

# Highly Accurate Protein Structure Prediction with AlphaFold

Developed by Google DeepMind and EMBL-EBI

歐任翔 Mark Ou, PhD

Sep 25, 2024



The slide is available here:

# Something About DeepMind

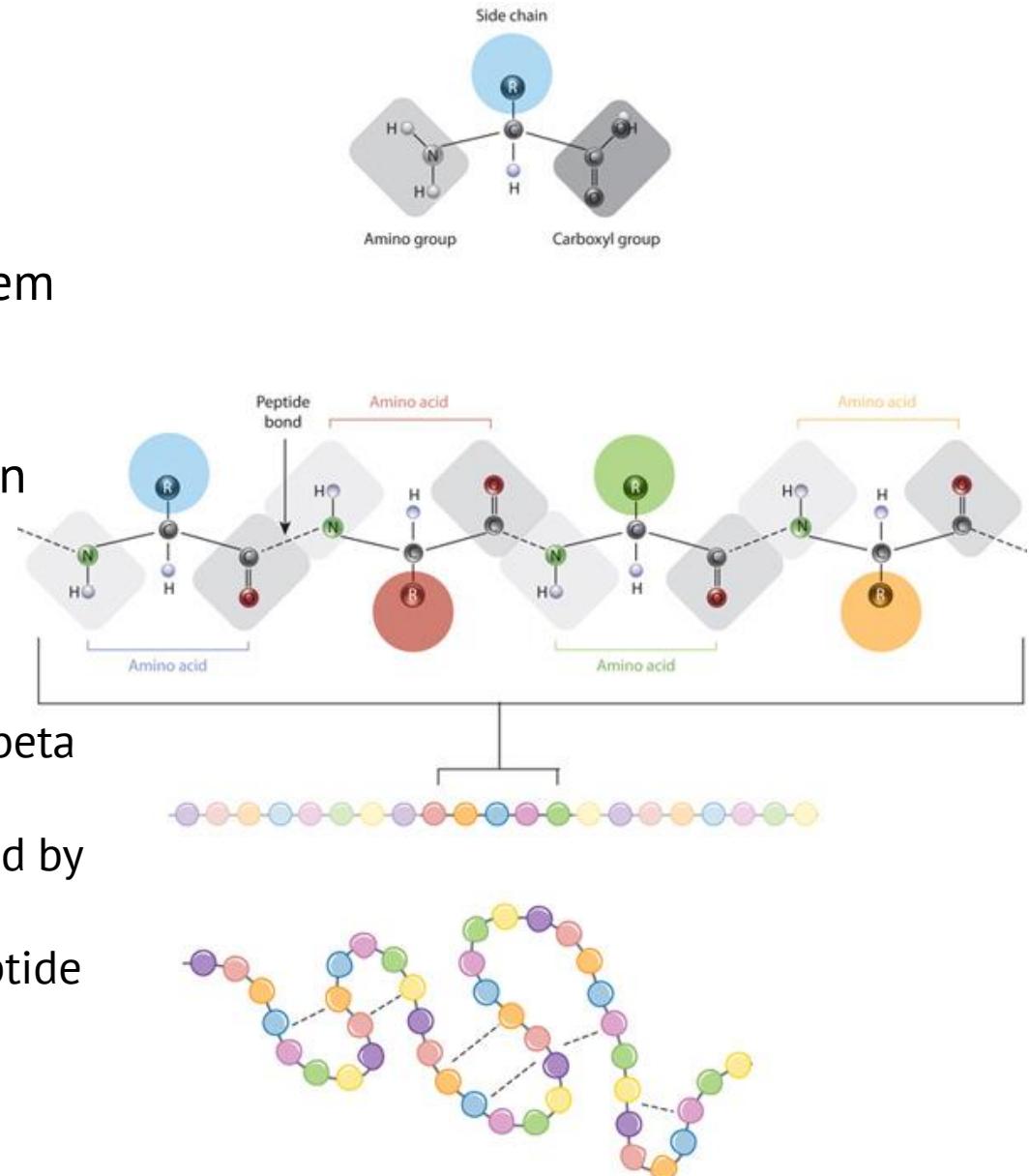
Central mission: solve fundamental problems with AI

=> Predicting the 3D structure of a protein from its amino acid sequence is one such challenge

<https://www.nature.com/articles/s41586-021-03819-2>

# What are proteins?

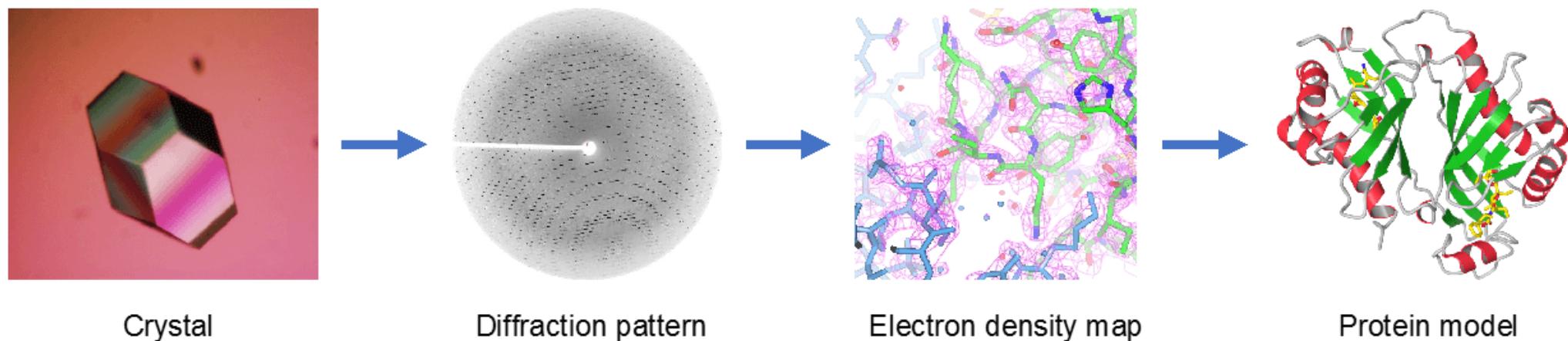
- Molecular machines which are essential to life
- Have many functions, from hair to the immune system
- Consist of *chains of amino acids* that fold into a 3D structure  
=> The 3D shape is important for a protein's function
- Protein structures
  - **Primary structure:** Linear sequence of amino acids.
  - **Secondary structure:** Patterns like alpha helices and beta sheets formed by hydrogen bonds.
  - **Tertiary structure:** The 3D shape of the protein formed by side-chain interactions.
  - **Quaternary structure:** Complexes of multiple polypeptide chains or subunits.



# Why and how to predict protein structure?

- Experimental structure determination takes months to years
- Structure prediction can provide actionable information faster

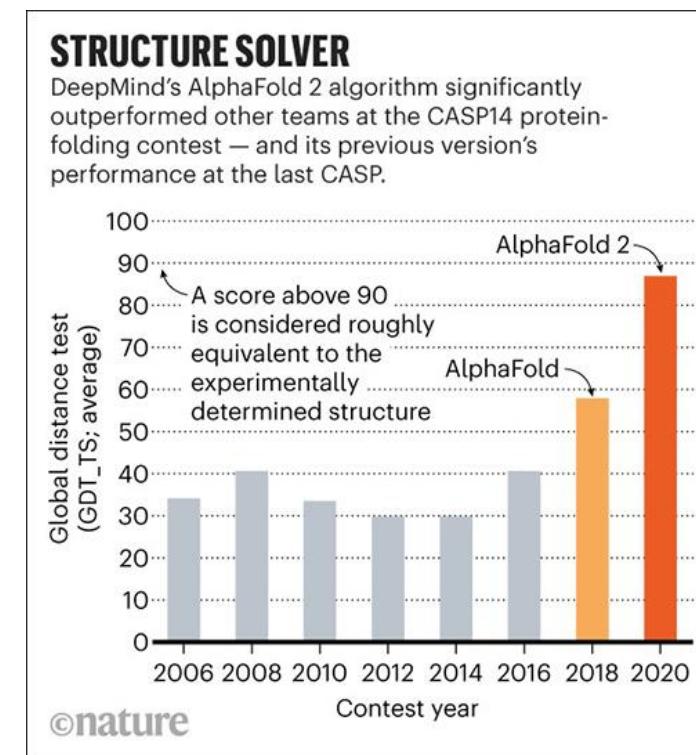
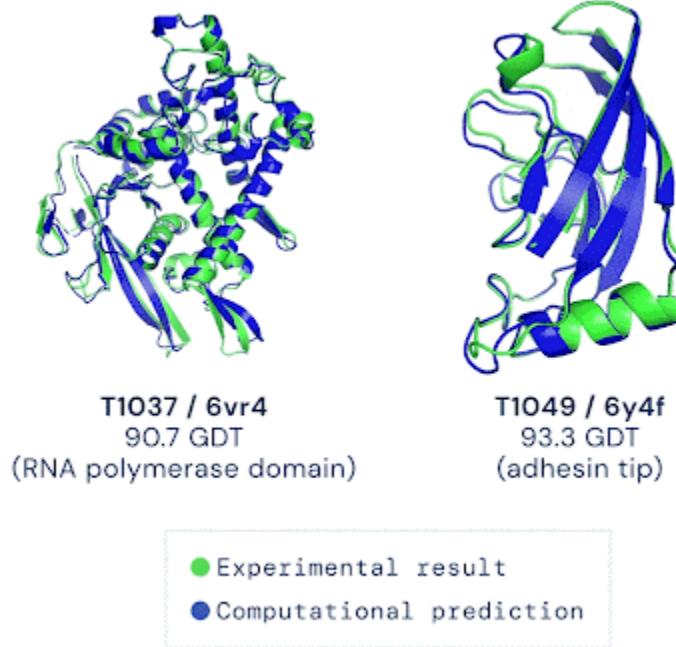
X-ray Crystallography:



Other methods: NMR spectroscopy, Cryo-EM, EM

# How well is the prediction by AlphaFold

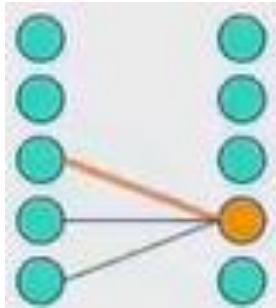
- Protein structure prediction community established CASP (Critical Assessment of Protein Structure Prediction)
- CASP assessment involves the prediction of recently solved structures that are not public
- From CASP 14 (2020), AlphaFold is the top-ranked method achieving consistently high accuracy



# How AlphaFold works: INDUCTIVE BIAS FOR DEEP LEARNING MODELS

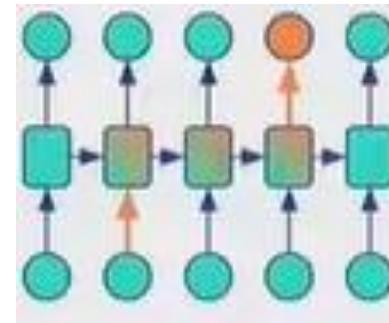
Examples:

## 卷積神經網路 | Convolutional Networks (CNN)



- Used to process structured data with locality (such as sequence data)
- Extract features from local regions in the protein sequence (interaction between amino acids)
- Local sequence segments creates secondary structures

## 循環神經網路 | Recurrent Networks (RNN)



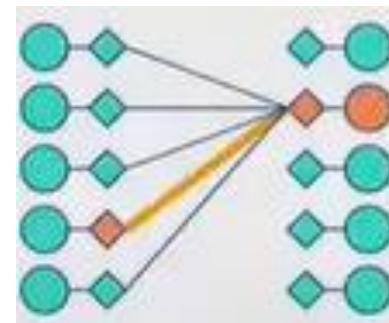
- Used to process sequence data, particular for capturing long-range dependencies.
- Capture interactions between distant amino acid → overall stability and folding
- The 3D structures of a protein is not only determined by adjacent amino acid but also by long-range interactions (hydrogen bonds or hydrophobic interactions).

## 圖神經網路 | Graph Networks (GNN)



- Used to process graph-structure data.
- Considering amino acids as nodes and edges as their interactions, GNNs help learn the complex relationships between amino acids and convert this into structural information
- Consider global information about the protein structure, not just local fragments.

## 注意力機制 | Attention Modules

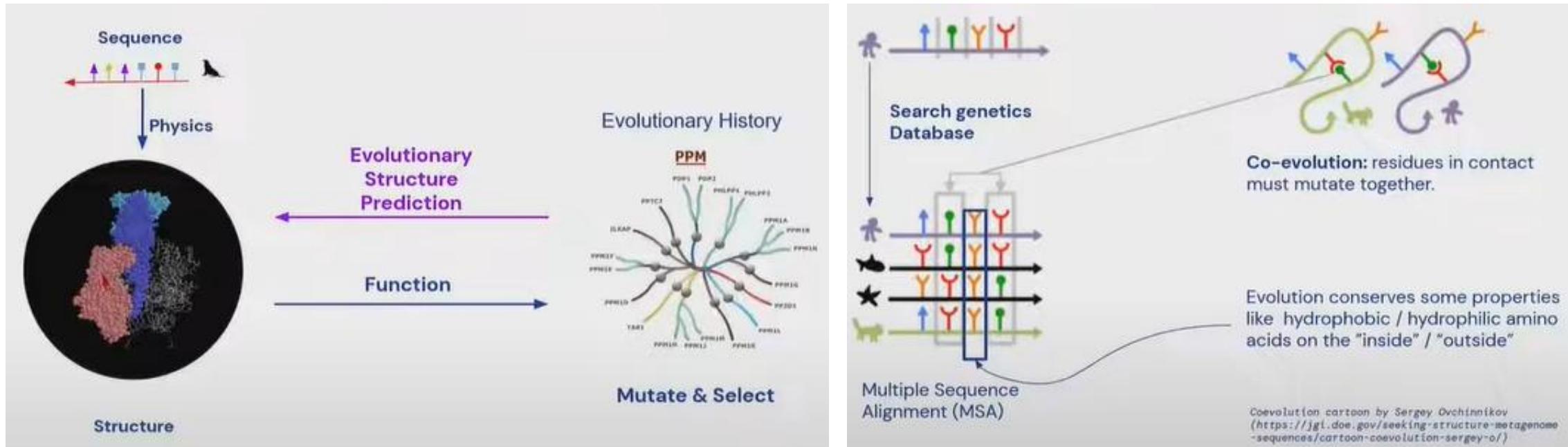


- It allow the model to focus on different part of a sequence based on the importance of each amino acid.
- Certain amino acids are more crucial than others, especially in forming structural cores or functional regions.
- Enable model to flexibly focus more attention on these important amino acids.

# Putting protein knowledge into the model

- Physical and geometric insights are built into the network structure, not just a process around it.
- Inductive biases reflect our knowledge of protein physics and geometry
  - De-emphasized sequential order of amino acids  
(any amino acid can talk to any amino acid in that protein)
  - Instead, residues that are close in space need to communicate
  - Iteratively learning a graph of which residues are close while reasoning over this implicit graph as it is being built

# Determining structure from evolution

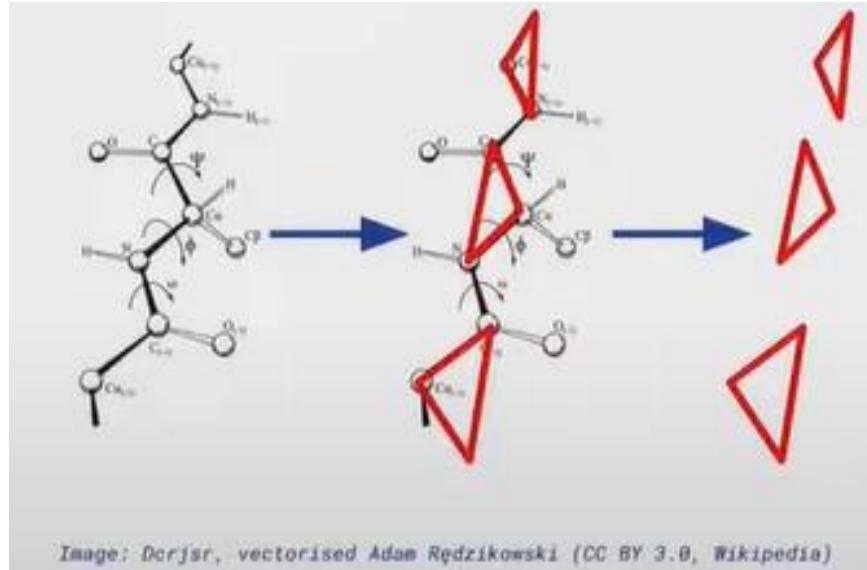


The structure needs to be similar to carry out the same function.

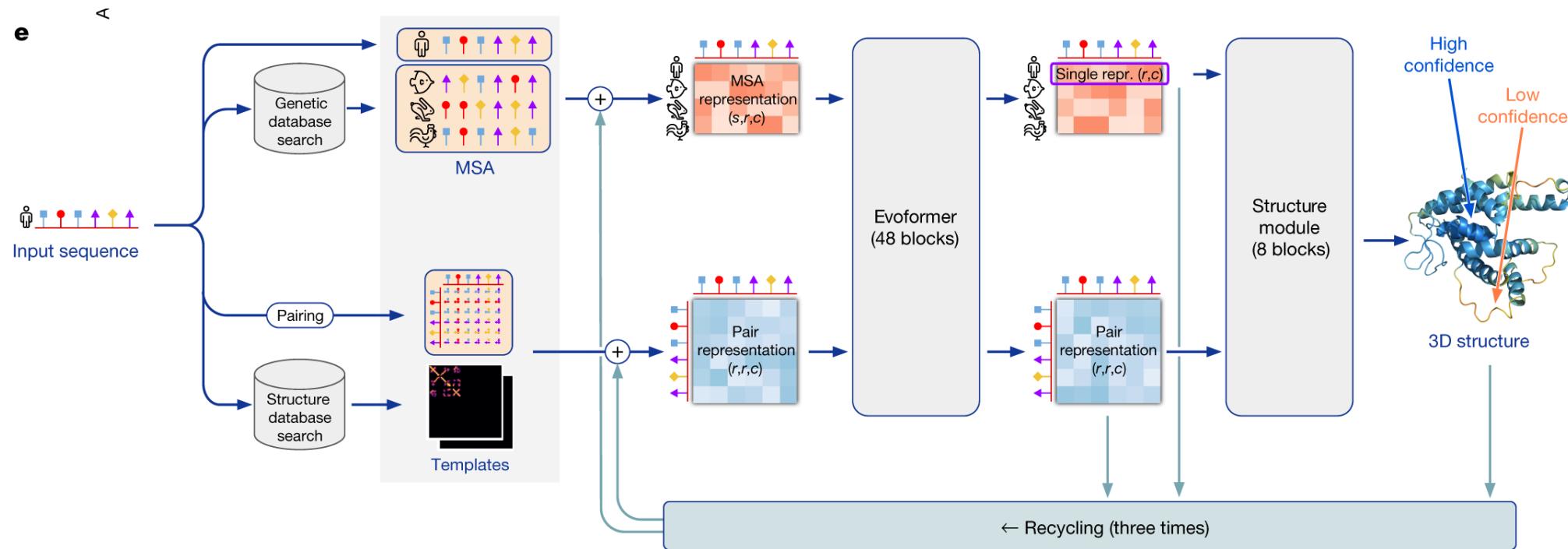
→ Given an evolutionarily related sequence, we can try to computationally infer the structure.

# Considering other important factors

- Structure module
  - End-to-end folding instead of gradient descent
- Noisy student distillation (bootstrapping oneself for a better performance)
  - Make use of unlabeled sequences (didn't have a known experimental structure)
  - Train AlphaFold on just PDB data → Predict structure on a large set of unlabeled sequences → Train second model where training set is enriched by confidently predicted structure of first model



# Overview

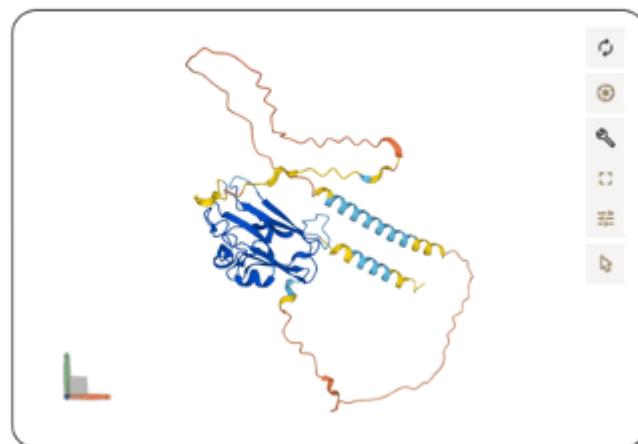


# How to interpret predictions?

## Predicted LDDT (local distance difference test)

- A metric used to assess the accuracy of predicted protein structures by comparing them to the true (experimental) structures
- Specifically, it measures how well the predicted atomic distance, both in backbone and side chains, matches the actual distance in the native structure
- **Ranging 0~100. The higher, the better.**
- Compared with the old criteria RMSD (root mean square deviation), which focuses on the overall alignment, LDDT emphasized the accuracy of local regions within the protein structure.

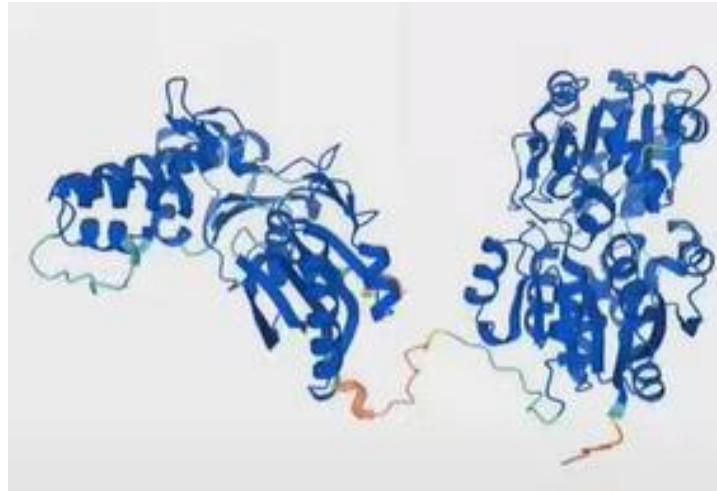
Ephrin-B2



- pLDDT < 50: don't trust the structure or not taking any structure
- pLDDT > 70: May work with the backbone predictions but may not want to trust the side chains in the area
- pLDDT > 90: Reasonable to investigate side chains / active site details

# Pitfalls of pLDDT

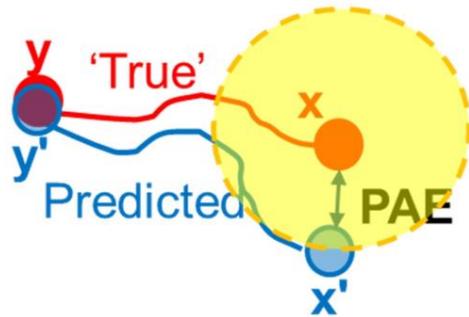
High pLDDT on all domains does **NOT** imply AlphaFold is confident of their relative positions



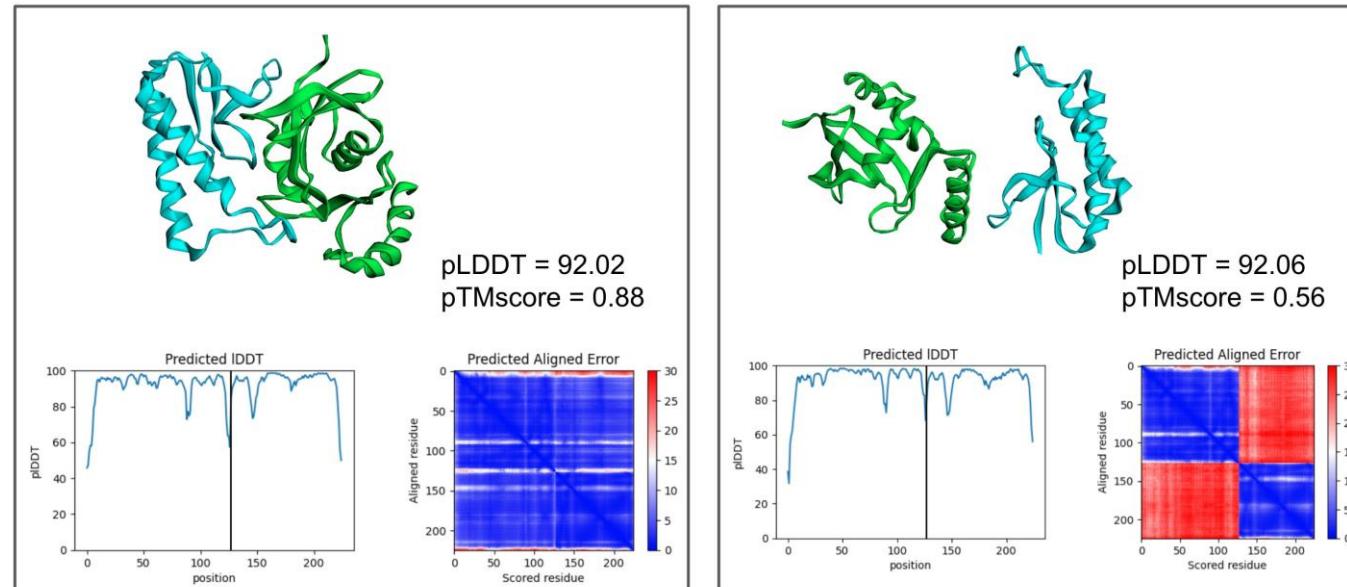
Assessing inter-domain confidence requires a different metric

# Predicted aligned error (PAE)

- PAE measures how well AlphaFold predicts the distance between two residues (amino acids) in the protein, along with the uncertainty in the alignment of those residues
- PAE score:
  - values often range from 0 to 30 Å (Ångströms;  $10^{-10}$  m) or more
  - Lower PAE score: high confidence in the relative position between those residues
  - Higher PAE score: greater uncertainty in their relative positioning



casp:H1065 - pdb:7M5F



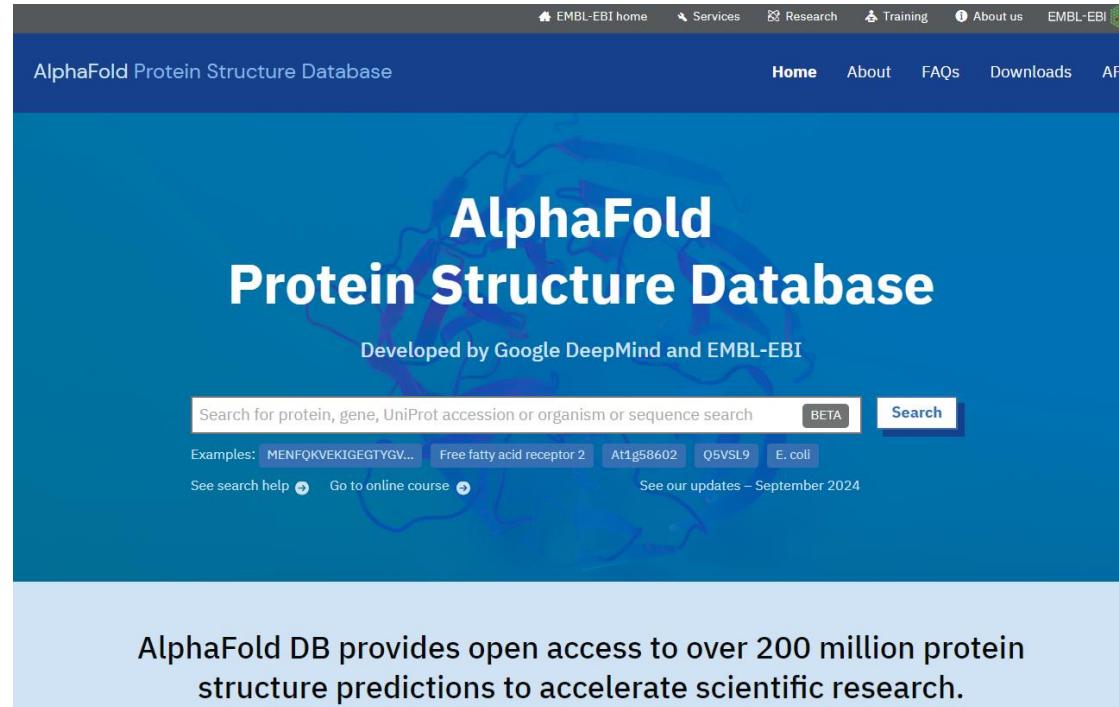
Don't trust how  
AlphaFold positions  
these two domains!

# Limitations of AlphaFold

- Only accepts the 20 standard amino acids in its input
- (By default) predicts 5 models per run:
  - However, models are generally very similar
  - Usually cannot predict conformational variability in a protein
- AlphaFold was not trained for
  - Predict assemblies (AlphaFold-Multimer was trying to do this)
  - Predict the effects of mutations (What AlphaMissense is trying to do)
  - Predict the binding of ligand molecules (some recent research is trying to achieve this using AlphaFold as the basic)
  - Predict nucleic acid structures

# AlphaFold protein structure database

- Website developed and hosted by EMBL-EBI



<https://alphafold.ebi.ac.uk/>

# An example

## Sodium/potassium-transporting ATPase subunit alpha-3

AF-P13637-F1-v4

Download   

PDB (protein data bank)

mmCIF (macromolecular crystallographic information file)

Share your feedback on structure with Google DeepMind

Looks great

Could be improved

### Information

Protein Sodium/potassium-transporting ATPase subunit alpha-3

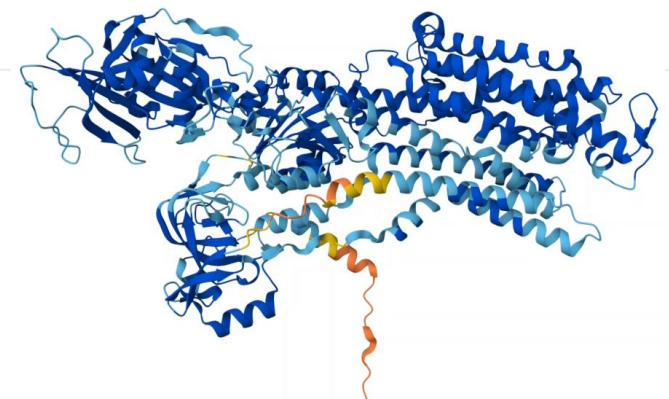
Gene ATP1A3

Source organism Homo sapiens (Human) [go to search](#)

UniProt P13637 [go to UniProt](#)

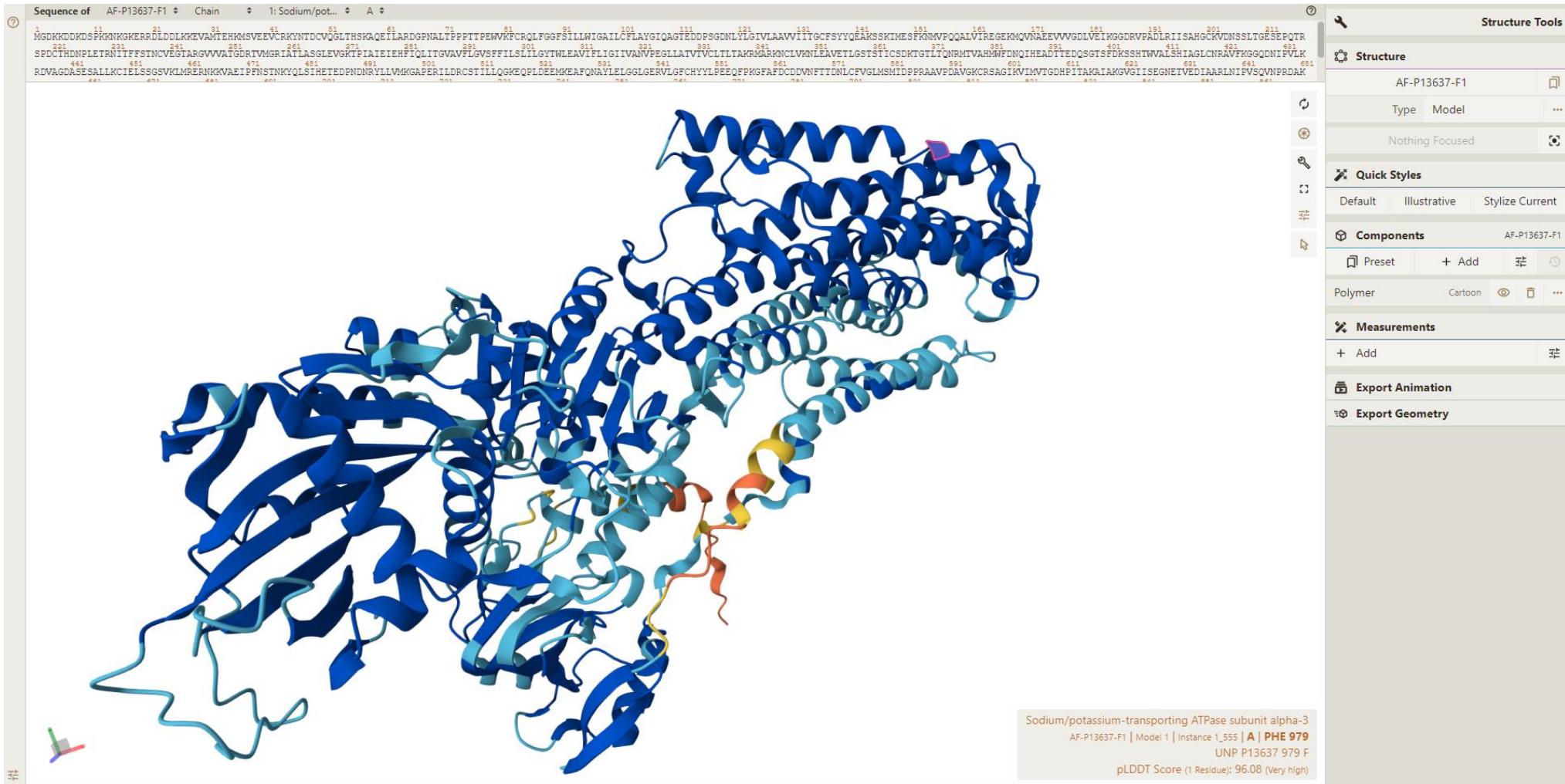
Experimental structures 5 structures in PDB for P13637 [go to PDBe-KB](#)

Biological function This is the catalytic component of the active enzyme, which catalyzes the hydrolysis of ATP coupled with the exchange of sodium and potassium ions across the plasma membrane. This action creates the electrochemical gradient of sodium and potassium ions, providing the energy for active transport of various nutrients. [go to UniProt](#)



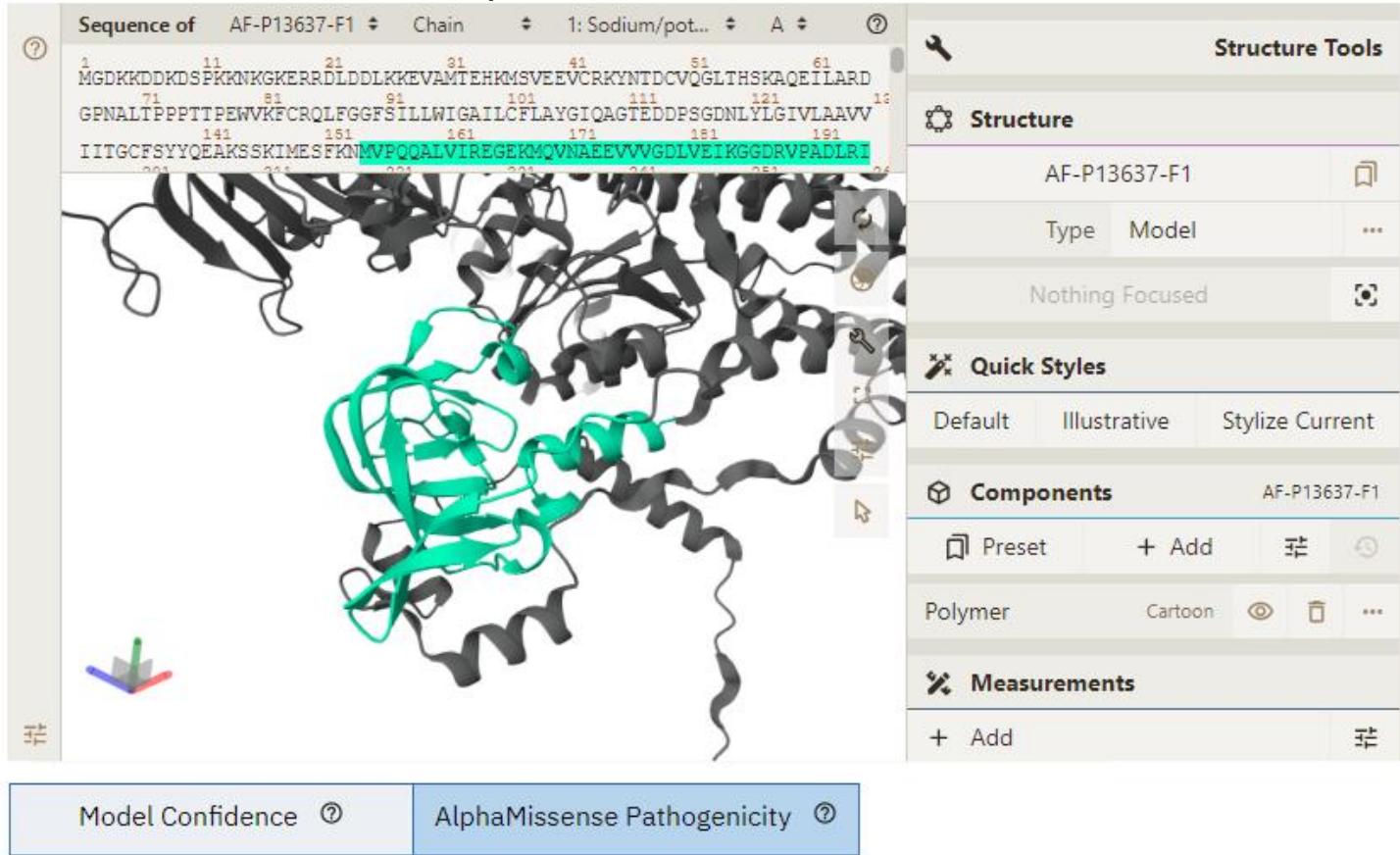
<https://alphafold.ebi.ac.uk/entry/P13637>

# Structure viewer

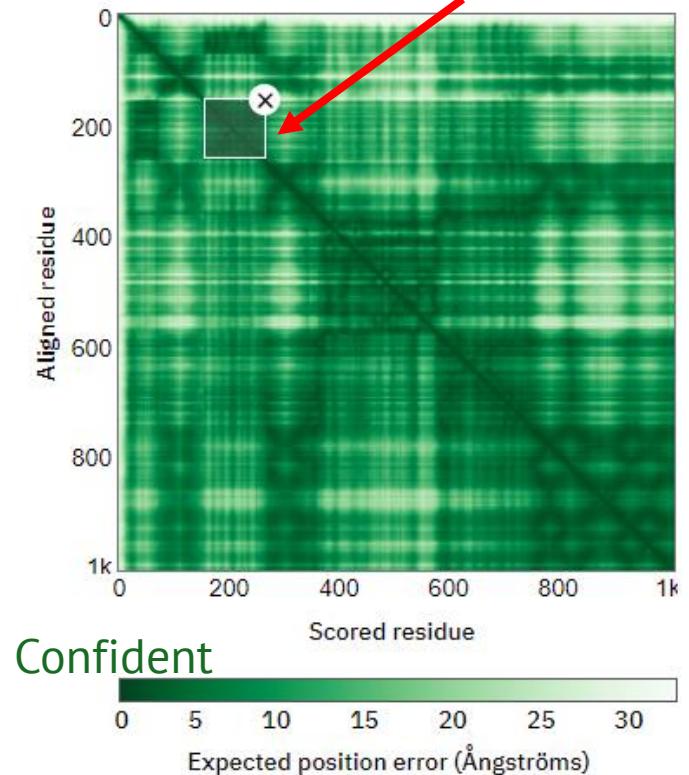


# Predicted aligned error (PAE)

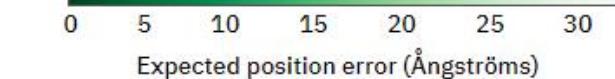
Assess relative domain positions.



The heatmap on the website is interactive!



Confident



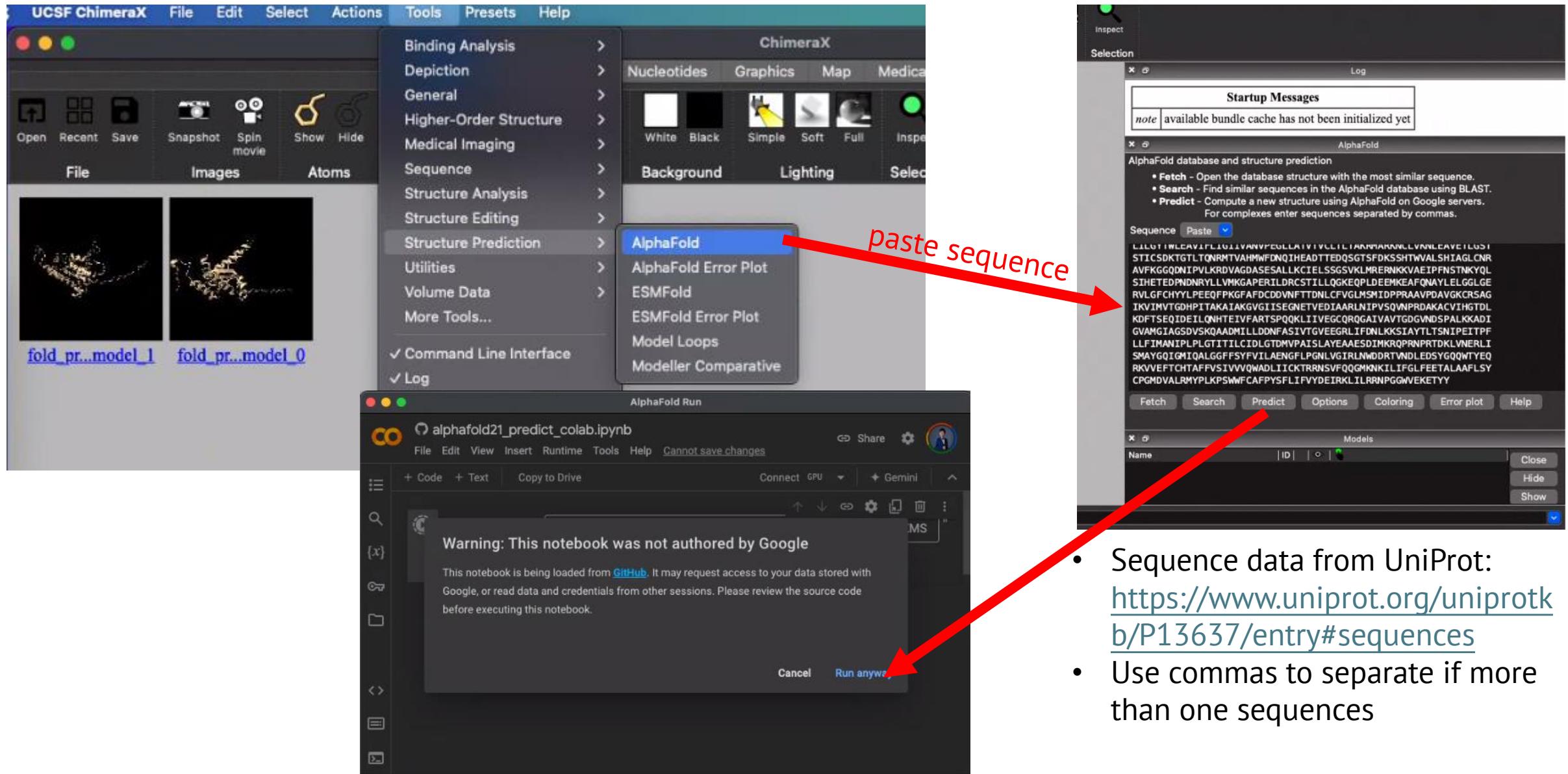
Predicted aligned error (PAE)

# Predict own data

Main ways of accessing predicted protein structures from AlphaFold:

- The open-source code is publicly accessible at <https://github.com/google-deepmind/alphafold>
  - Total control over predictions
  - Need large storage space and a modern GPU
- Interactive Google Colab notebooks: <https://bit.ly/alphafoldcolab>
  - More limited in terms of configuration
  - Easier to use, harder to break

# UCSF ChimeraX + Colab (google) + AlphaFold



# Options

For complexes enter sequences separated by commas.

Sequence  Paste ▾

```
LIGTIWLEAVIFLIGIIVANVPEGLLATIVVCLTLIAKRMARNNCLVNNEAVEILC  
STICSDKTGTLTQNRMVTVAHMWF DNQIHEADTTEDQSGTSFDKSSHTVALSHIAGL  
AVFKGGQDNIPVLKRDVAGDASESALLKCIELSSGSVKLMRERNKKVAEIPFNSTNKY  
SIHETEDPNDNRYLLVMKGAPERILDRCSTILLQGKEQLDDEEMKEAFQNAYLELGGI  
RVLGFCHYYLPEEQFPKGFAFD CDDVNFTTDNLCFVGLMSMIDPPRAAVPDAVGKCRS  
IKVIMVTGDHPITAKAIAKGVGIISEGNETVEDIAARLNIPVSQVNPRDAKACVIHGT  
KDFTSEQIDEILQNHTIEIVFARTSPQQKLIIIVEGCQRQGAIVAVTGDGVNDSPALKKA  
GVAMGIAGSDVSKQAADMILLDDNFASIVTGVEEGRLIFDNLKKSIAYLT SNIPEIT  
LLFIMANIPLPLGTITILCIDLGTD MVPAISLAYEAESDIMKRQPRN PRTDKLVNEP  
SMAYGQIGMIQALGGFFSYFVILAENGFLPGNLVGIRLNWDDRTVNDLED SYGQQWTV  
RKVVEFTCTA FFVSVVVQWADLIICKTRRNSVFQQGMKNKILIFGLFEETALAAFL  
CPGMDVALR MYPLKPSWWFC AFPYSFLIFVYDEIRKLILRRNPGGWVEKETYY
```

Fetch Search Predict Options Coloring Error pl

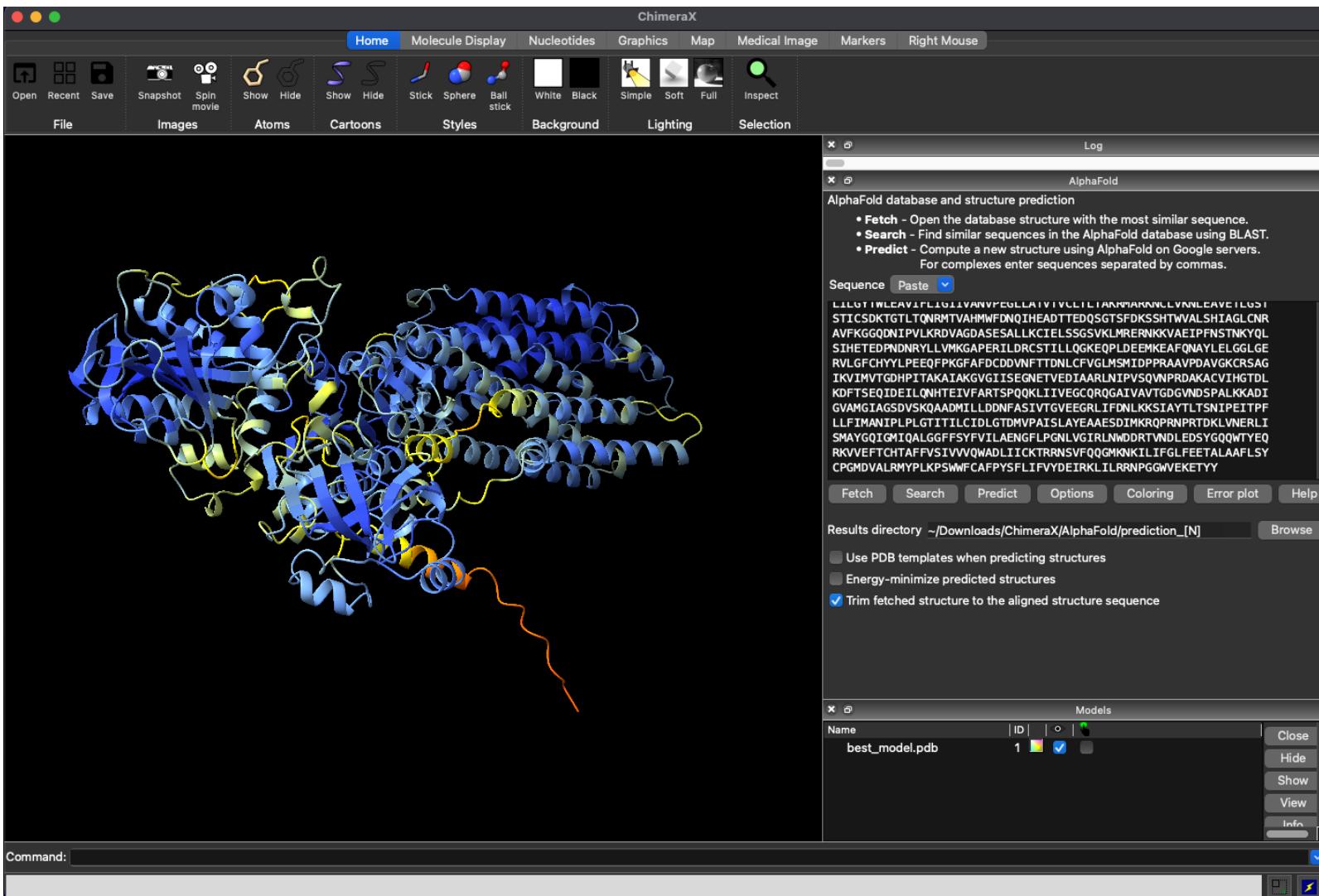
Results directory ~/Downloads/ChimeraX/AlphaFold/prediction\_[N]

Use PDB templates when predicting structures

Energy-minimize predicted structures

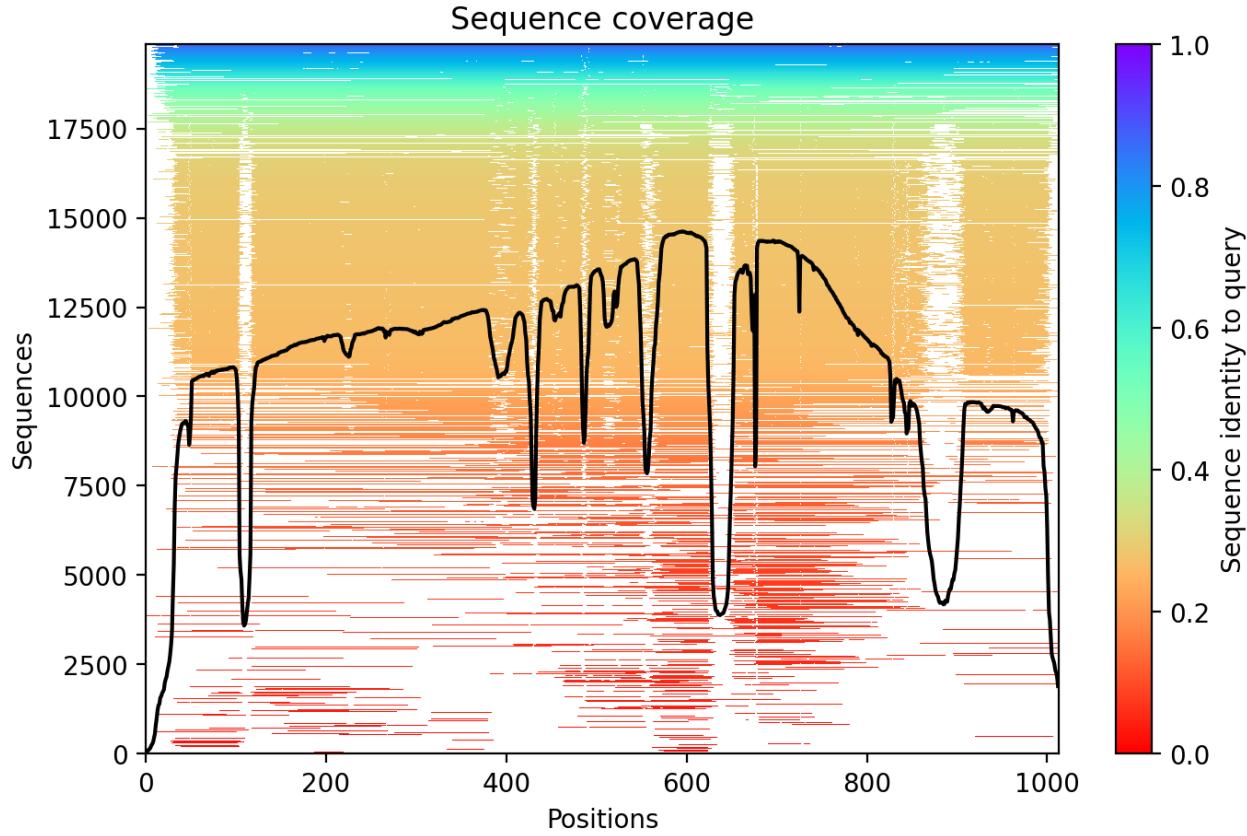
Trim fetched structure to the aligned structure sequence

# Result



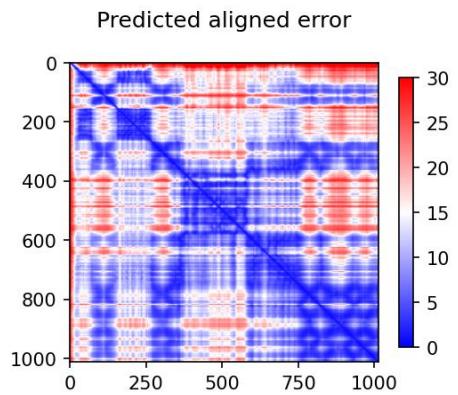
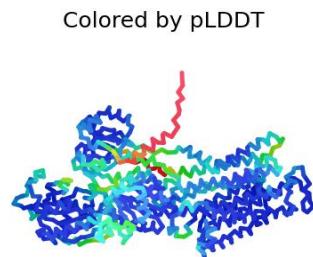
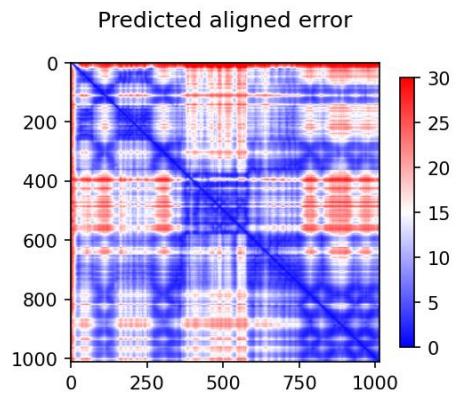
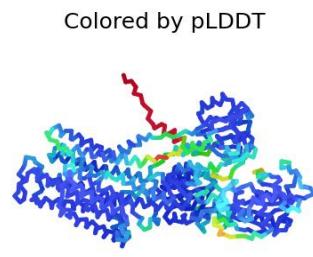
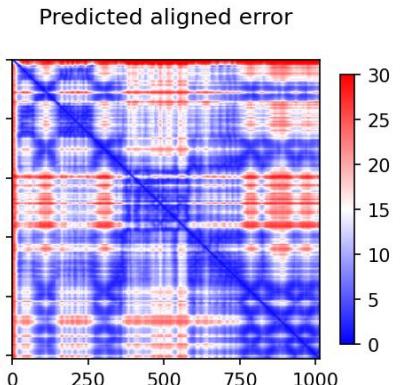
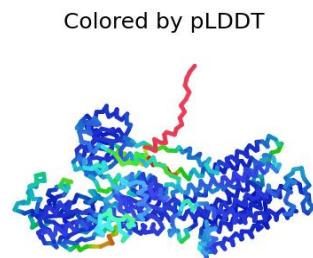
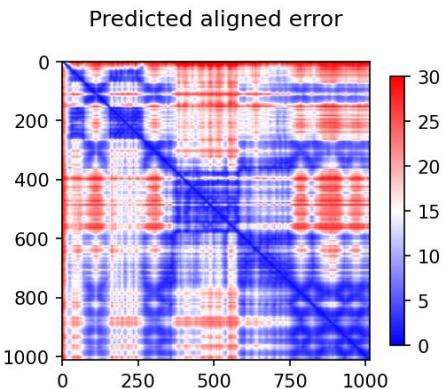
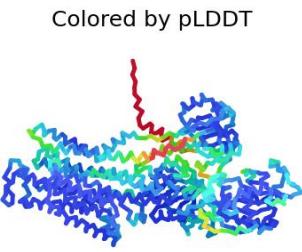
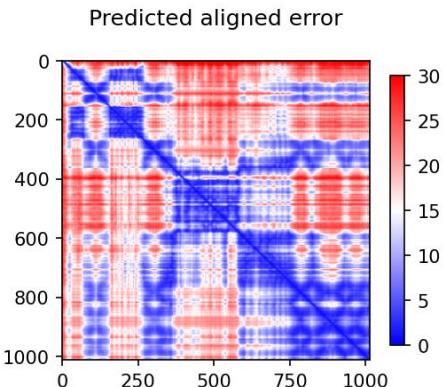
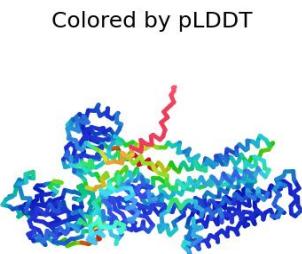
Output files available here: [[prediction](#)]

# Sequence coverage

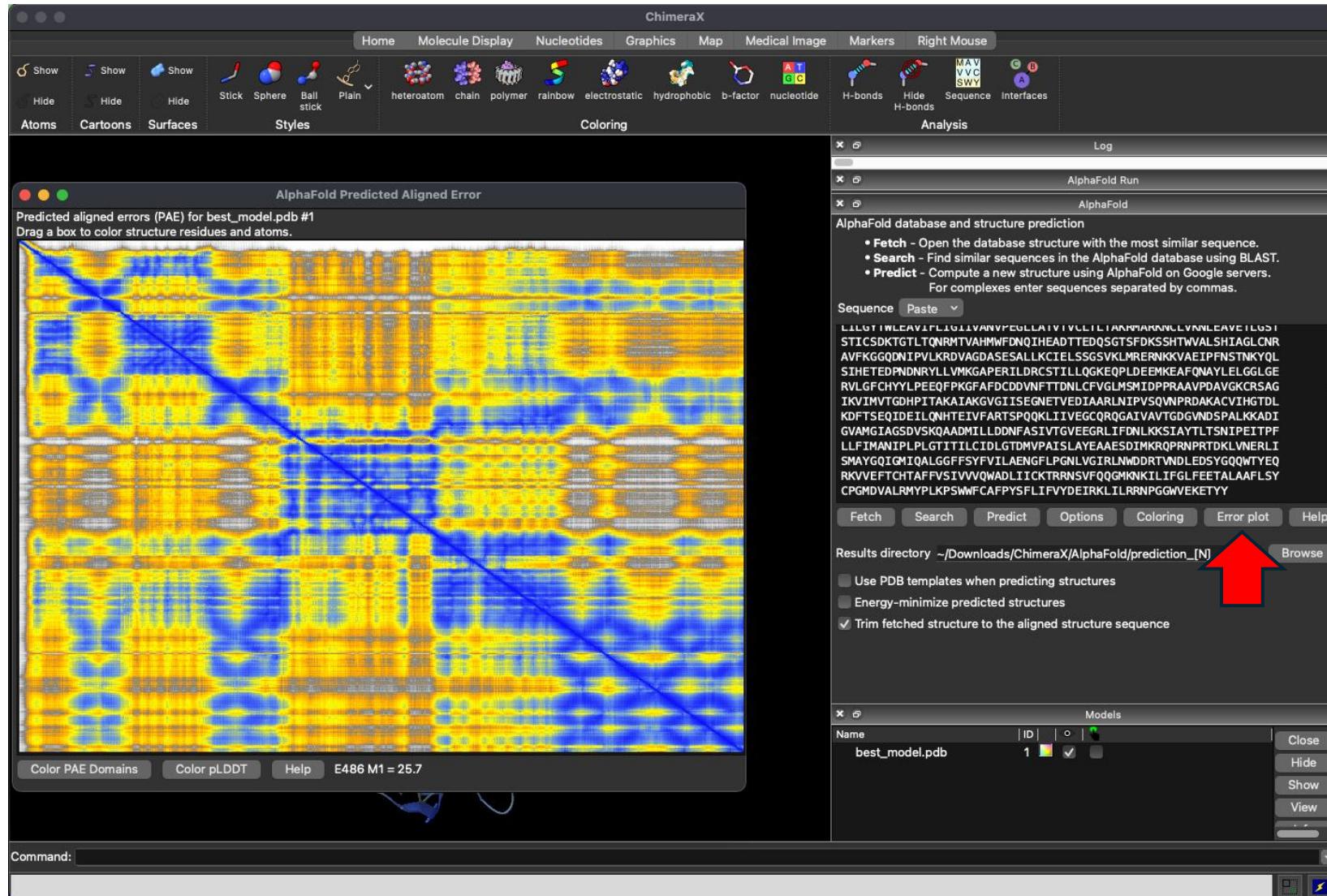


- It refers to how well the target protein sequence is covered by other sequences in the multiple sequence alignment.
- High coverage:
  - Many sequences in the MSA align to the target sequence at each position
  - Informed by a large amount of evolutionary data

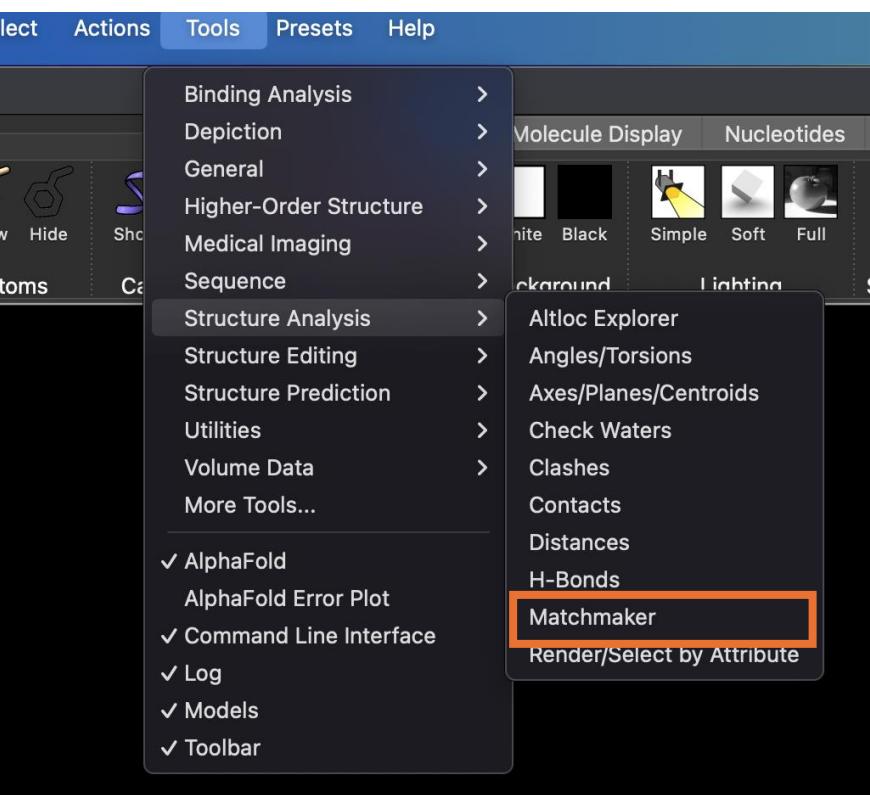
# Different predictions



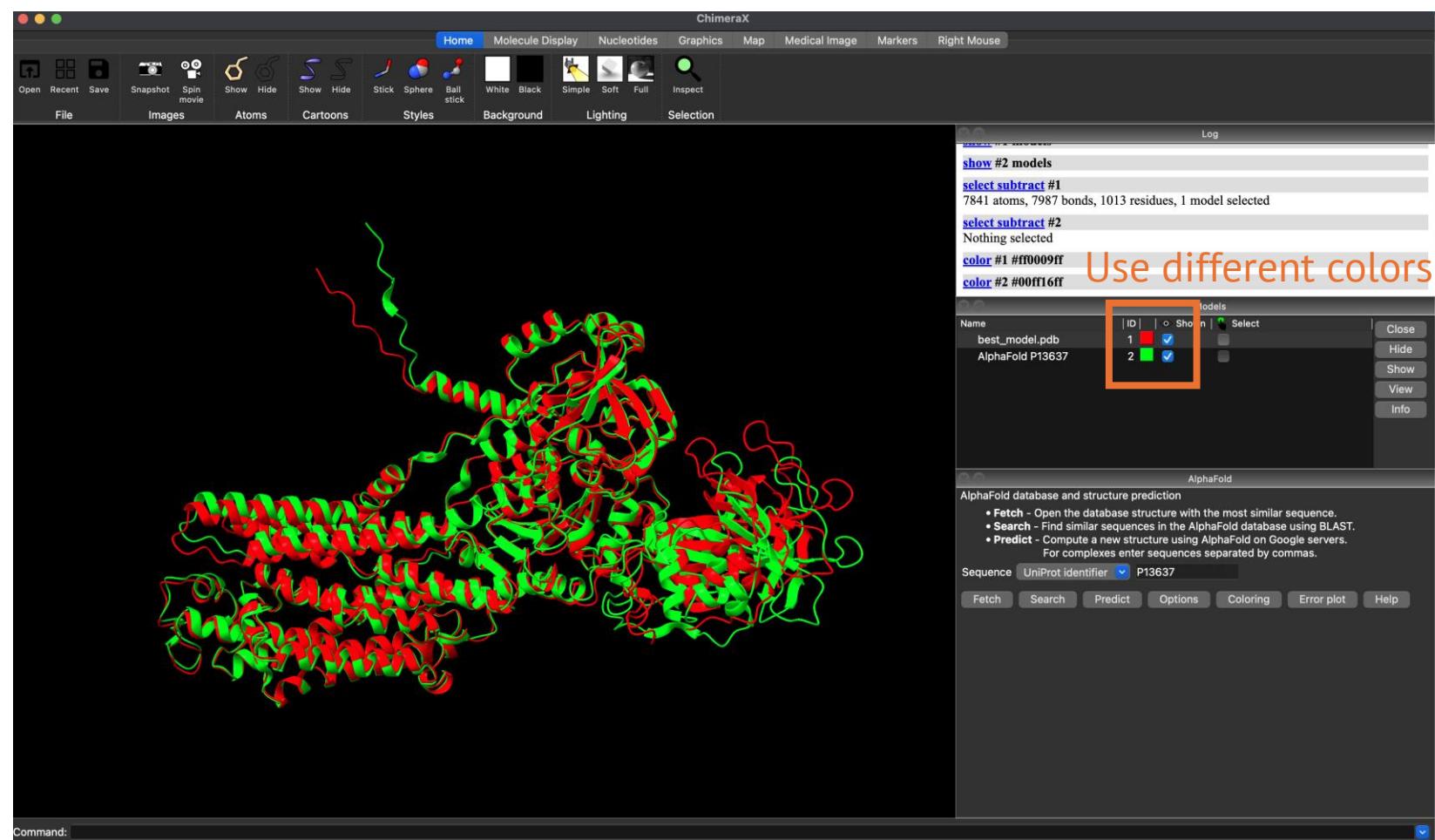
# Show error plot in ChimeraX



# Align the predicted structure with the experimental result

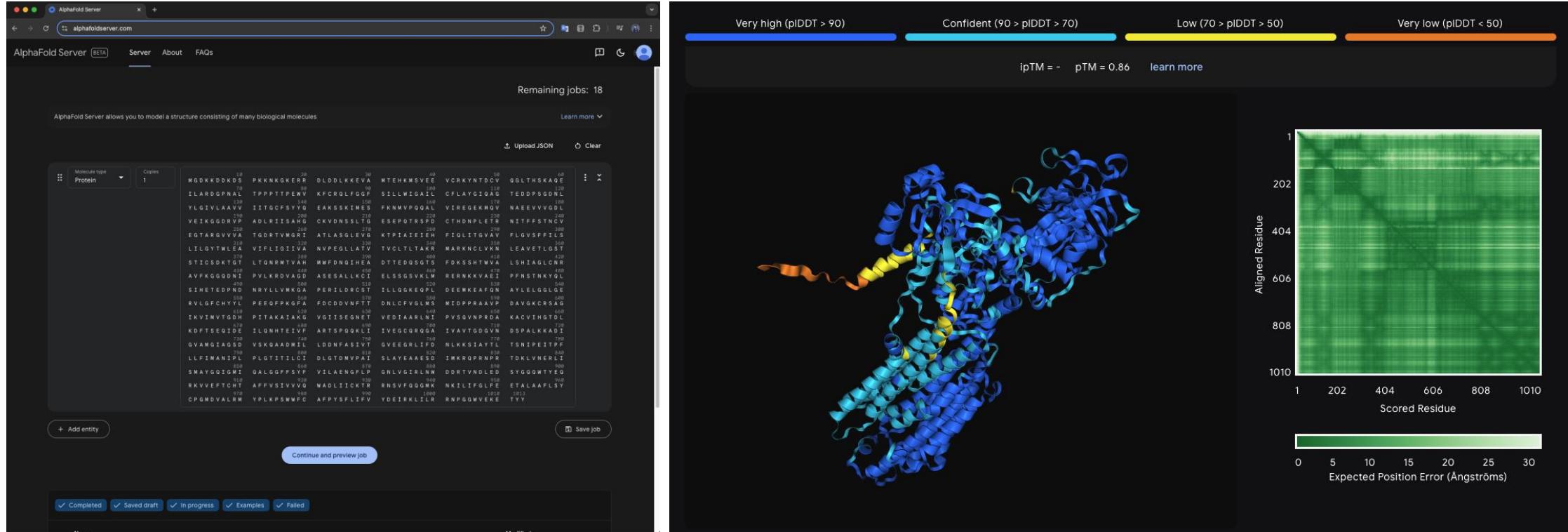


Use matchmaker

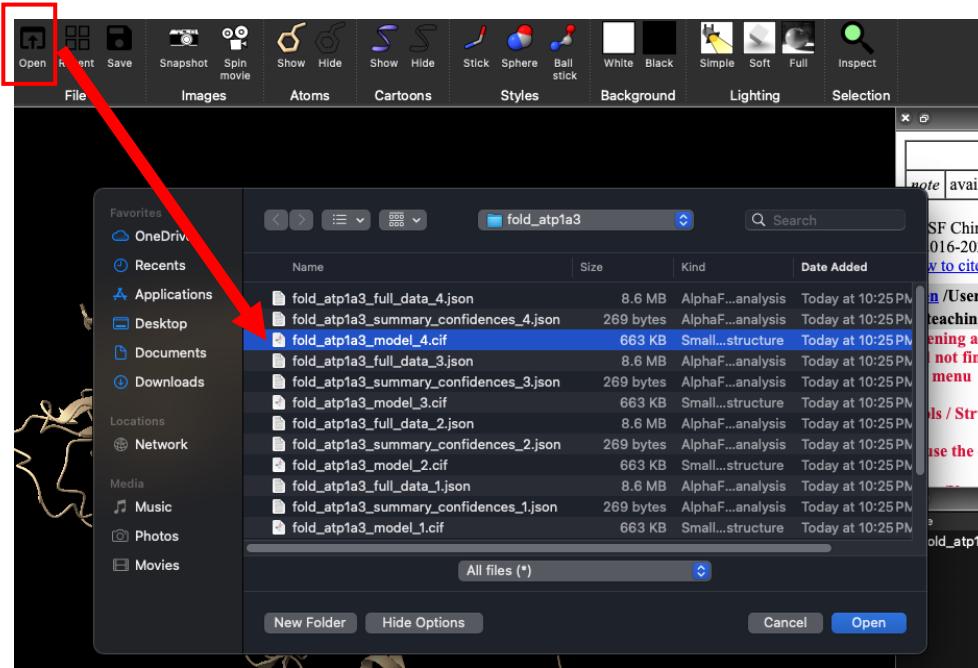


Use different colors

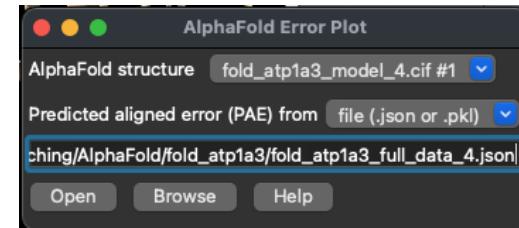
# An alternative for Colab: AlphaFold server



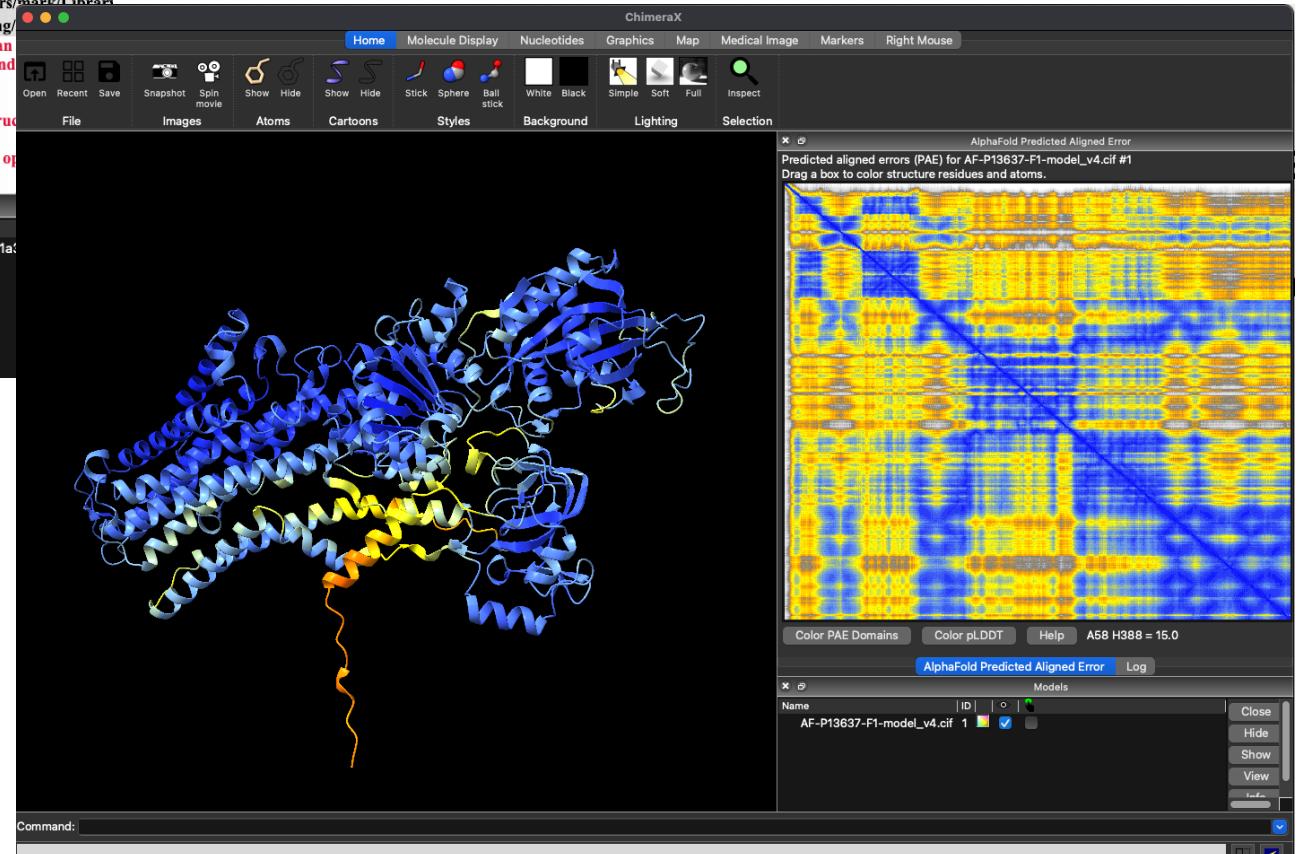
# Open the results from the AlphaFold server in ChimeraX



Results available [[HERE](#)]



Error plot info was safe as json format



# Other visualization tools

- PyMOL: <https://github.com/schrodinger/pymol-open-source>
  - The open-source version has same functions as the commercial version
  - Installed in conda environment: `conda install conda-forge::pymol-open-source`
- Mol\* 3D Viewer
  - Hosted by RCSB protein data bank
  - <https://www.rcsb.org/3d-view>

