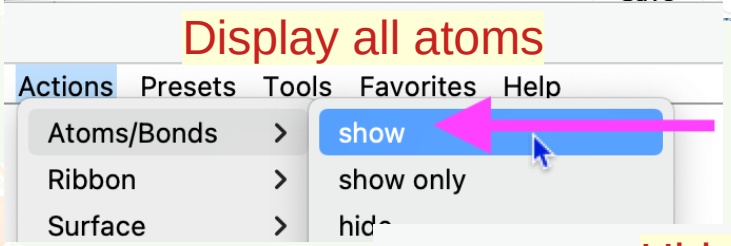


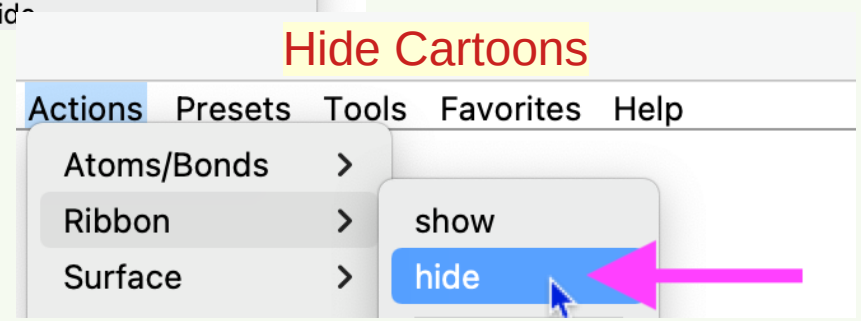


Open Log & see the RMSD value

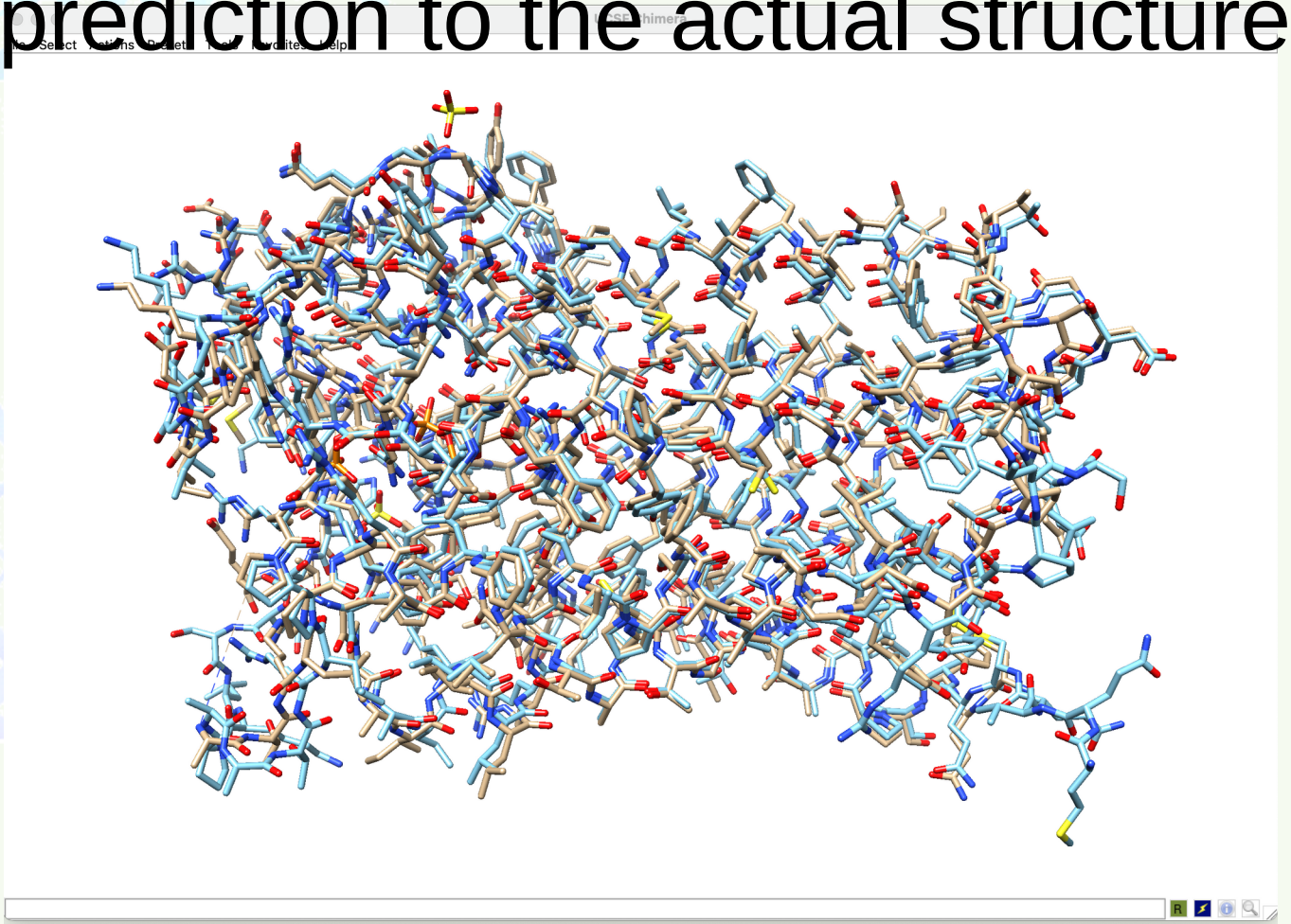


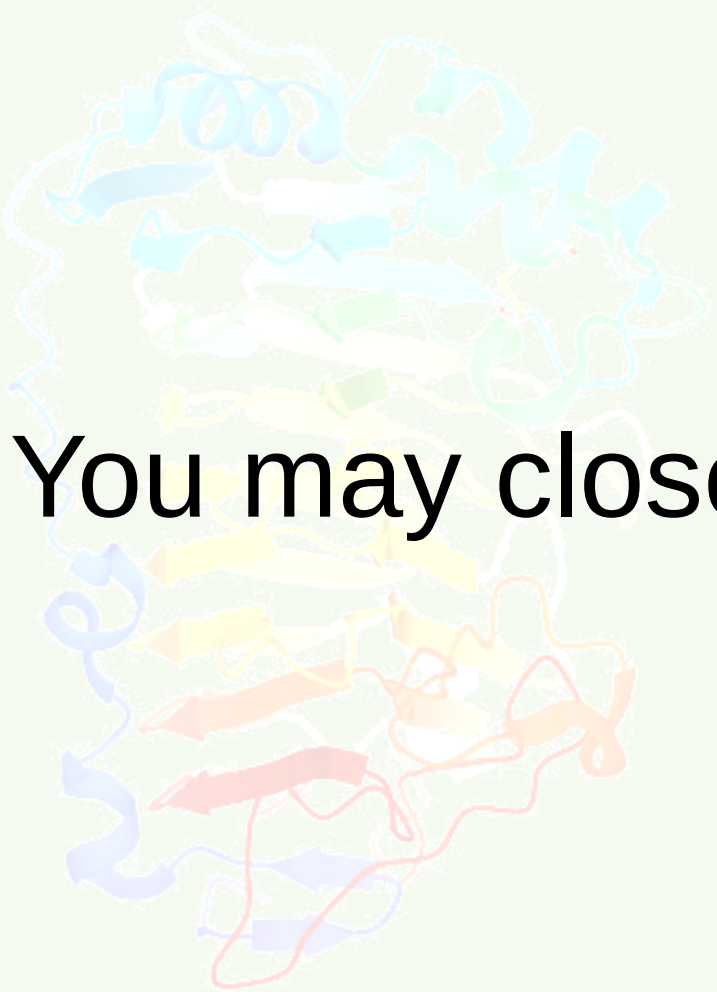






We can inspect how close is the prediction to the actual structure





You may close UCSF Chimera

Let’s check isoform Nº 1

I’m Nº

3-oxo-5-alpha-steroid 4-dehydrogenase 1

P18405 (S5A1_HUMAN)

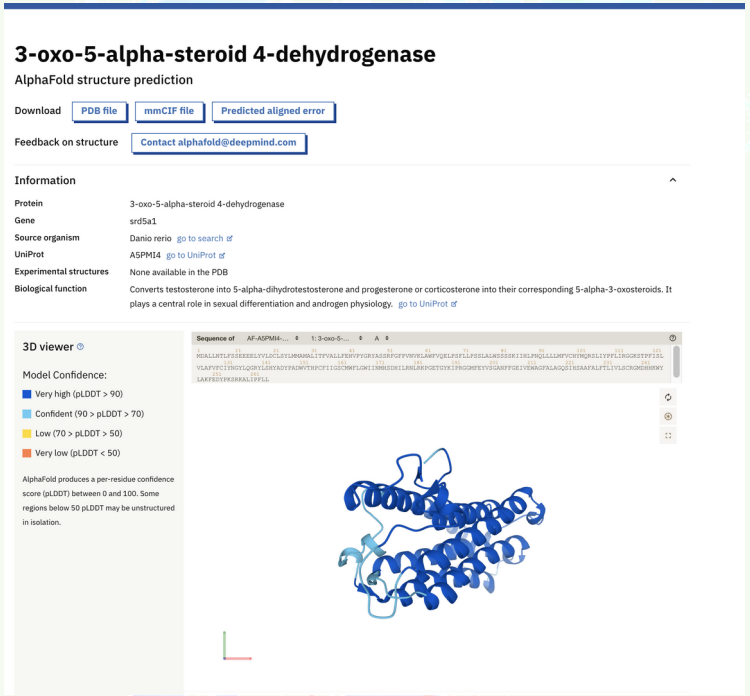
Protein	3-oxo-5-alpha-steroid 4-dehydrogenase 1
Gene	SRD5A1
Source Organism	Homo sapiens search this organism
UniProt	P18405 go to UniProt

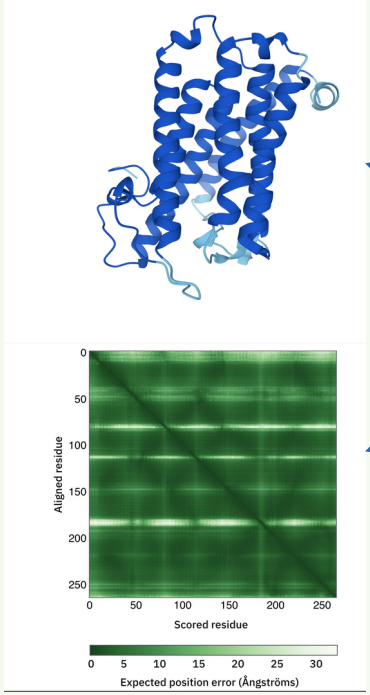
Well, isoform-1’s structure has not been solved, but Alpha Fold 2.0 has surely predicted it.

Information No PDB data in this case

Protein	3-oxo-5-alpha-steroid 4-dehydrogenase 1
Gene	SRD5A1
Source organism	Homo sapiens (Human) go to search
UniProt	P18405 go to UniProt
Experimental structures	None available in the PDB

We get this:





- This is the prediction view
- Below we have the “alignment error” plot

See what the Predicted error plot indicates

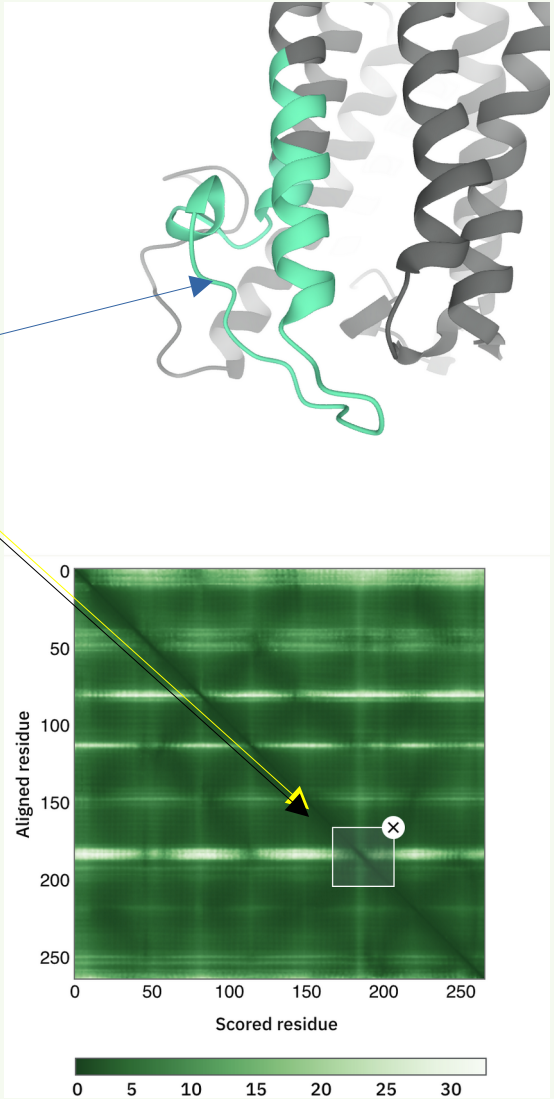
We now select a region on the error plot

- Click and drag on the diagonal of the plot below

1. The plot represents the alignment error produced when the select region is used as reference for the (structural) alignment.

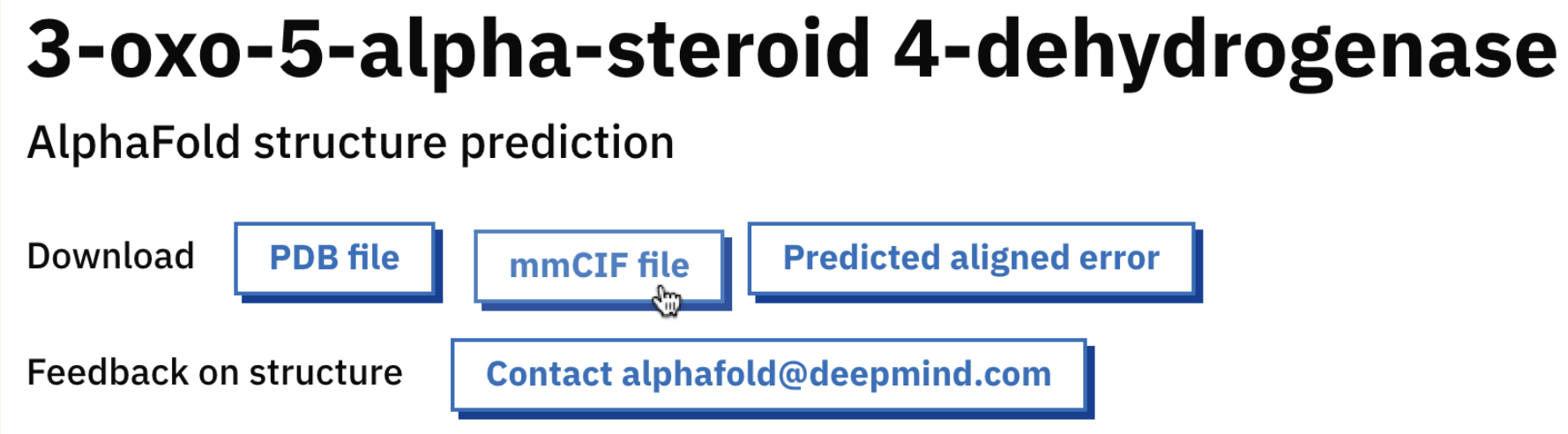
2.Here the region chosen is highlighted in green, but its central portion corresponds to a white area in the plot.

That is grossly around the loop, and is a segment with low confidence score.



Download the prediction

- use the mmCIF file or the PDB file link
- PDB format is currently more portable



AF-P18405-F1-model_v1.cif — the middle code is the Uniprot accesion code

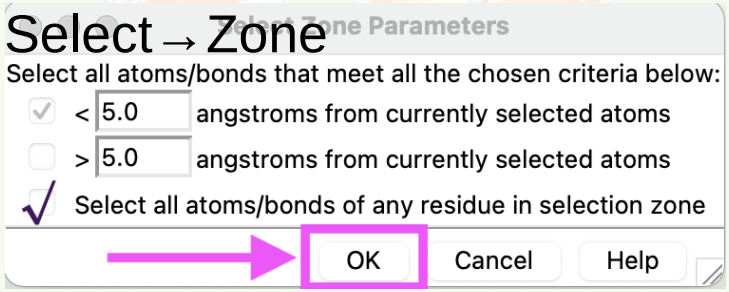
We may use Tools→Structure Comparison→MatchMaker again to compare against the experimental data for isoform 2

- file→fetch→ structure 7bw1 from PDB
- file→open (browser opens) AF-P18405-F1-model_v1.cif

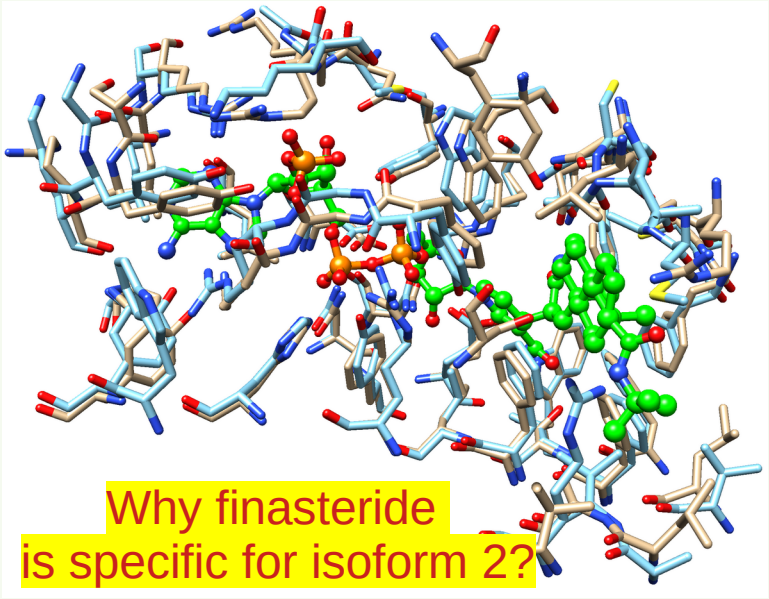
  downloaded from Alpha Fold DB
- Tools menu→Structure Comparison→MatchMaker
  - chose the Xray structure as reference
  - choose the prediction as target
  - Favorites→Replay Log
    - Check the RMSD of “best match” atoms, and “all-residues” RMSD
  - Now look at the aligned result
  - Actions→Atom/Bonds→show
  - Compare the orientation of side chains

Compare the residues around the experimental ligand

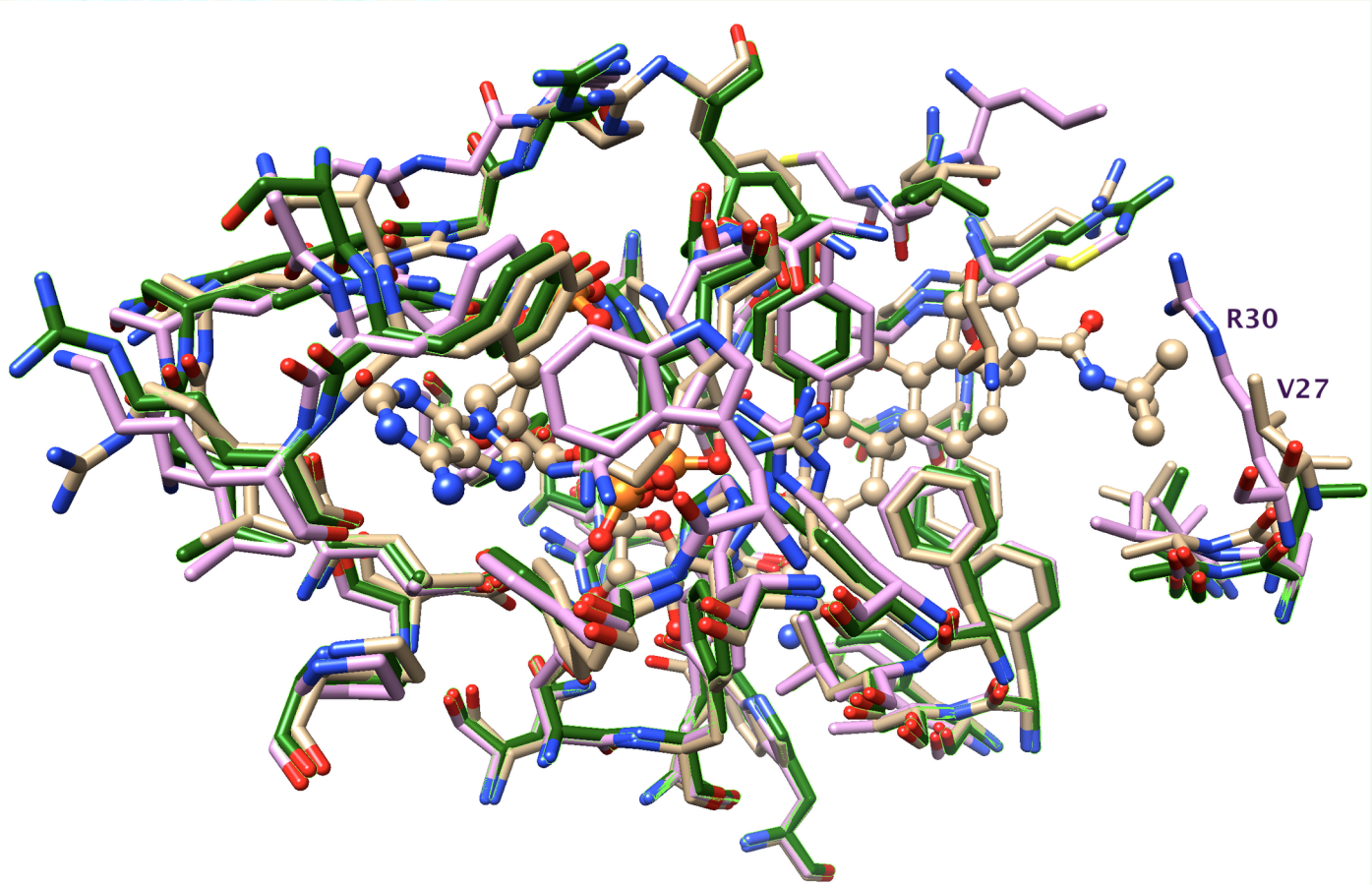
- Select → Select all
- Action → Atoms/Bonds → hide
- Action → Ribbon → hide
- Select → Clear selection
- Select → Residue → NDX
- Action → Atoms/Bonds → Ball & Stick
- Select → Zone



- Action → Atoms/Bonds → show
- Select → Clear selection



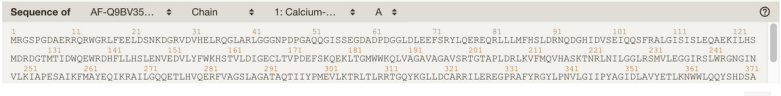
One change V → R explains the difference in specificity

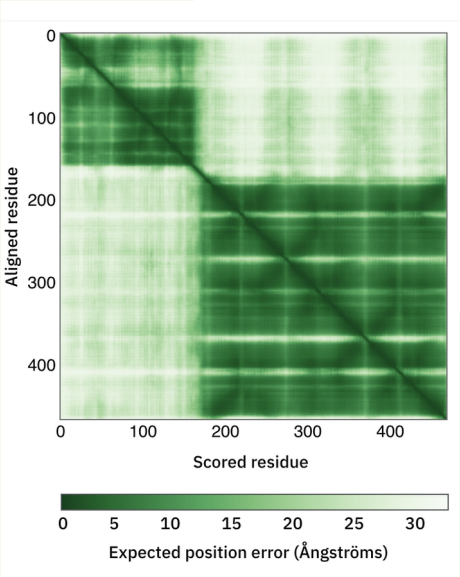


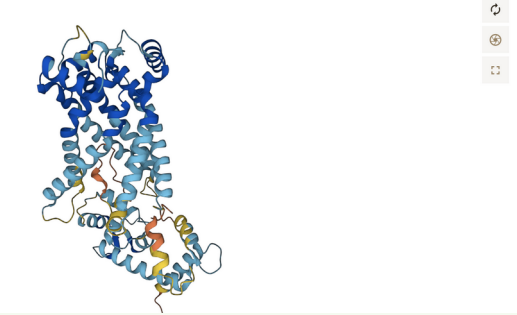
Judge the quality of a model

We will judge the quality of the model for:

- **Calcium-binding mitochondrial carrier protein SCaMC-3**
- **uniprot/Q9BV35**





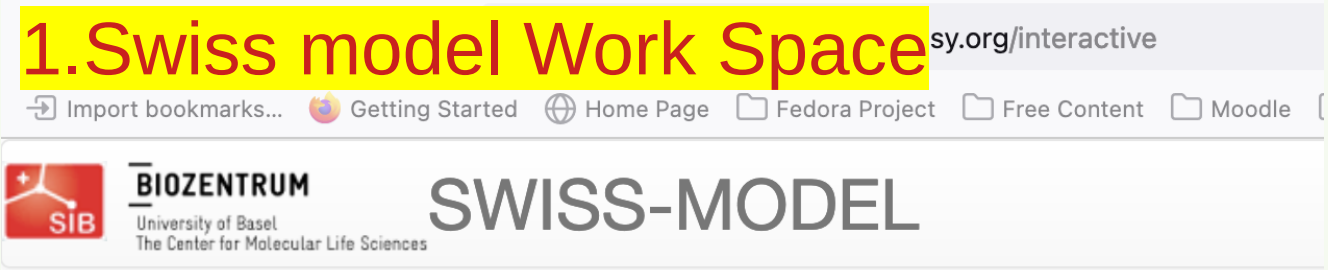


The PAE plot

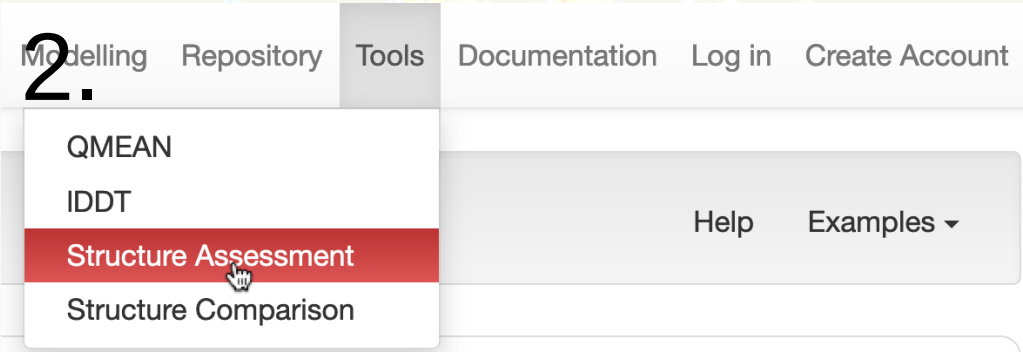
- In this case, the PAE plot has white areas between AA 1-167 & 171-468
- This means:
  - AF has some confidence in the way each domain folds.
  - But it has low confidence in the contact between the domains

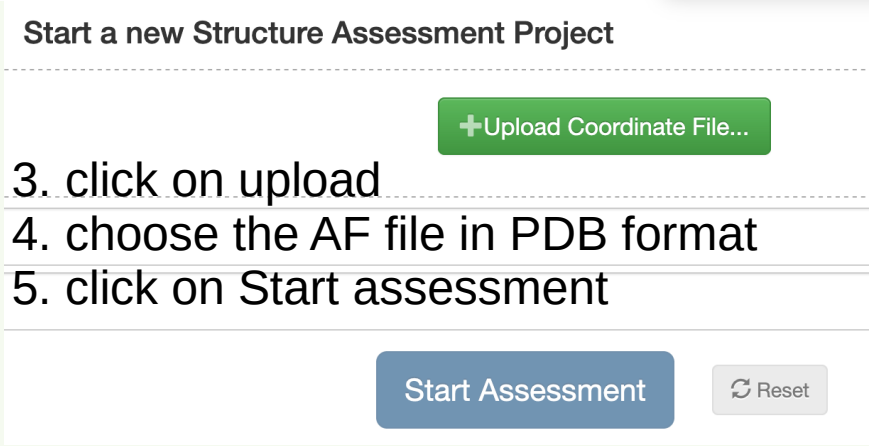
Let us visit: Swiss model work space

1.Swiss model Work Space



2.





3. click on upload
4. choose the AF file in PDB format
5. click on Start assessment

Upload your file and check the results

- Check the Ramachandra plot
  - how many residues lay in optimal (green) regions?
  - Are there residues in “disallowed” (white) areas?
- Is the protein a membrane protein?
- Review the QMeanDisCo global score and the local plot
  - Check the local quality (blue is good, red is bad)
- Are there any suspicious regions?
- Is the model correct?
- Is the model well refined?

You may get more data at UCLA SAVES

- Errat: Statistics of non-bonded interactions by a comparison with statistics from highly refined structures.
- PROVE: Volumes of atoms like hard spheres and calculates a statistical Z-score deviation from PDB-deposited structures.
- WHATCHECK: Extensive checking of many sterochemical parameters on model's residues
- PROCHECK: Quality by analyzing residue-by-residue geometry and overall structure geometry.
- Verify3D: Compatibility of the atomic model (3D) with its own amino acid sequence (1D) and comparing the results to good structures.