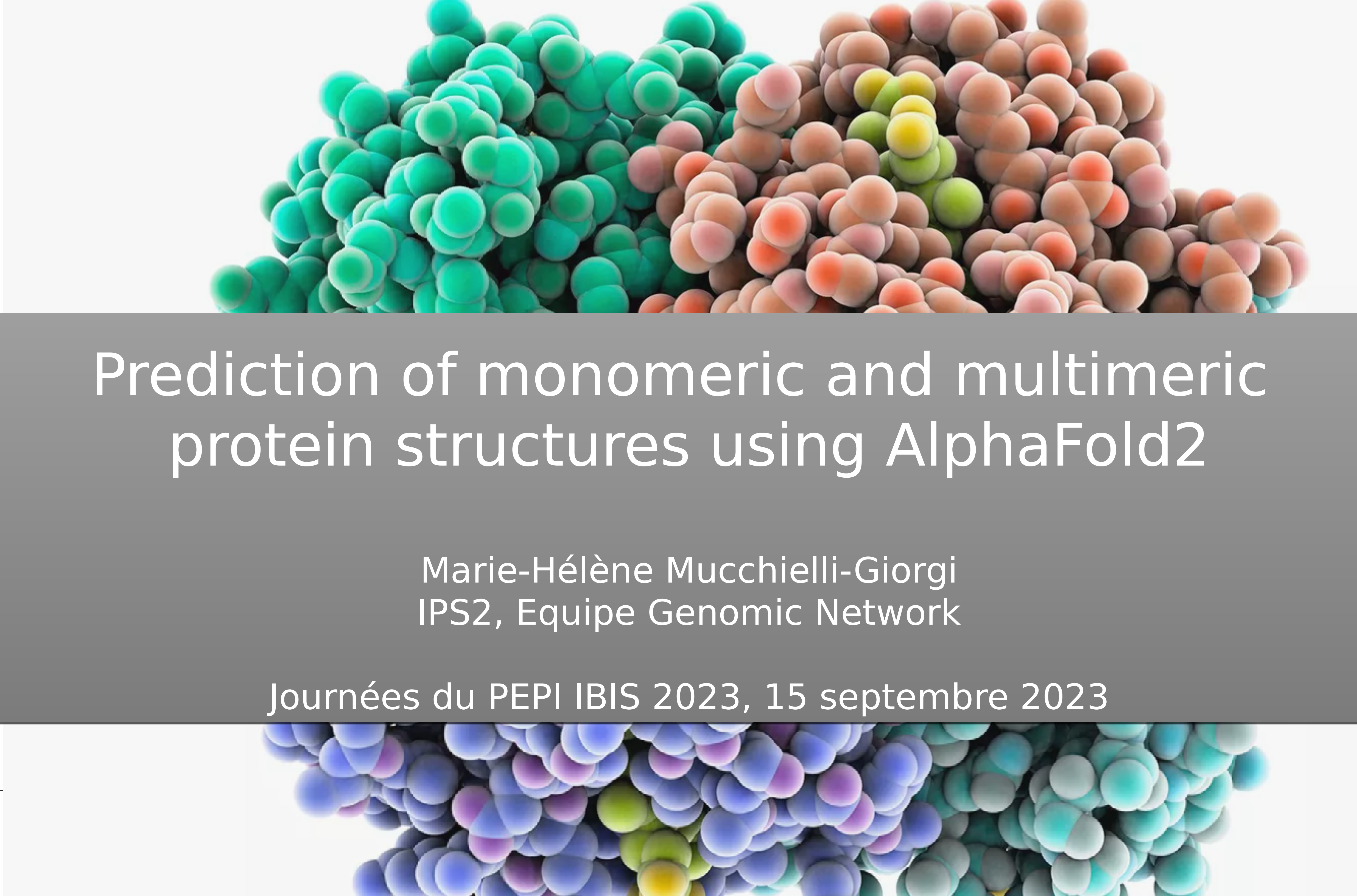
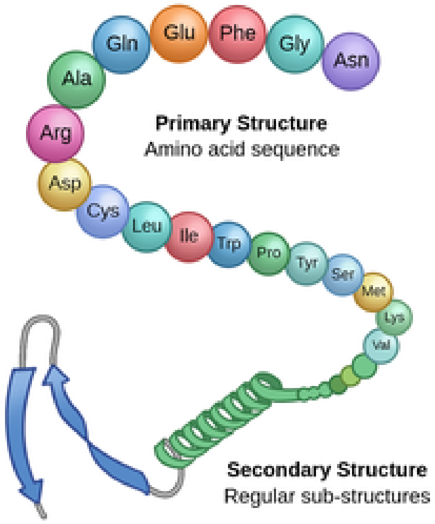


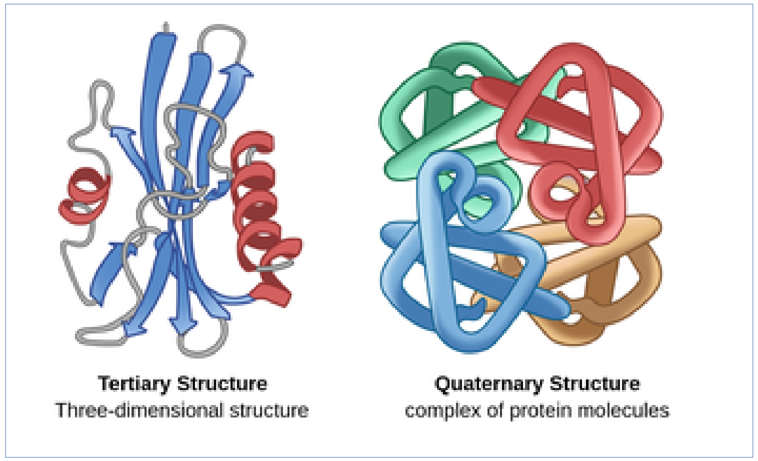
Prediction of monomeric and multimeric protein structures using AlphaFold2

Marie-Hélène Mucchielli-Giorgi
IPS2, Equipe Genomic Network

Journées du PEPI IBIS 2023, 15 septembre 2023

From sequence to 3D structure





Experimental determination :

- X-ray crystallography
- Nuclear Magnetic Resonance (NMR)
- Cryo-electron microscopy

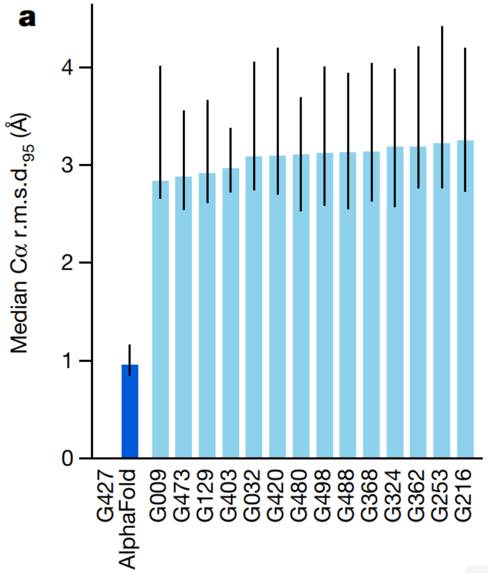
Slow and costly

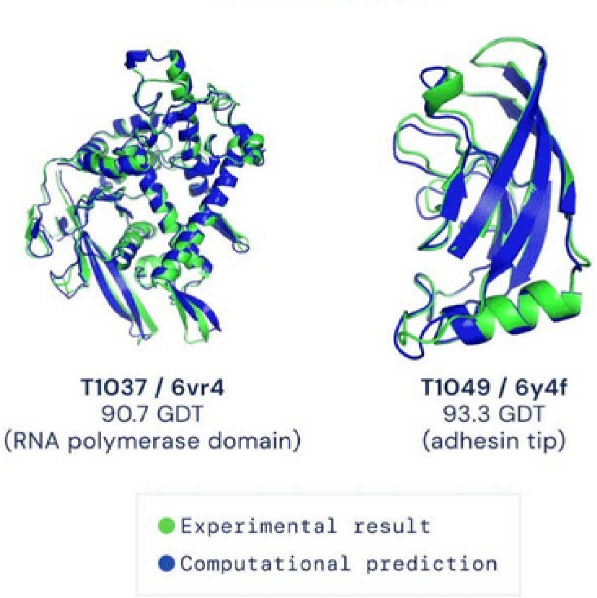
in silico **prediction**

Alphafold2 : a revolutionary approach for structure prediction

- May_August 2020 : Casp14 (14th Critical Assessment of Techniques for Protein Structure Prediction)

AlphaFold2, a tool developed by DeepMind, beats the competition hands down.





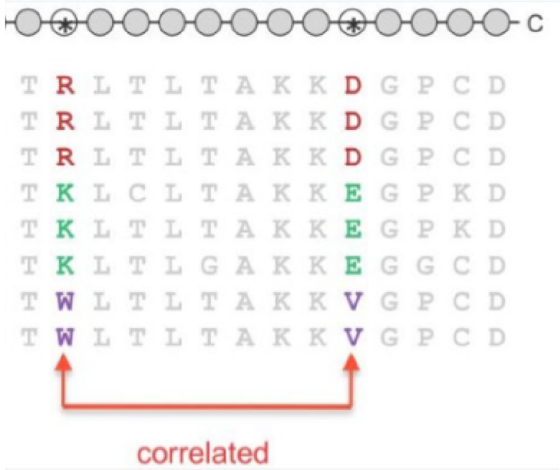
Highly accurate protein structure prediction with AlphaFold
https://www.nature.com/articles/s41586-021-03819-2

Alphafold2 : a revolutionary approach for monomeric structure prediction

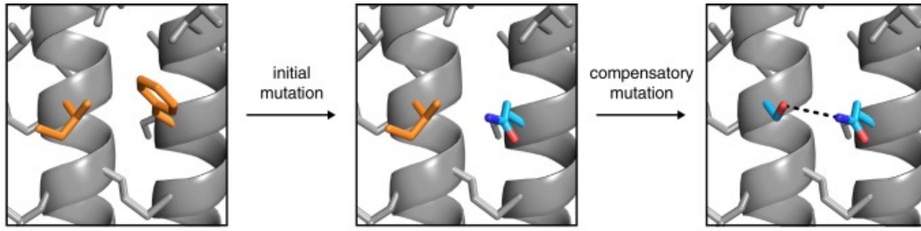
Search for sequences similar to input sequence in other species

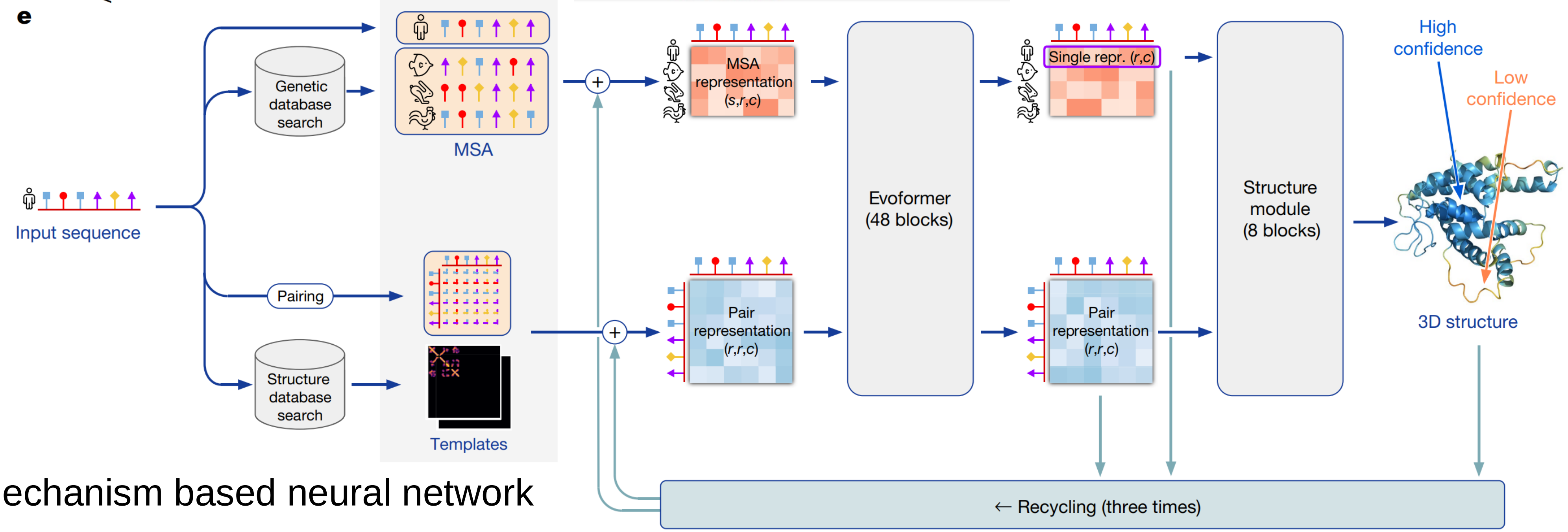
Multiple alignment of sequences (MSA)

Prediction based on amino acid co-variations



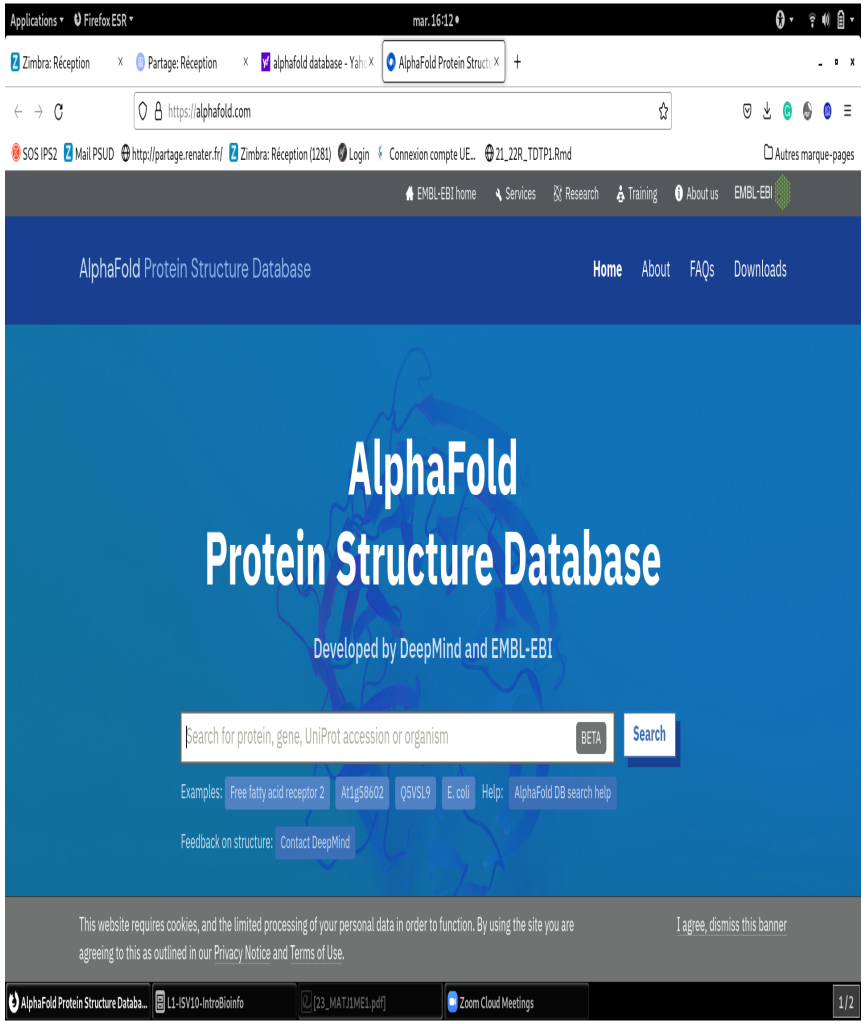
T R L T L T A K K D G P C D
T R L T L T A K K D G P C D
T R L T L T A K K D G P C D
T K L C L T A K K E G P K D
T K L T L T A K K E G P K D
T K L T L G A K K E G G C D
T W L T L T A K K V G P C D
T W L T L T A K K V G P C D




e
Input sequence
Genetic database search
MSA
Pairing
Structure database search
Templates
+
MSA representation (s,r,c)
Pair representation (r,r,c)
Evoformer (48 blocks)
Single repr. (r,c)
Pair representation (r,r,c)
Structure module (8 blocks)
High confidence
Low confidence
3D structure
← Recycling (three times)

Attention mechanism based neural network

AlphaFold Protein structure DataBase (https://alphafold.com)



This website requires cookies, and the limited processing of your personal data in order to function. By using the site you are agreeing to this as outlined in our Privacy Notice and Terms of Use.

Organism	Number of protein structures (without isoforms)	Number of protein structures (including isoforms)
Arabidopsis thaliana	132903	136433
Phaseolus vulgaris	31910	30501
Medicago trunctula	88614	90399
Brachypodium distachyon	44494	45408
Triticum aestivum	139466	168967
Oryza sativa	145703	148907

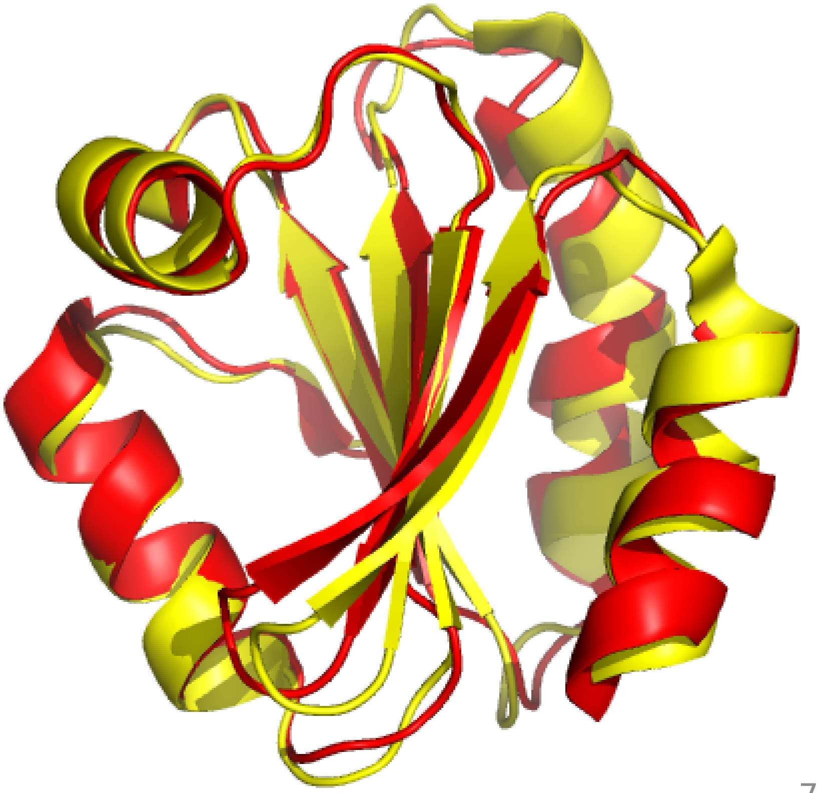
How much confidence can we place in prediction?

Measures of prediction quality produced by Alphafold

- pTM (predicted Template Modeling score)
- pLDDT (predicted Local Distance Difference Test)
- PAE (Predicted Alignment Error)
- ipTM (interface predicted Template Modeling score)

pTM (predicted Template Modeling score)

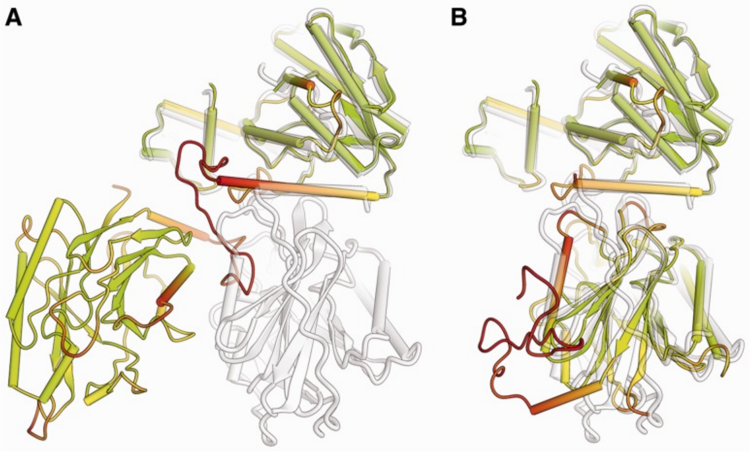
- The TM score measures the difference between the experimental structure and the predicted structure, normalized by protein length.
- Varies from 0 to 1 (1 being a perfect match)
- Le pTM is a predicted TM score



pLDDT (predicted Local Distance Difference Test)

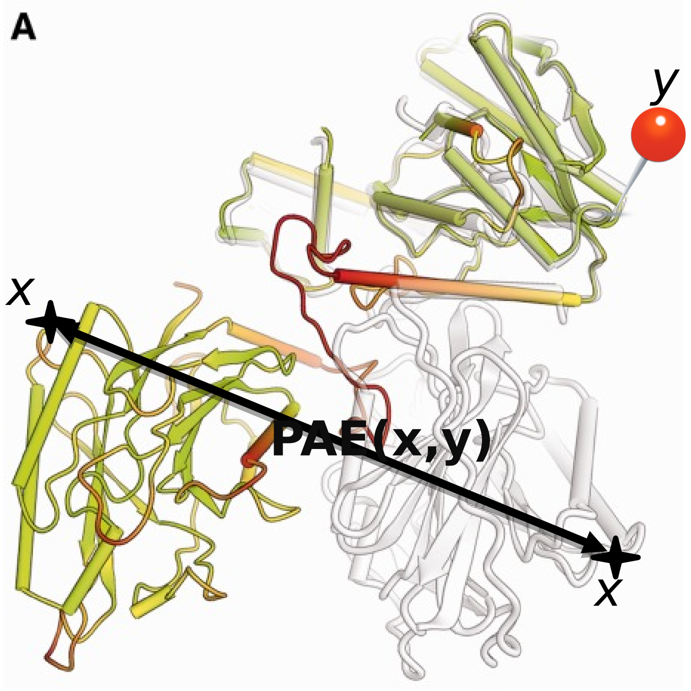
- LDDT locally compares experimental structure and prediction
- Gives a measure of the quality of the prediction of each amino acid's environment
- The pLDDT is a predicted LDDT.

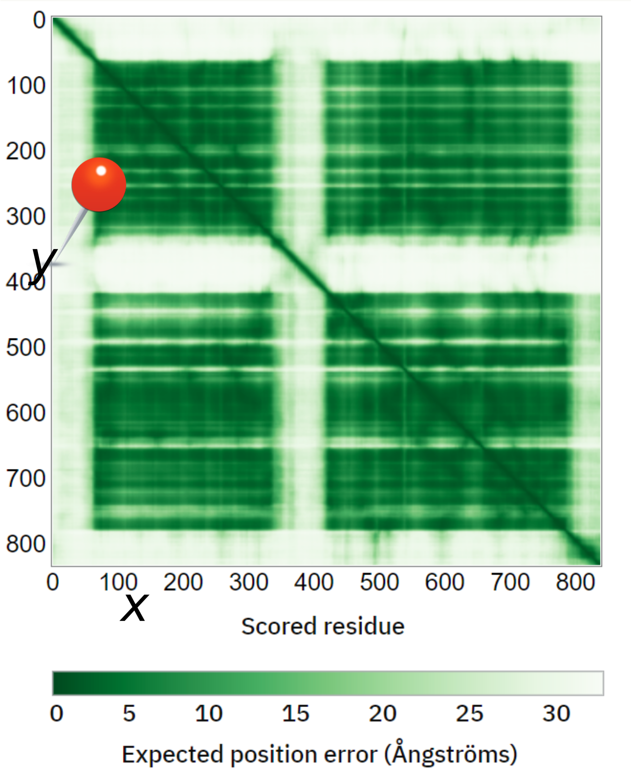
pLDDT > 90: regions modeled with high precision
pLDDT between 70 and 90: well-modeled regions
pLDDT between 50 and 70: regions predicted with low accuracy
pLDDT less than 50: strong predictor of disordered regions



PAE (Predicted Alignment Error)

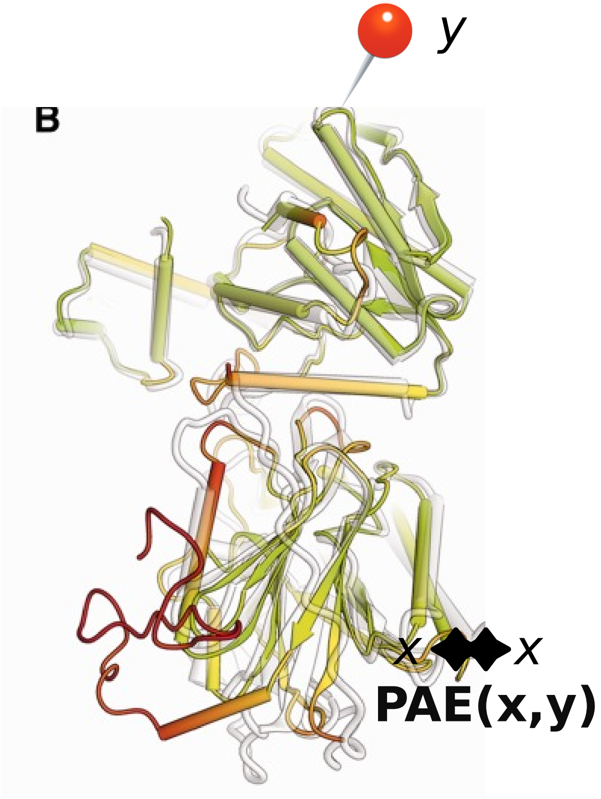
Indicates, **for each *x* position**, the difference between the experimental structure and the predicted structure **when the two structures are aligned at the *y* position**.

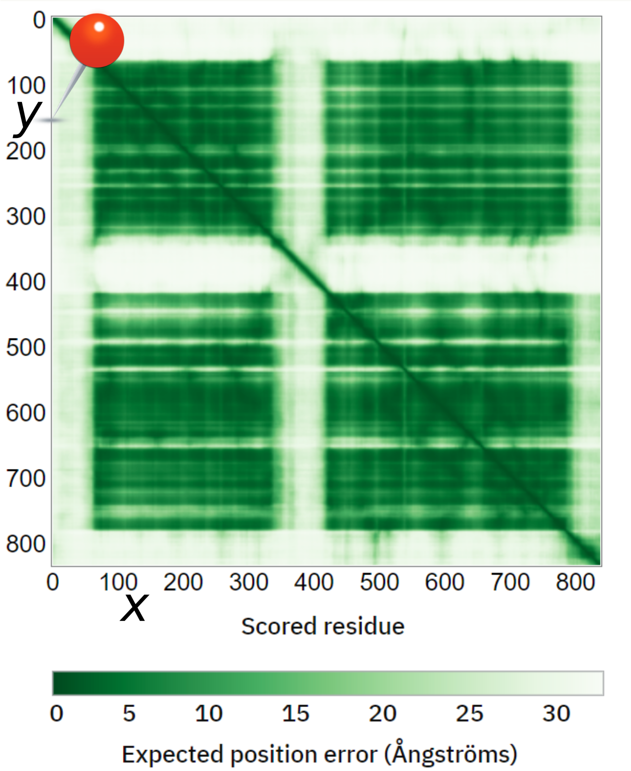




PAE (Predicted Alignment Error)

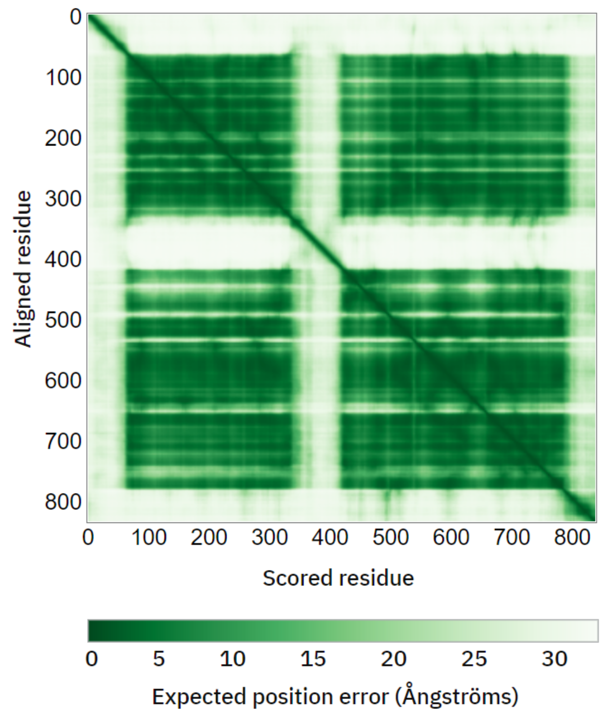
Indicates, **for each *x* position**, the difference between the experimental structure and the predicted structure **when the two structures are aligned at the *y* position**.





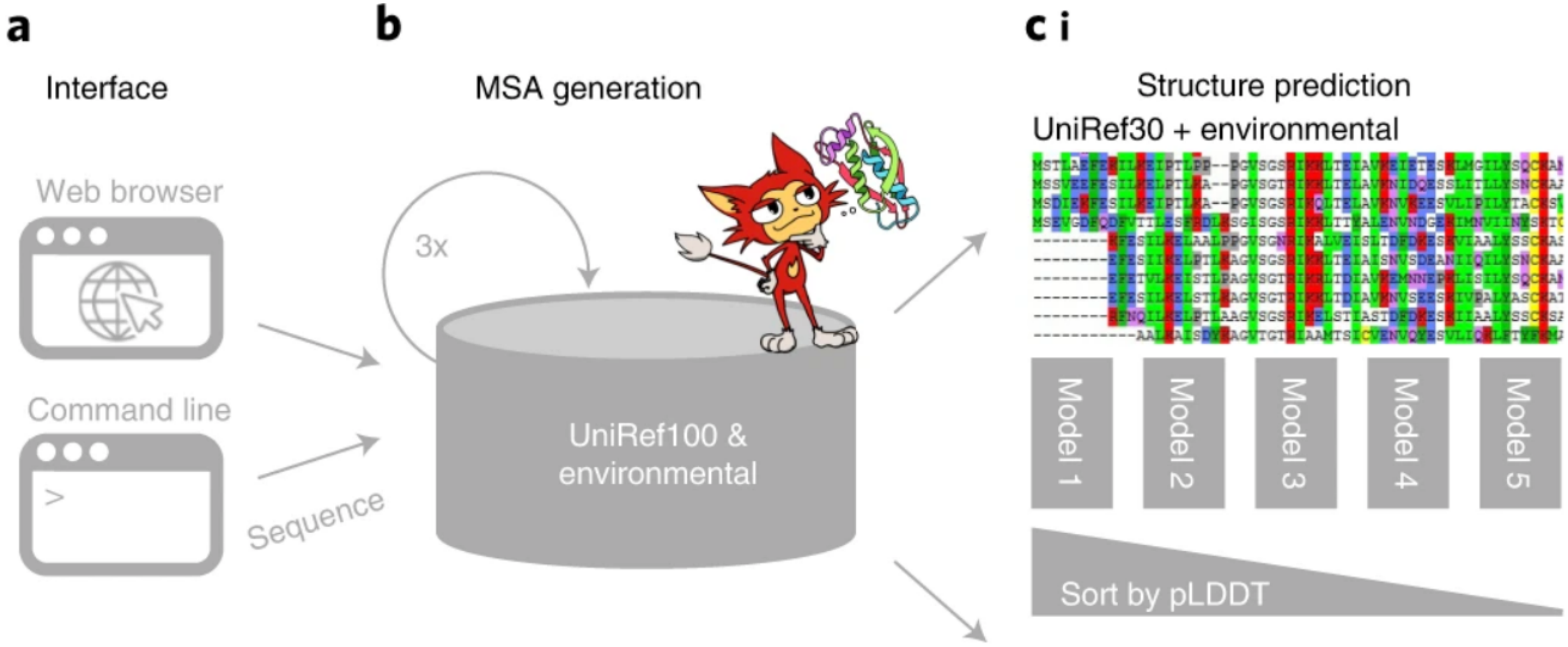
PAE (Predicted Alignment Error)

PAE is a **measure of the quality of the prediction of domains position relative to each othe**r in the predicted structure.



ColabFold : accelerated Alphafold2

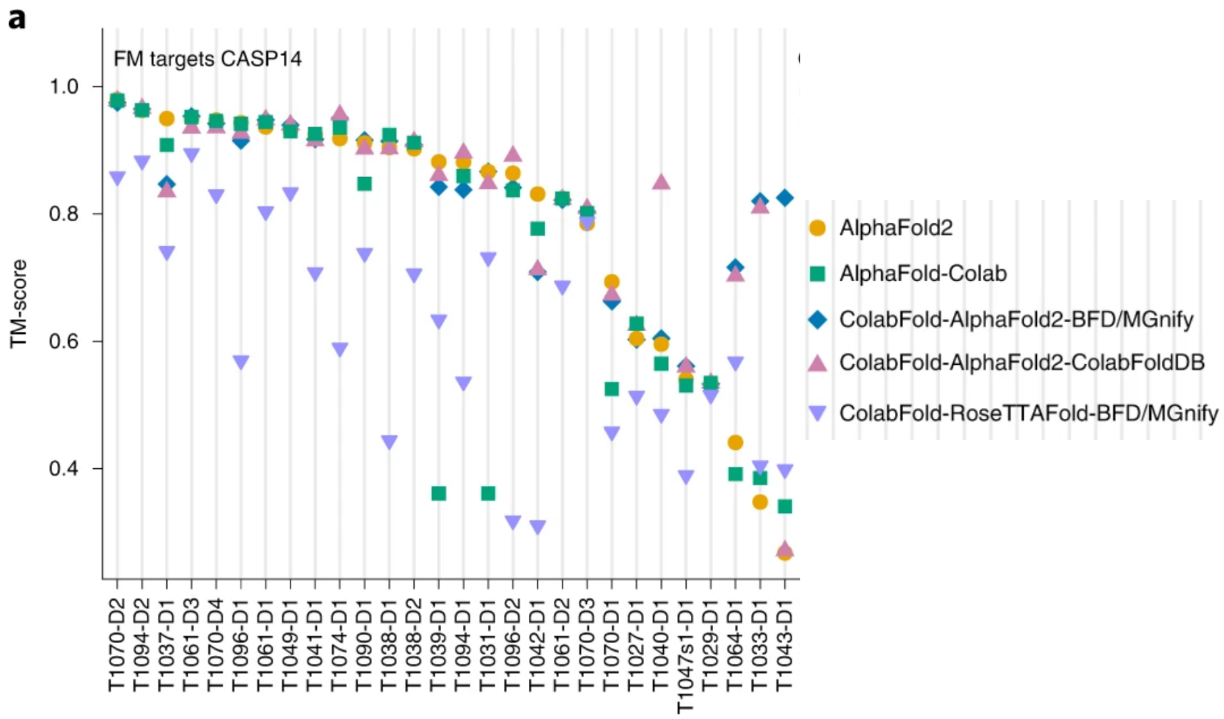
Accelerated MSA generation using the MMseqs2 algorithm on databases where redundancy has been reduced to a minimum

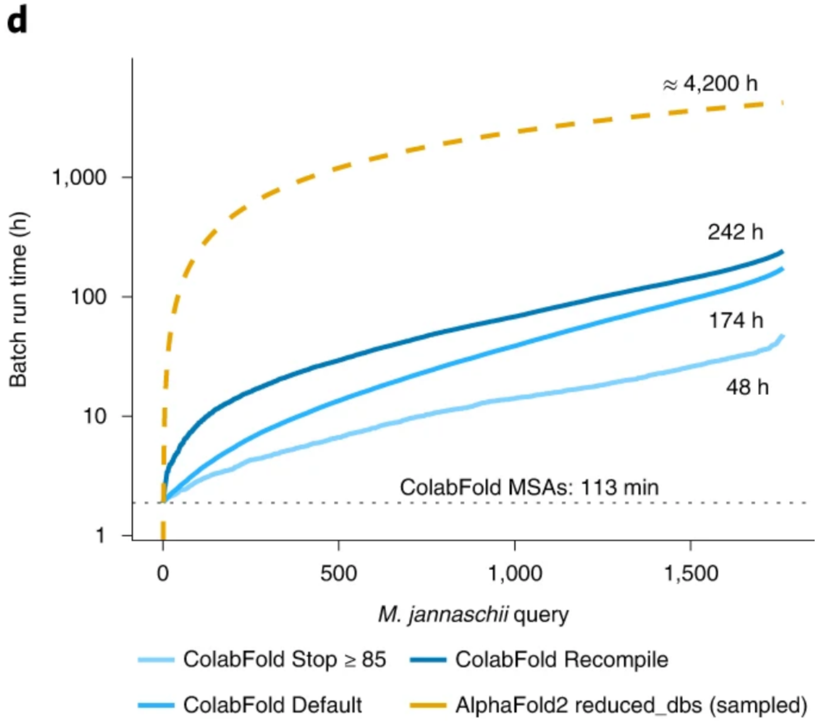


Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: making protein folding accessible to all. Nat Methods. 2022 Jun;19(6):679-682. doi: 10.1038/s41592-022-01488-1

ColabFold : accelerated Alphafold2

Pipeline 40 to 60 times faster with very little loss of quality





Launch AlphaFold2 with Colabfold

- Launching ColabFold online with google colab ici

https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/batch/AlphaFold2_batch.ipynb

- On a cluster: node with graphics card strongly recommended

- Ordering:

- module load colabfold cuda

- Colabfold_batch *input output_path options*

Launch AlphaFold with Colabfold

- Colabfold_batch *input* *output_path* *options*

- Input :
  - A fasta file :
    - Multiple alignment is done on a public server (shared ressources but very fast)
    - To predict a complex separate the protein sequences with ":"
  - A multiple alignment file (*.a3m)
    - Do not use the public server
    - Can be generated without using the public alignement server but very long

Launch AlphaFold with Colabfold

- Colabfold_batch *input* *output_path* *options*

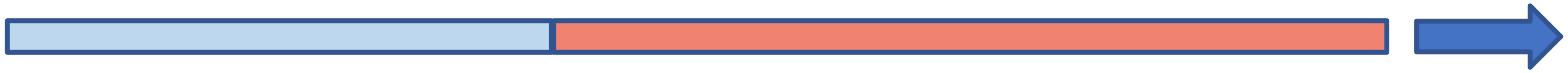
- Options :
  - --stop-at-score (stop the recycling when the score is reach)
  - --num-recycle (max number of recycling)
  - --random-seed (to fix the randomness)
  - --num-models (number of models predicted)
  - --model-type (which version of Alphafold2 should be used)
  - --rank (which metric is used to rank the models)
  - --amber (last step of structure refining)
  - --use-gpu-relax (faster amber calculation by using gpu)
  - --save-all (save all raw data produced by Alphafold pipeline)

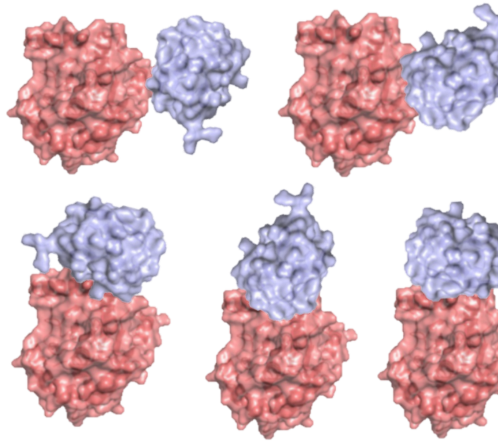
Prediction of multimeric protein complexes with AlphaFold-Multimer

AlphaFold-Multimer = AlphaFold2 trained specifically for multimeric inputs

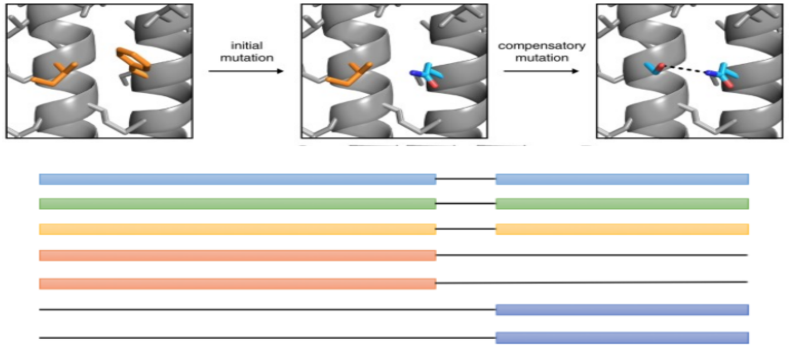
BioRxiv, March 10, 2022

doi: https://doi.org/10.1101/2021.10.04.463034





5 predicted complexes



Paired MSA

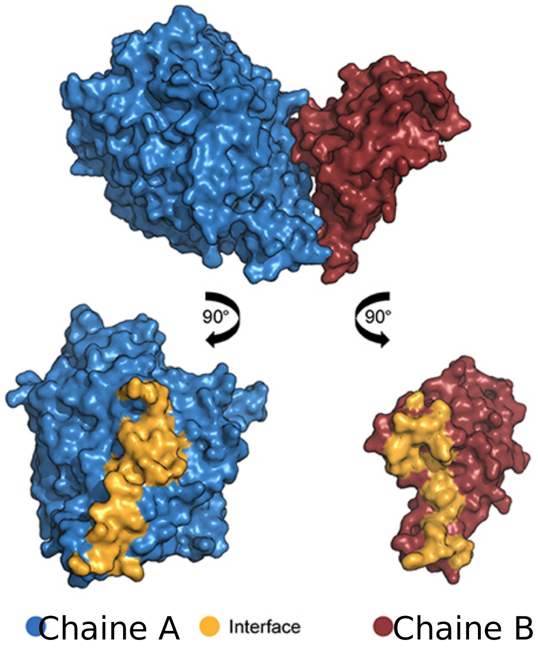
Figure S2. Diagram for paired multiple sequence alignments.

ipTM (interface predicted Template Modeling score)

- Measurement of the quality of protein-protein interface prediction

- Alignment of the experimental structure and the predicted structure on residue i belonging to the interface. Calculation of TM score using residues not belonging to the same chain at the interface.

- The ipTM is a predicted value.

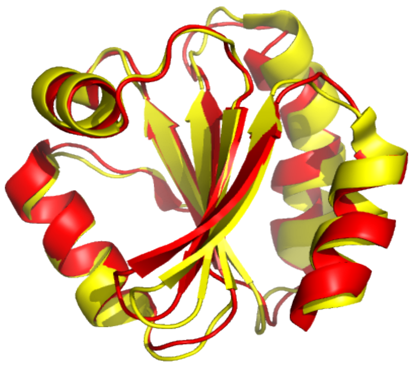


Ranking the predicted complexes

$$\text{model confidence} = 0.8 \cdot \text{ipTM} + 0.2 \cdot \text{pTM}$$

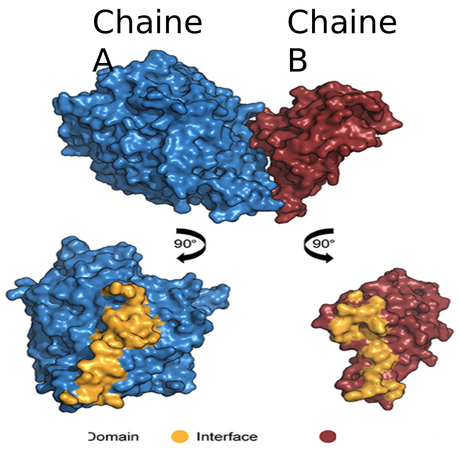
- pTM (predicted Template Modeling score)

  TM : deviation between experimental and predicted structures, normalized to be in the range [0, 1]



- ipTM (interface predicted Template Modeling score)

  ITM : Alignement of experimental and predicted structures on residue *i* of the interface. Compute TM score using residues of the interface of other chain



Results of AlphaFold-Multimer

- DockQ score = protein-protein docking model quality measure in the range [0, 1]
  = combination of two terms :
  - The fraction of native interfacial contacts preserved in the interface of the predicted complex.
  - The RMS deviation calculated for the backbone of the shorter chain of the model after superposition of the longer chain

- $0 \leq$ DockQ < 0.23 : Incorrect prediction
  $0.23 \leq$ DockQ < 0.49 : Acceptable prediction
  $0.49 \leq$ DockQ < 0.80 : Medium prediction
  $0.80 \leq$ DockQ : High prediction

- Prediction results :
  - 67 % (resp. 23%) of heteromeric interfaces are correctly (resp. High accuratly) predicted
  - 69% (resp. 34%) of homomeric interfaces are correctly (resp. High accuratly) predicted
- High computing time consuming