# Analysis of AlphaFold2 Predictions for MPOX-22 Proteins
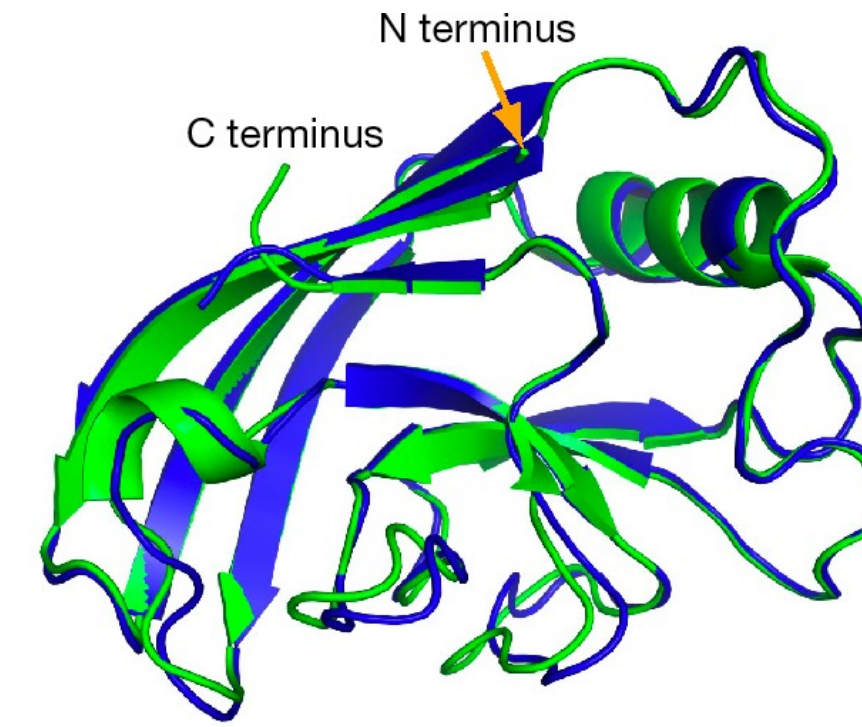
Devon J. Boland
Norman Borlaug Endowed Research Scholar
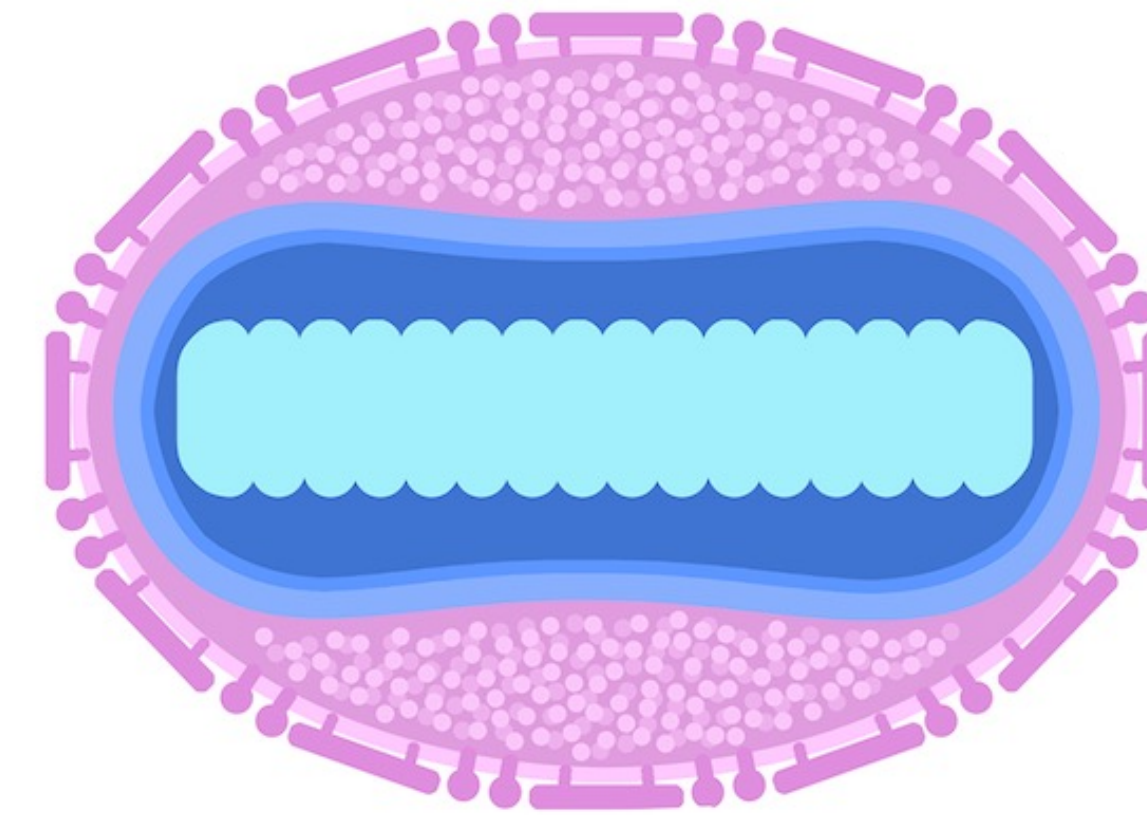
- **We used AF2 to predict protein 3D structure**

- **Each assigned a protein from the recently assembled MPOX-22' outbreak strain**

- **Today we are going to analyze our predicted structures and even infer the expected function**



N terminus

C terminus

AlphaFold   Experiment

r.m.s.d.$_{95}$ = 0.8 Å; TM-score = 0.93

Jumper, *et. al.* 2021 Nature.



50 nanometers    Monkeypox Virus

https://en.wikipedia.org/wiki/Mpox
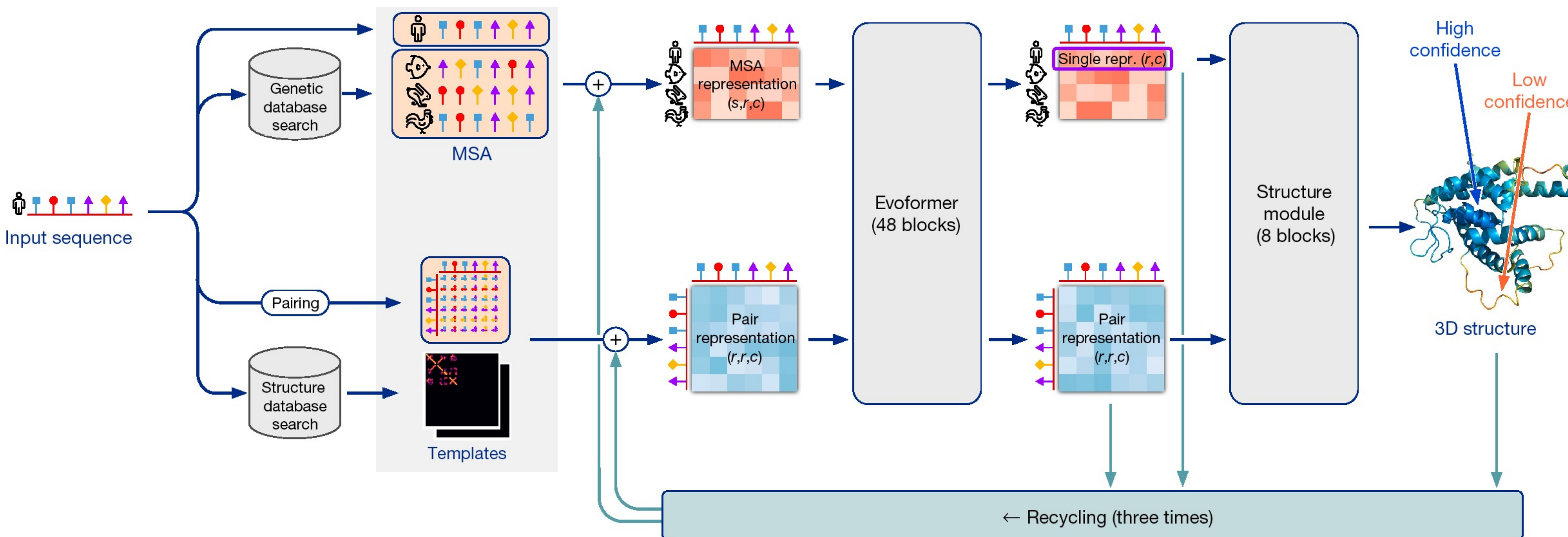


Jumper, *et. al.* 2021 Nature.

# Extracting Our Output For Analysis

- **Our data is stored here, we need to offload it so that we can analyze it**
  - **You will download the files ranked_0.pdb, ranked_debug.json**
  - **and any file ending in the .pkl extension**
  - **You will upload the entire output folder to the class drive folder**
  - **You also must install ChimeraX for us to view the structures**
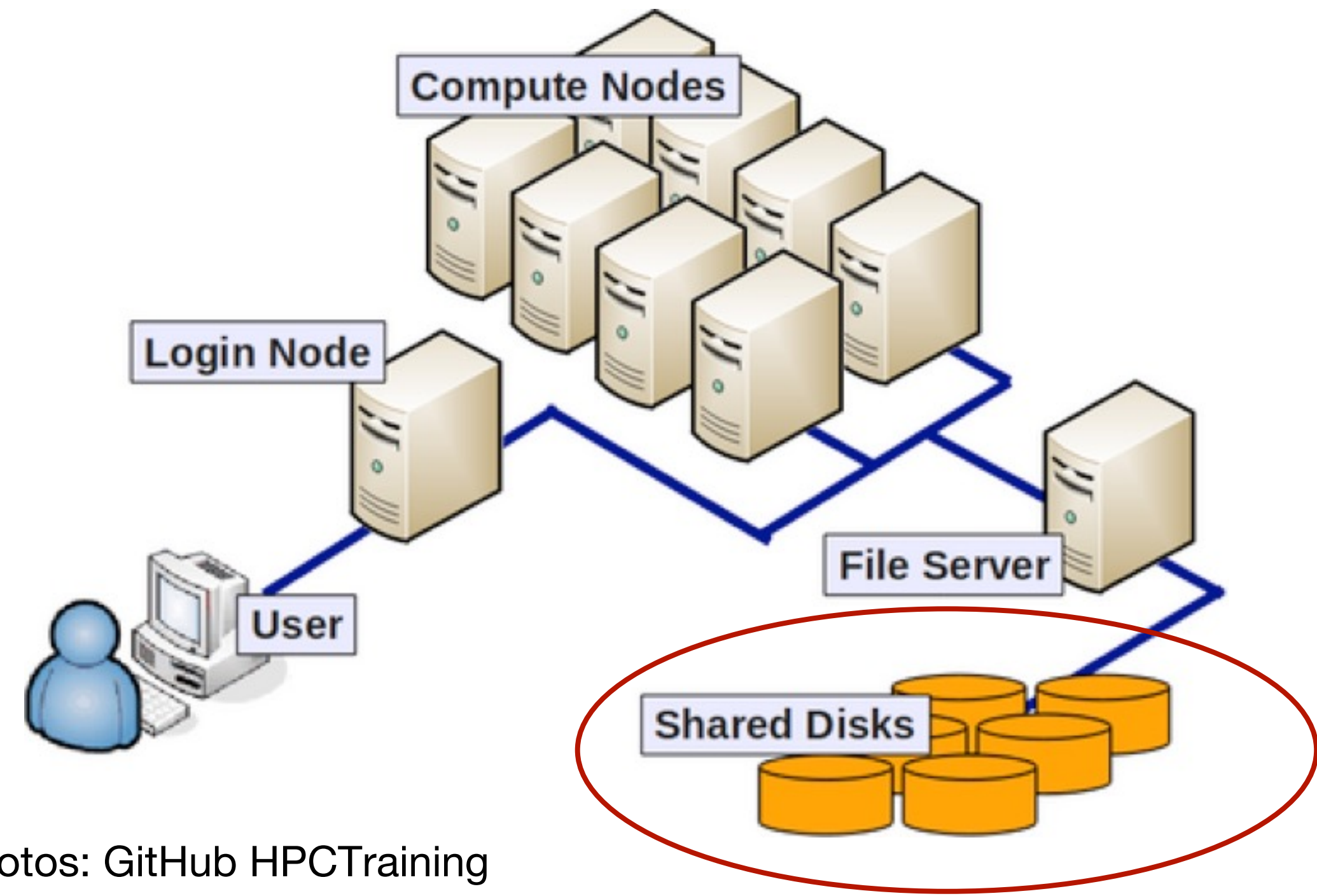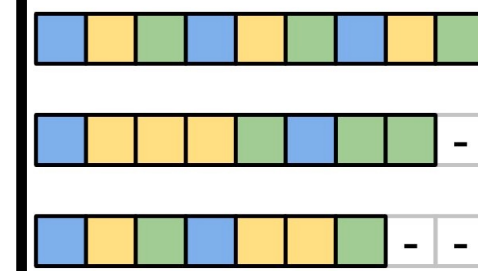


Photos: GitHub HPCTraining

[https://github.com/deepmind/alphafold#alphafold-output](https://github.com/deepmind/alphafold#alphafold-output)

**Sequence Homology**

**MS**
(Multiple sequences)

**MSA**
(Multiple sequence alignment)

Align

hhblits, blast,
jackhmmer,
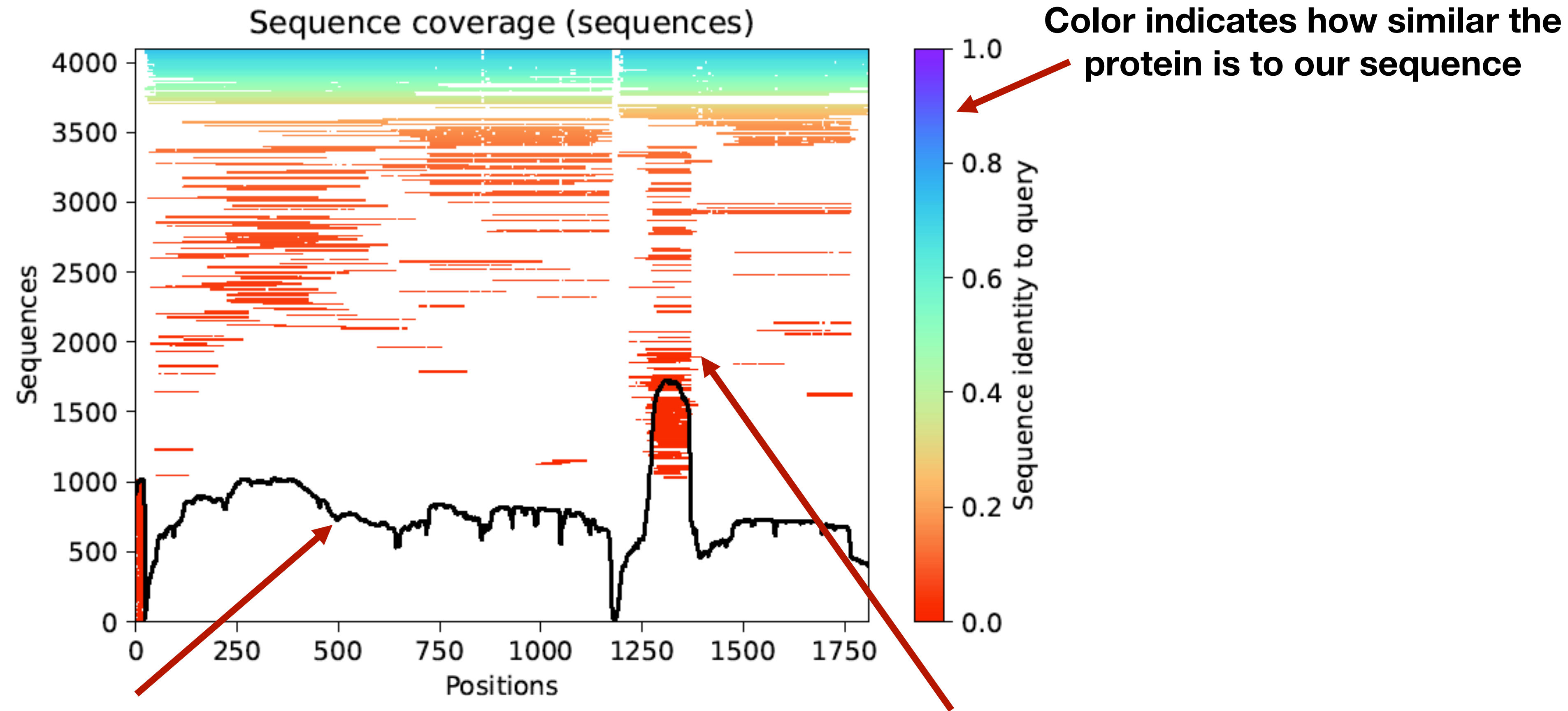clustal, etc.

Petti S, *et. al. Bioinformatics*. 2023;39(1)

- **THE limiting factor**
  - **sequence coverage depth**
  - **>30 sequences/residue**
- **Typically areas of low coverage:**
  - **random disordered coils**
  - **low pLDDT**
  - **High pAE**

# Sequence Coverage Plot

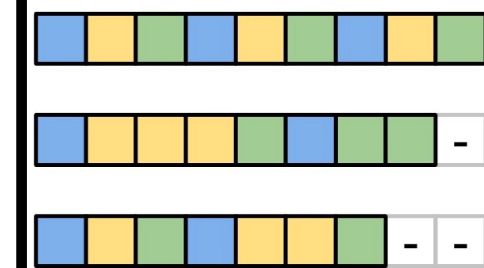**Sequence coverage plot from FLS2-BAK1-flg22 Receptor Complex - From AlphaFold 2.1.3**



**Color indicates how similar the protein is to our sequence**

**Black line shows coverage depth (i.e. how many homologous sequences at given residue)**

**Each line here represents a single sequence (protein)**

## Sequence Homology

**MS**
(Multiple sequences)

**MSA**
(Multiple sequence alignment)
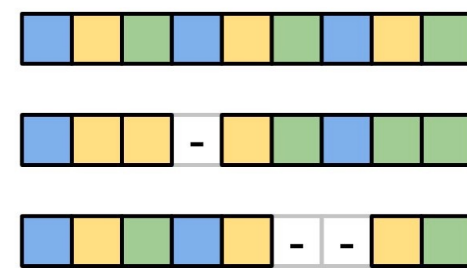
Align

hhblits, blast,
jackhmmer,
clustal, etc.

Petti S, *et. al. Bioinformatics*. 2023;39(1)

- **THE limiting factor**
  - **sequence coverage depth**
  - **>30 sequences/residue**
- **Typically areas of low coverage:**
  - **random disordered coils**
  - **low pLDDT**
  - **High pAE**

## Side-Chain Cα Confidence

Mariani V, *et. al., Bioinformatics*. 2013;29(21)

- **Local Distance Difference Test**
  - **3D-structure depedent**
  - **(LDDT)**
- **R-group "feasibility"**
  - **Very High (pLDDT > 90)**
  - **Confident (90 > pLDDT < 70)**
  - **Low (70 > pLDDT > 50)**
  - **Very Low (pLDDT < 50)**

**\*suitable for experimental design**

**pLDDT plot for 25 models of FLS2-BAK1-flg22 Receptor Complex - From AlphaFold 2.1.3**



Predicted LDDT per position (sequences)

**Each line represents a different predicted model for the structure (Your's will only have 5)**

**Remember we want this value to be as >90 for experimental design**

# How Can We Evaluate Our Confidence In The Model?

## Sequence Homology



**MS**
(Multiple sequences)
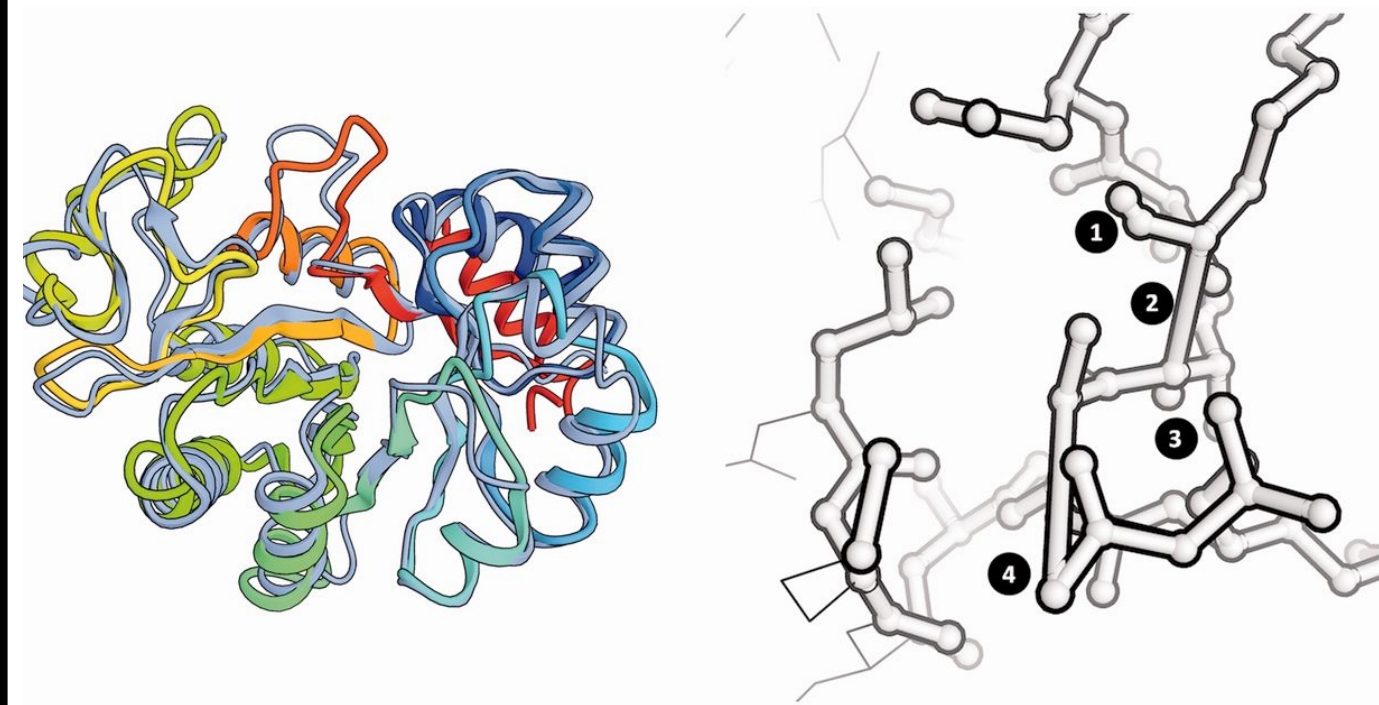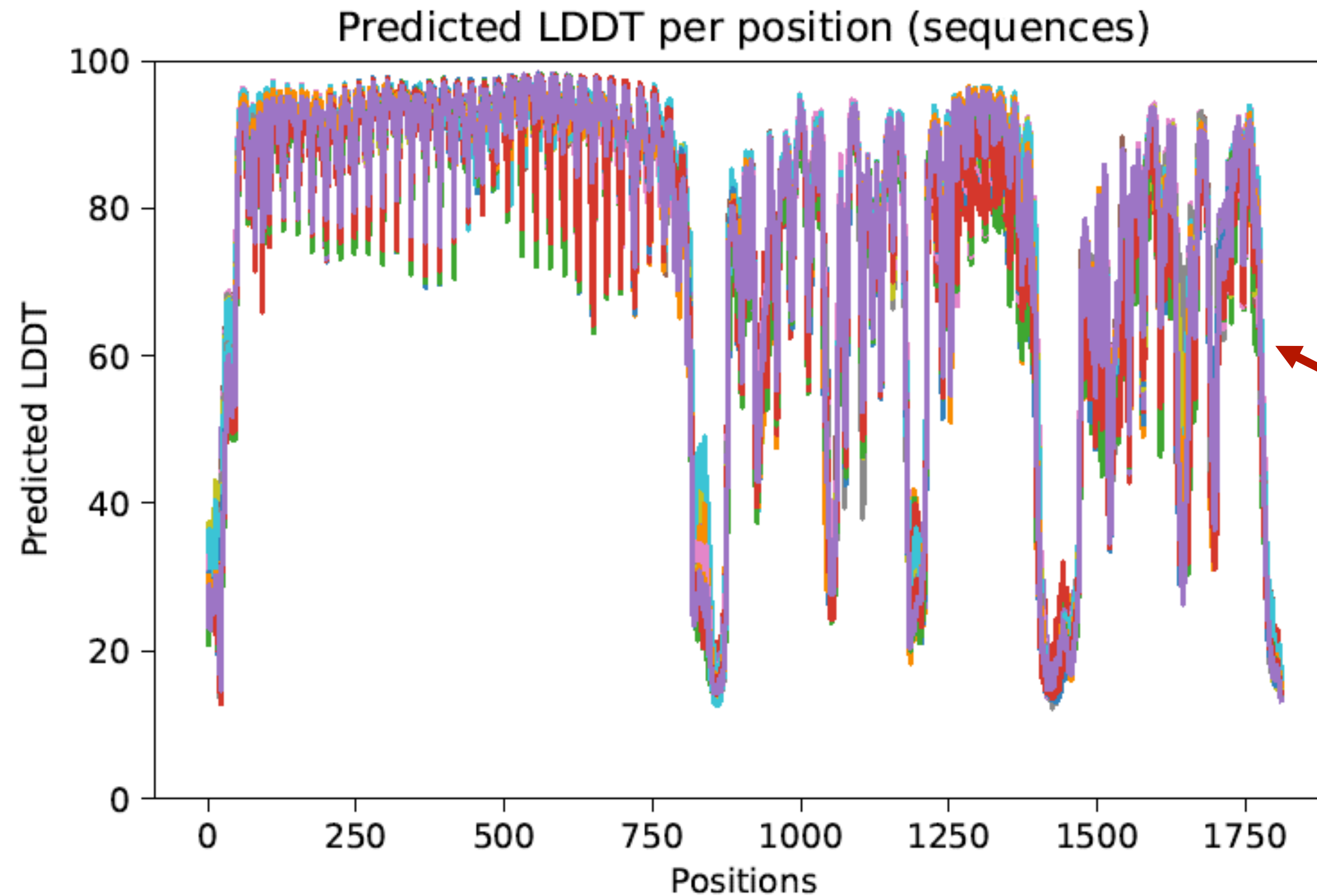
Align
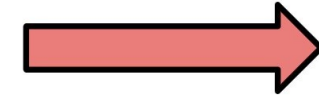
hhblits, blast, jackhmmer, clustal, etc.

**MSA**
(Multiple sequence alignment)

Petti S, *et. al. Bioinformatics*. 2023;39(1)

- **THE limiting factor**
  - **sequence coverage depth**
  - **>30 sequences/residue**
- **Typically areas of low coverage:**
  - **random disordered coils**
  - **low pLDDT**
  - **High pAE**

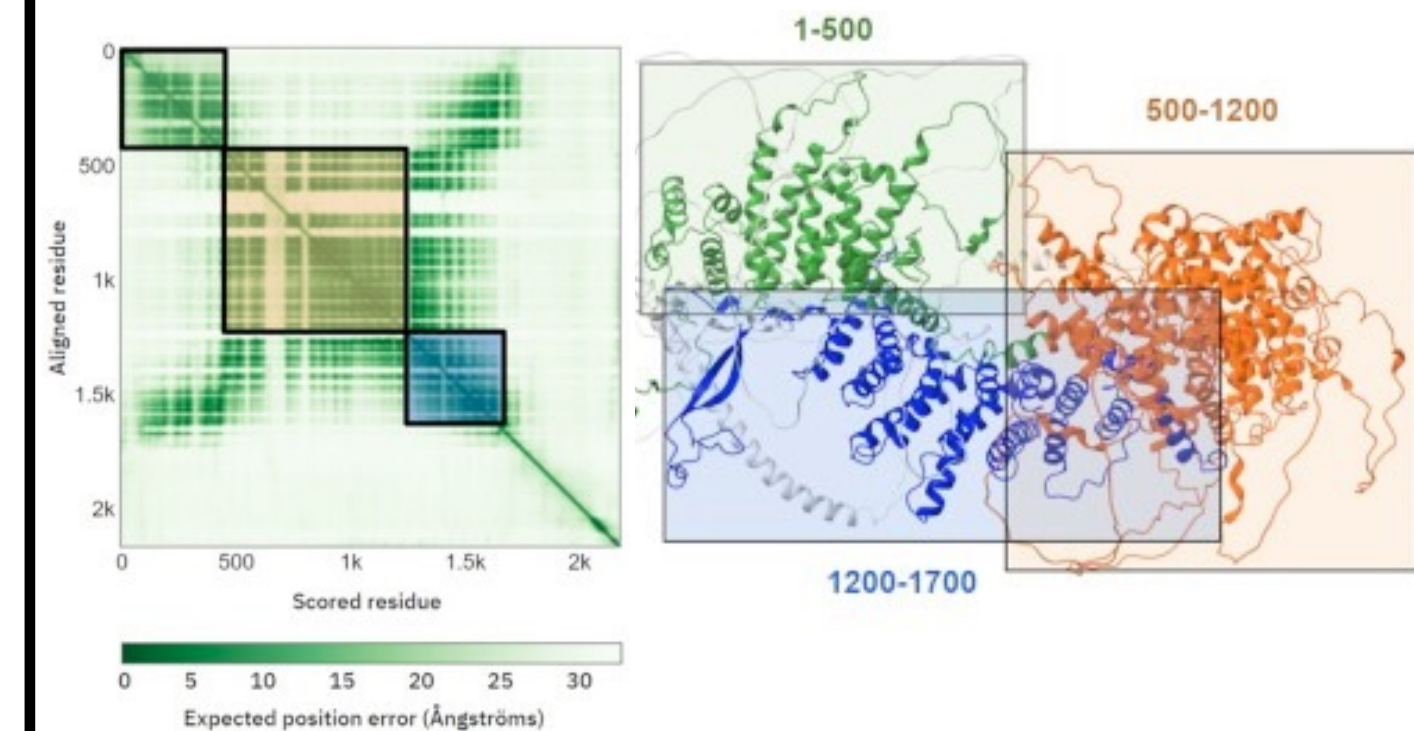## Side-Chain Cα Confidence
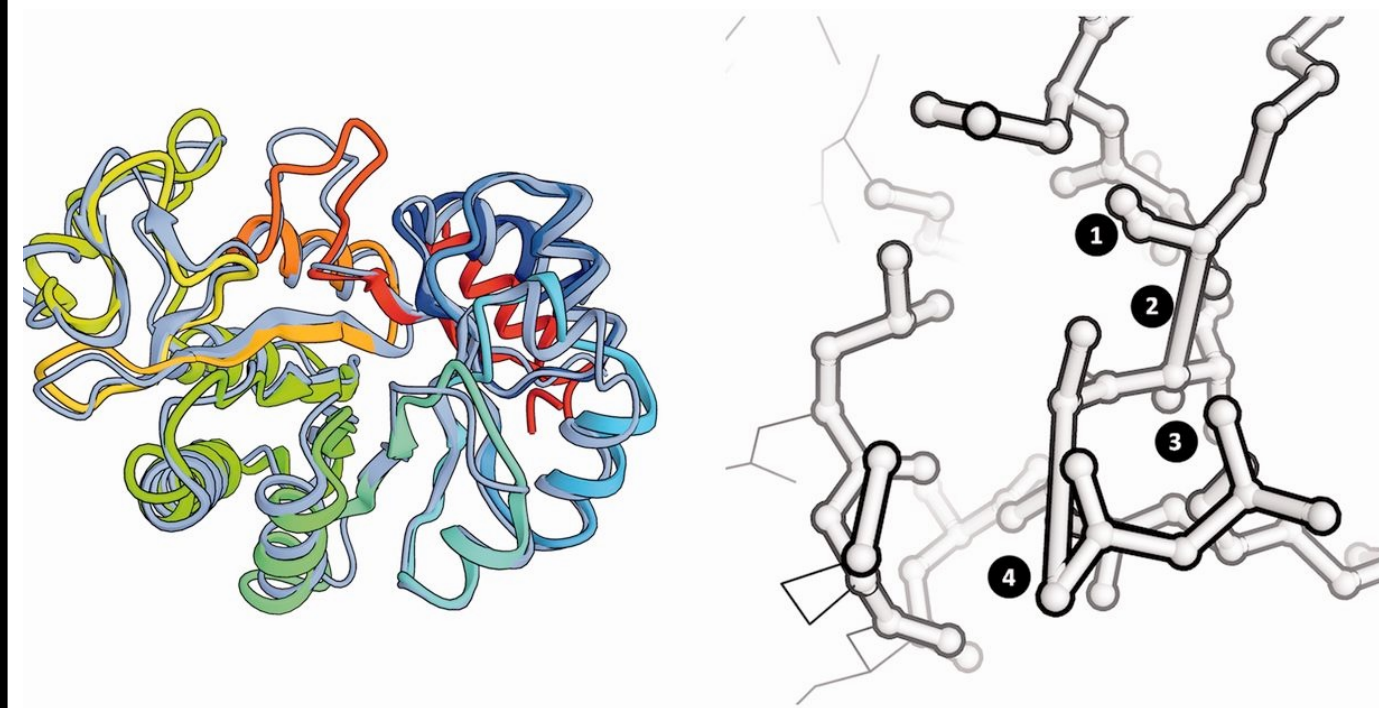


Mariani V, *et. al., Bioinformatics*. 2013;29(21)

- **Local Distance Difference Test**
  - **3D-structure depedent**
  - **(LDDT)**
- **R-group "feasibility"**
  - **Very High (pLDDT > 90)**
  - **Confident (90 > pLDDT < 70)**
  - **Low (70 > pLDDT > 50)**
  - **Very Low (pLDDT < 50)**

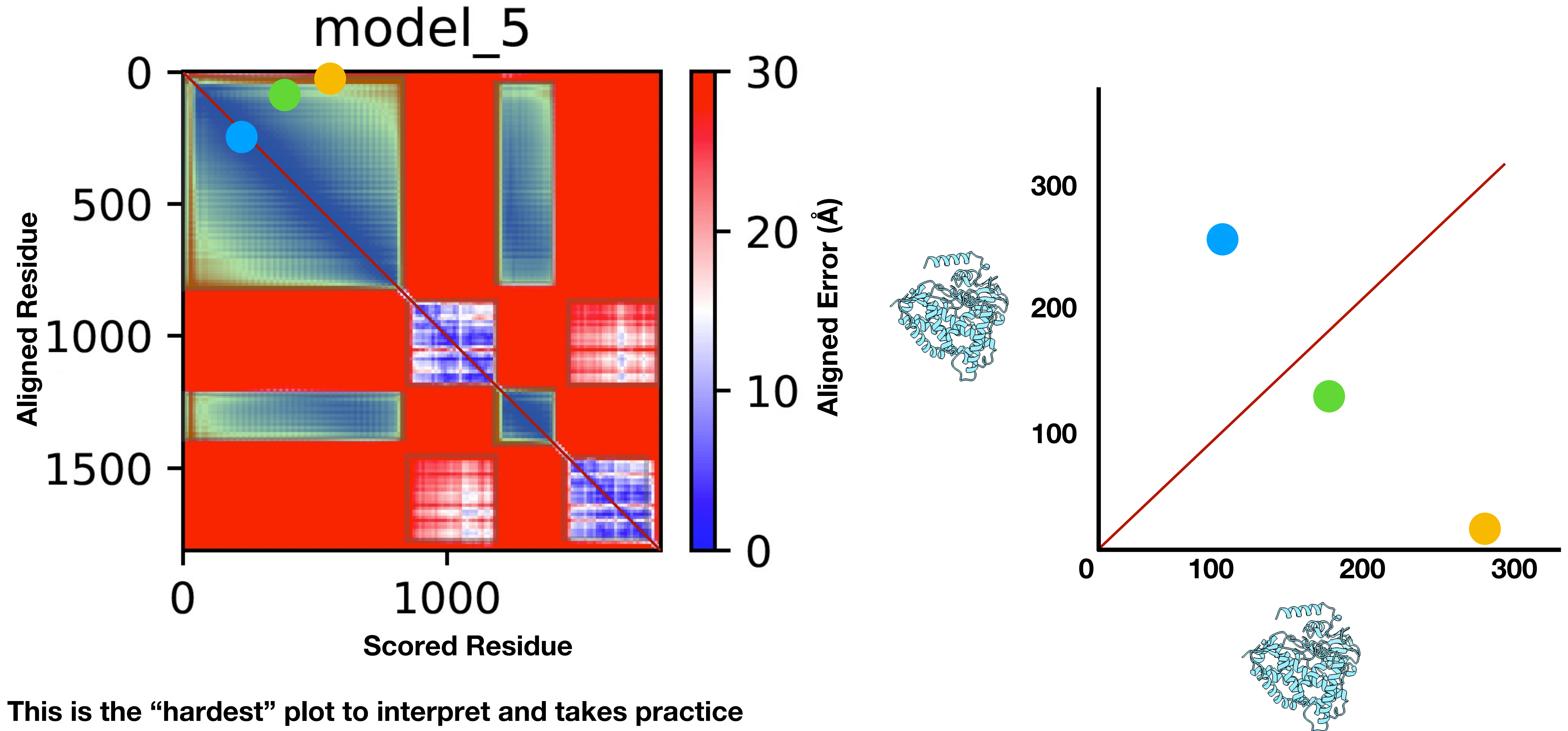**\*suitable for experimental design**

## Inter-domain Accuracy



Varadi M, *et. al., Nucleic Acids Res*.2022;50

- **Aligned Error**
  - **3D-structure indepedent**
  - **(AE)**
- **Relative position of domains**
- **Mutual location of domains**

# Predicted Aligned Error Plot

**pAE plot for model 5 of FLS2-BAK1-flg22 Receptor Complex - From AlphaFold 2.1.3**



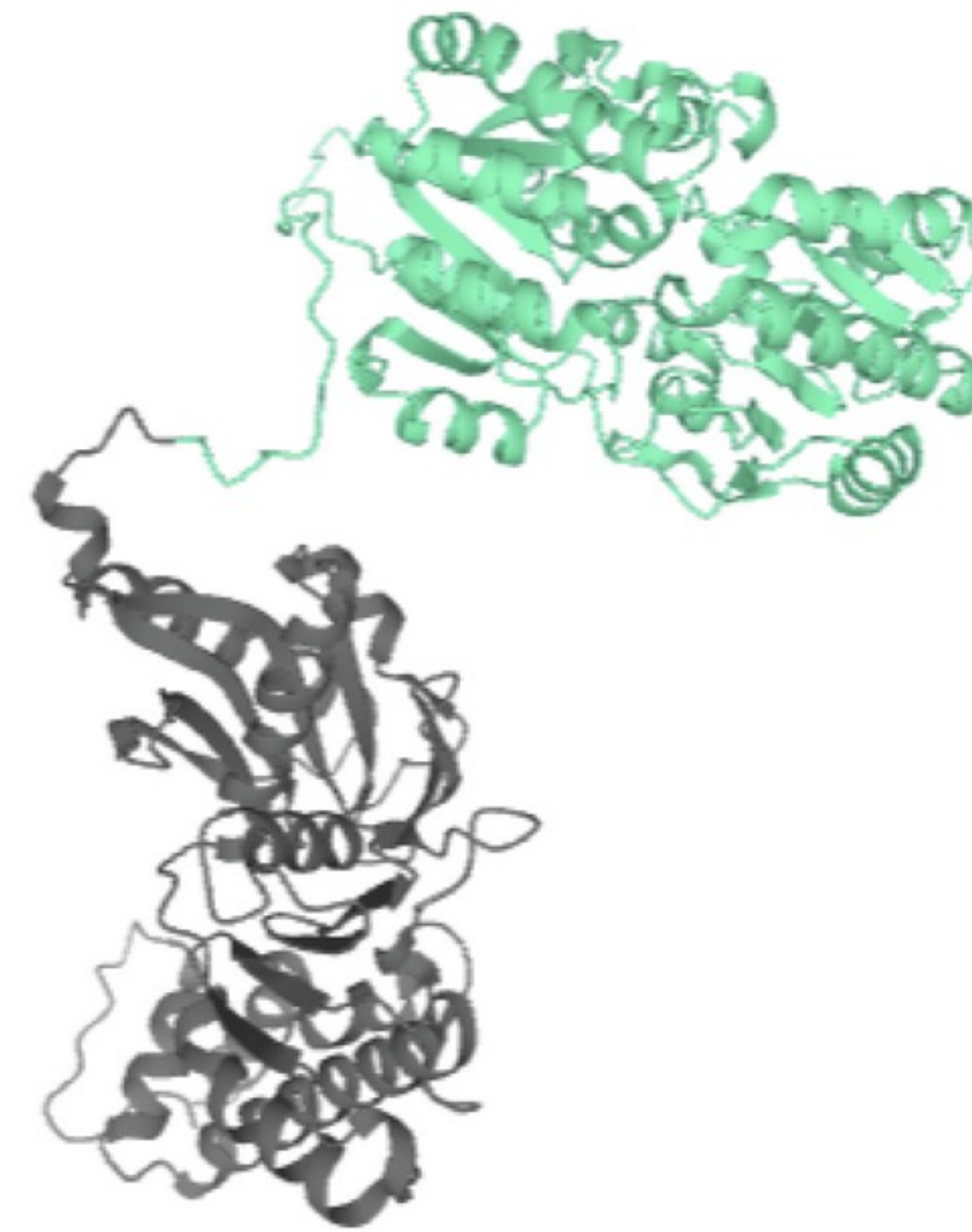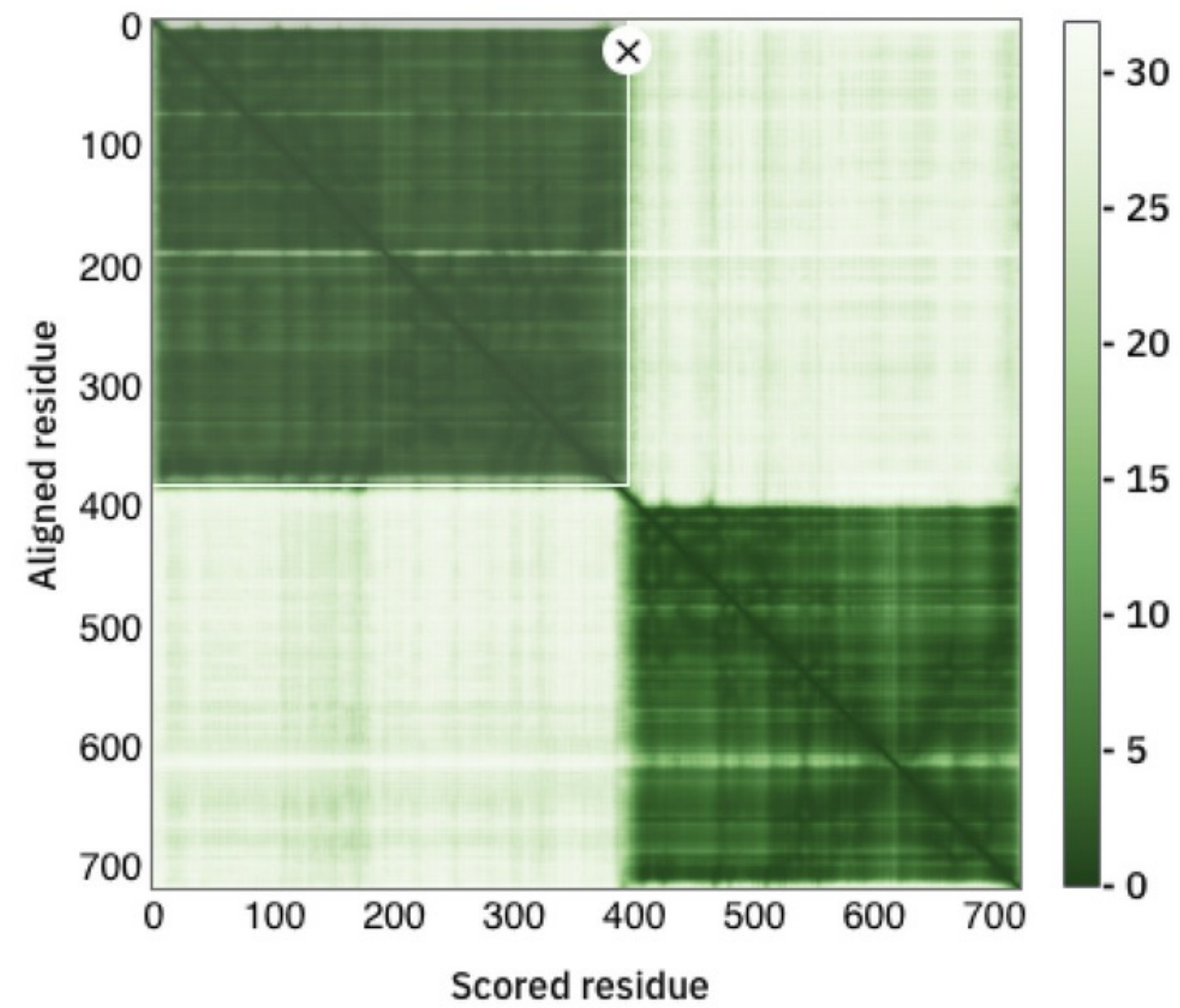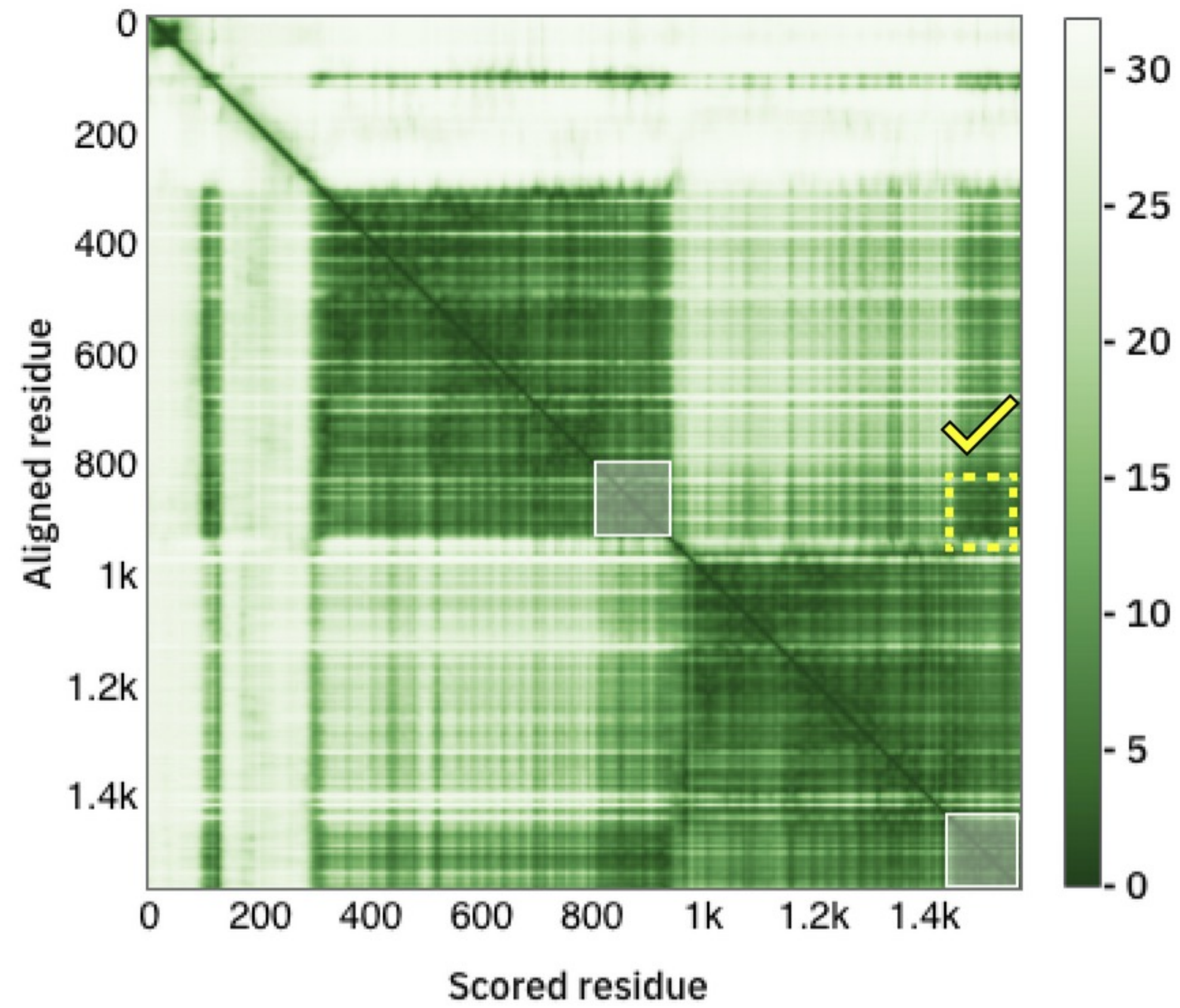**This is the "hardest" plot to interpret and takes practice**
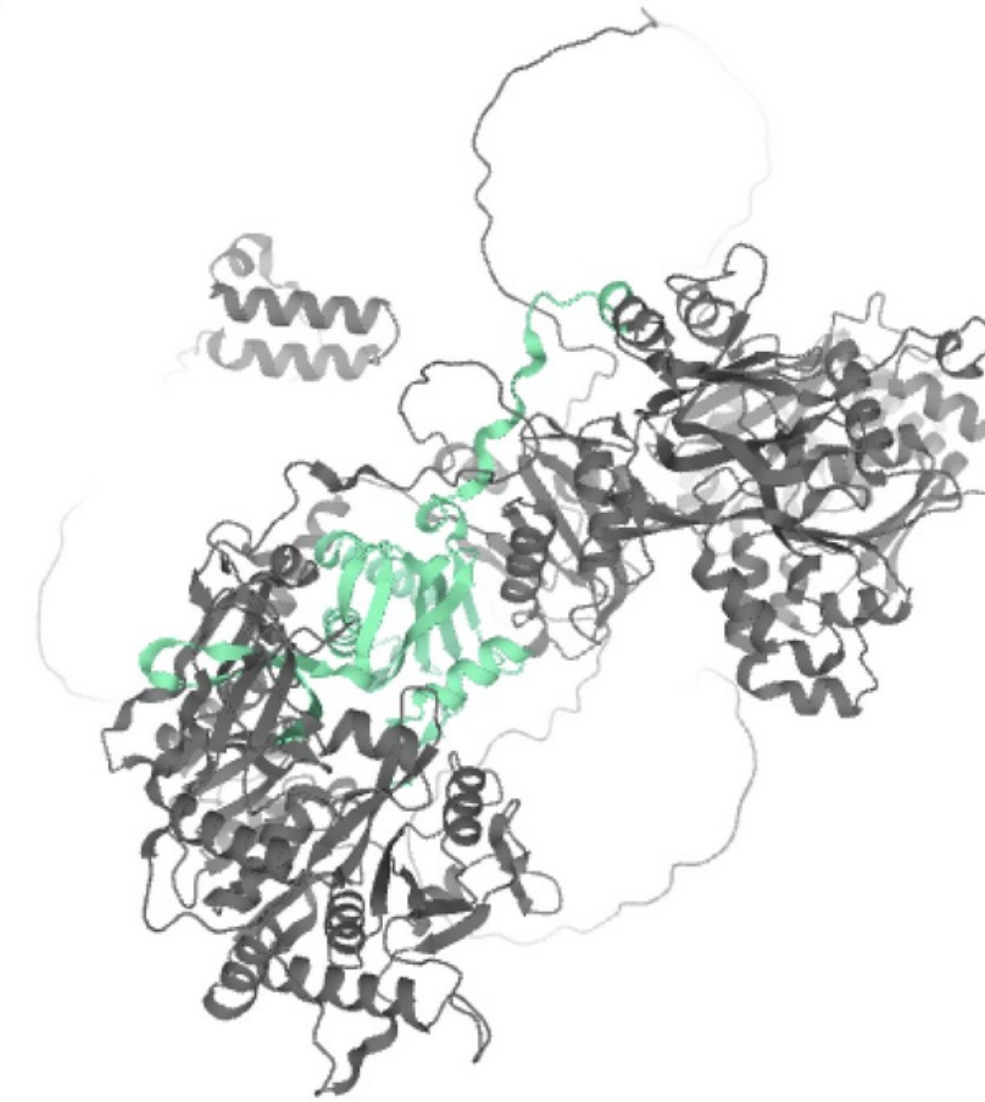
Photo: AlphaFold2 Database, PAE Tutorial
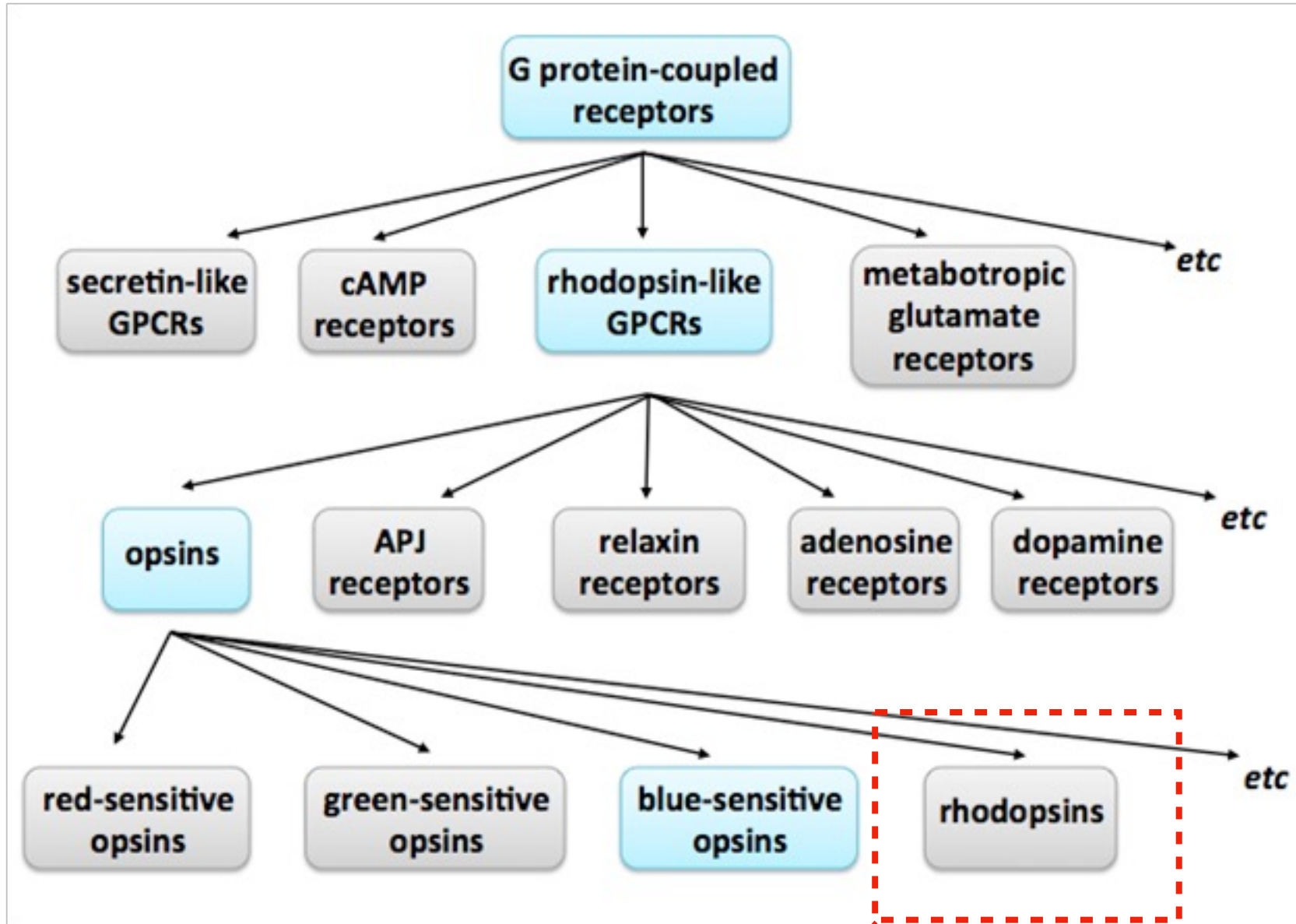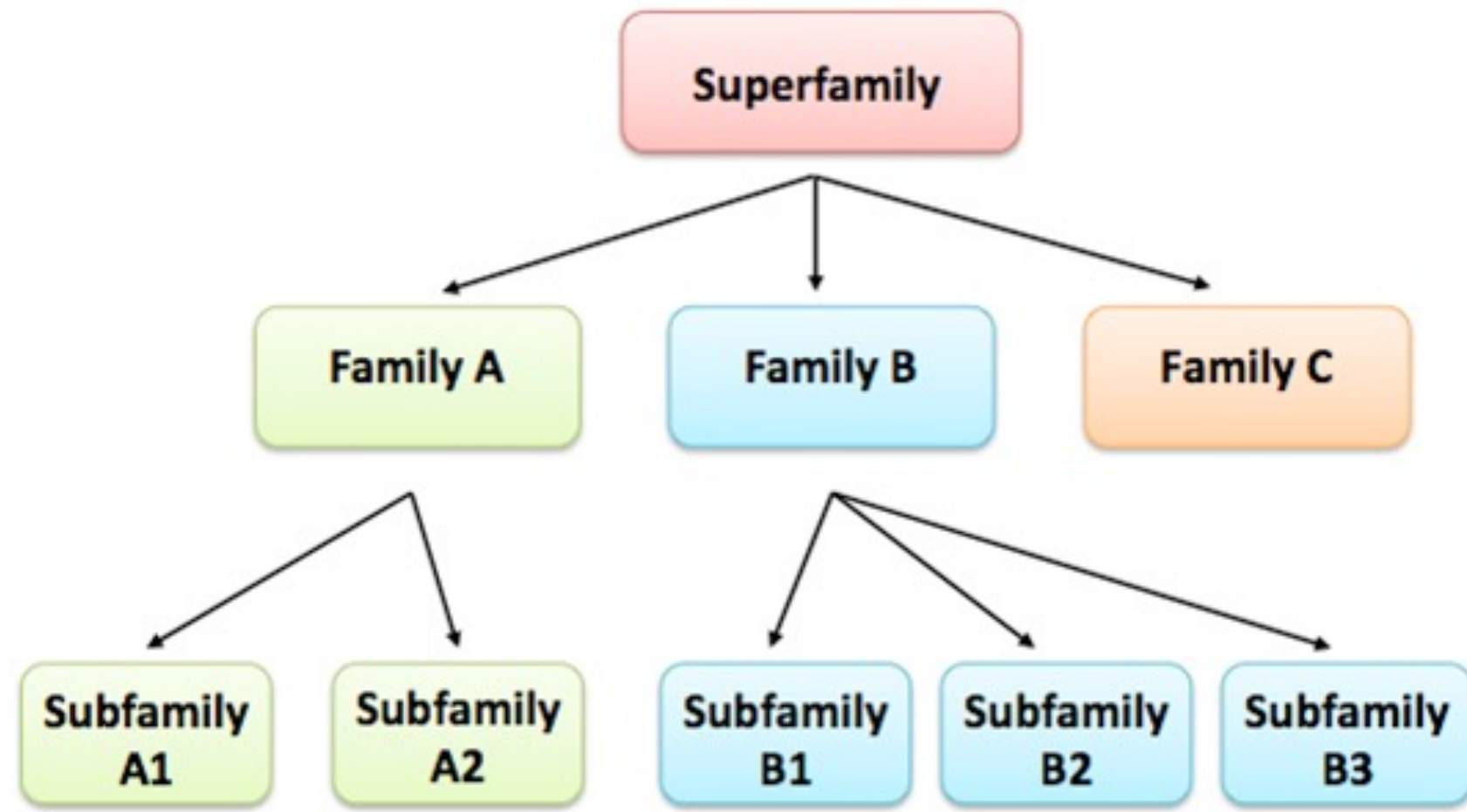
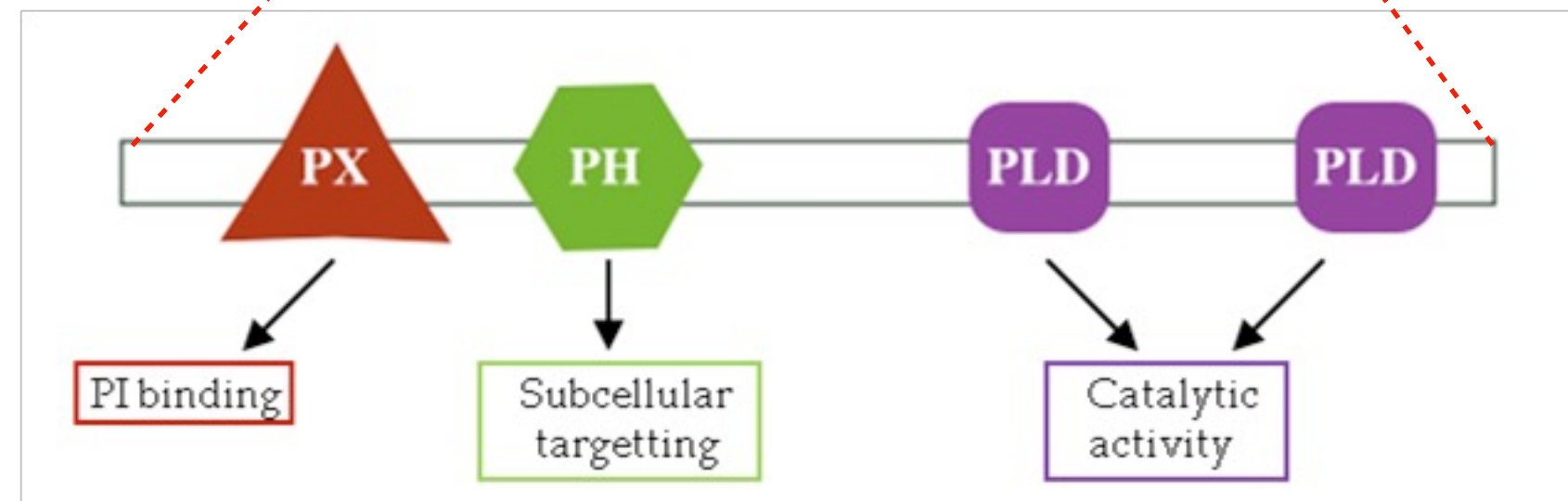Photo: AlphaFold2 Database, PAE Tutorial

# Using Protein Conserved Domains to Probe Function



**We can leverage this classification system to probe potential function of our protein sequences.**

Photos: EMBL-EBI Rescources, Protein Classification

# Combining Predicted Structures & Conserved Domain For Sequence Annotation

- So far you have….

  - predicted 3D structure

  - analyzed sequence for conserved domains

- Combine all of this information

  - structure confidence

  - predicted protein function

- Can you think of a way we could test our structure for cofactor/substrate binding?

- Conclusions:

  - What metrics do we have for scoring AF2 model confidence?

    - pLDDT (3D structure depedent, R chain placement)

    - pAE (3D structure indepedent, inter-domain placement)

  - What situtations is AF2 evaluated for?

    - Single monomeric, naturally occuring protein chains

  - What is the largest limitation in AF2?

    - Homologous sequence coverage