ColabFold

Making protein folding accessible to all

github.com/sokrypton/ColabFold

ColabFold - Making protein folding accessible to all, M. Mirdita et al., (2021), *biorxiv*

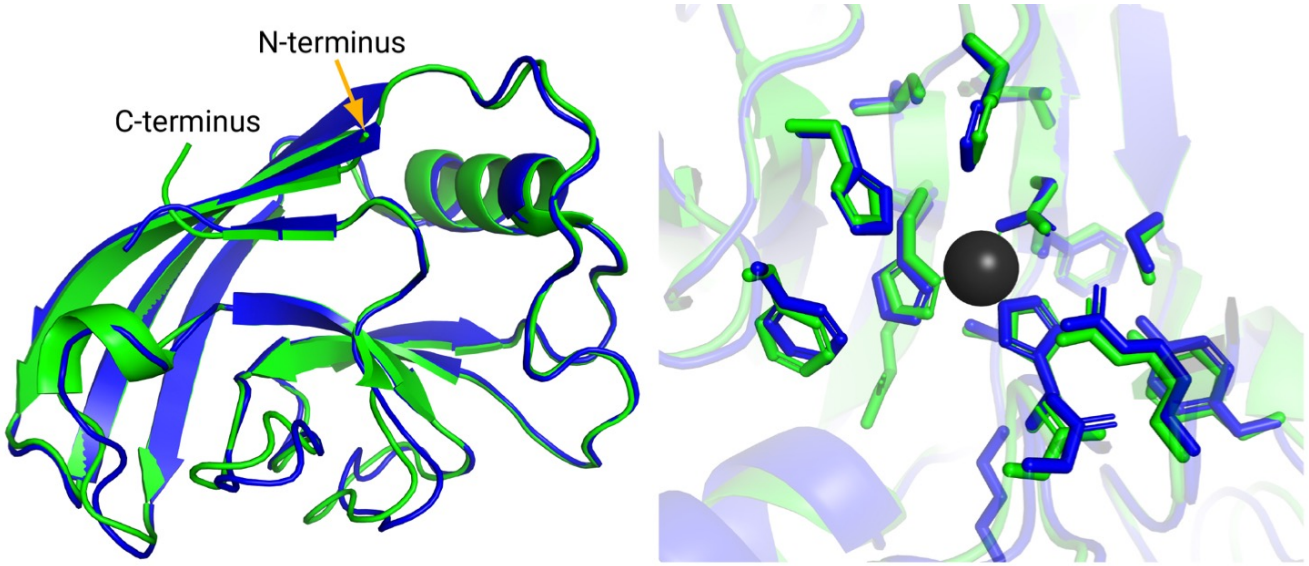
Milot Mirdita
Söding Lab

@milot_mirdita

AlphaFold2 will revolutionize protein bioinformatics

"Everything that relies on a protein sequence, we can now do with protein structure" Mohammed AlQuraishi, Columbia U.

By the end of this year, EMBL EBI will hold structural models of >100 million proteins



DEEPMIND'S AI PREDICTS STRUCTURES FOR A VAST TROVE OF PROTEINS

AlphaFold neural network produced 'transformative' database of more than 350,000 structures.

Article | Open Access | Published: 15 July 2021

Highly accurate protein structure prediction with AlphaFold

Martin Steinegger

John Jumper, Richard Evans, [...] Demis Hassabis

Nature **596**, 583–589 (2021) | Cite this article

399k Accesses | **2804** Altmetric | Metrics

Article | Open Access | Published: 22 July 2021

Highly accurate protein structure prediction for the human proteome

Kathryn Tunyasuvunakool, Jonas Adler, [...] Demis Hassabis

Nature **596**, 590–596 (2021) | Cite this article

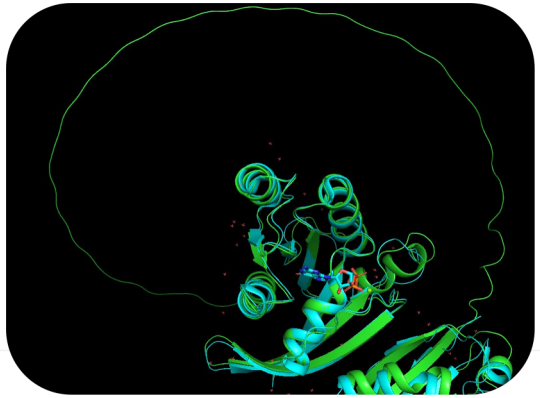
155k Accesses | **8** Citations | **1367** Altmetric | Metrics

AlphaFold2: science happening live online

Yoshitaka Moriwaki @Ag_smith · Jul 19

AlphaFold2 can also predict heterocomplexes. All you have to do is input the two sequences you want to predict and connect them with a long linker.

G-linker!

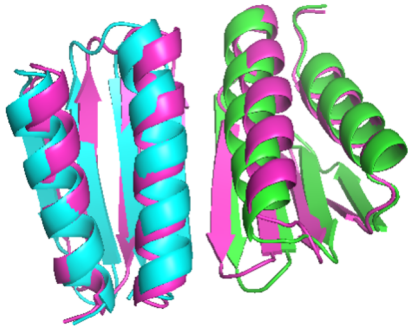


Minkyung Baek
@minkbaek

Don't actually need a G-linker!

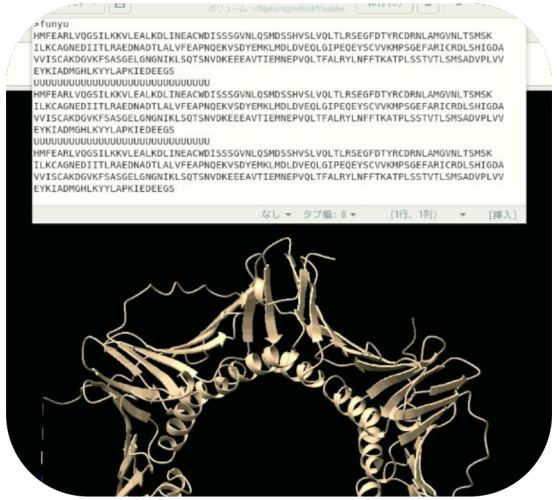
Adding a big enough number for "residue_index" feature is enough to model hetero-complex using AlphaFold (green&cyan: crystal structure / magenta: predicted model w/ residue_index modification).
#AlphaFold #alphafold2

```
# add big enough number to residue index to indicate chain breaks
idx_res = feature_dict['residue_index']
L_prev = 0
# Ls: number of residues in each chain
for L_i in Ls[:-1]:
    idx_res[L_prev+L_i:] += 200
    L_prev += L_i
feature_dict['residue_index'] = idx_res
```



Hiroki Onoda
@onoda_hiroki

Unknown linker may be useful for multimer prediction on the local Alphafold2!!



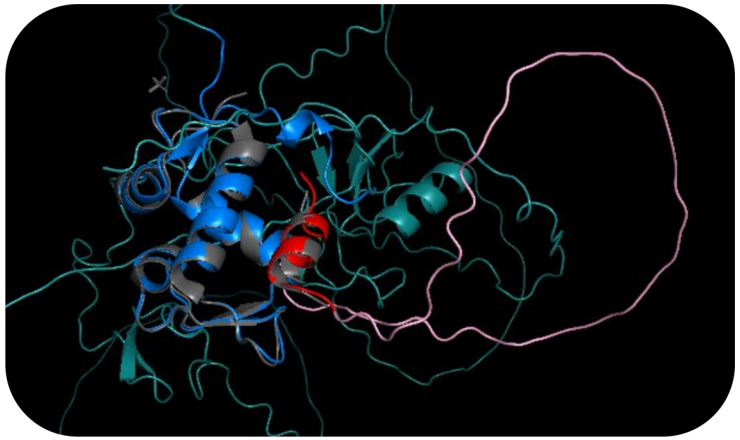
UNK-linker!

大上雅史 | Ohue M 2.5G
@tonets

あ、AlphaFold2でペプチドドッキングでき

Translated from Japanese by Google

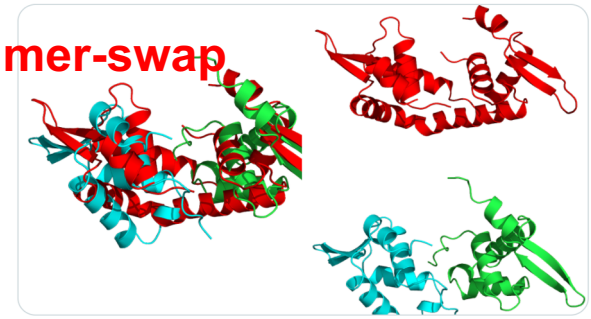
Oh, I was able to dock the peptide with AlphaFold2



Protein-peptide interaction

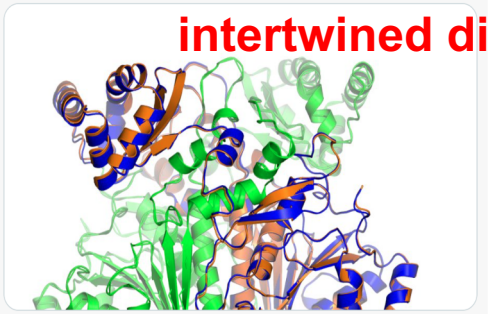
Community finding creative uses for AlphaFold2 in real time

Cesar Ramirez-Sarmiento @cxarramirez · Jul 22
Replying to @cxarramirez @sokrypton and 3 others
OMG😱 the monomers in the predicted "homodimer" of FoxP1 (cyan, green) show very similar orientations when compared to the monomers in the domain-swapped structure (red)🤯



dimer-swap

James Murray @jwm_imperial · Jul 24
Replying to @drpetermoody
There are already notebooks to predict heterodimers and homooligomers. github.com/sokrypton/Cola... For my unpublished intertwined dimer, the monomer and dimer predictions were essentially identical, also matching the crystal structure exactly except for a few rotamers.

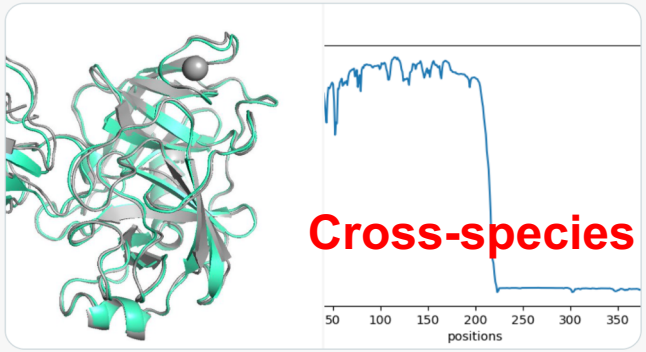


intertwined dimer

Eugene Valkov @eugenevalkov · Jul 26
After days of running #alphafold2, I am still astounded by insights it provides. Here, it predicted mode of peptide binding *completely consistent* with biochemical data and mutagenesis and gave additional clues which we will explore!

consistent w/ biochem data

padhorny @padhorny · Jul 20
Replying to @sokrypton and @minkbaek
Amazing stuff. Seemingly can even do cross-species complexes (at least the strong binders). Here is what it gave me for rcsb.org/structure/1AVX (not the top model though):



Cross-species

Preprints rolling in...

Can AlphaFold2 predict protein-peptide complex structures accurately?

Junsu Ko, Juyong Lee

bioRxiv 2021.07.27.453972; doi: https://doi.org/10.1101/2021.07.27.453972

Harnessing protein folding neural networks for peptide-protein docking

Tomer Tsaban, Julia Varga, Orly Avraham, Ziv Ben-Aharon, Alisa Khramushin, Ora Schueler-Furman

bioRxiv 2021.08.01.454656; doi: https://doi.org/10.1101/2021.08.01.454656

Improved Docking of Protein Models by a Combination of Alphafold2 and ClusPro

Usman Ghani, Israel Desta, Akhil Jindal, Omeir Khan, George Jones, Sergey Kotelnikov, Dzmitry Padhorny, Sandor Vajda, Dima Kozakov

bioRxiv 2021.09.07.459290; doi: https://doi.org/10.1101/2021.09.07.459290

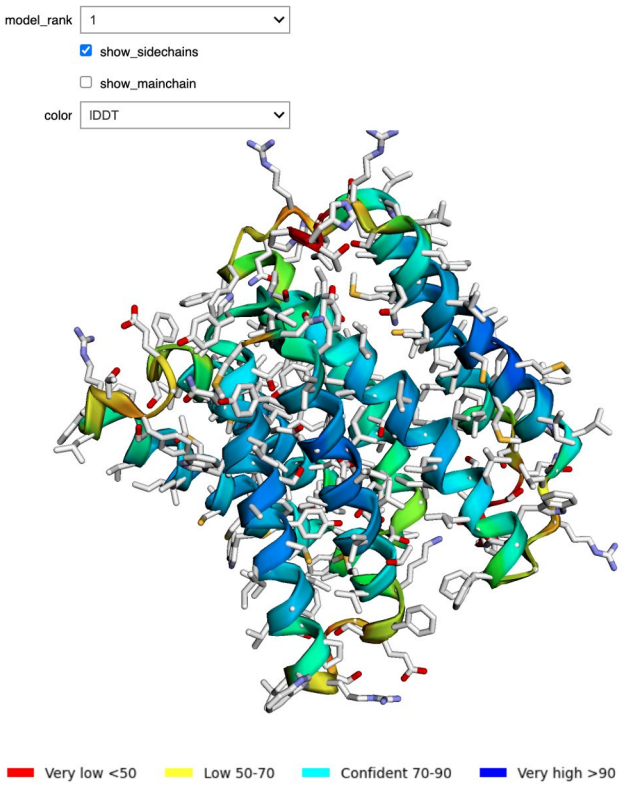
ColabFold is a system to make protein structure prediction available to everyone

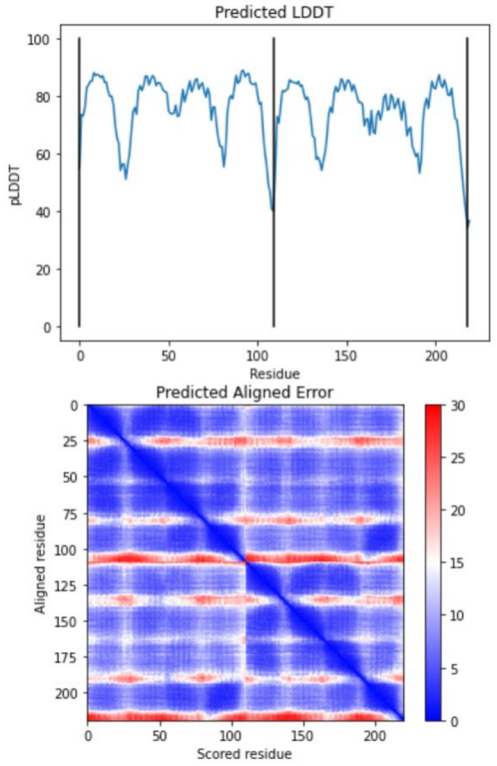
AlphaFold2 + colab + MMseqs2

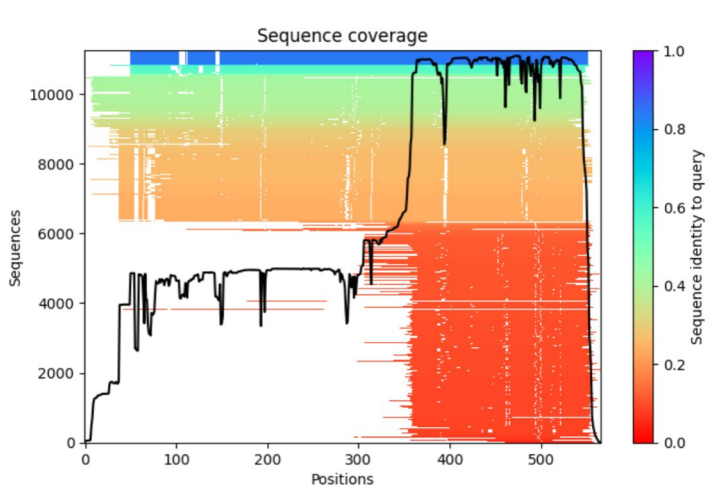
Input protein sequence, then hit `Runtime` -> `Run all`

query_sequence: PIAQIHILEGRSDEQKETLIREVSEAISRSLDAPLTSVRVIITEMAKGHFGIGGELASK

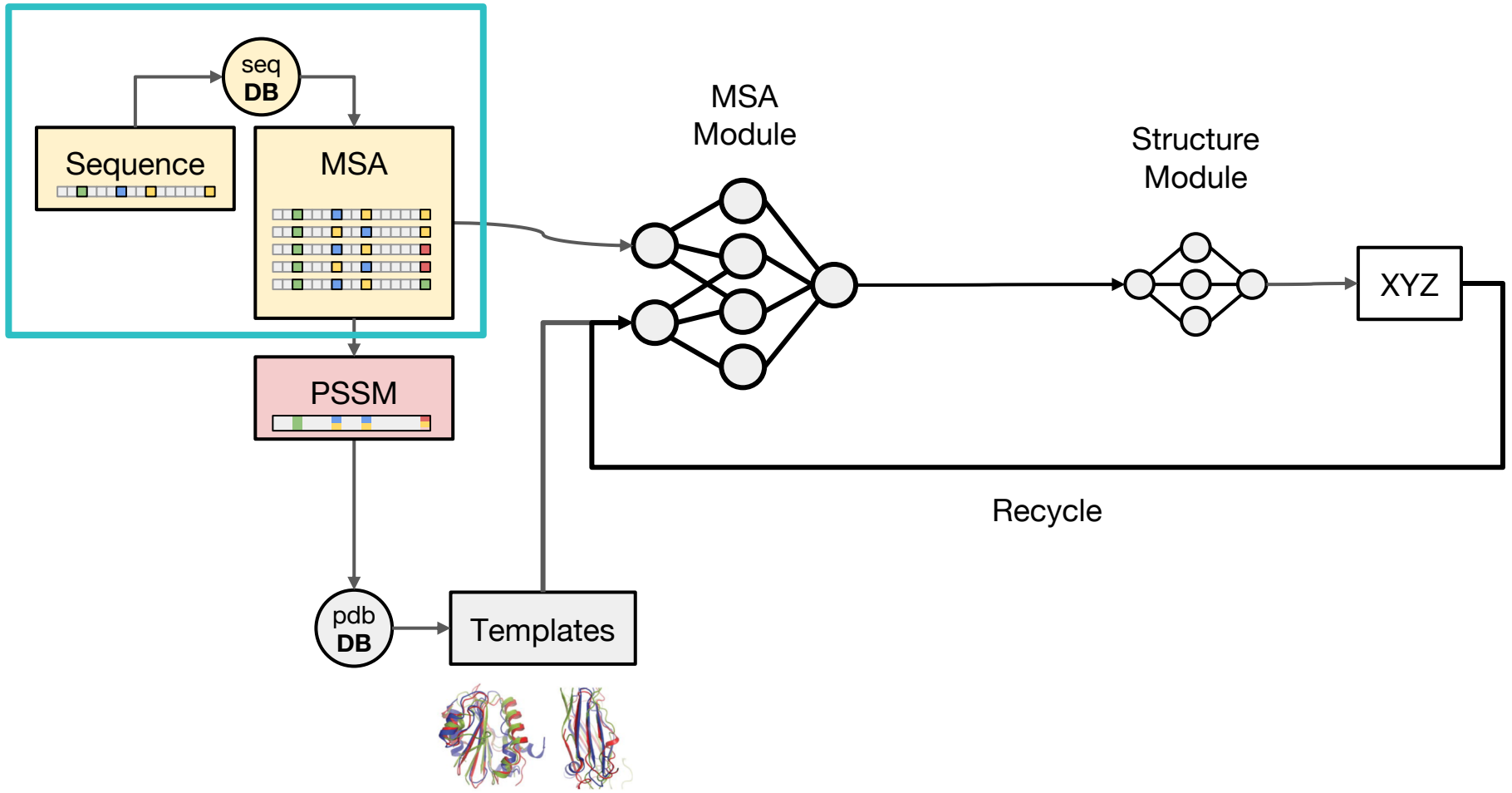
jobname: test



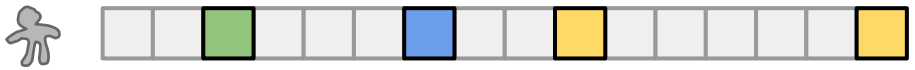




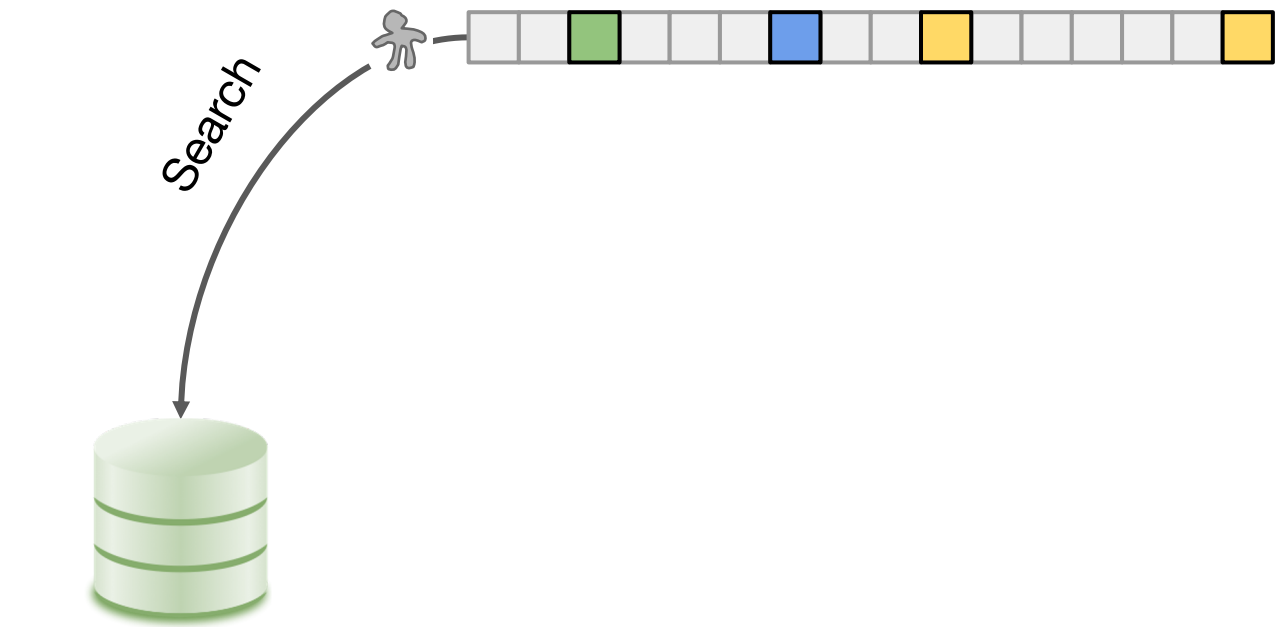
Alphafold2 structure prediction is only as good as the input **multiple sequence alignment (MSA)**



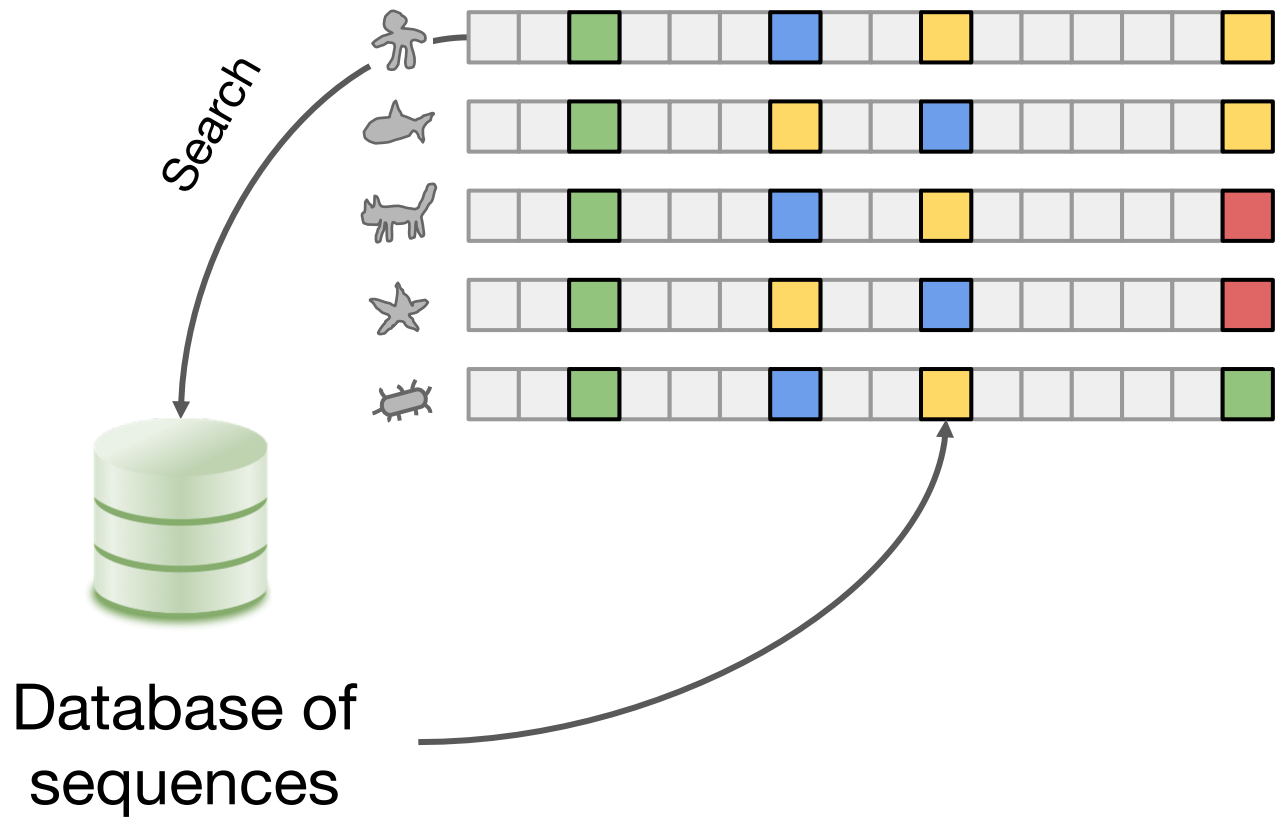
What is a multiple sequence alignment (MSA)?



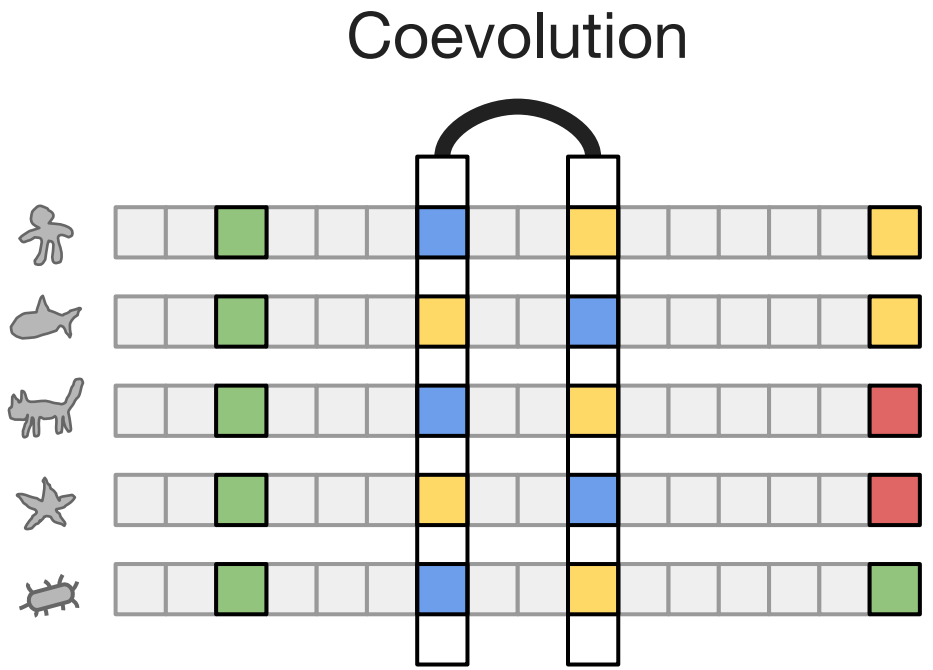
Search against a database of sequences



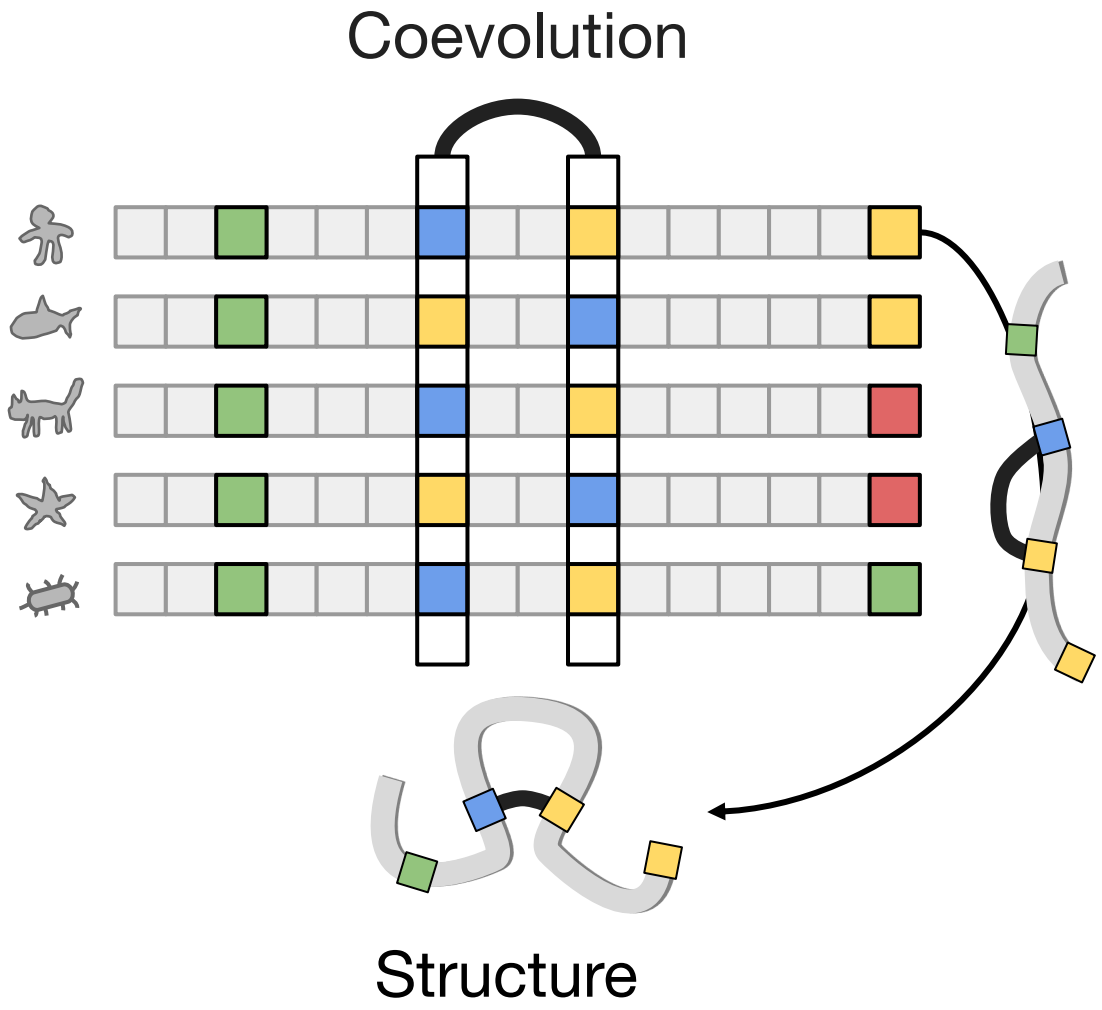
Generate a multiple sequence alignment of homologs proteins



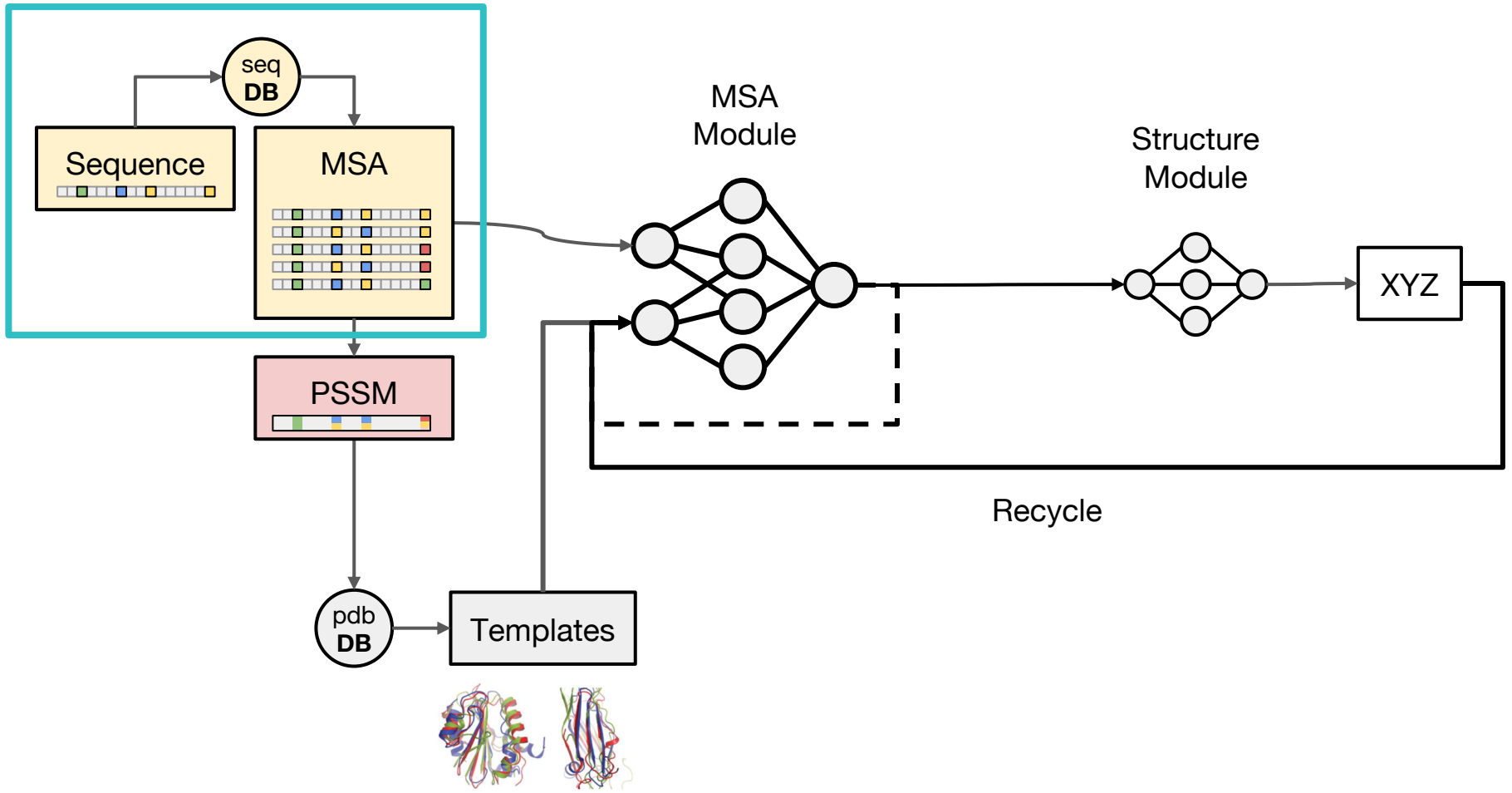
Detect co-evolving residues (columns) in the MSA



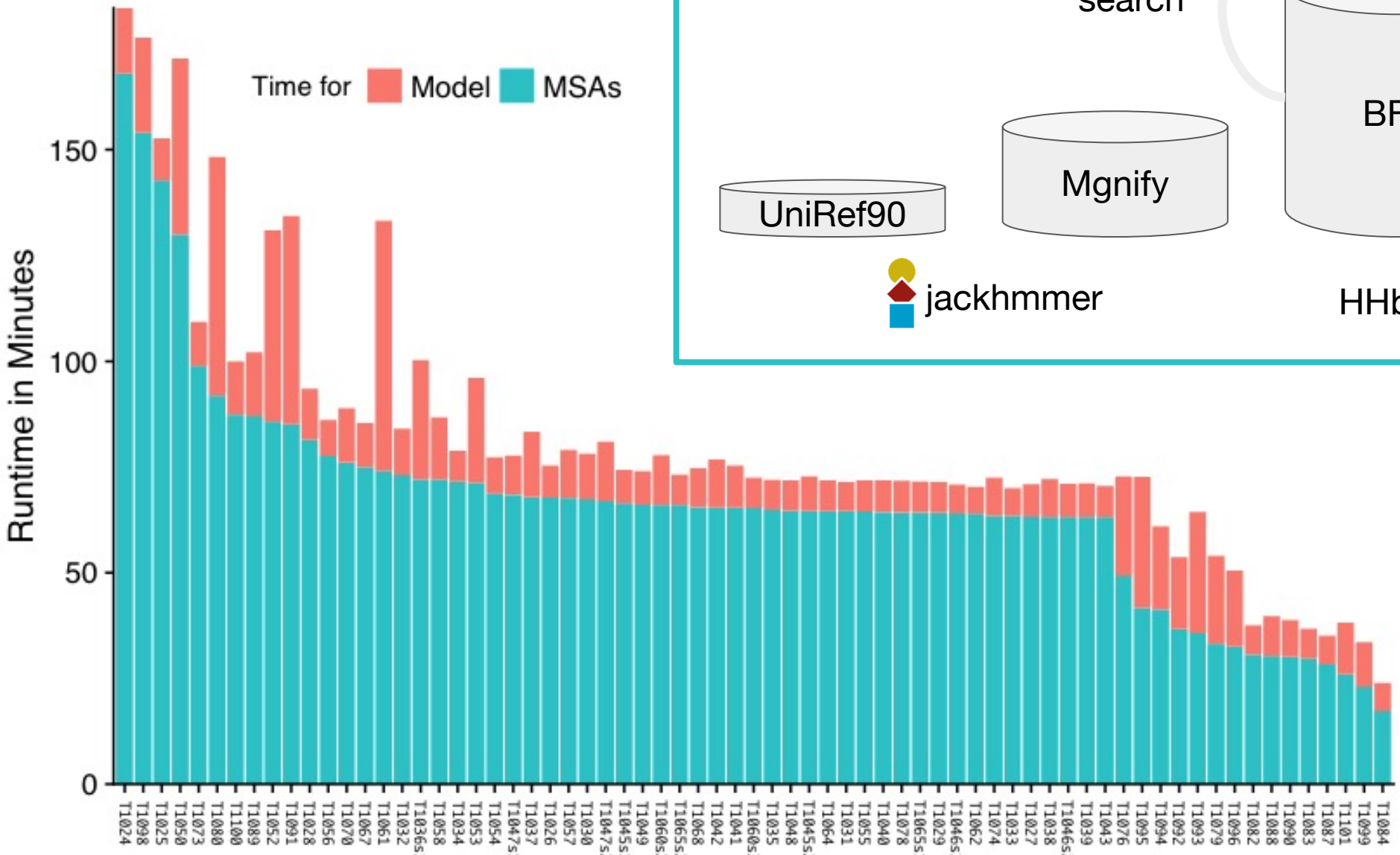
Co-evolving residues have a high chance to be in physical contact



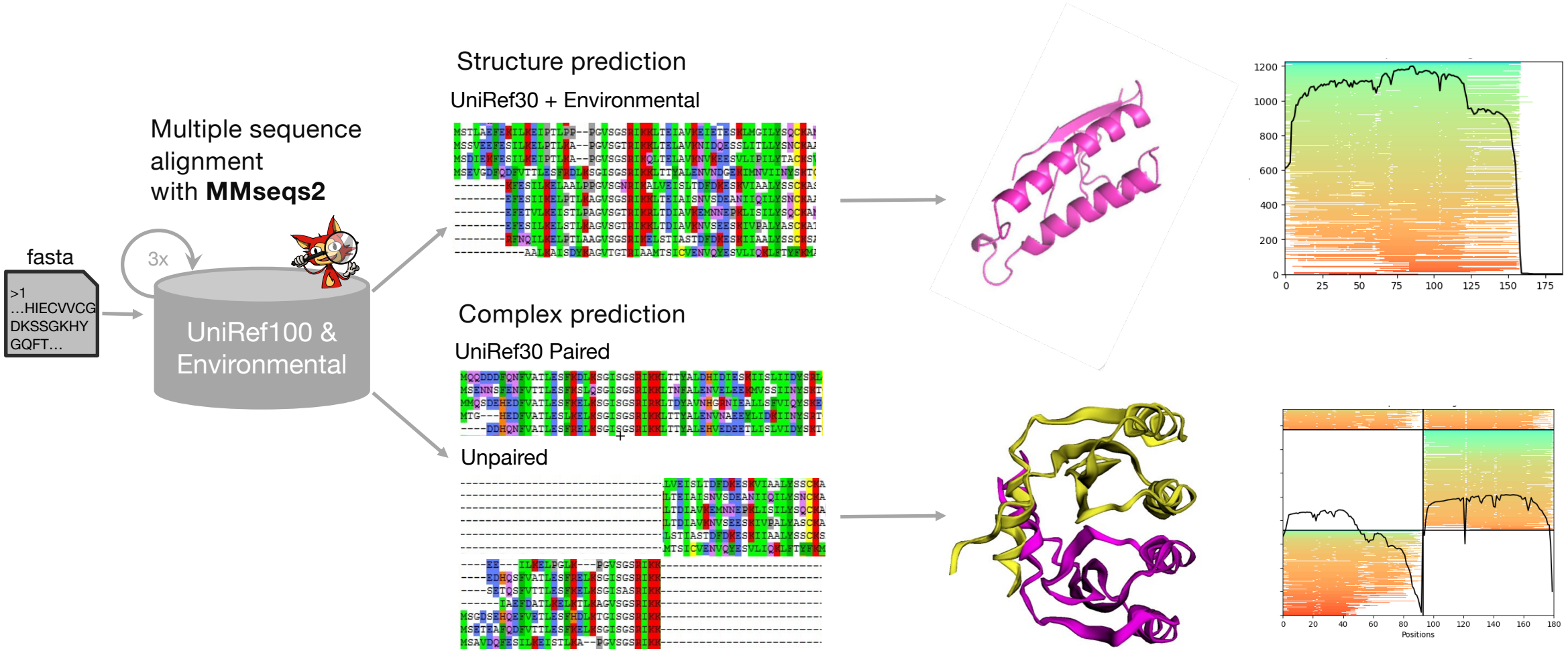
Alphafold2 structure prediction is only as good as the input **MSA**



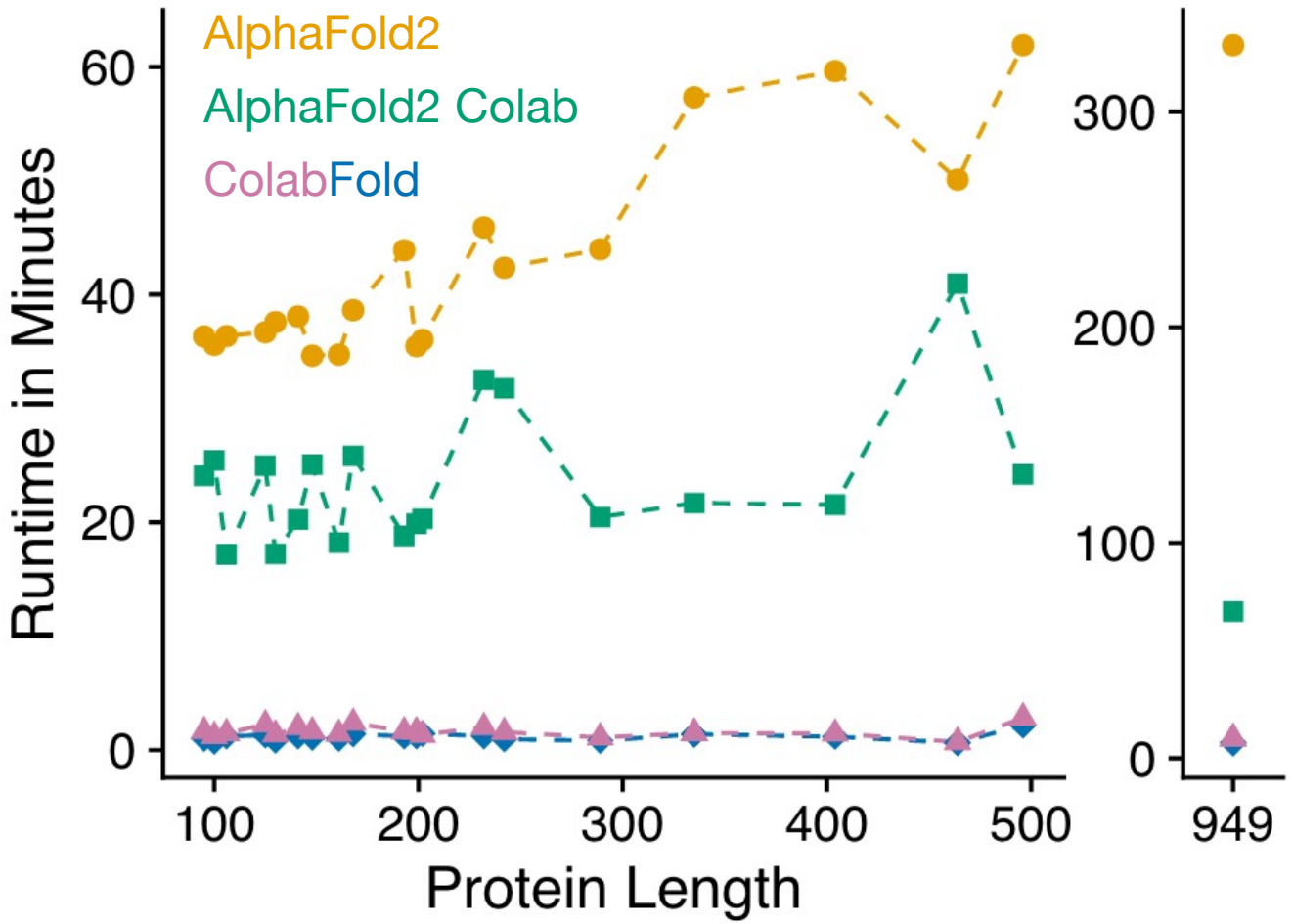
Speed of AlphaFold2



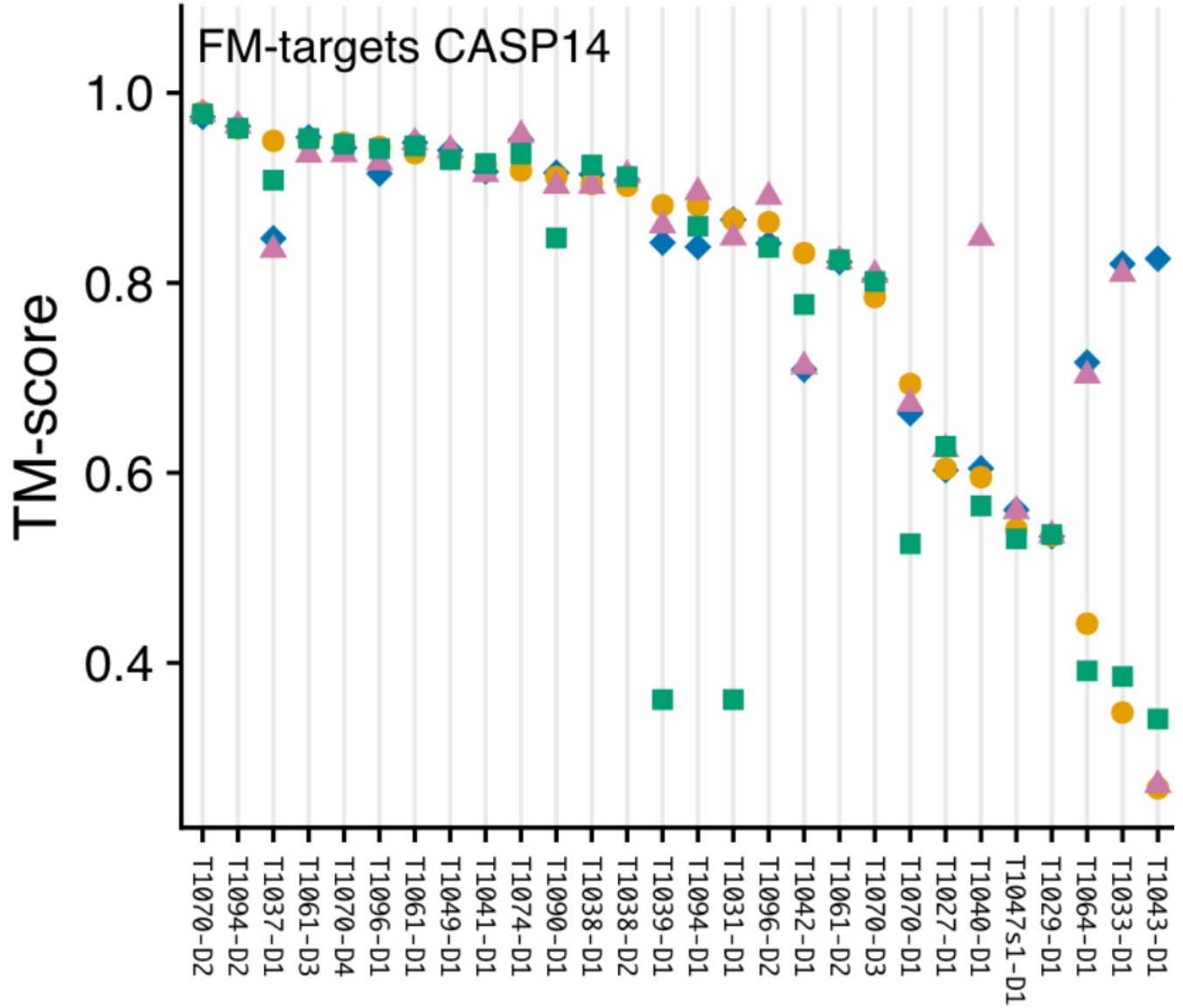
ColabFold uses **MMseqs2** to speed up homology search



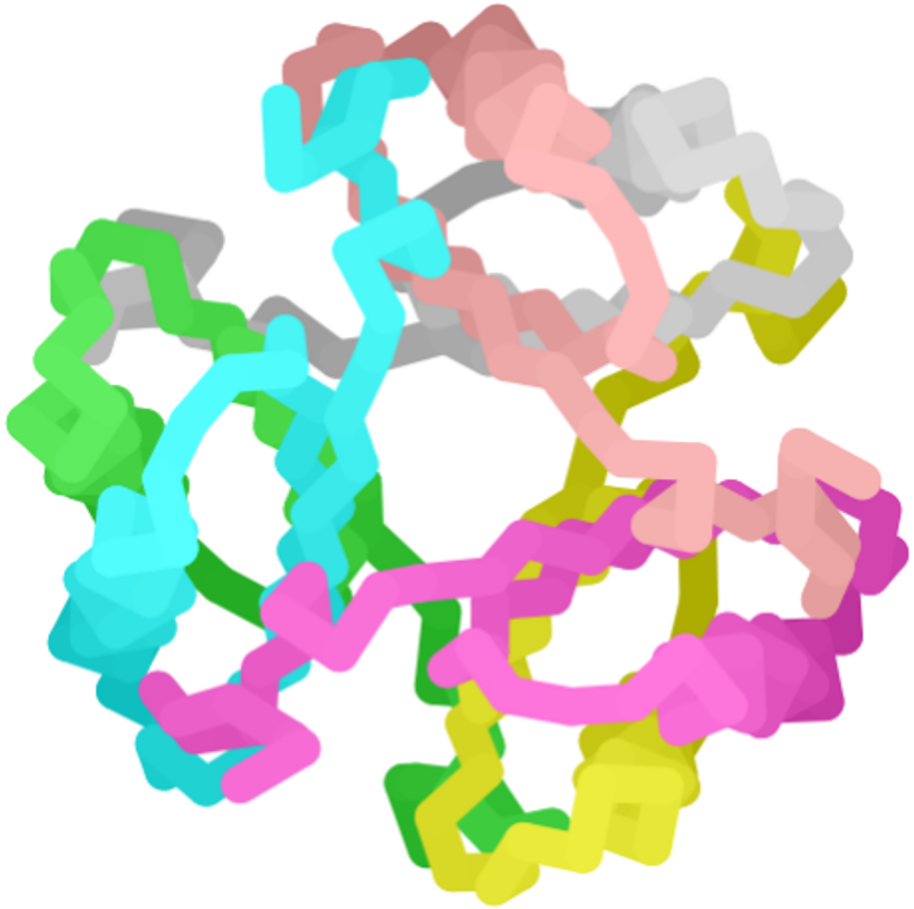
Speed of MMseqs2 **20-30x** faster than Alphafold2's homology search



ColabFold produces structures at the quality of AlphaFold2 and is more precise than AlphaFold2 Colab

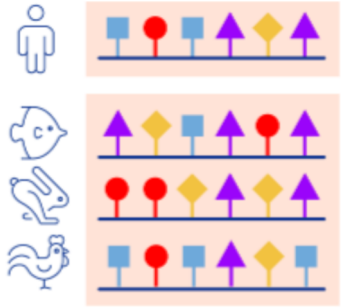


Predicting homo-oligomers



Modeling a protein given an MSA

Residue index `[1,2,3,4,5,6]`

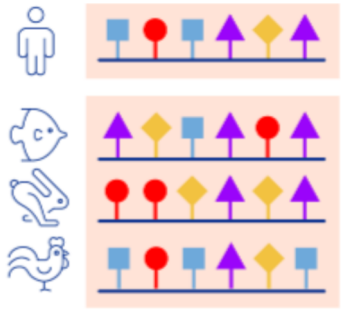


Multiple Sequence Alignment

MSA image borrowed from Kathryn T. (Deepmind)

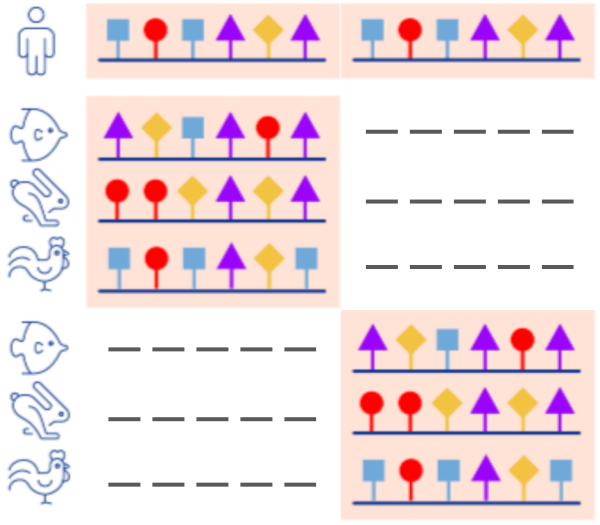
Modeling homo-oligomeric interactions by duplicating, padding and concatenating the MSAs

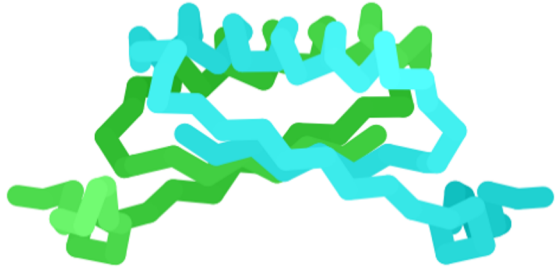
Residue index `[1,2,3,4,5,6]`



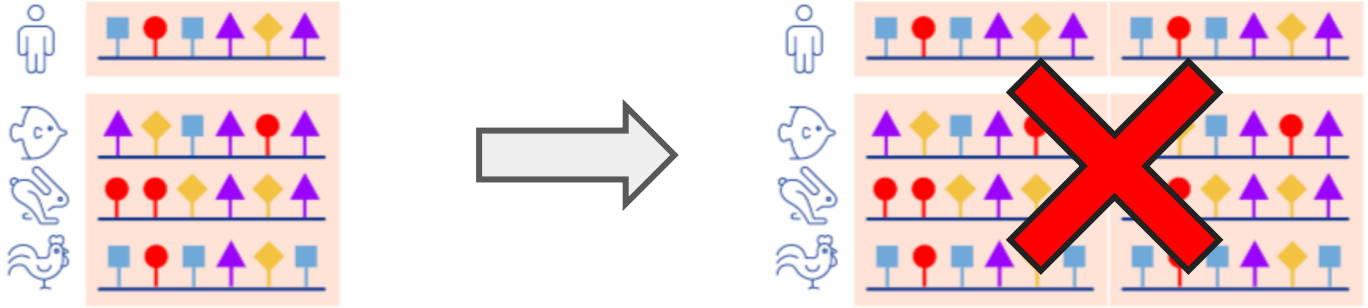
Multiple Sequence Alignment

`[1,2,3,4,5,6,207,208,209,210,211,212]`





Just duplicating often does **not** work.



Multiple Sequence Alignment

Complexes - monomer

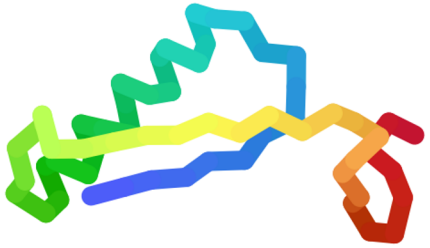
Enter the amino acid sequence to fold ⬇️

sequence: " PIAQIHILEGRSDEQKETLIREVSEAISRSLDAPLTSVRVIITEMAKGHFGIGGELASK

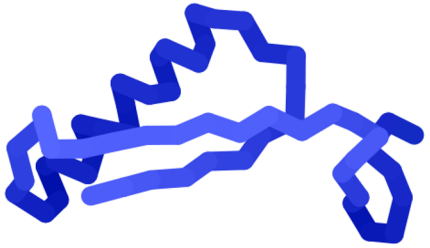
jobname: " test

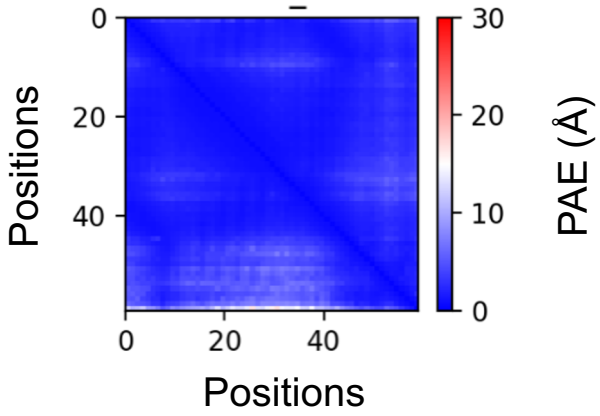
homooligomer: 1

colored by N→C



colored by pLDDT





Complexes - homodimer

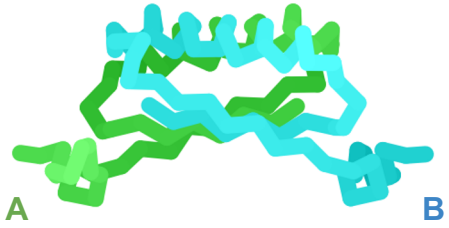
Enter the amino acid sequence to fold ⬇️

sequence: " PIAQIHILEGRSDEQKETLIREVSEAISRSLDAPLTSVRVIITEMAKGHFGIGGELASK

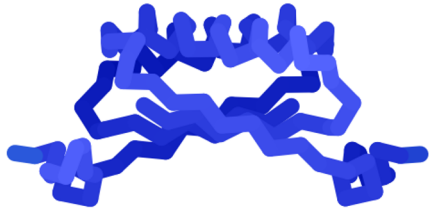
jobname: " test

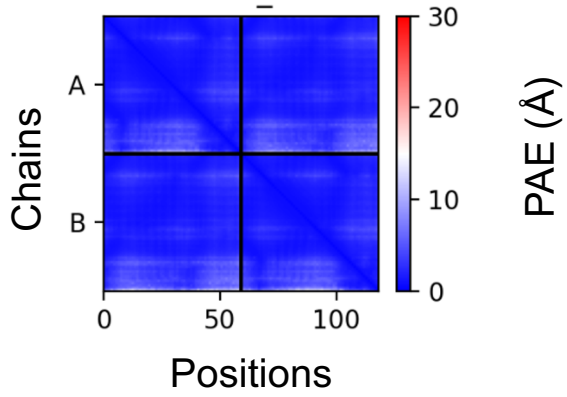
homooligomer: 2

colored by chain



colored by pLDDT





Complexes - homo-6-mer

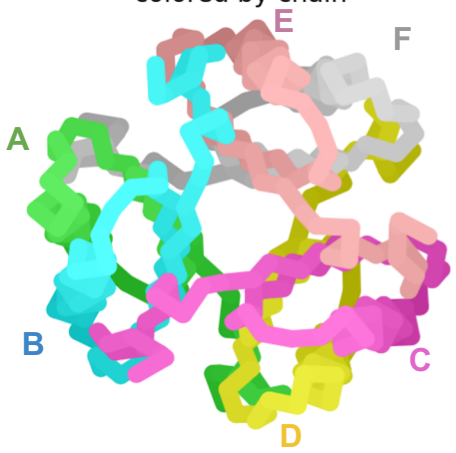
Enter the amino acid sequence to fold ⬇️

sequence: " PIAQIHILEGRSDEQKETLIREVSEAISRSLDAPLTSVRVIITEMAKGHFGIGGELASK

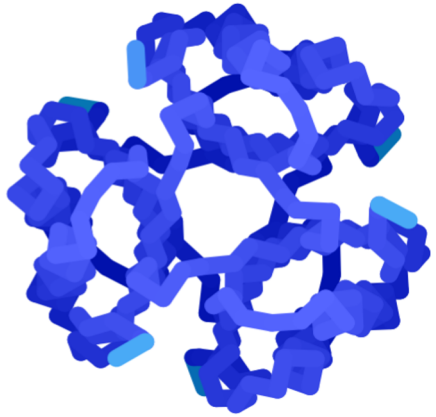
jobname: " test

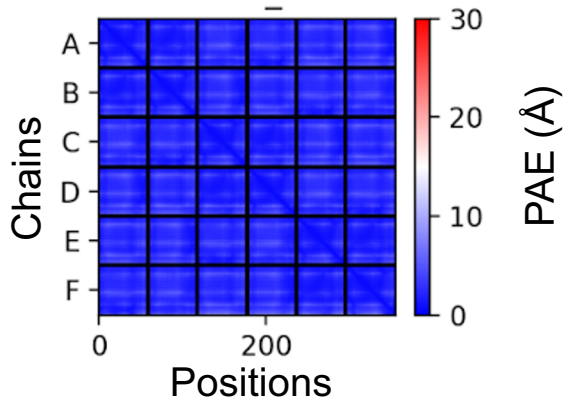
homooligomer: 6

colored by chain



colored by pLDDT





Complexes - homo-8-mer?

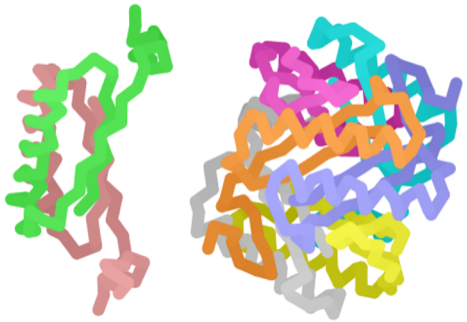
Enter the amino acid sequence to fold ⬇️

sequence: PIAQIHILEGRSDEQKETLIREVSEAISRSLDAPLTSVRVIITEMAKGHFGIGGELASK

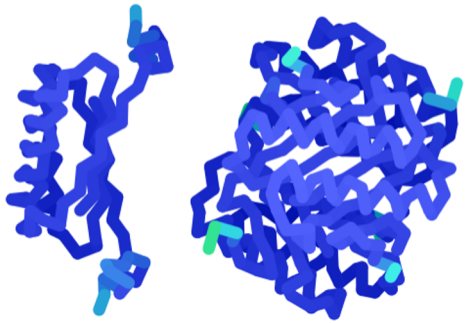
jobname: test

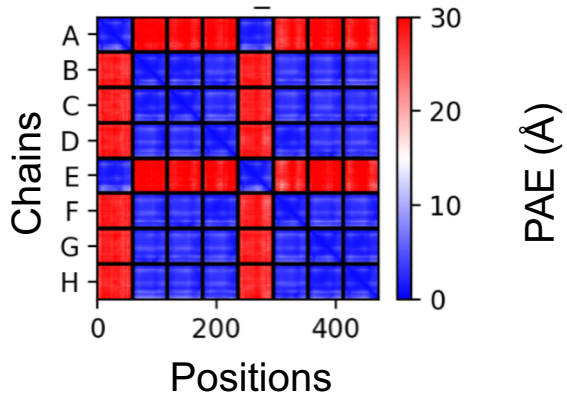
homooligomer: 8

colored by chain

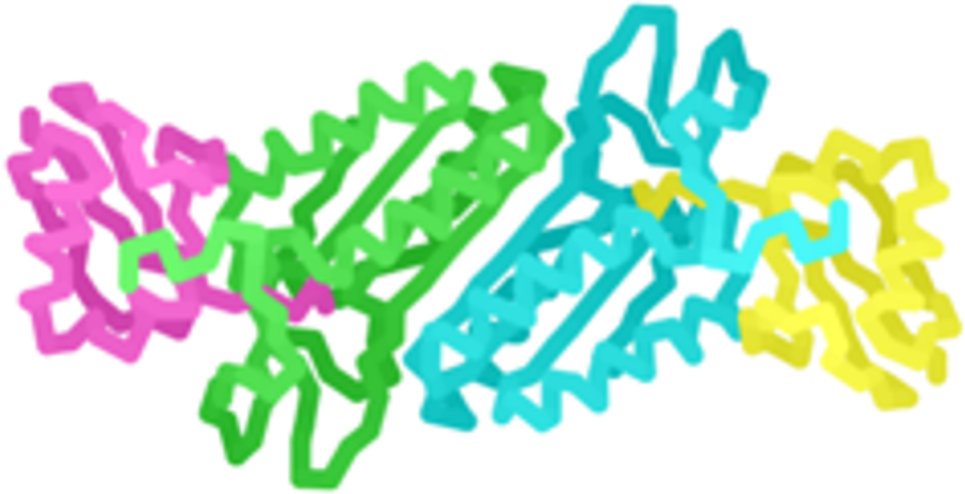


colored by pLDDT

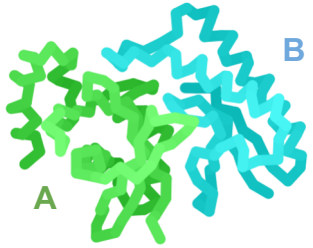


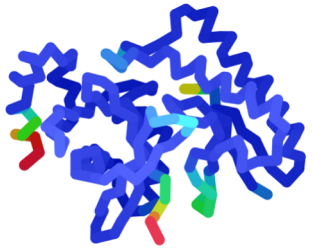


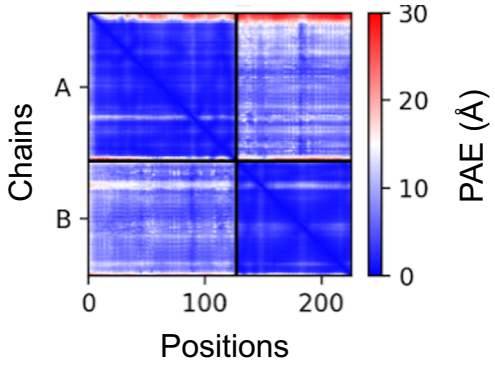
How about hetero-oligomers?



Hetero-dimer (1:1) - CASP target H1065







Enter the amino acid sequence to fold ⬇️

sequence: PIAQIHILEGRSDEQKE::::::::::::::TLIREVSEAISRSLDAPLTSVR

jobname: test

homooligomer: 1:1

- `sequence` Specify protein sequence to be modelled.
  - Use `/` to specify intra-protein chainbreaks (for trimming regions within protein).
  - Use `:` to specify inter-protein chainbreaks (for modeling protein-protein hetero-complexes).

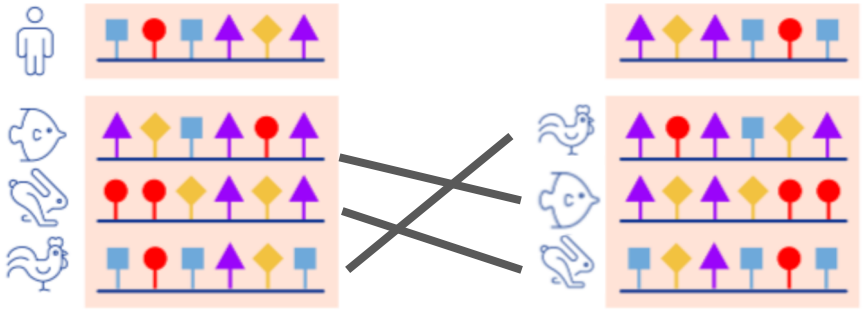
pair msa options

Experimental option for protein complexes. Pairing currently only supported for proteins in same operon (prokaryotic genomes).

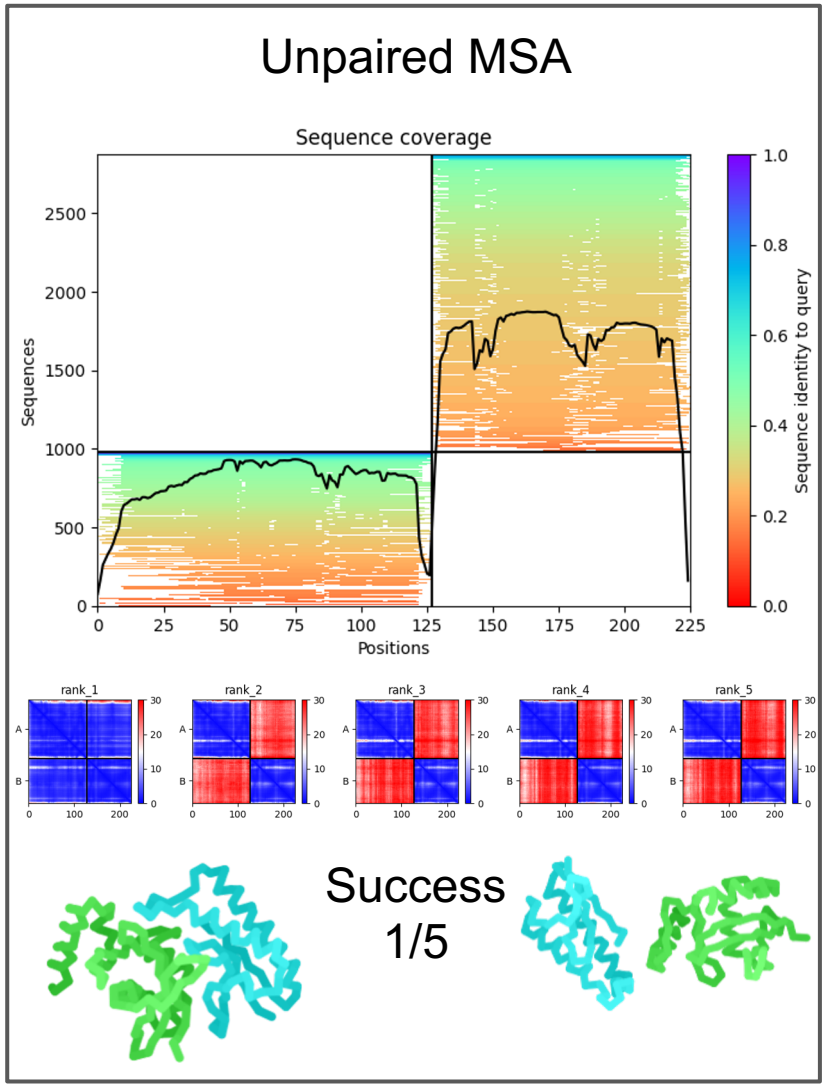
pair_mode: unpaired

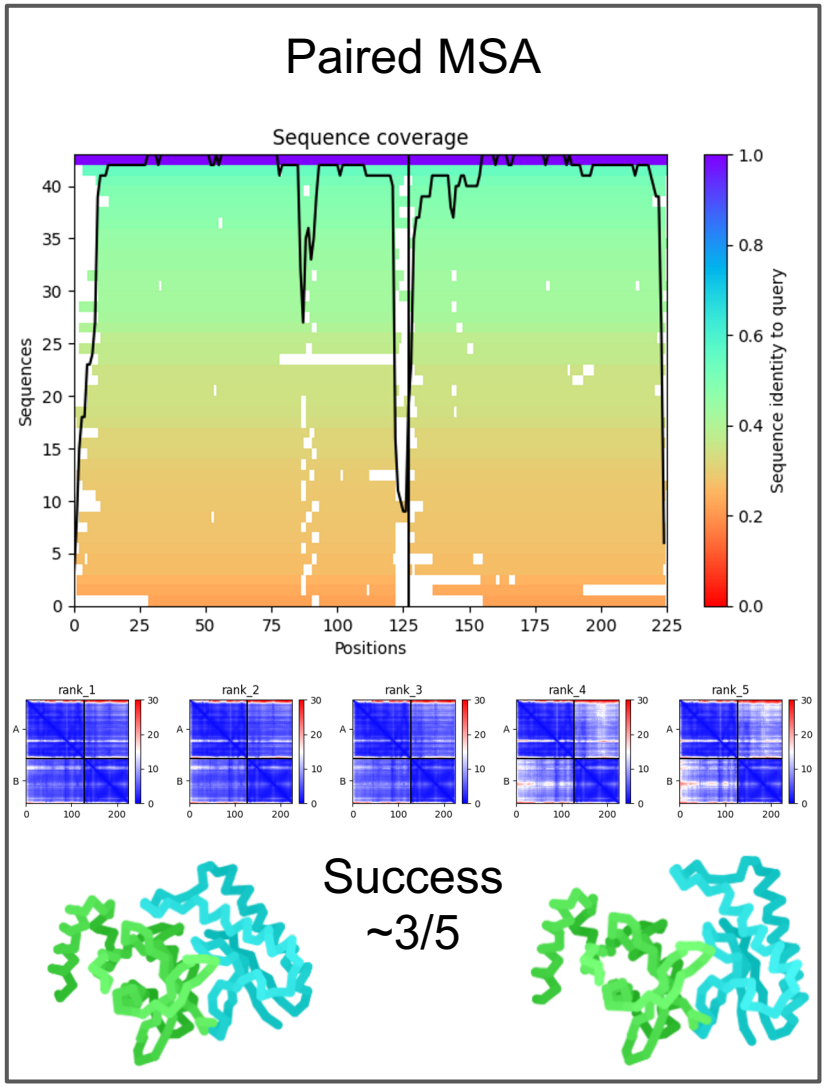
- `unpaired` - generate seperate MSA for each protein.
- `unpaired+paired` - attempt to pair sequences from the same operon within the genome.
- `paired` - only use sequences that were sucessfully paired.

paired msa (currently only works for prokaryotic operons)



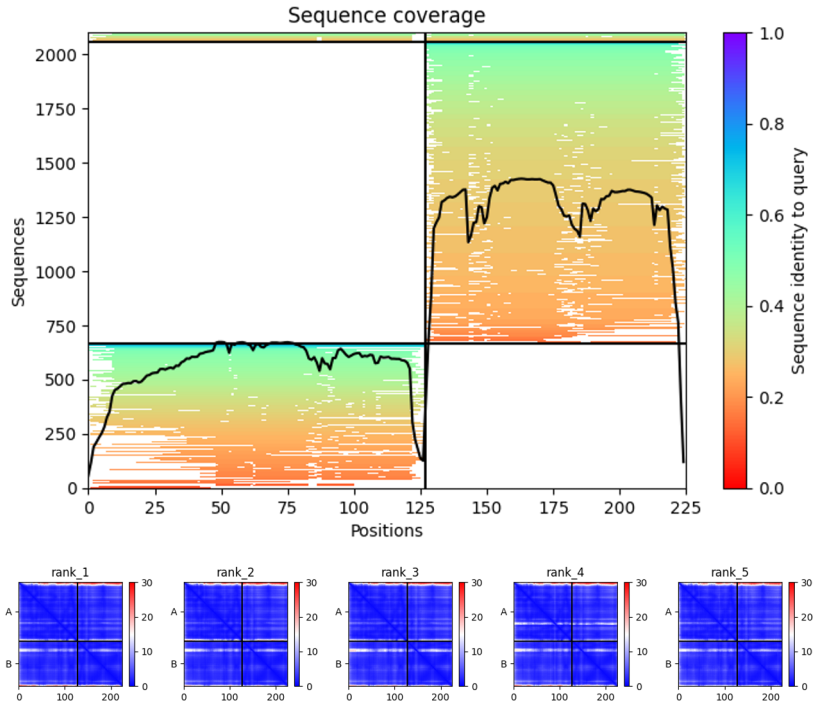
Sometimes unpaired MSA works (example: CASP target H1065)





Combining paired+unpaired helps (example: CASP target H1065)

Unpaired+Paired MSA



Success
5/5

DeepMind

2021-10-04

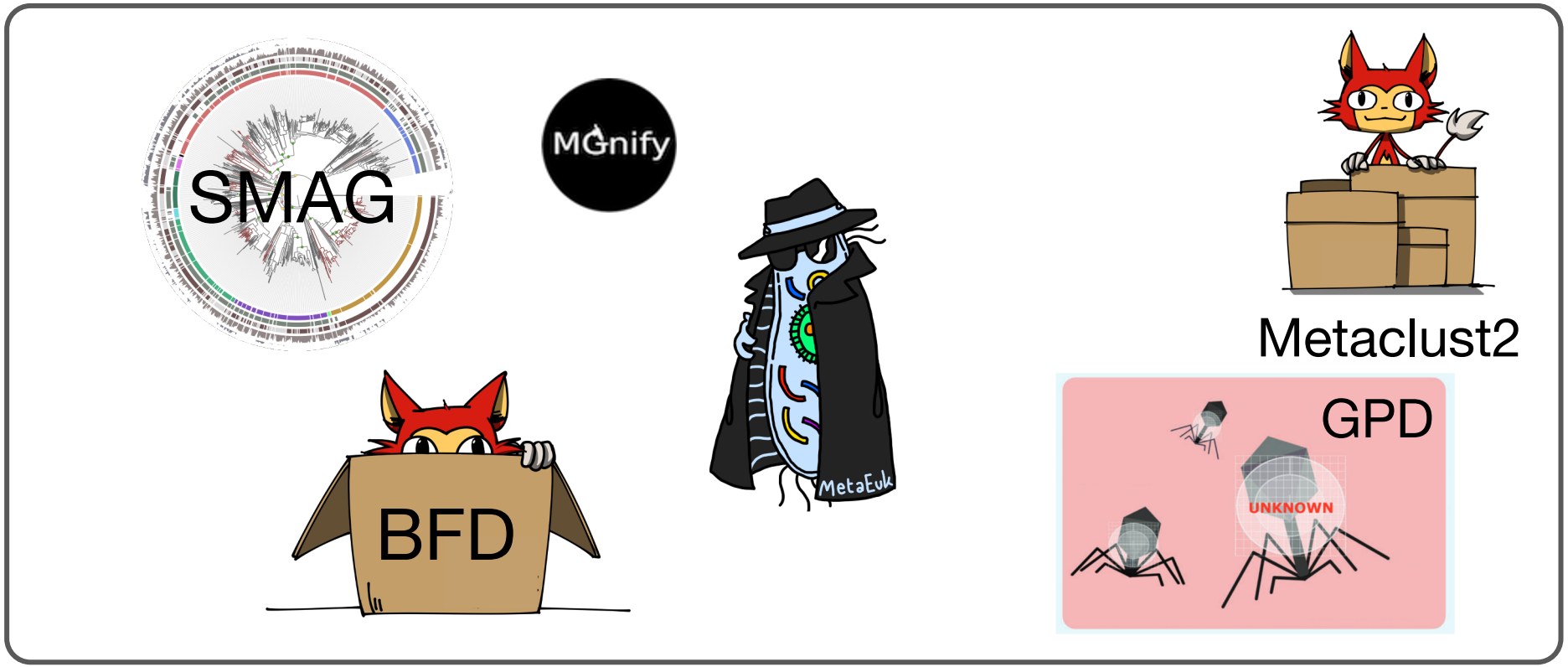
Protein complex prediction with AlphaFold-Multimer

Richard Evans[1*], Michael O'Neill[1*], Alexander Pritzel[1*], Natasha Antropova[1*], Andrew Senior[1], Tim Green[1], Augustin Žídek[1], Russ Bates[1], Sam Blackwell[1], Jason Yim[1], Olaf Ronneberger[1], Sebastian Bodenstein[1], Michal Zielinski[1], Alex Bridgland[1], Anna Potapenko[1], Andrew Cowie[1], Kathryn Tunyasuvunakool[1], Rishub Jain[1], Ellen Clancy[1], Pushmeet Kohli[1], John Jumper[1*] and Demis Hassabis[1*]

[1]DeepMind, London, UK, [*]These authors contributed equally

Coming soon to ColabFold!

ColabFoldDB contains many new metagenomic reference catalogues



>2 Billion Proteins (BFD) Jumper et al., *Nature*, 2021
>1 Billion Proteins (Mgnify) Mitchell et al., *Nucleic Acids Research*, 2019
6 Mio Eukaryotic Proteins (Metaeuk) Levy Karin et al, *Microbiome*, 2020
10 Mio Proteins from SAGs and MAGs (SMAG) Delmont et al. *biorxiv*, 2021
12 Mio Eukaryotic Proteins (TOPAZ) Alexander et al. *biorxiv*, 2021
11,8 Mio Viral Proteins (MGV) Nayfach et al. *Nature Microbiology,* 2021
7.5 Mio Phage Viral Proteins (GPD) Camarillo-Guerrero et al. *Cell*, 2021
36 Billion Proteins Metaclust2, unpublished

ColabFold summary

ColabFold enables structures and complex prediction

- **Fast** structure and complex prediction
- Runs in the **browser** through Google Colaboratory
- Larger metagenomic database (ColabFoldDB)
- Enabling thousands of protein structures on a single GPU

ColabFold has predicted over 400,000 protein structures from researchers around the world

github.com/sokrypton/ColabFold

ColabFold - Making protein folding accessible to all, M. Mirdita et al., (2021), biorxiv

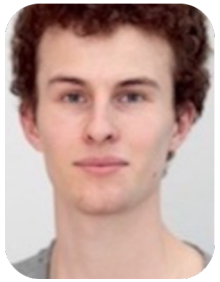
Acknowledgments

ColabFold

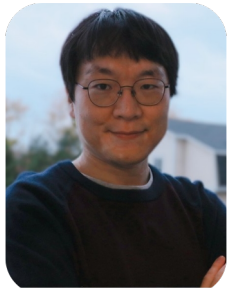
Martin Steinegger

Sergey Ovchinnikov

Yoshitaka Moriwaki



Konstantin Schütze



Lim Heo

Martin Steinegger

Milot Mirdita

GEFÖRDERT VOM Bundesministerium für Bildung und Forschung

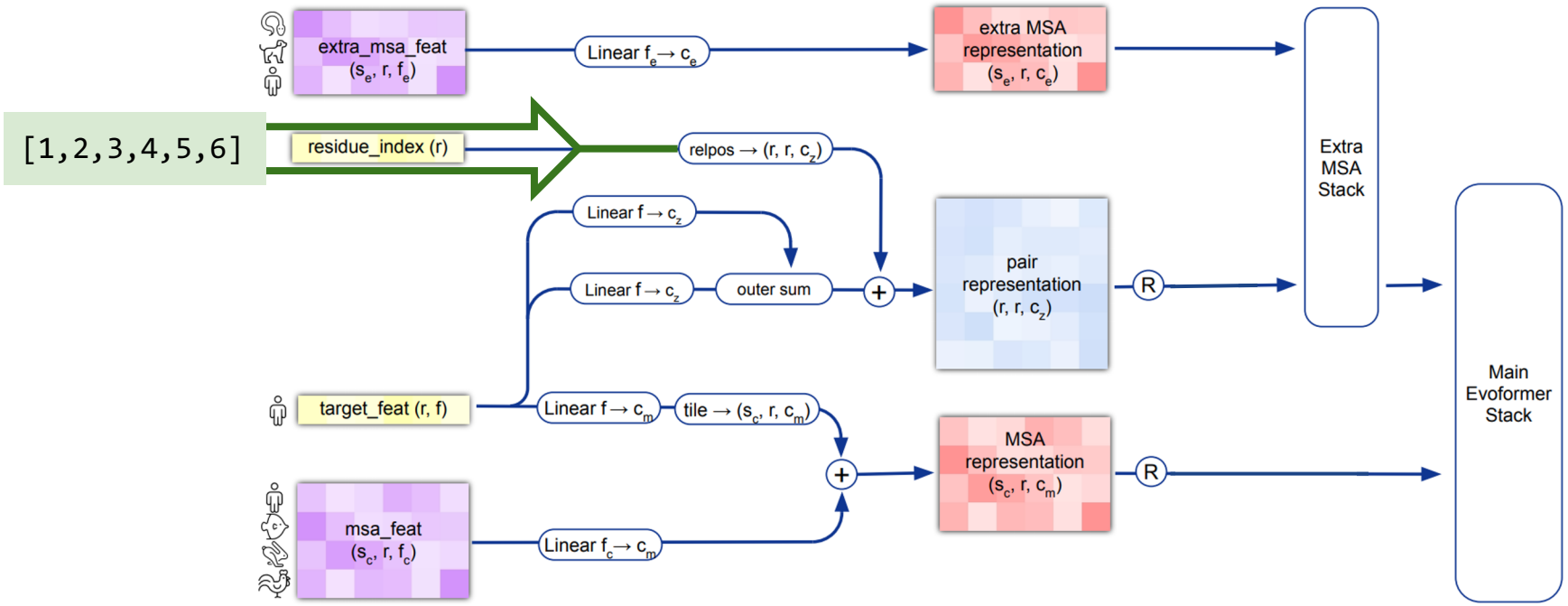
Sergey Ovchinnikov

ColabFold logo by Doyoon Kim

Söding Lab

Based on slides from Sergey and Martin. Also checkout our talk on ColabFold Tutorial presented at the Boston Protein Design and Modeling Club. [video] [slides].

`residue_index` in AlphaFold2 is used to create a relative positional encoding



Algorithm 4 Relative position encoding

def $\text{relpos}(\{f_i^{\text{residue\_index}}\}, \mathbf{v}_{\text{bins}} = [-32, -31, \ldots, 32]) :$

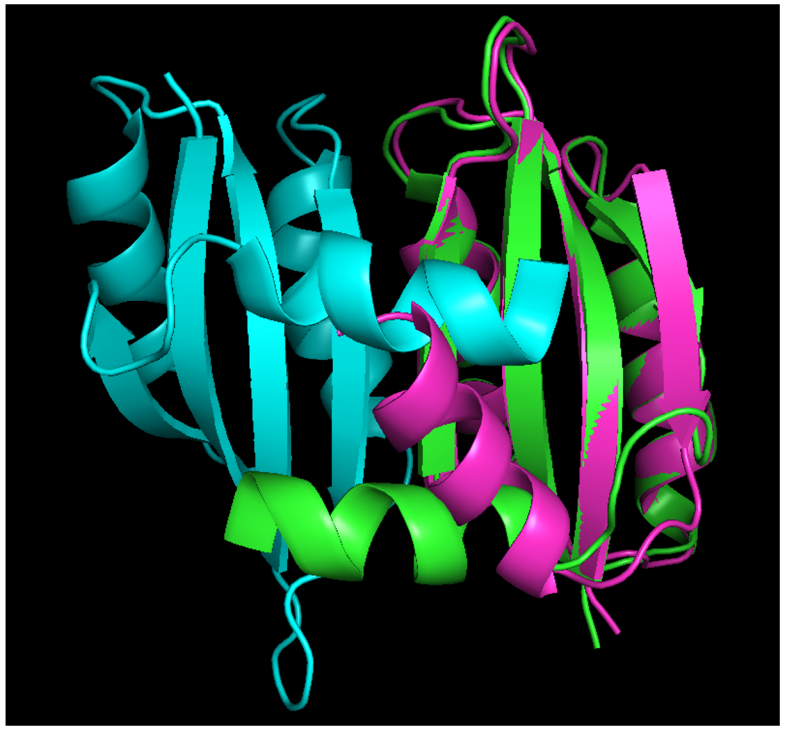
1: $d_{ij} = f_i^{\text{residue\_index}} - f_j^{\text{residue\_index}}$ $\quad d_{ij} \in \mathbb{Z}$

2: $\mathbf{p}_{ij} = \text{Linear}(\text{one\_hot}(d_{ij}, \mathbf{v}_{\text{bins}}))$ $\quad \mathbf{p}_{ij} \in \mathbb{R}^{c_z}$

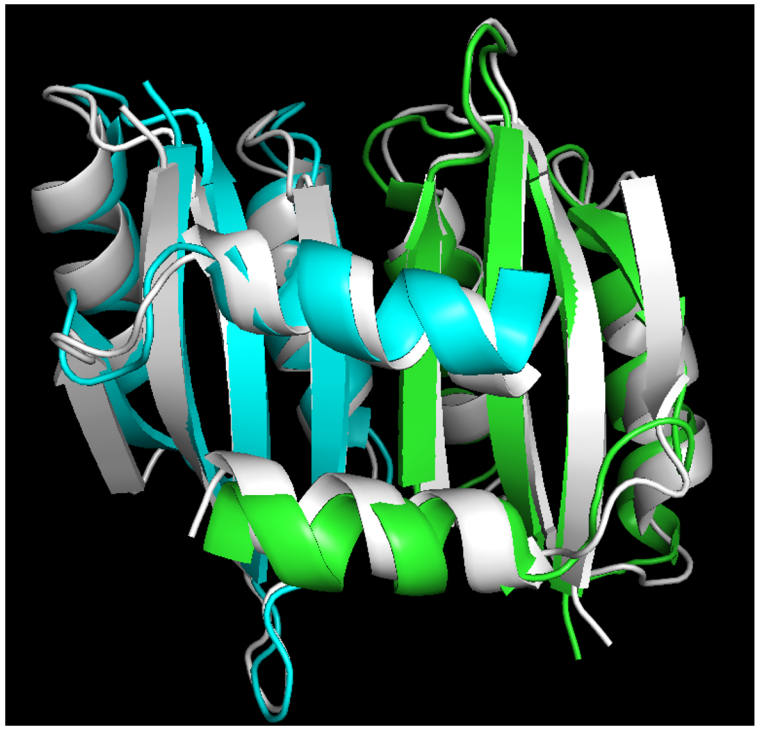
3: **return** $\{\mathbf{p}_{ij}\}$

Residue index difference capped at 32

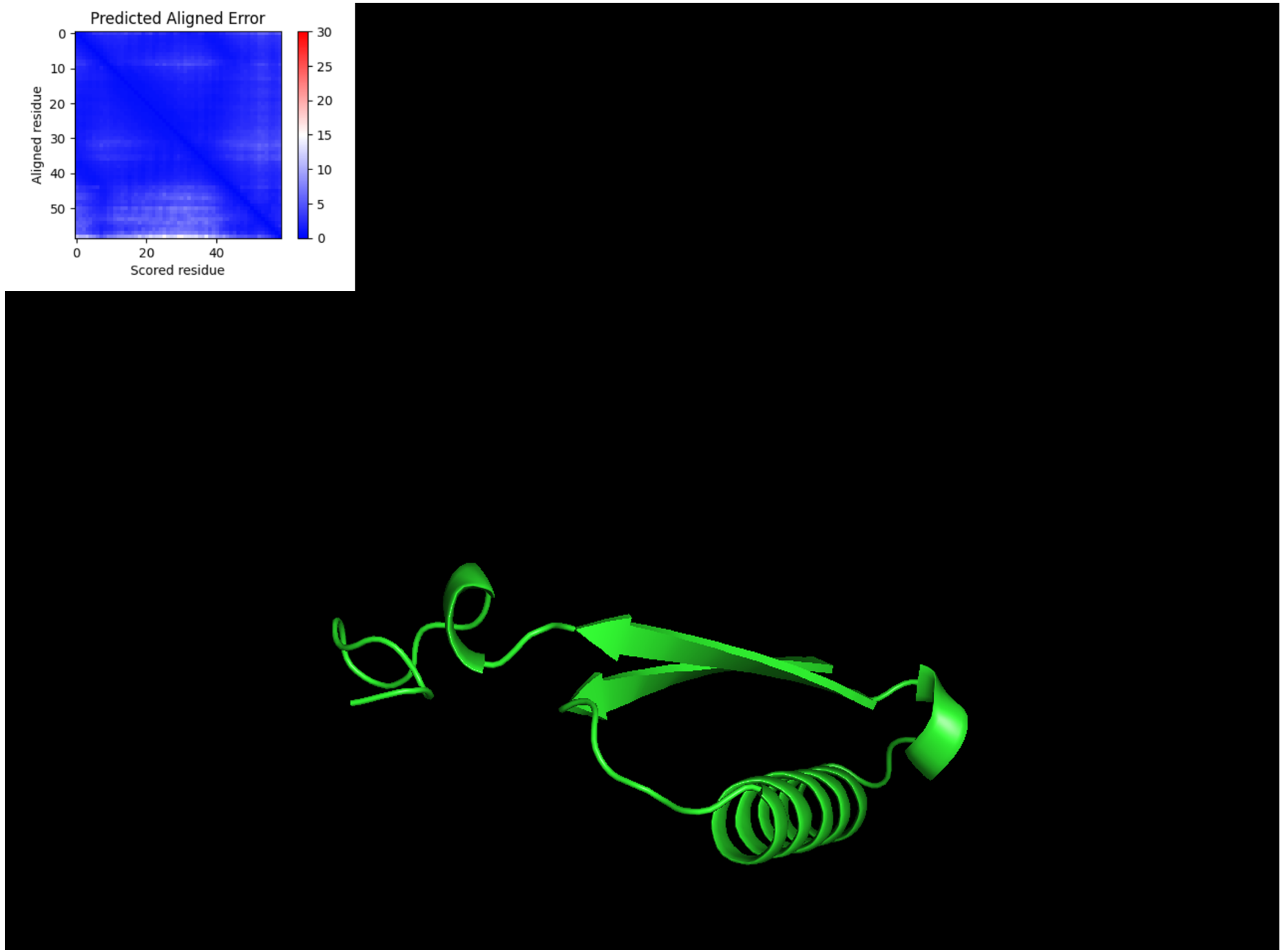
Modeling as complex can fix details in monomers

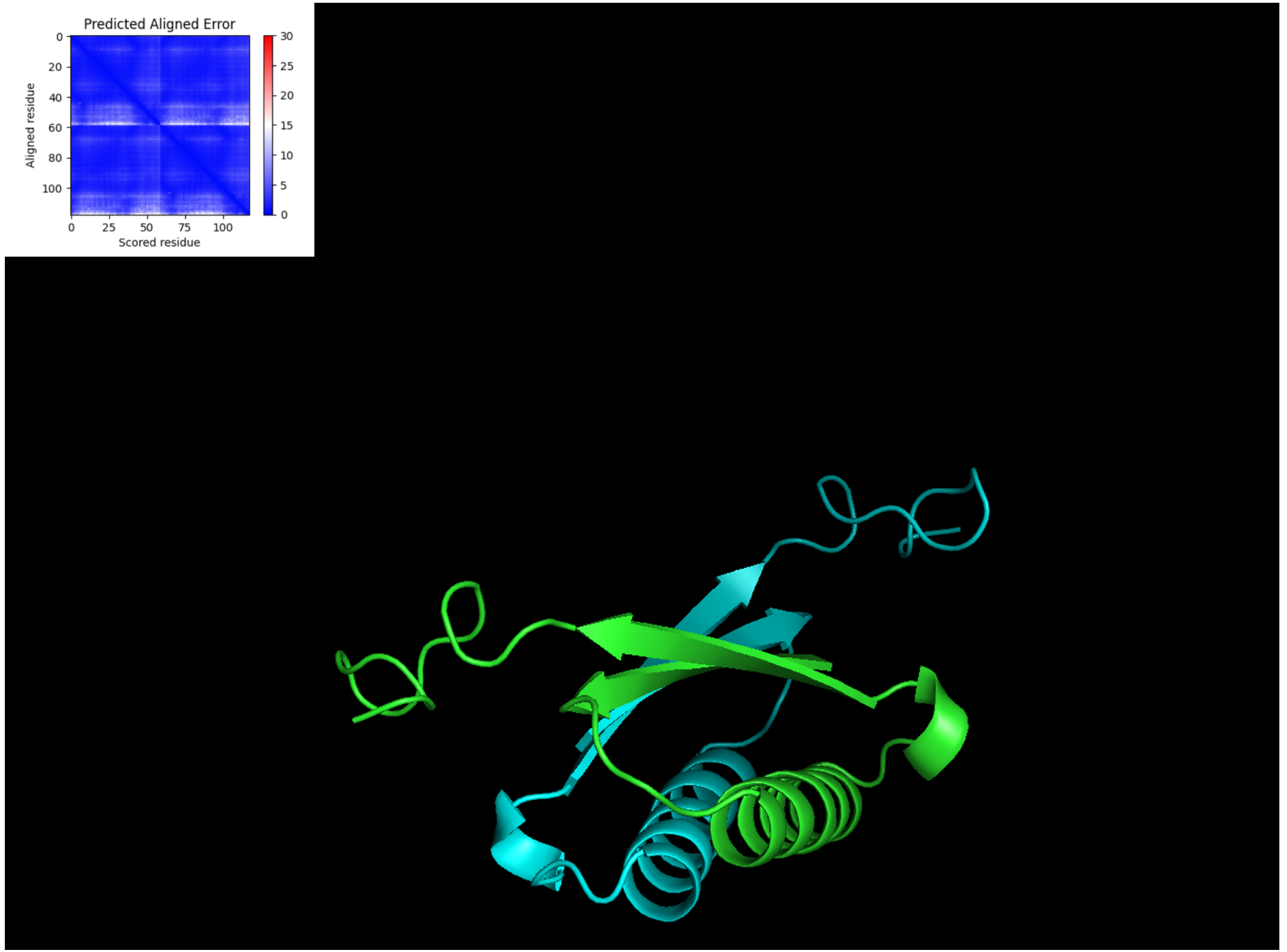


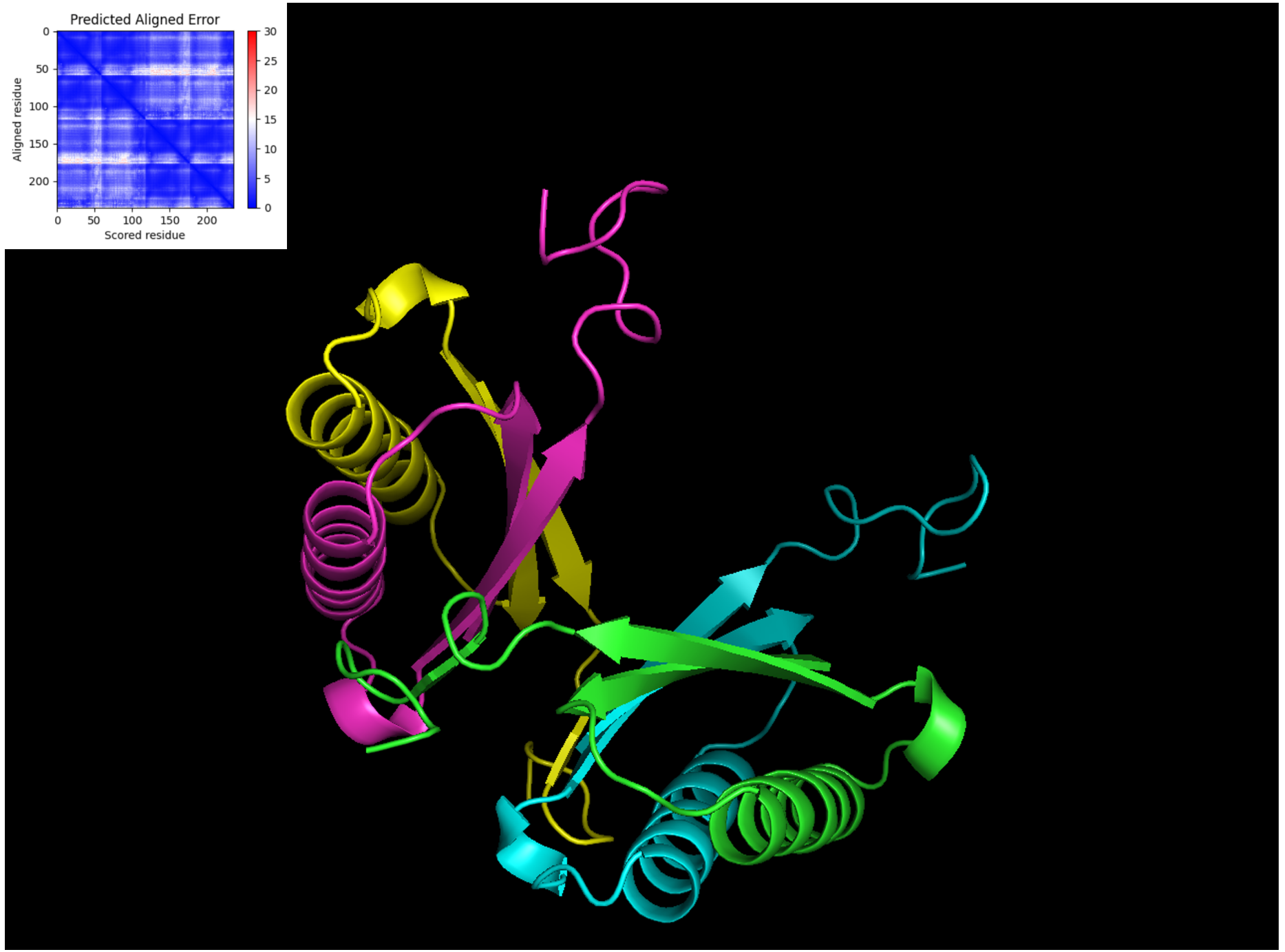
pink = monomer model

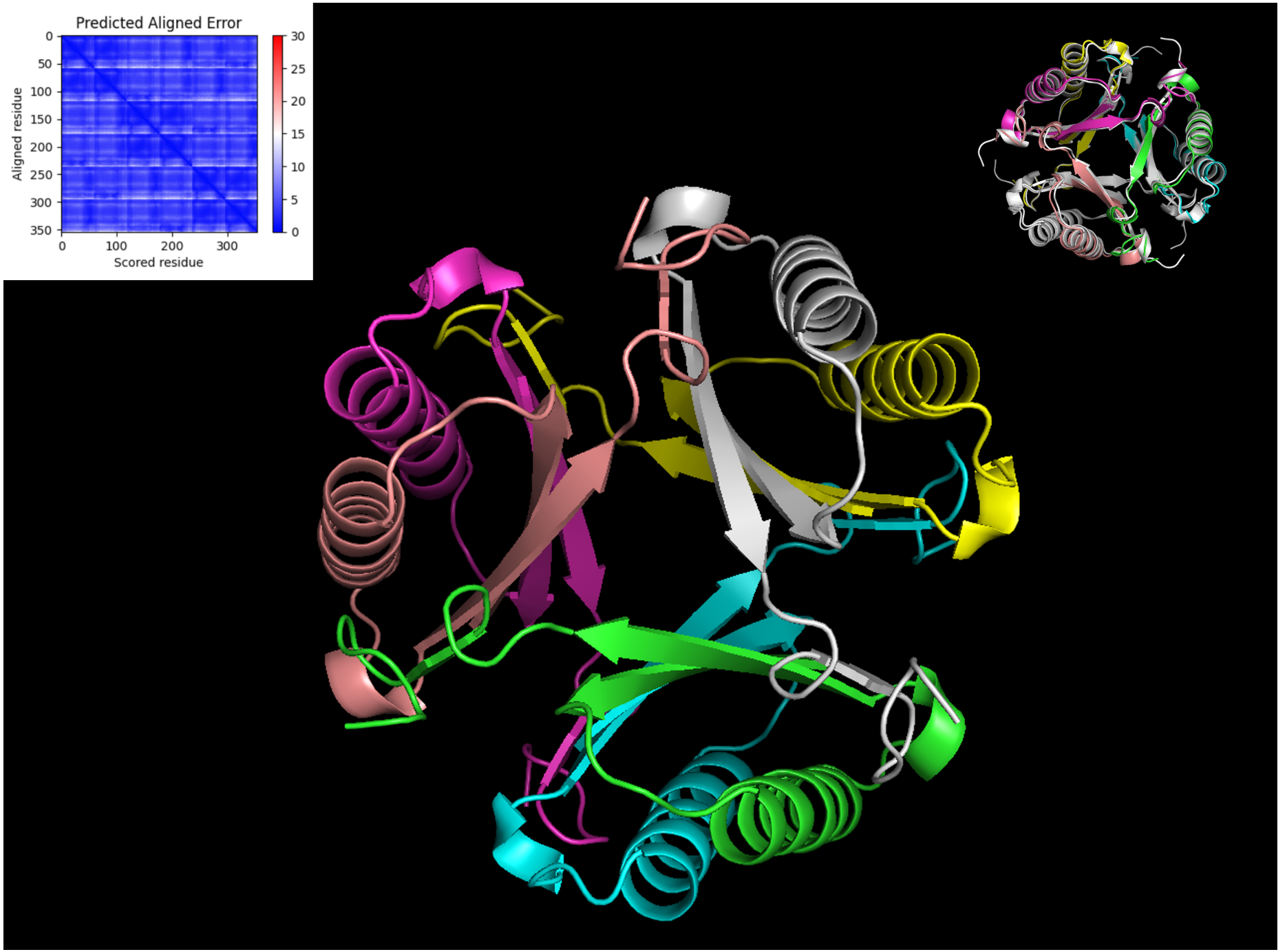


white = homo-dimer model


Predicted Aligned Error
Aligned residue
Scored residue
0
10
20
30
40
50
0
20
40
0
5
10
15
20
25
30

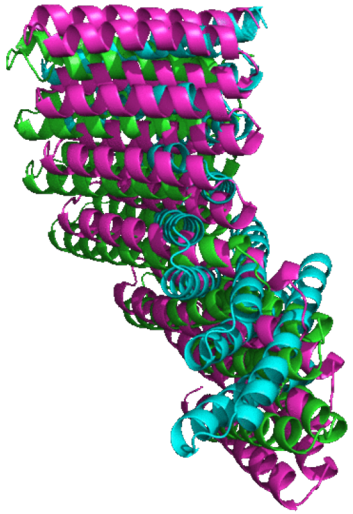

Predicted Aligned Error
Aligned residue
Scored residue
0
20
40
60
80
100
0
25
50
75
100
30
25
20
15
10
5
0


Predicted Aligned Error
Aligned residue
Scored residue
0
50
100
150
200
0
5
10
15
20
25
30


Predicted Aligned Error
Aligned residue
Scored residue
0
50
100
150
200
250
300
350
0
100
200
300
30
25
20
15
10
5
0

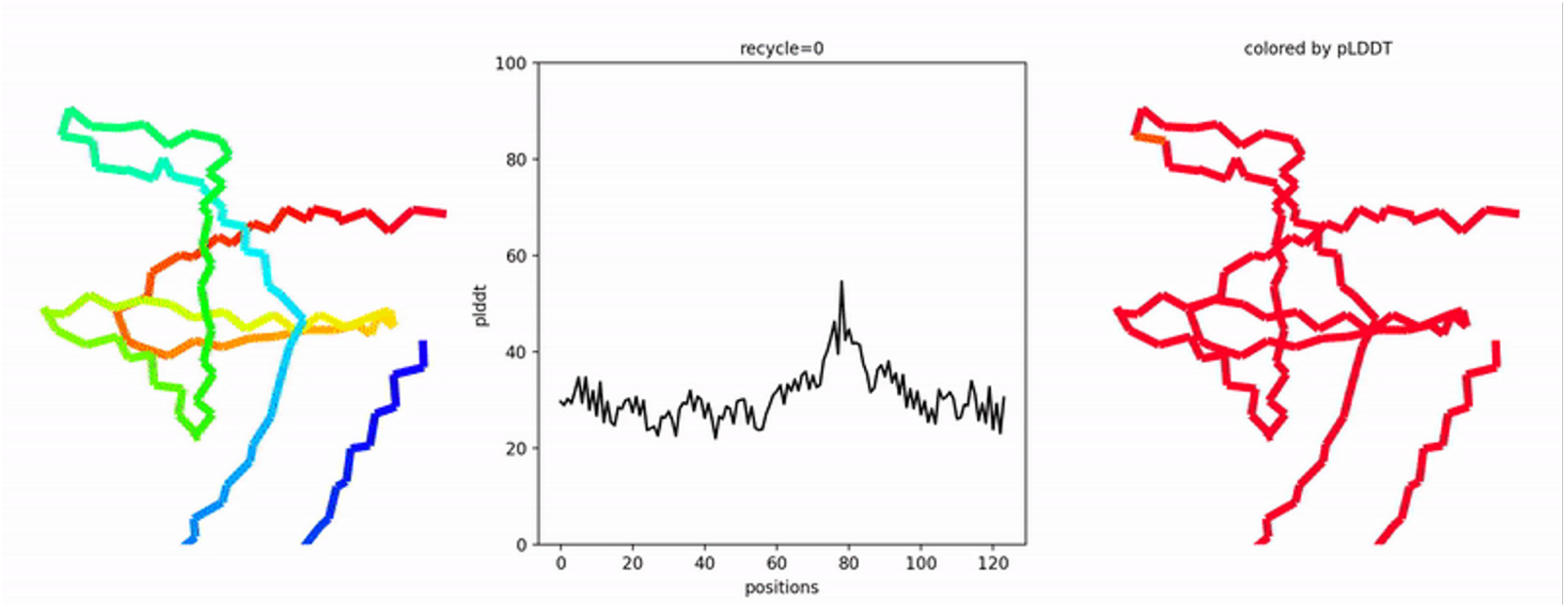

Predicted Aligned Error
Aligned residue
Scored residue
0
100
200
300
400
30
25
20
15
10
5
when given 8 copies,
predicted as a homo-8-mer
instead homo-6-mer

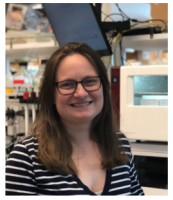
homooligomeric assembly with increased number of recycles



Credit Ryan Kibler

Another example: need up to 12 recycles to get the correct fold





Vorobieva, A.A., White, P., Liang, B., Horne, J.E., Bera, A.K., Chow, C.M., Gerben, S., Marx, S., Kang, A., Stiving, A.Q. and Harvey, S.R., 2021. De novo design of transmembrane β barrels. *Science*, *371*(6531).