

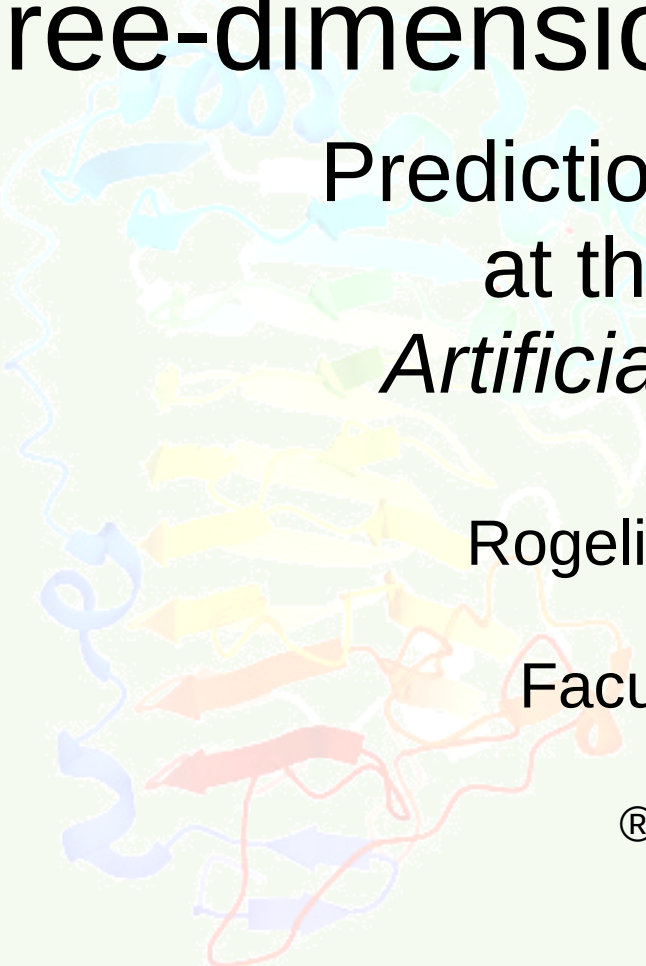
Three-dimensional protein structure

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

Rogelio Rodríguez Sotres

Facultad de Química,
UNAM

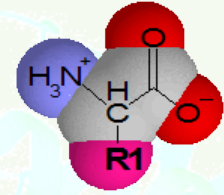
®all rights reserved



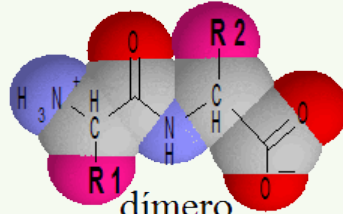
Las
proteínas
son

hetero-
polímeros
con gran
diversidad
estructural
y funcional

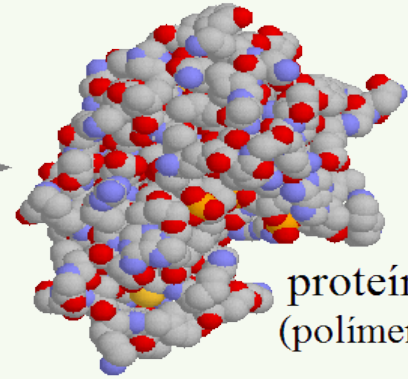
A



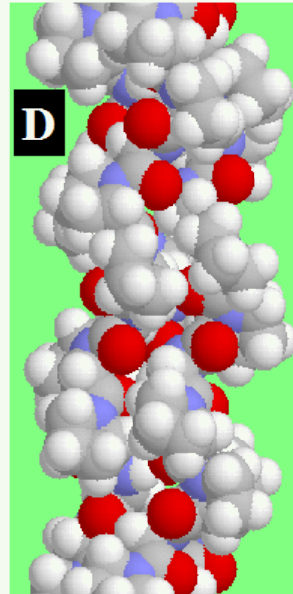
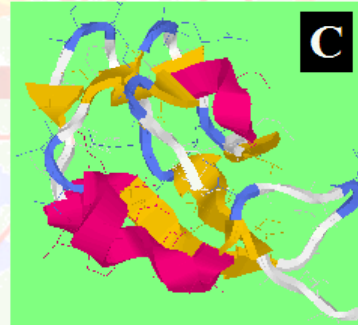
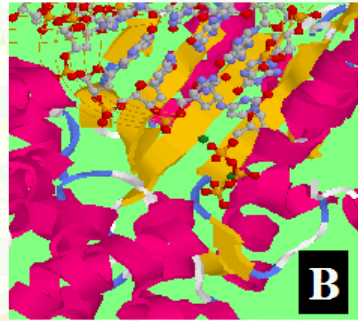
aminoácido
(monómero)



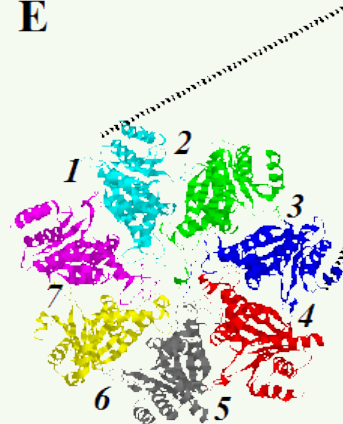
dímero
(unión de 2
monómeros)



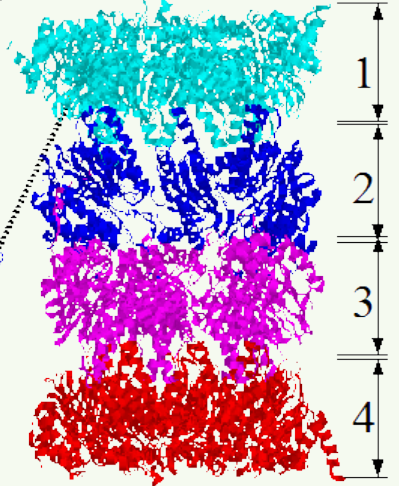
proteína
(polímero)



E



vista superior



vista lateral

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.**

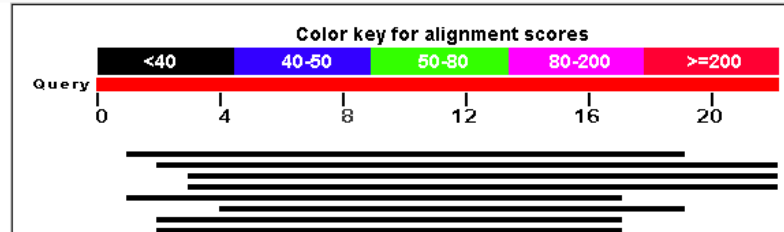
*Useful fact: An average protein has 345 aa for a MW of 38 KDa.

Example: BLAST some weird sequence:

A
VER
SI
ESTA
VEZ
LE
ATINARE

Distribution of 10 Blast Hits on the Query Sequence

Mouse-over to show define and scores, click to show alignments



```
Query 2   VERSI--ESTAVEZ----LEATIN 19
          VERSI E+T+VE+   LE TI+
Sbjct 184 VERSIVGEATSVEDGKQRLEDTID 207
```

Distance tree of res:

Sequences producing :

ref ZP_02014872.1 	initiation factor 2B related [Halorubrum l...	34.1	0.81
ref ZP_02883359.1 	methyl-accepting chemotaxis sensory transd...	33.7	1.1
ref NP_395816.1 	hypothetical protein VNG6320C [Halobacterium...	32.5	2.6
ref YP_702220.1 	sensor kinase, two-component system [Rhodoco...	31.2	6.4
ref XP_365972.1 	hypothetical protein MGG_10192 [Magnaporthe ...	31.2	6.4
ref XP_001606577.1 	PREDICTED: similar to oxidoreductase [Nas...	30.8	8.6
ref XP_001194598.1 	PREDICTED: similar to cadherin 23 [Strong...	30.8	8.6
ref XP_798827.2 	PREDICTED: similar to cadherin 23, partial [...	30.8	8.6
ref NP_044873.1 	hypothetical protein MuHV4gp34 [Murid herpes...	30.8	8.6
ref ZP_02161273.1 	thioredoxin (trxA) [Kordia algicida OT-1]	30.3	11

G
G
U
G
G
G
G
G
G

Alignments

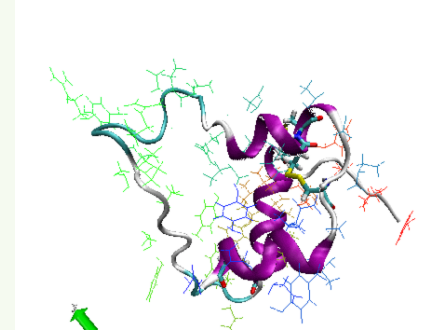
```
>|ref|ZP_02014872.1| initiation factor 2B related [Halorubrum lacusprofundi ATCC 49239]
Length=394
```

```
Score = 34.1 bits (73), Expect = 0.81
Identities = 13/24 (54%), Positives = 17/24 (70%), Gaps = 6/24 (25%)
```

```
Query 2   VERSI--ESTAVEZ----LEATIN 19
          VERSI E+T+VE+   LE TI+
Sbjct 184 VERSIVGEATSVEDGKQRLEDTID 207
```


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](https://doi.org/10.1073/pnas.89.1.20)

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

Highly accurate protein structure prediction with AlphaFold

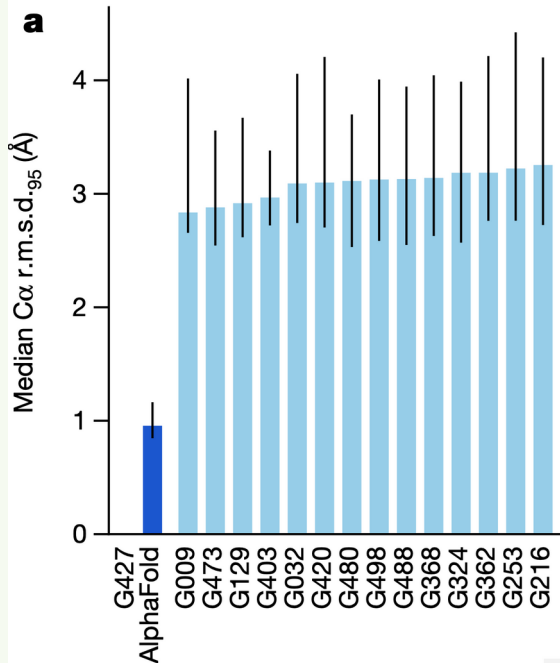


Jumper J , Evans R , Pritzel A , Green T , Figurnov M , Ronneberger O , Tunyasuvunakool K , Bates R , Žídek A , Potapenko A , Bridgland A , Meyer C , Kohl SAA , Ballard AJ , Cowie A , Romera-Paredes B , Nikolov S , Jain R , Adler J , Back T , Petersen S , Reiman D , Clancy E , Zielinski M , Steinegger 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

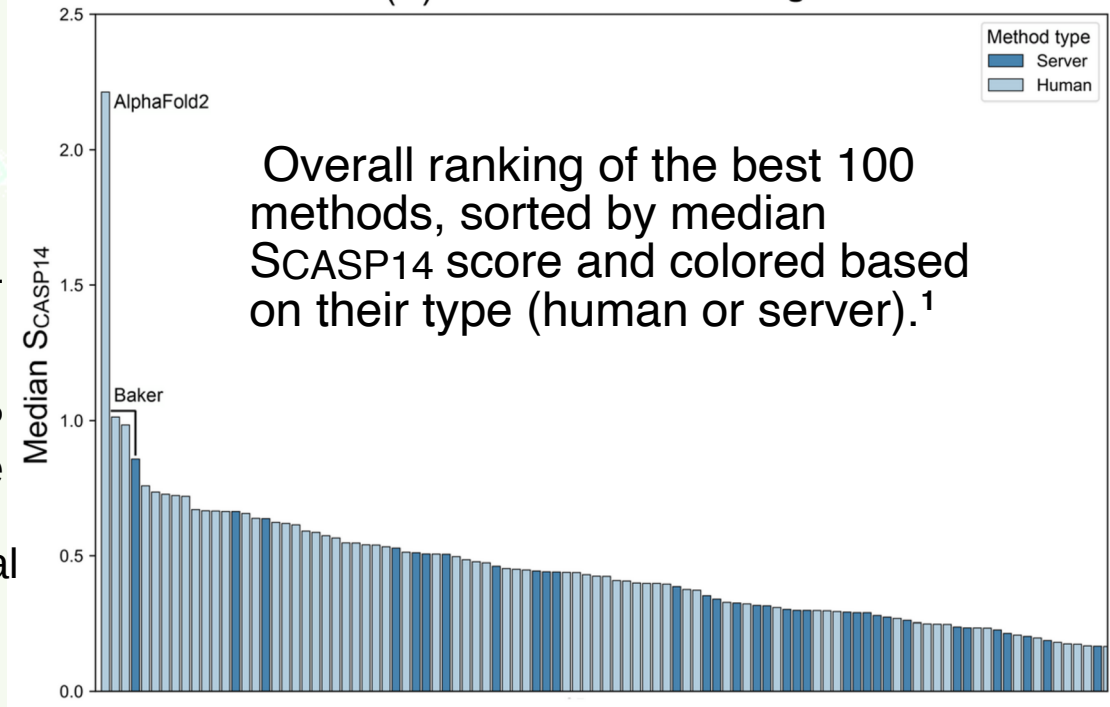
IN 2021 SOMETHING REMARKABLE HAPPENED:

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 \pm 95% confidence interval (10,000 bootstrap samples).²



¹Pereira, J. et al. Proteins (2021), DOI [10.1002/prot.26171](https://doi.org/10.1002/prot.26171)

²Jumper, J. et al. Nature (2021) 596: 583—589, DOI [10.1038/s41586-021-03819-2](https://doi.org/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

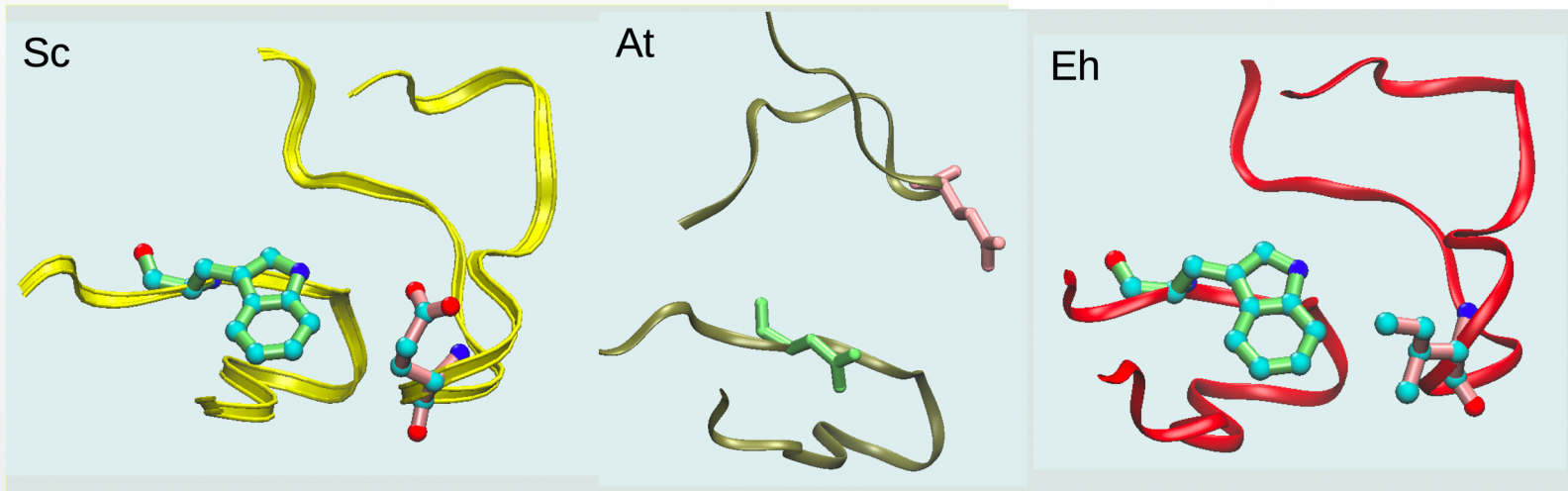
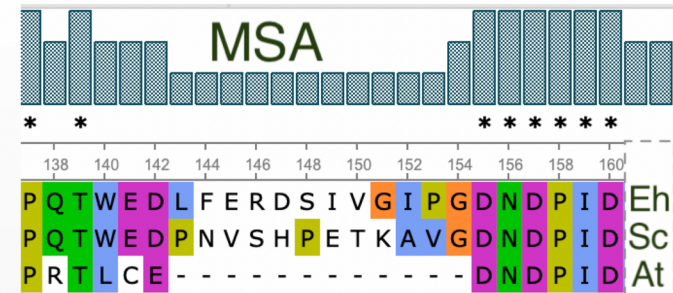


Figure 1. Example of change in contact constraints for inorganic pyrophosphatases from *Enthamoeba histolytica*, *Saccharomyces cereviciae* and *Arabidopsis thaliana*.



Deep Mind promise[‡]

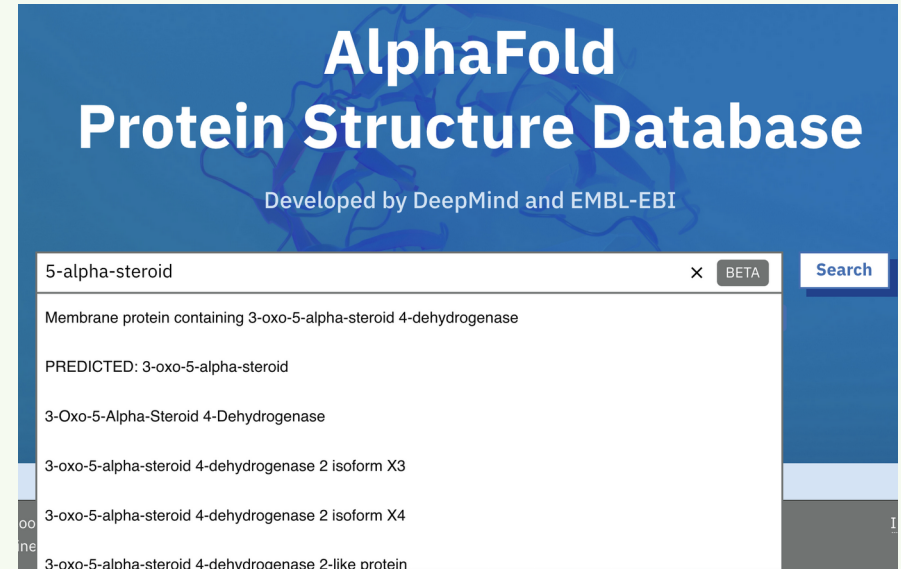
- Deep Mind promised to predict the nearly complete proteomes of all organisms with sequenced genomes at:
- UNIREF genomic data base
- As for now, prediction for the entire UniprotKB database are already available:

  <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

Developed by DeepMind and EMBL-EBI

5-alpha-steroid × BETA Search

Membrane protein containing 3-oxo-5-alpha-steroid 4-dehydrogenase

PREDICTED: 3-oxo-5-alpha-steroid

3-Oxo-5-Alpha-Steroid 4-Dehydrogenase

<input type="checkbox"/> Bacteroides uniformis (6)	3-oxo-5-alpha-steroid 4-dehydrogenase A0A3B4FU30 (A0A3B4FU30_9CICH) Protein 3-oxo-5-alpha-steroid 4-dehydrogenase Gene Unknown Source Organism Pundamilia nyererei search this organism ↗ UniProt A0A3B4FU30 go to UniProt ↗
<input type="checkbox"/> Cottoperca gobio (6)	
<input type="checkbox"/> Halieaceae bacterium (6)	
<input type="checkbox"/> Holophagales bacterium (6)	
<input checked="" type="checkbox"/> Homo sapiens (6)	
<input type="checkbox"/> Mycobacteroides franklinii (6)	
<input type="checkbox"/> Myripristis murdjan (pinecone soldierfish) (6)	

Chose the isoform N° 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

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

AlphaFold structure prediction

Download

PDB file

mmCIF file

Predicted aligned error

Note: We have recently **Download the PDB file** on of the updated format.

Protein 3-oxo-5-alpha-steroid 4-dehydrogenase 2
Gene SRD5A2
Source organism Homo sapiens (Human) [go to search](#)
UniProt P31213 [go to UniProt](#)
Experimental structures 1 structure in PDB for P31213 [go to PDB-KB](#)
Biological function Converts testosterone (T) into 5-alpha-dihydrotestosterone (DHT) and progesterone or corticosterone into their corresponding 5-alpha-3-oxosteroids. It plays a central role in sexual differentiation and androgen physiology. [go to UniProt](#)

3D viewer

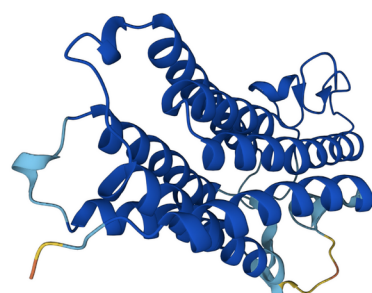
Model Confidence:

- Very high (pLDDT > 90)
- Confident (90 > pLDDT > 70)
- Low (70 > pLDDT > 50)
- Very low (pLDDT < 50)

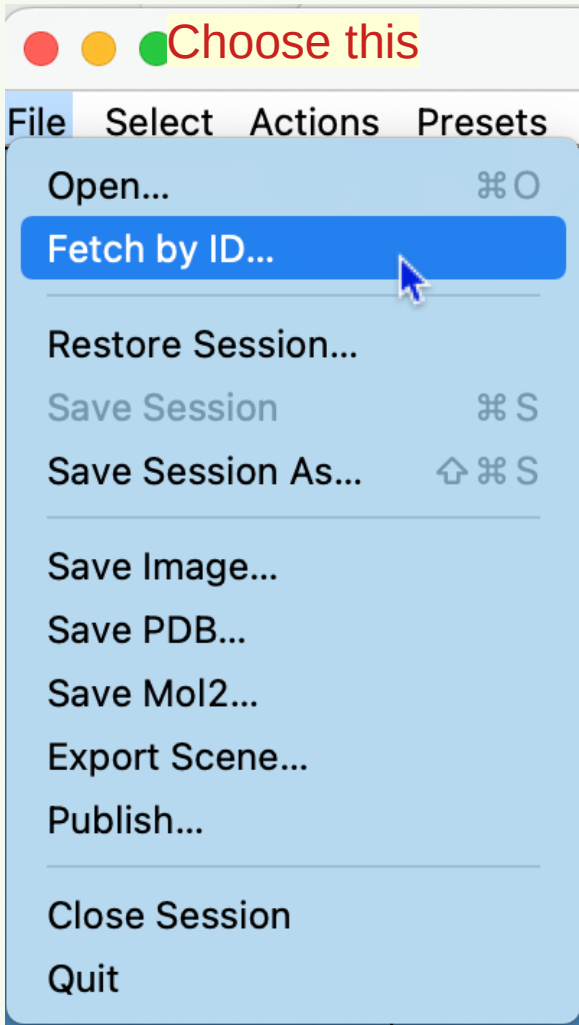
AlphaFold produces a per-residue confidence score (pLDDT) between 0 and 100. Some regions below 50 pLDDT may be unstructured in isolation.

Sequence of AF-P31213-... Chain 1: 3-oxo-5-... A

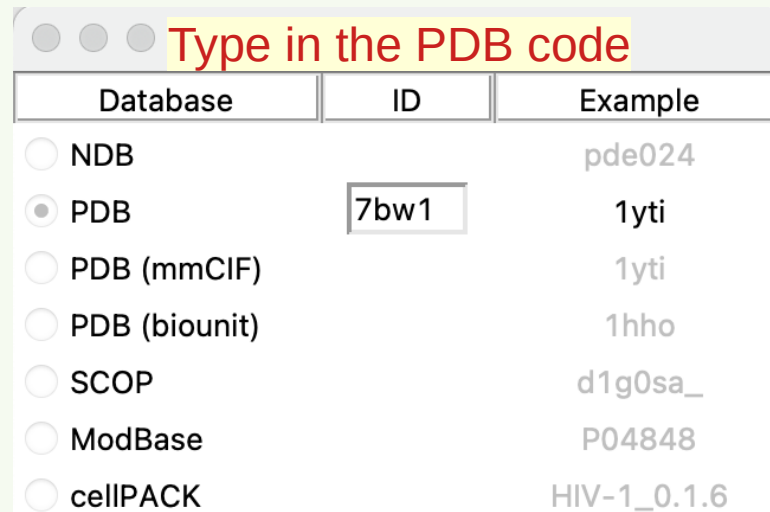
```
1 11 21 31 41 51 61 71 81 91 101 111
MQVCCQSPVLGASATLVALGALALYVAKPSGYGRHTESLKPATRLPARAAWFLQELPFAVPAGILARQLSLFPGPSTVLLGLFCVHYFHRFTVYSLNNGREYPAILLRGTAF
101 111 121 131 141 151 161 171 181 191 201 211 221 231 241
LOGYLLYCAEYFDGWTDIRFSLGVELFLLGMSINIRSDYILLRQLRKEGISYRIPQGGGLFTYVSGANFLGEIIEWIGYALATNSLPALAFAPFSLGLRAFHHRFYLKMFEDYI
251
LIPFIF
```



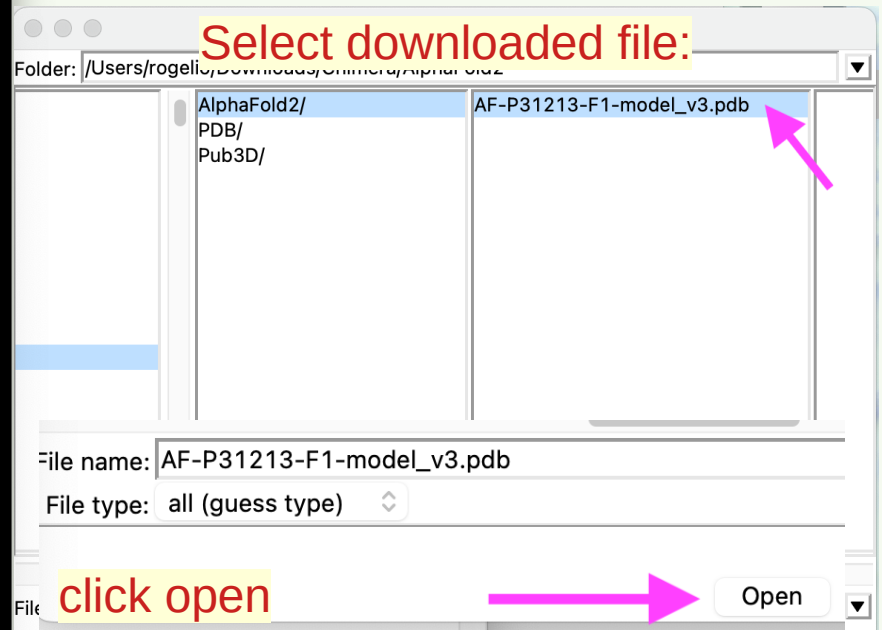
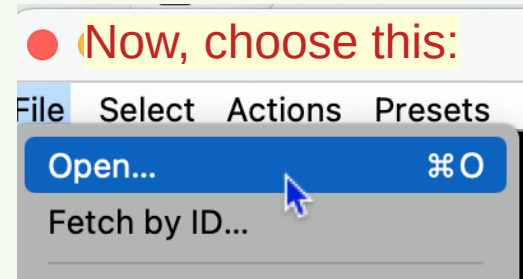
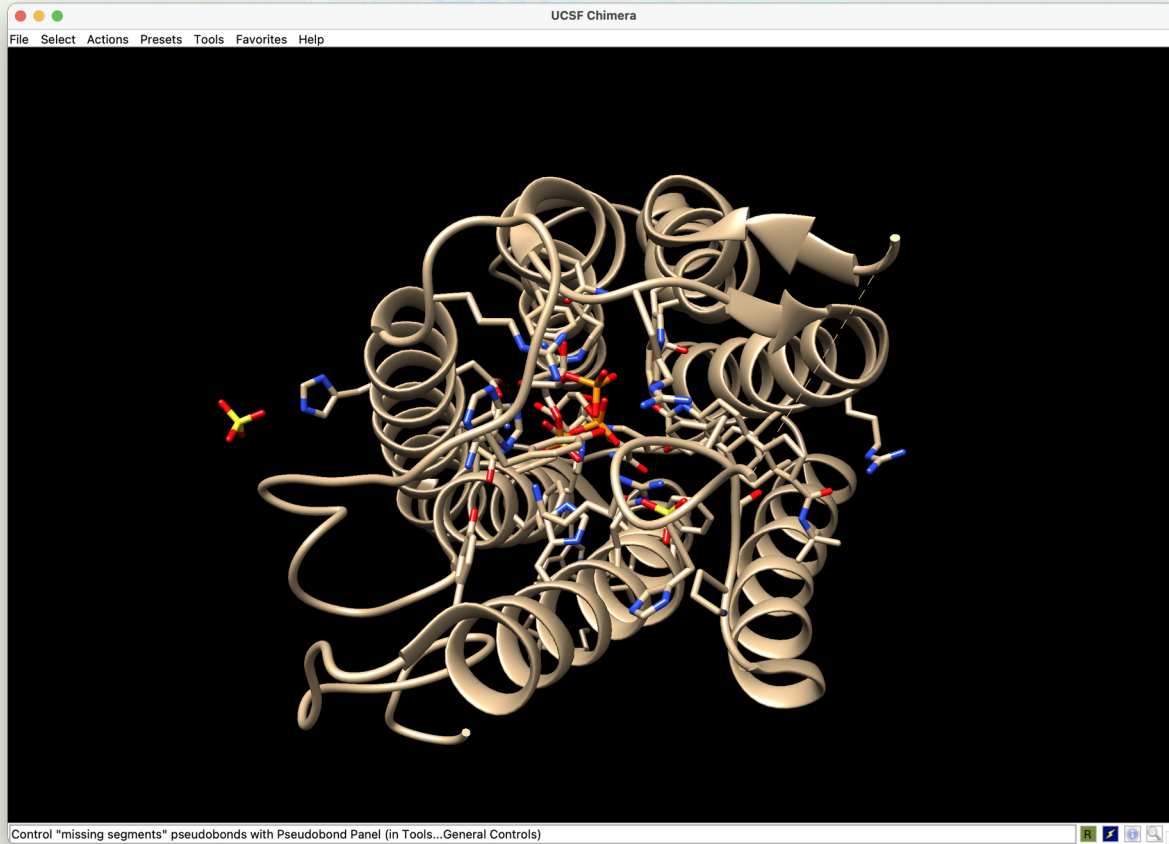
and there is an experimental structure of this protein, linked here



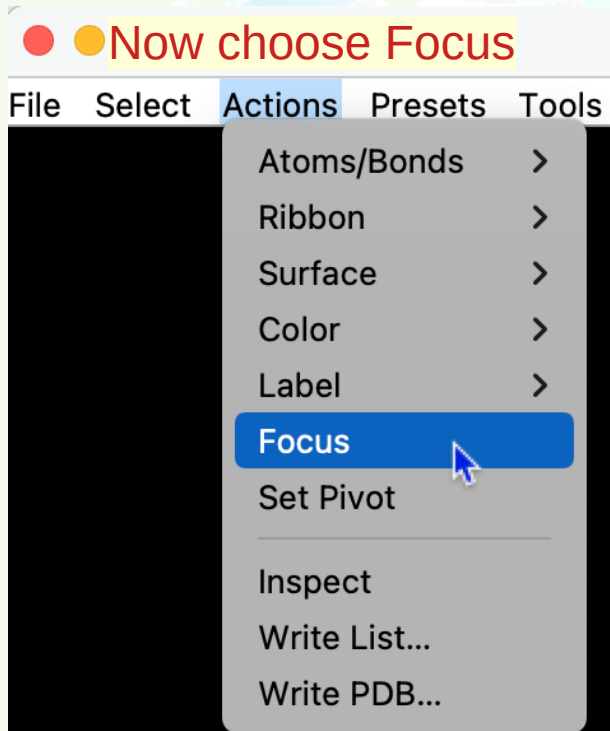
Open Chimera



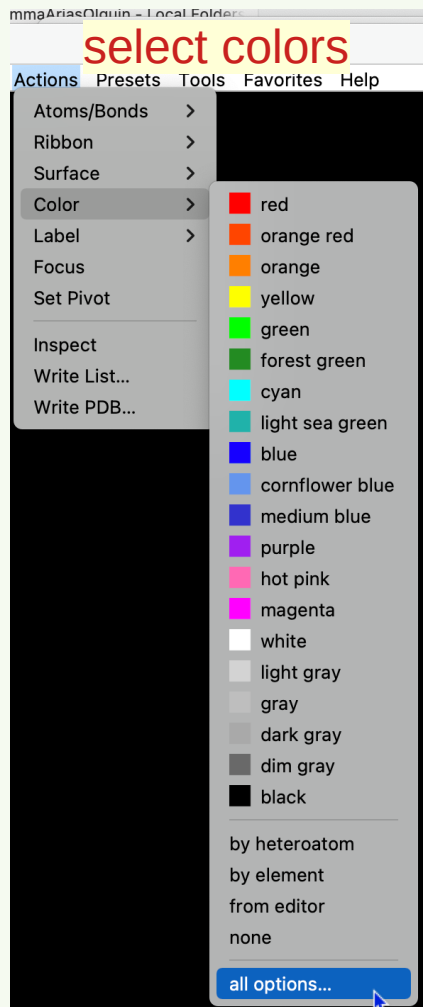
We get this image



Get loaded the model also



Make some decorations



Superimpose the structures

Choose Match Maker

Tools Favorites Help

- General Controls >
- Viewing Controls >
- Depiction >
- Structure Analysis >
- Structure Comparison >
 - MatchMaker**
 - Match -> Align
 - Morph Conformations
 - RR Distance Maps
 - Ensemble Cluster
 - Ensemble Match
 - Tile Structures
 - Minrms Plot
- Sequence >
- Surface/Binding Analysis >
- Structure Editing >
- Amber >
- MD/Ensemble Analysis >
- Higher-Order Structure >
- Volume Data >
- Demos >
- Movement >
- Utilities >
- Additional Tools >

MatchMaker

Reference structure: 1
7bw1 (#0)
AF-P31213-F1-model_v3.pdb (#1)

Structure(s) to match:
7bw1 (#0)
AF-P31213-F1-model_v3.pdb (#1)

Further restrict matching to current selection

Chain pairing

- Best-aligning pair of chains between reference and match structure
- Specific chain in reference structure with best-aligning chain in match structure
- Specific chain(s) in reference structure with specific chain(s) in match structure

Alignment algorithm: Needleman-Wunsch Matrix: BLOSUM-62

Gap opening penalty 12 Gap extension penalty 1

Include secondary structure score (30%) Show parameters

Compute secondary structure assignments

Show pairwise alignment(s)

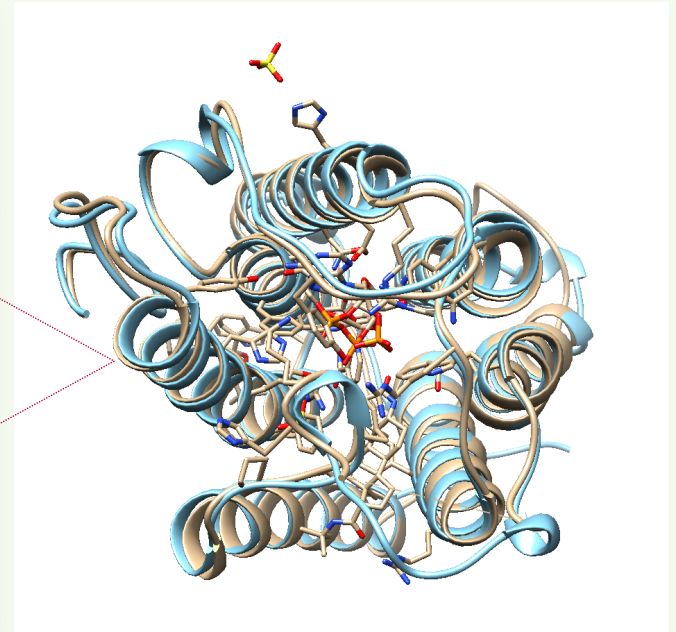
Matching

Iterate by pruning long atom pairs until no pair exceeds: 2.0 angstroms

After superposition, compute structure-based multiple sequence alignment

Save settings Reset to defaults

4 OK Apply Cancel Help



Open Log & see the RMSD value

Open Log

- Model Panel
- Side View
- Command Line
- Sequence
- Reply Log**
- Add to Favorites/Toolbar...
- Preferences...

RMSD: selected atoms vs. all atoms

Needleman-Wunsch using BLOSUM-62
ss fraction: 0.3
gap open (HH/SS/other) 18/18/6, extend 1
ss matrix: (O, S): -6 (H, O): -6 (H, H): 6 (S, S): 6 (H, S): -9 (O, O): 4
iteration cutoff: 2
RMSD between 237 pruned atom pairs is 0.630 angstroms; (across all 245 pairs: 1.064)

Clear Copy Search: Forward Back Save Close Help

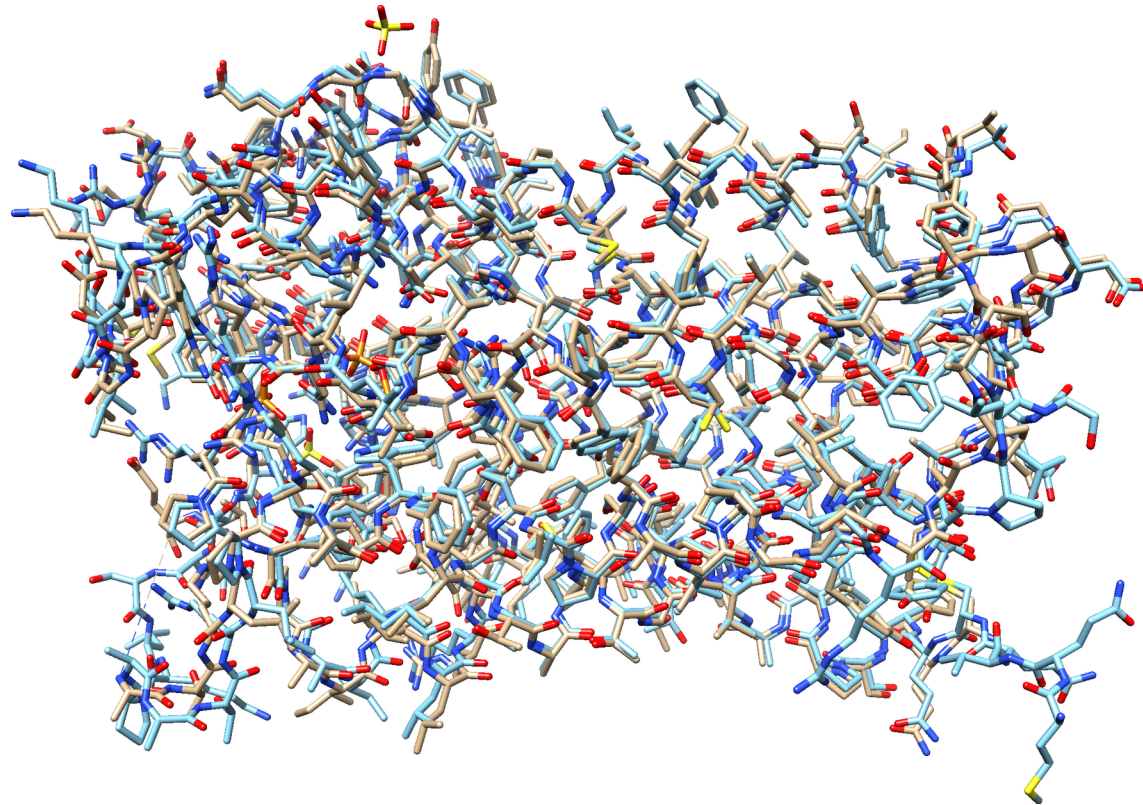
Display all atoms

- Atoms/Bonds > **show**
- Ribbon > show only
- Surface > hid-

Hide Cartoons

- Atoms/Bonds >
- Ribbon > show
- Surface > **hide**

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





You may close UCSF Chimera

Let's check isoform N° 1

I'm
N°

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:

3-oxo-5-alpha-steroid 4-dehydrogenase
AlphaFold structure prediction

Download [PDB file](#) [mmCIF file](#) [Predicted aligned error](#)

Feedback on structure Contact alphafold@deepmind.com

Information

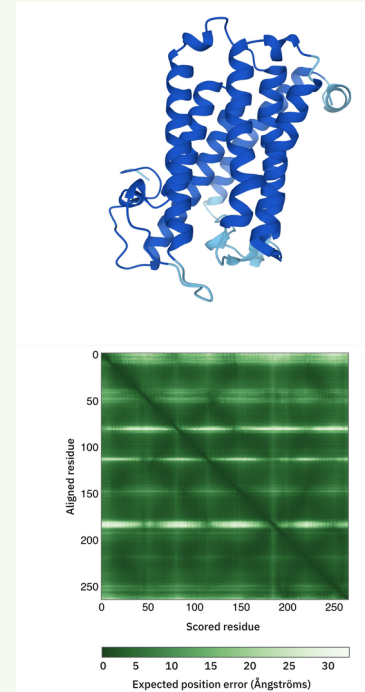
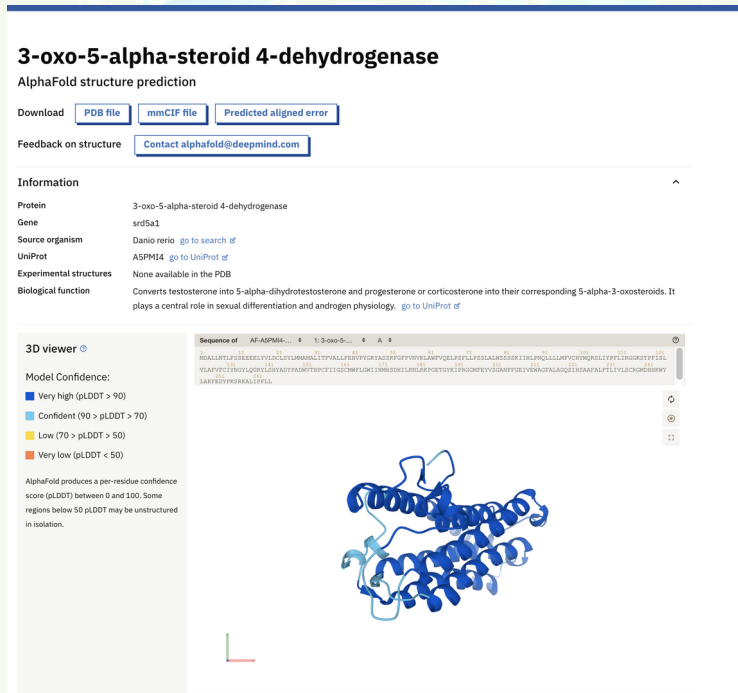
Protein 3-oxo-5-alpha-steroid 4-dehydrogenase
Gene srd5a1
Source organism [Danio rerio go to search of](#)
UniProt [ASPM14 go to UniProt of](#)
Experimental structures None available in the PDB
Biological function Converts testosterone into 5-alpha-dihydrotestosterone and progesterone or corticosterone into their corresponding 5-alpha-3-oxosteroids. It plays a central role in sexual differentiation and androgen physiology. [go to UniProt of](#)

3D viewer

Model Confidence:

- Very high (pLDDT > 90)
- Confident (90 > pLDDT > 70)
- Low (70 > pLDDT > 50)
- Very low (pLDDT < 50)

AlphaFold produces a per-residue confidence score (pLDDT) between 0 and 100. Some regions below 50 pLDDT may be unstructured in isolation.



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

See what the Predicted error plot indicates

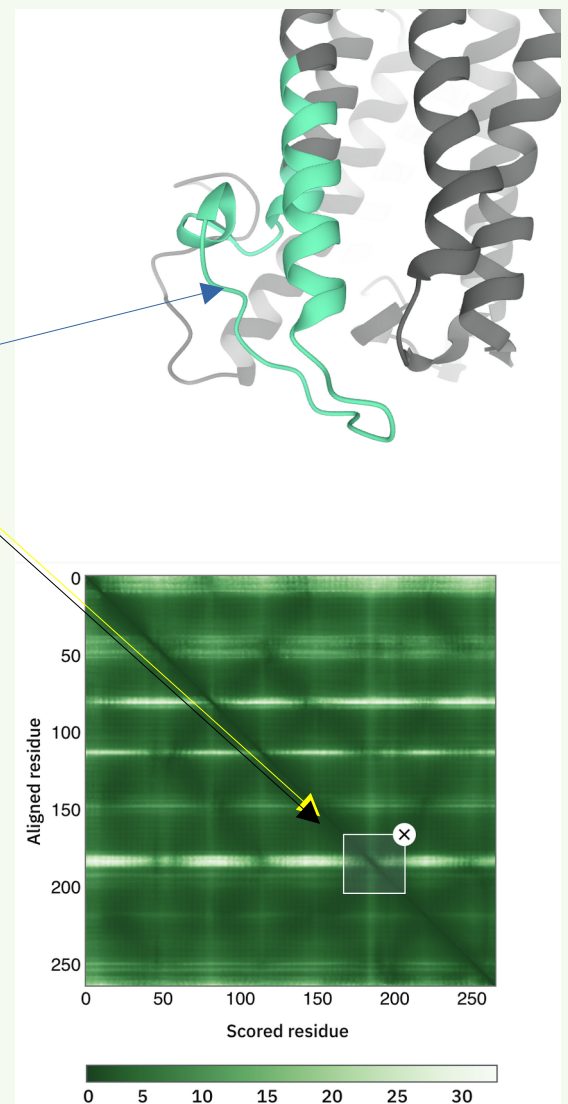
We now select a region on the error plot

- 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

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

AlphaFold structure prediction

Download

[PDB file](#)

[mmCIF file](#)

[Predicted aligned error](#)

Feedback on structure

[Contact alphafold@deepmind.com](mailto:alphafold@deepmind.com)

[AF-P18405-F1-model_v1.cif](#) — the middle code is the Uniprot accession code

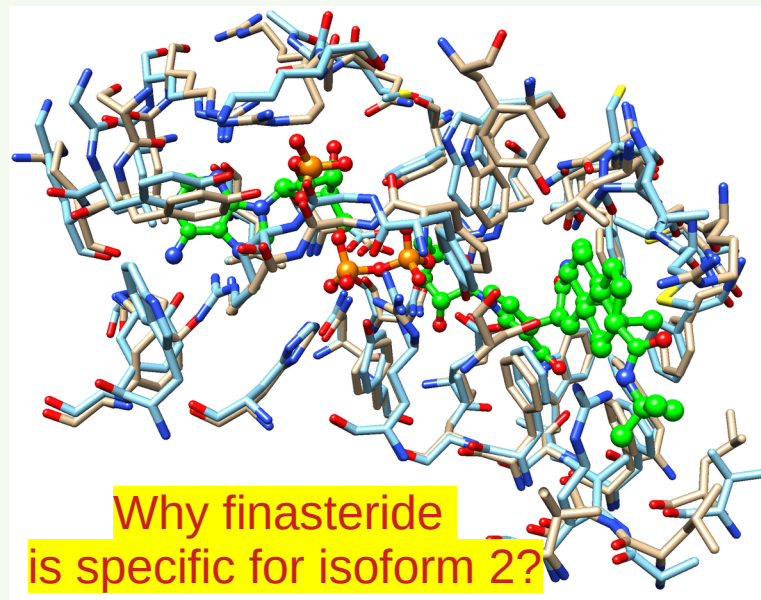
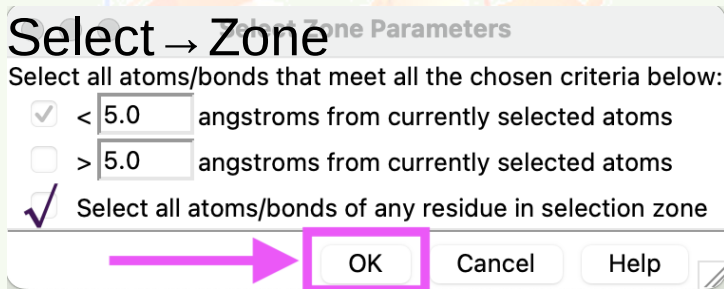
We may use Tools → Structure Comparison → MatchMaker again to compare against the experimental data for isoform 2

- 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

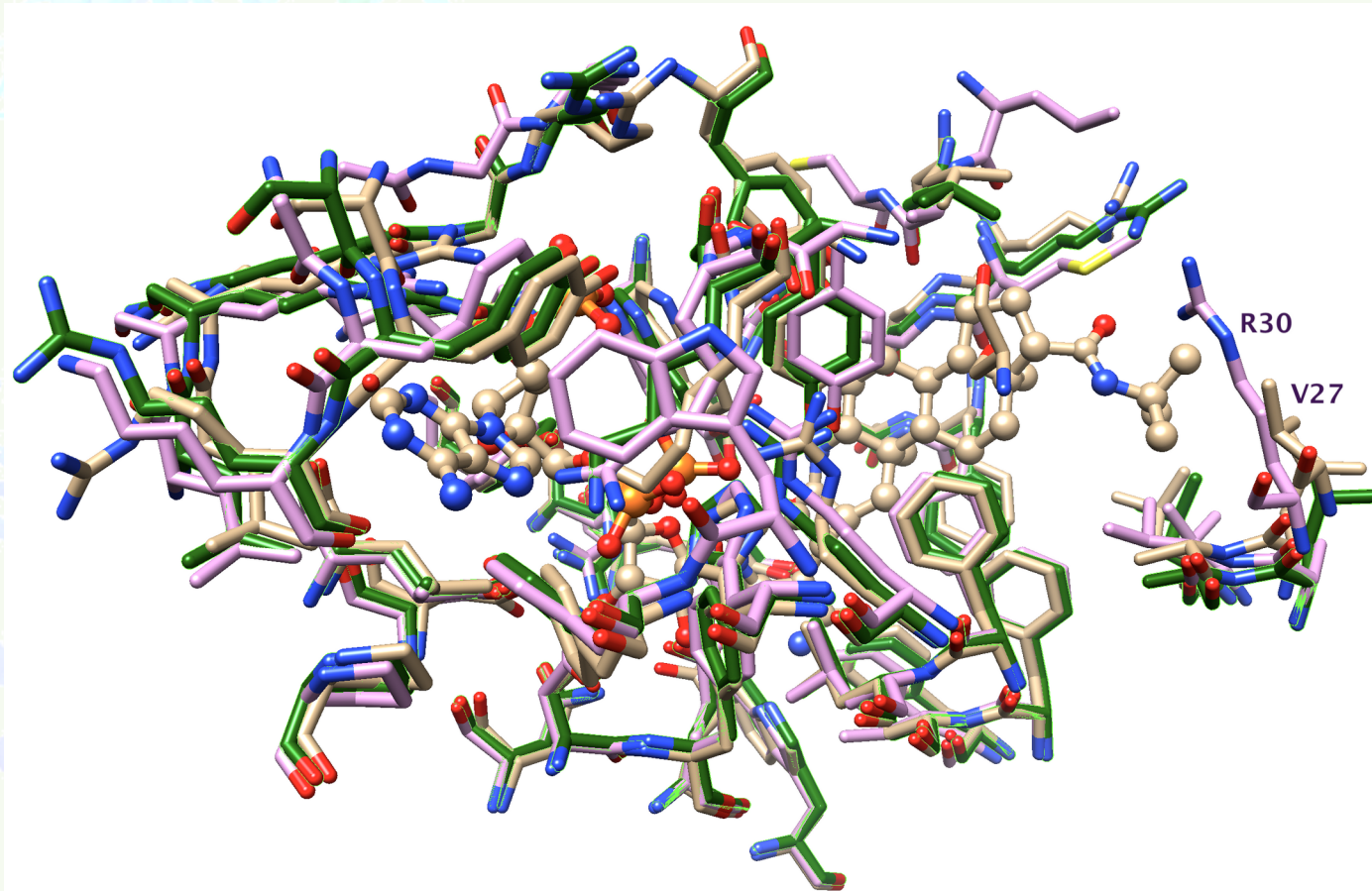
Compare the residues around the experimental ligand

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

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



One change $V \rightarrow R$ explains the difference in specificity

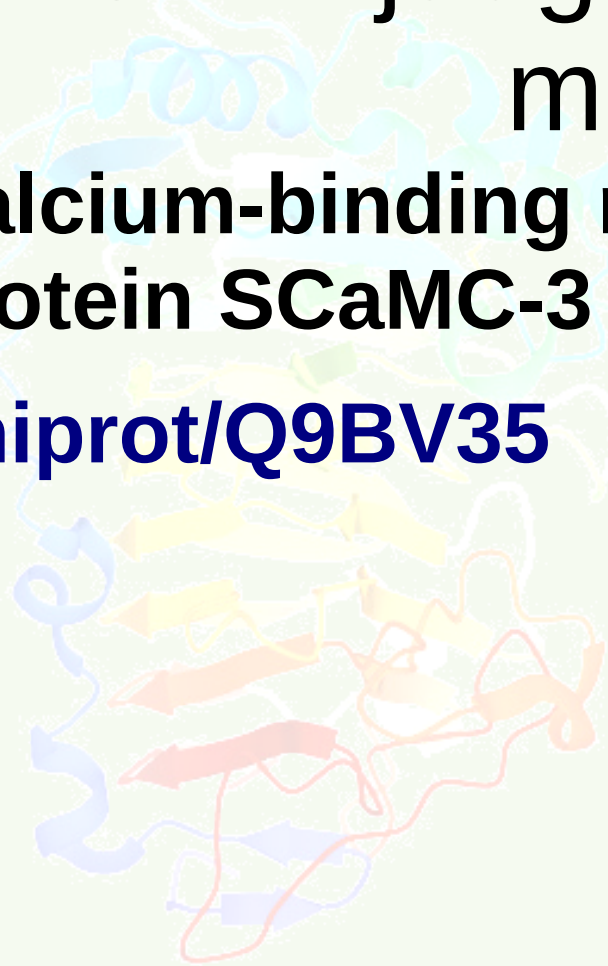




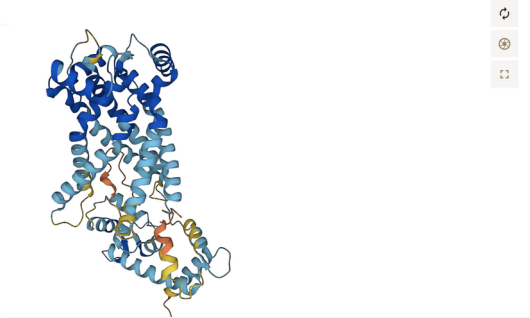
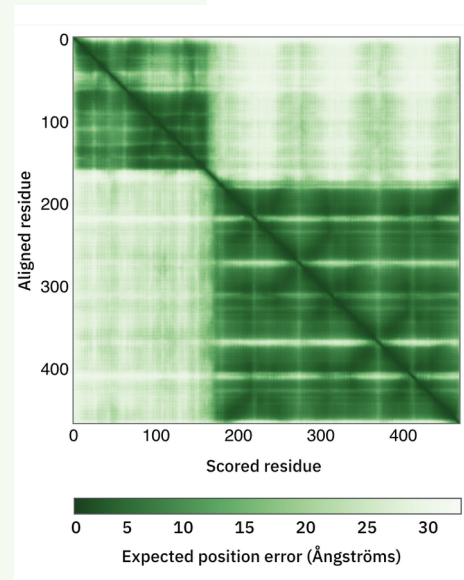
Judge the quality of a model

We will judge the quality of the model for:

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



```
Sequence of AF-Q9BV35... Chain 1: Calcium-... A
MRGSPGDAERQKHWGLPEELDSNKGKRVVHLEAQLALGSGMFPDPAQQQISSEGDAPDGGDLLEEFKRVLQEREQRLMLMFHSLDRWQDGHIDVSEYQOSFRALGISILEQAEKLIHS
131 141 151 161 171 181 191 201 211 221 231 241
MDRDCGTMITDQKWRDHFLLHSLNVEDVLYFKHSTVLDIGECLTVDFEFSKQEKLTGMWVKQLVAGAVAGVSRVTTAPLDRKVFVMVHASKNRLNLLGGLRSMVLEGGIRSLWRGNGIN
251 261 271 281 291 301 311 321 331 341 351 361
VLKIAPESAIFMAYEQIKRAILGQETLHVQERFVAGSLAGATAQTIIYPMEVLRTRTLRRTQYKGLLDCARRILEREPRAFYRGLPNVLGIIPYAGIDLAVYETLKNWVQQYSHDSA
```



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 [swiss-model.org/interactive](https://www.swiss-model.org/interactive)

Import bookmarks... Getting Started Home Page Fedora Project Free Content Moodle



2. Modelling Repository **Tools** Documentation Log in Create Account

QMEAN

IDDT

Structure Assessment

Structure Comparison

Help Examples ▾

Start a new Structure Assessment Project

+ Upload Coordinate File...

3. click on upload

4. choose the AF file in PDB format

5. click on Start assessment

Start Assessment

Reset

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 stereochemical 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.