

ML Based Structural Modeling and Prediction with AlphaFold2 and RoseTTAFold



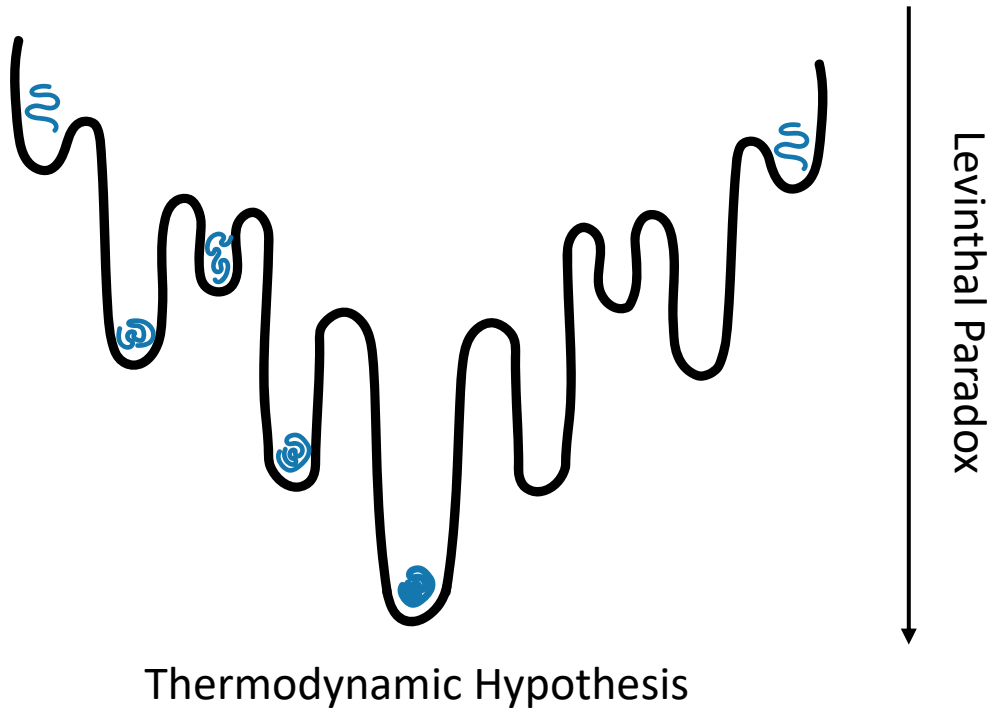
VANDERBILT
UNIVERSITY

Eli F. McDonald

12/04/2023

Rosetta Workshop '23, Vanderbilt University

Protein folding problem



- Consider a 100 residue protein w/ 2 DOF/residue
- # of possible conformations
 $= 2^{100} \sim 10^{30}$
- vibration rate $\sim k_v = 10^{13}$ conformations/sec
- Time to sample all confs:
- $T = j^N / Nk_v$
 $= 10^{30} / (100 * 10^{13})$
 $= 10^{17} \text{ sec} = 10^{10} \text{ years}$
- Age of universe = $13.4 * 10^9$ years



Experimental structure approaches

- NMR spectroscopy
 - peptides and flexible proteins
- EPR labeling
 - interdomain distances
- X-ray crystallography
 - both soluble and membrane proteins
- Cryo-EM
 - larger proteins
 - complexes and membrane proteins

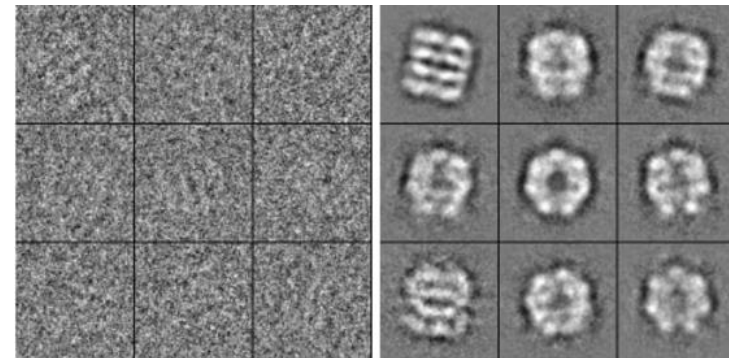
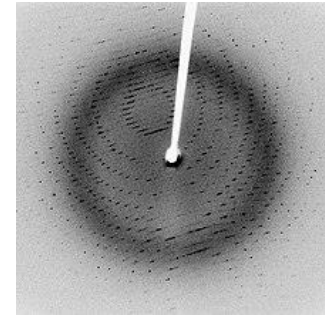
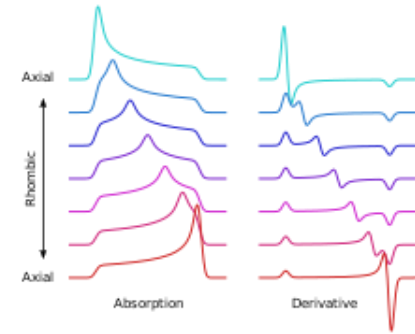
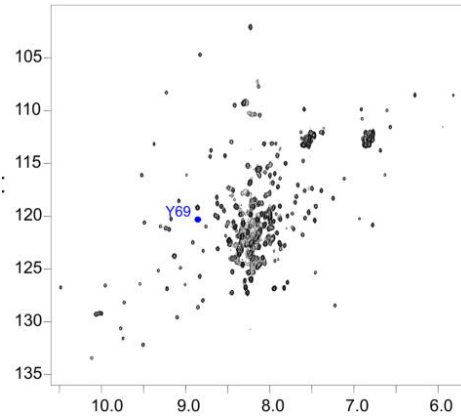
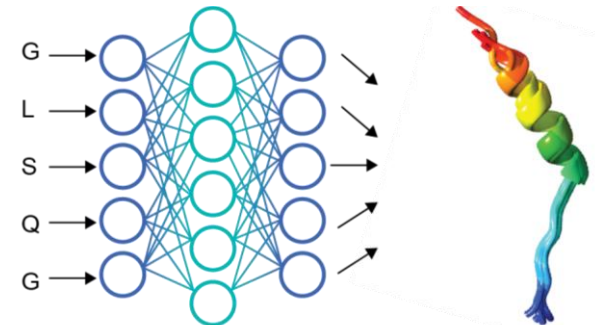
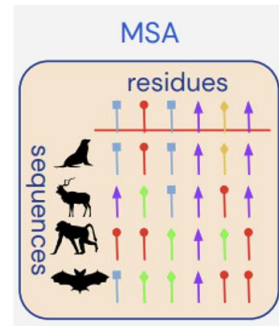
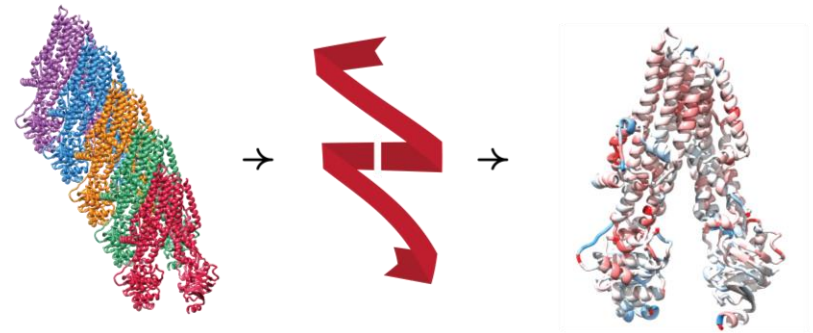
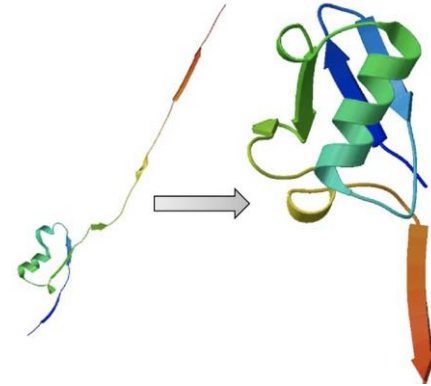


image sources: Wikipedia, jiang.bio.purdue.edu/



Computational structure approaches

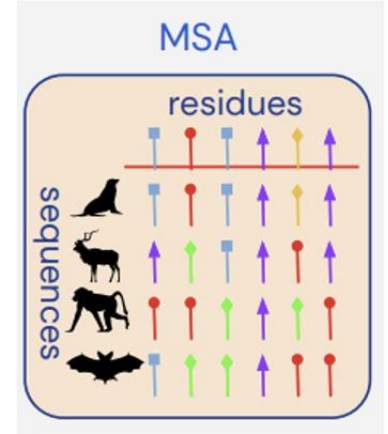
- De novo folding
 - sequence information
 - fragment libraries
 - physics/statistics based
- Homology modeling
 - structural information
 - templates + fragment libraries
- Deep learning methods
 - evolutionary information
 - predict residue distances
 - structural training data



Deep learning methods

- MSA-based methods

- use MSAs to infer information about residue pairs
- predicts 3D structures based on pairs
- AlphaFold2 and RoseTTAFold



- MSA-independent methods

- usually based on transformer protein language models
- run with a single sequence without the need to calculate MSAs
- OmegaFold and ESMFold



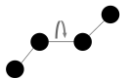
What is AlphaFold2 (AF2)

- AF2: a convolutional neural network trained on most of the PDB structures to predict protein structure from sequence.
- Only uses sequence as an input (no template needed)
- Generated models are "relaxed" further using a classical mechanical force field and analyzed for accuracy



$$E_{\text{Bond}} = \frac{1}{2} k_B (R - R_0)^2$$

1) bond length



$$E_{\text{Tor}} = k_{\phi} [1 + \cos(n\phi - \gamma)]$$

3) torsion angle

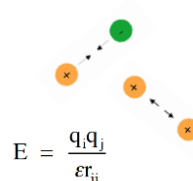
Covalent Interactions



$$E_{\text{Angle}} = \frac{1}{2} k_{\lambda} (\theta - \theta_0)^2$$

2) bond angle

Non-covalent Interactions



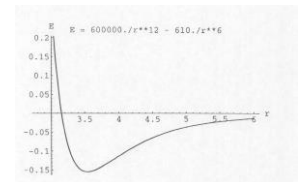
$$E = \frac{q_i q_j}{\epsilon r_{ij}}$$

1) electrostatic



$$E = 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right]$$

2) Van der Waals



AF2 Inputs

Sequence Databases

- Uniref90 (JackHMMER)
- BFD (Hhblits)
- MGnify clusters (JackHMMER)

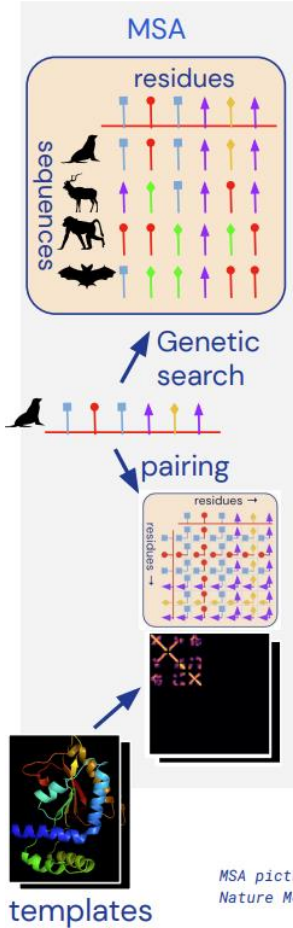
Structural Databases

- PDB (training)
- PDB70 clustering (hhsearch)

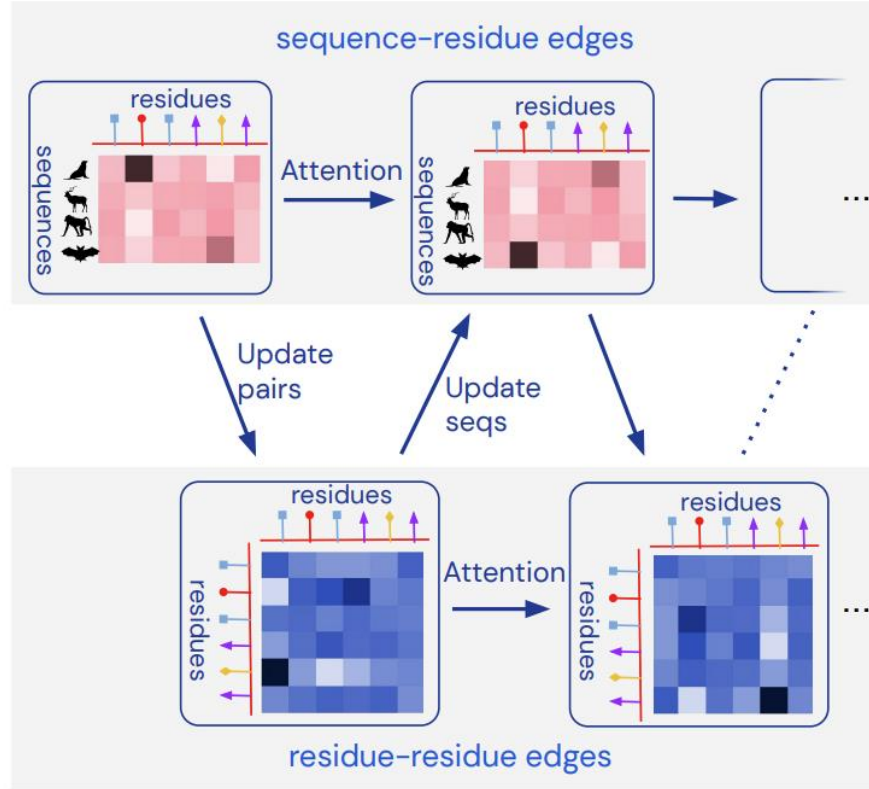


AF2 general idea

Embedding

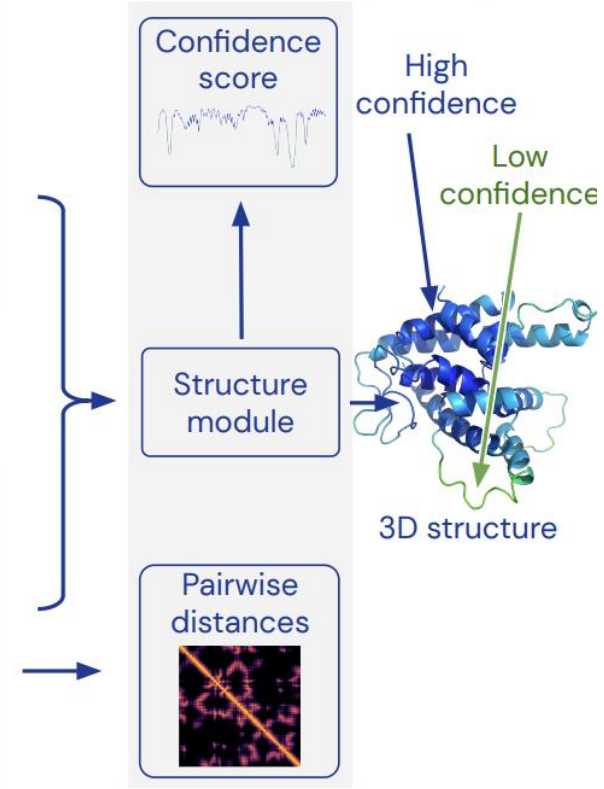


Trunk



Heads

© 2020 DeepMind Technologies Limited



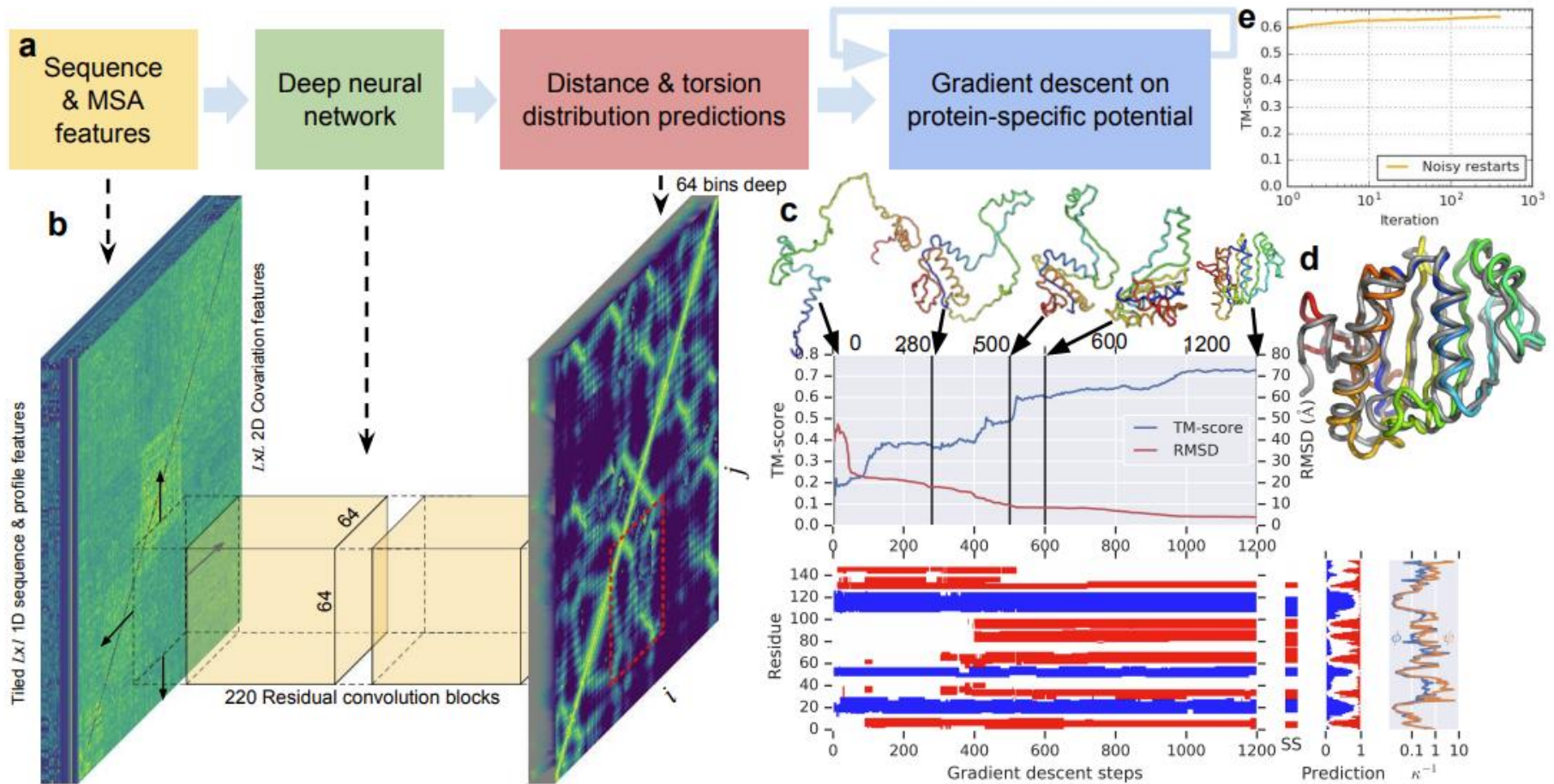
MSA picture inspired by: Riesselman, A.J., Ingraham, J.B. & Marks, D.S.,
Nature Methods (2018) doi:10.1038/s41592-018-0138-4



Adapted from

https://predictioncenter.org/casp15/doc/presentations/2020_12_01_TS_predictor_AlphaFold2.pdf

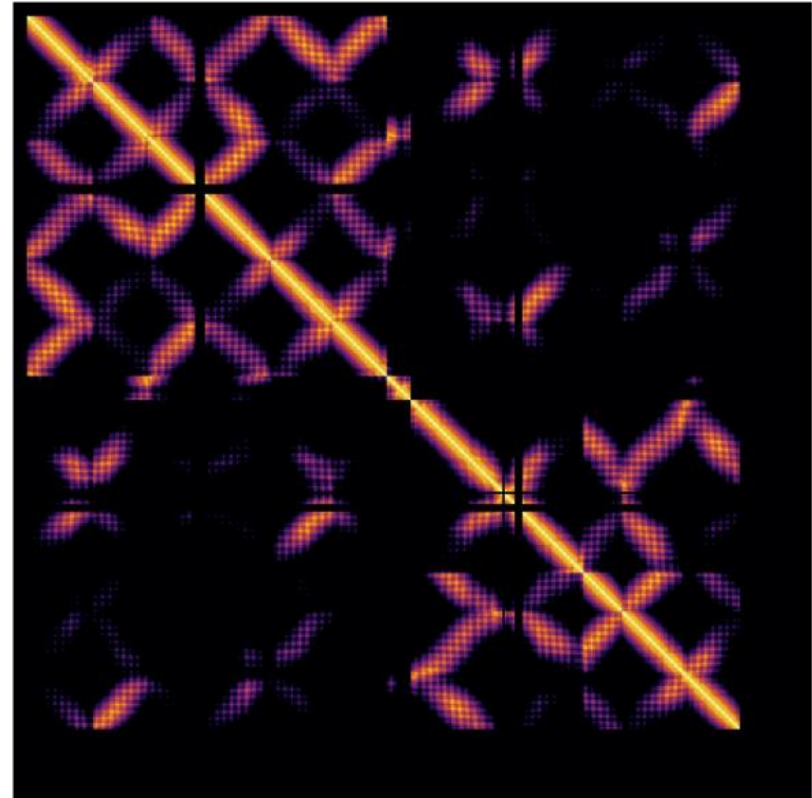
Structure prediction



Adapted from Figure 2

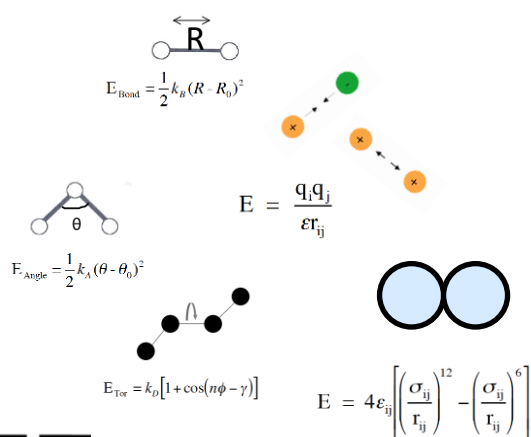
Template embedding

- 4 templates used from PDB70 clusters.
- Input features are sequences, side chains, and distograms.
- Templates are processed in the same way as the residue-residue representation.
- Templates are basically used to adjust the backbones for the given sequence.

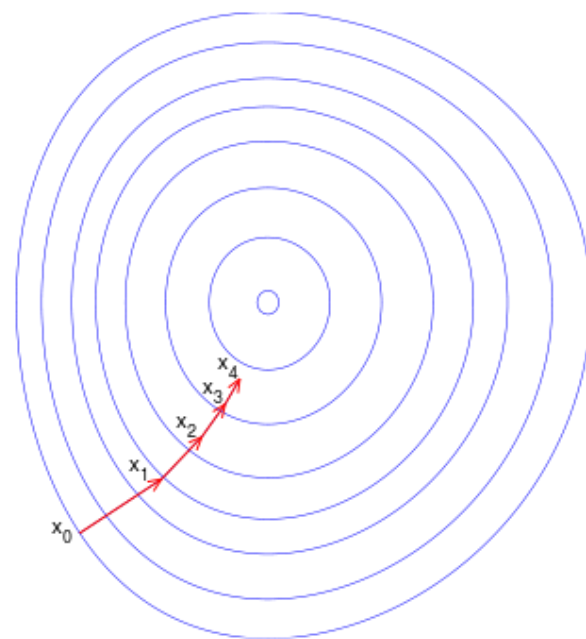


Relaxation

- The predicted models may not have good stereochemical properties
- A gradient-descent relaxation step w/ Amber ff99SB force field and coordinate restraints
- note: AF2 models lack hydrogens



$$\begin{aligned}
 V(\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N) = & \sum_{\text{bonds}} \frac{1}{2} K_b (R - R_0)^2 \\
 & + \sum_{\text{angles}} \frac{1}{2} K_a (\theta - \theta_0)^2 \\
 & + \sum_{\text{dihedrals}} K_d [1 + \cos(n\phi - \gamma)] \\
 & + \sum_{ij} 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right] + \frac{q_i q_j}{\epsilon r_{ij}}
 \end{aligned}$$



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



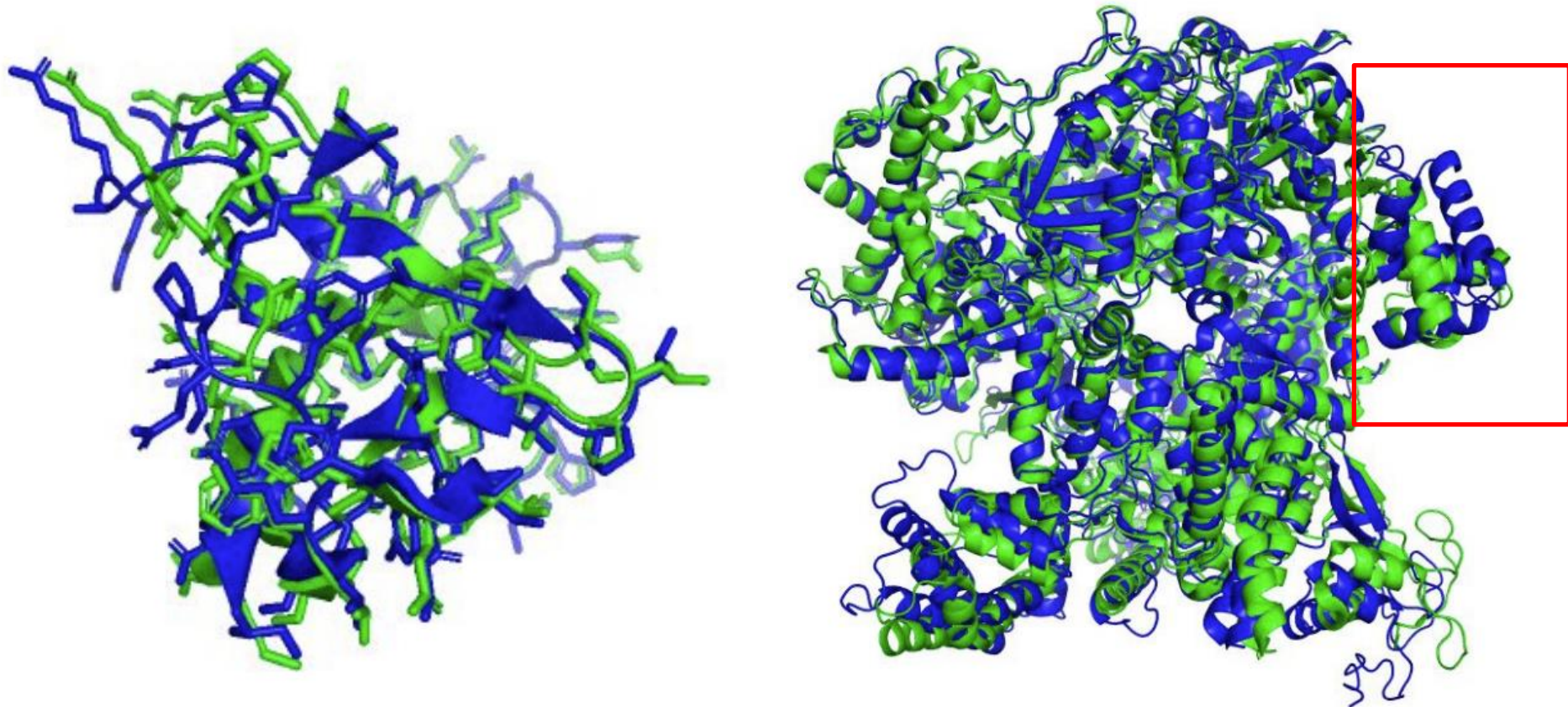
Confirmation

- pLDDT (Local Distance Difference Test) and/or IDDT-Ca are used as the metrics to select the “best” structure among the multiple structures generated by AF2.
- What these methods measure is roughly a “likelihood to be a realistic structure” based on all the protein structures in the PDB.
- These values also indicate lack of structure in some cases: low pLDDT values typically indicate lack of secondary structure rather than a poor prediction.

$$lDDT_r = \frac{100}{4L} \sum_{t \in \{0.5, 1, 2, 4\}} \sum_{i=1}^L \frac{\sum_{j, |i-j| \geq r, D_{ij} < 15} \mathbf{1}(|D_{ij} - d_{ij}| < t)}{\sum_{j, |i-j| \geq r, D_{ij} < 15} 1}.$$



AF2 prediction examples



Ground truth
Prediction



Adapted from

https://predictioncenter.org/casp15/doc/presentations/2020_12_01_TS_predictor_AlphaFold2.pdf

How can you run AF2?

- Local installations: Install it on your computer and run it from there. This is the only “full” version whereby the MSA and template searches are done on all the libraries (<https://github.com/deepmind/alphafold>).
- ChimeraX plugin: ChimeraX has an AF2 plugin by default, which you can use to run AF2 directly (<https://www.cgl.ucsf.edu/chimerax/>).
- Colab: There is also a colab version of AF2 that can be used to run it remotely on a server without the need to install it locally (<https://colab.research.google.com/github/deepmind/alphafold/blob/main/notebooks/AlphaFold.ipynb>)

ChimeraX and Colab use truncated MSAs and either eliminate or limit template searches, which may reduce the accuracy of the results.

These two methods also suffer from size limitations due to the limited availability of CPU and GPUs on the remote servers.



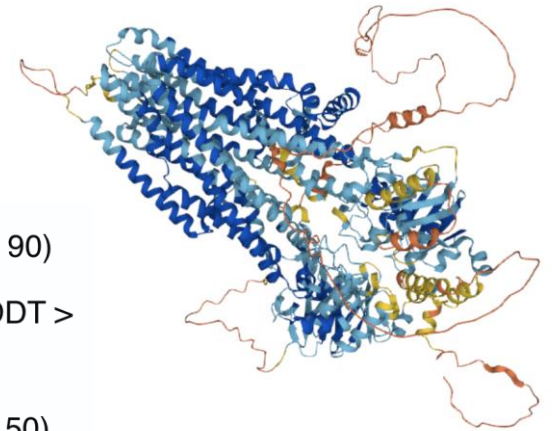
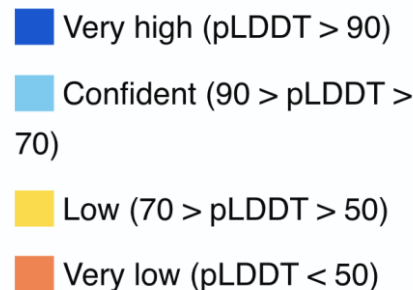
Additional features – oligomeric structures

- AF2 can also model oligomeric proteins using the AlphaFold Multimer model.
- workflow is the same except that more than one sequence are provided separated by a " : "
- AF2 determines the relative coordinates of the individual subunits automatically and generates a model for the complex.

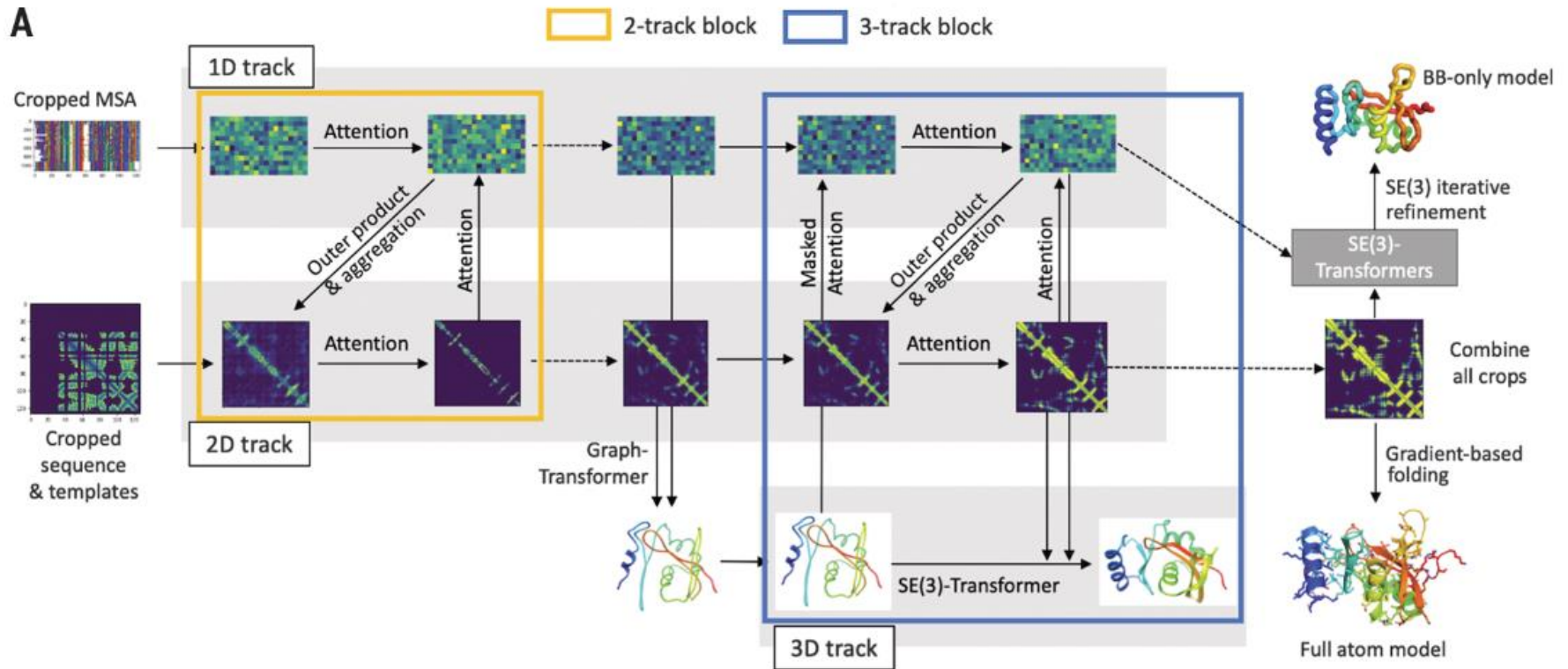


AF2 shortcomings

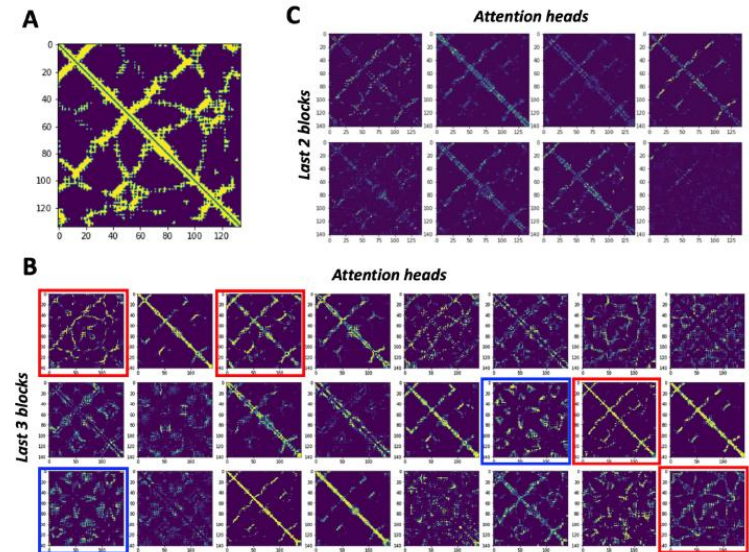
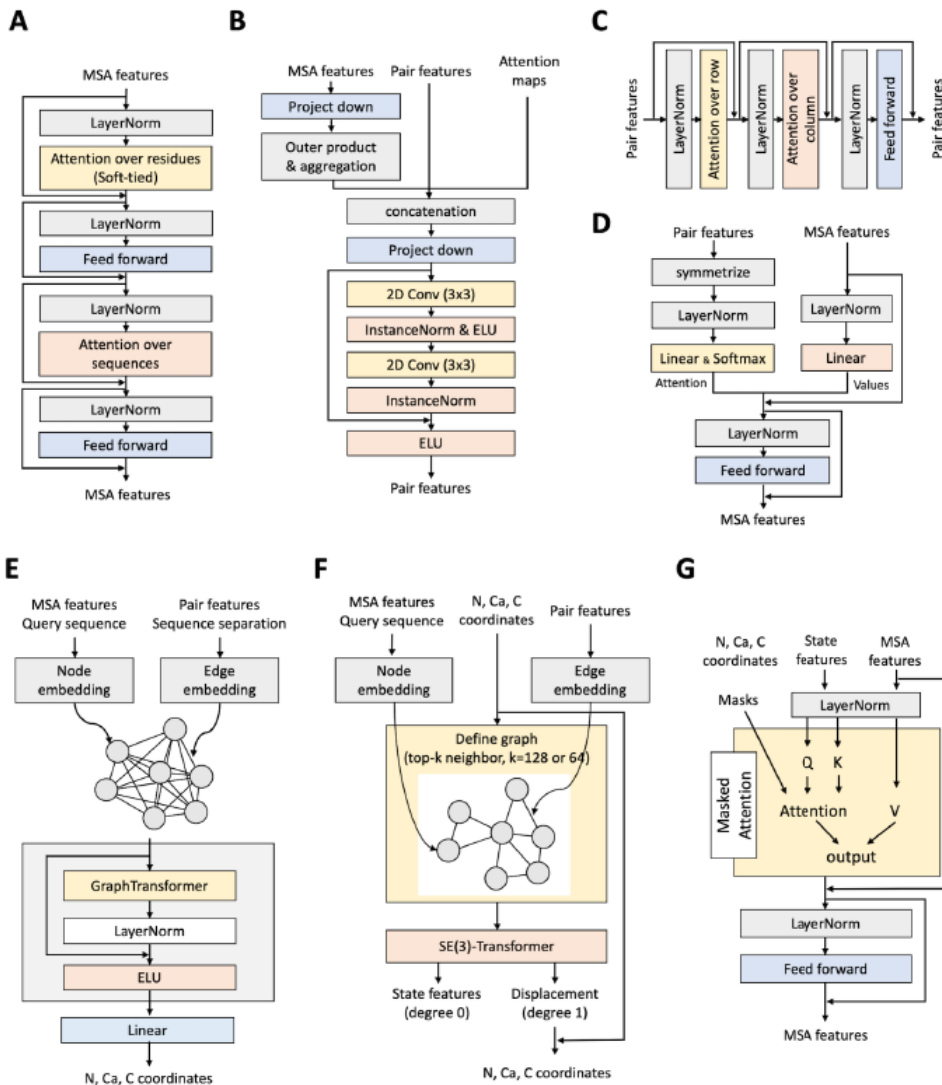
- the predictions are not always perfect
- AF2 is being constantly improved both by DeepMind and other scientists in the field
- The prediction accuracy of AF2 may vary depending on your system
- AF2 offers little for disordered systems



What is RoseTTAFold?



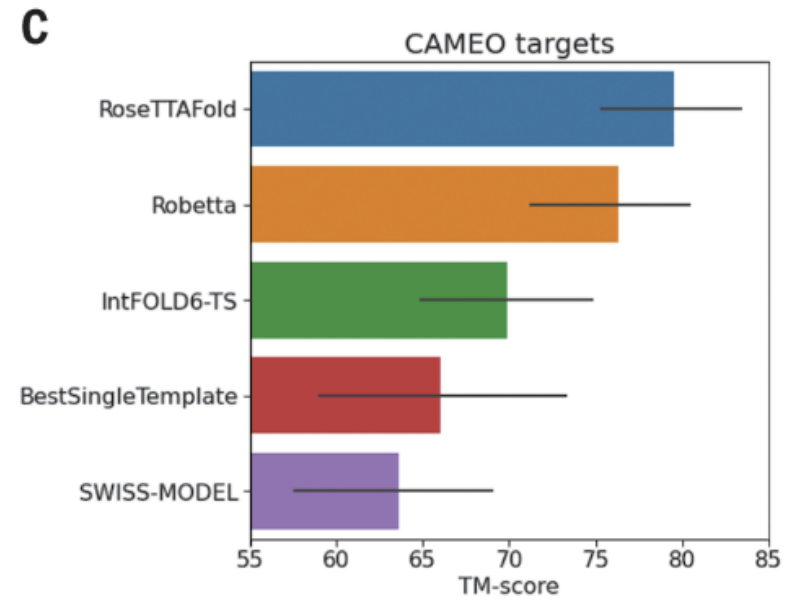
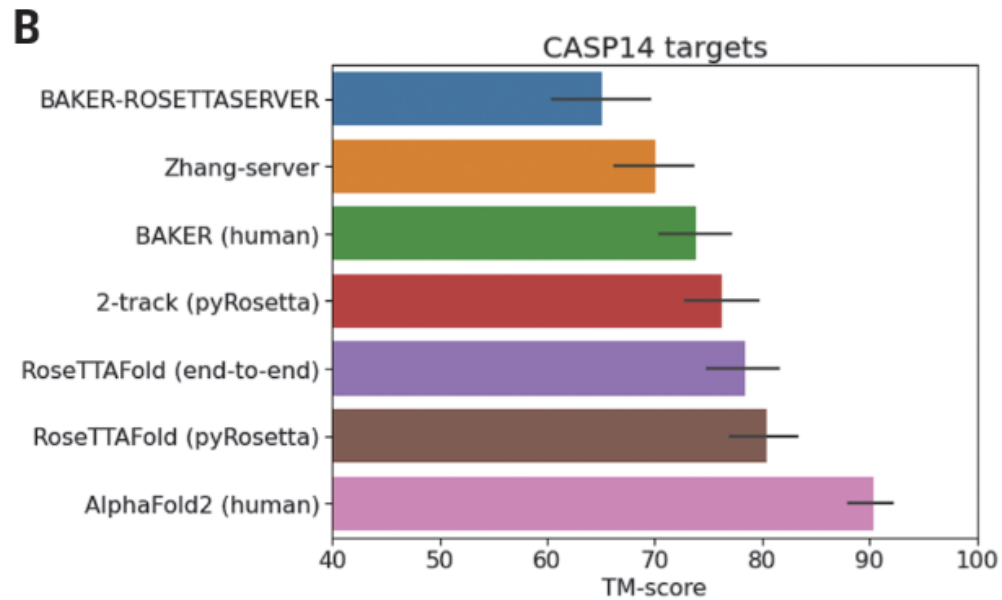
RoseTTAFold method



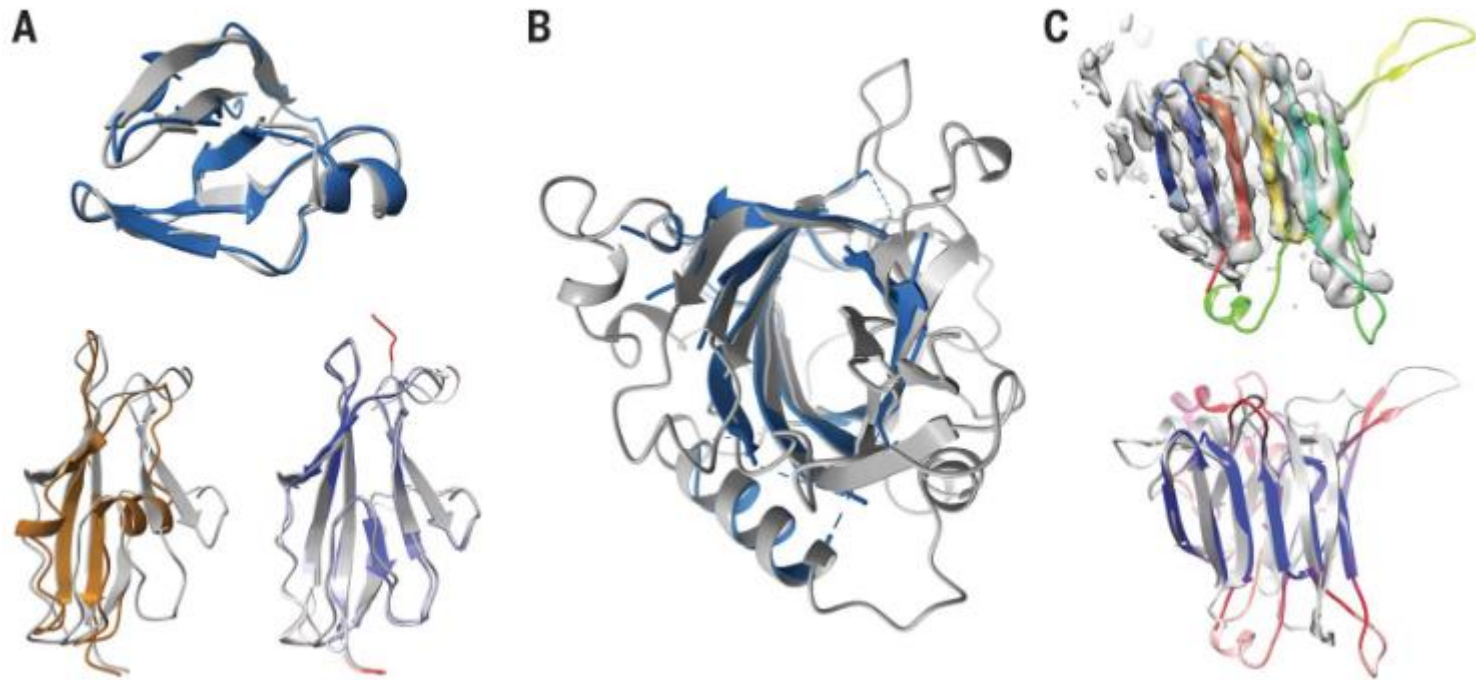
Red box: resembles true contact map
 Blue box: positions not making direct contacts



RoseTTAFold versus other methods



RoseTTAFold monomer predictions



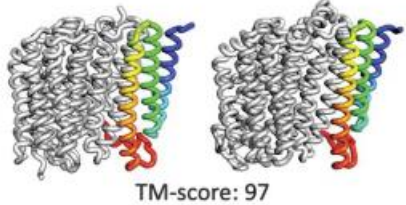
Additional features – oligomeric structures and all atom

- RoseTTAFold can also model oligomeric proteins using the 3-track algorithm
- There is a separate script under the RoseTTAFold folders that can be used for this purpose.
- Use of the complex application requires paired alignments. There is a script under the RoseTTAFold folders for prokaryotic proteins, but there is no trivial way for eukaryotic proteins.

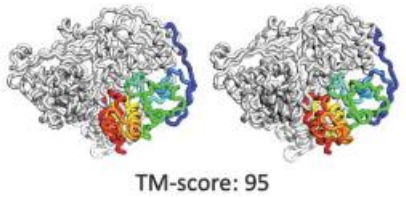


RoseTTAFold complex predictions

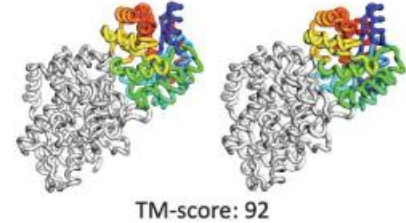
A NADH-quinone oxidoreductase (nuoJ/nuoK)



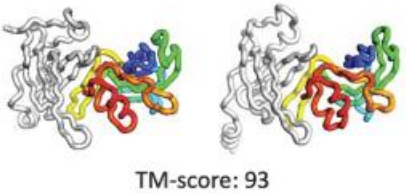
Aldehyde oxidoreductase (paoA/paoB)



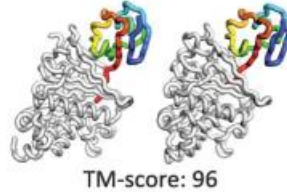
Tryptophan synthase (trpA/trpB)



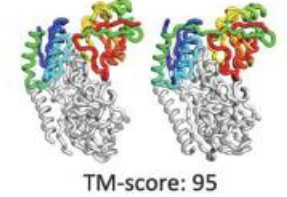
50S ribosome (rplS/rplN)



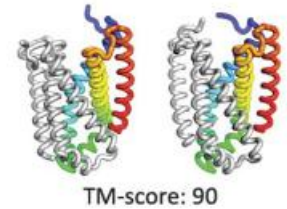
thiS/thiF complex



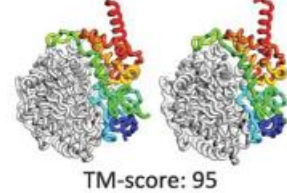
5-oxoprolinase (pxpB/pxpC)



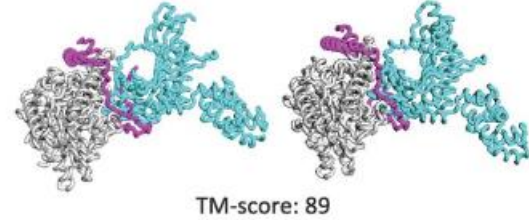
Succinate dehydrogenase (sdhD/sdhC)



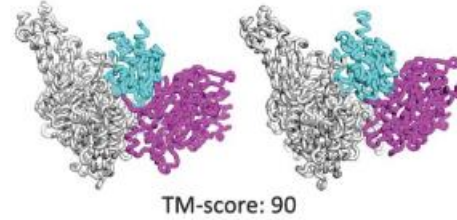
Fumarate reductase (frdA/frdC)



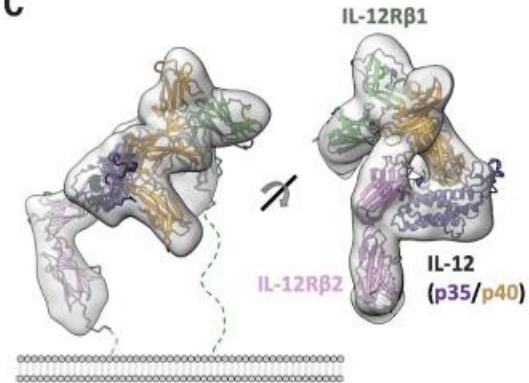
B tRNA-dependent amidotransferase



4-hydroxybenzoyl-CoA reductase



C



How to run RoseTTAFold

- Local installations: Install it on your computer and run it from there (<https://github.com/RosettaCommons/RoseTTAFold>).
- Colab: There is also a colab version of RoseTTAFold that can be used to run it remotely on a server without the need to install it locally (<https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/RoseTTAFold.ipynb>).



Applications and Outlook

- Alternative conformation predictions and MSA clustering

Sampling alternative conformational states of transporters and receptors with AlphaFold2

Diego del Alamo^{1,2†}, Davide Sala^{3†}, Hassane S Mchaourab^{1*}, Jens Meiler^{2,3*}

Predicting multiple conformations via sequence clustering and AlphaFold2

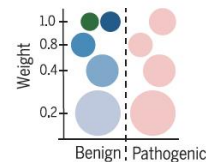
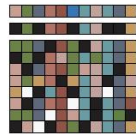
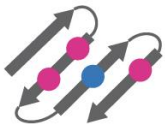
Received: 7 July 2023

Accepted: 3 November 2023

Hannah K. Wayment-Steele, Adedolapo Ojoawo, Renee Otten, Julia M. Apitz, Warintra Pitsawong, Marc Hömberger, Sergey Ovchinnikov, Lucy Colwell & Dorothee Kern

- Predict mutation effects

- 1 Structure context
- 2 Protein language modeling
- 3 Training variants



- Protein universe prediction

Article

Uncovering new families and folds in the natural protein universe

<https://doi.org/10.1038/s41586-023-06622-3>

Received: 24 March 2023

Accepted: 7 September 2023

Published online: 13 September 2023

Janani Durairaj^{1,2}, Andrew M. Waterhouse^{1,2}, Toomas Mets^{3,4}, Tatlana Brodlazhenko⁵, Minhaj Abdullah^{1,4}, Gabriel Studer^{1,2}, Gerardo Tauriello^{1,2}, Mehmet Akdel⁶, Antonina Andreeva⁶, Alex Bateman⁶, Tanel Tonson⁷, Vasilii Hauryliuk^{3,4,2,8}, Torsten Schwede^{1,2,8} & Joana Pereira^{1,2,8}

Accurate proteome-wide missense variant effect prediction with AlphaMissense

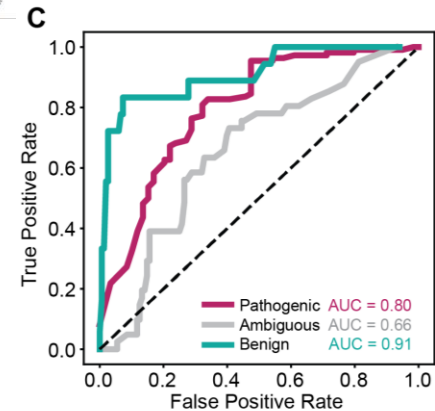
Jun Cheng^{*}, Guido Novati, Joshua Pan[†], Clare Bycroft[†], Akvilė Žemgulytė[†], Taylor Applebaum[†], Alexander Pritzel, Lai Hong Wong, Michal Zielinski, Tobias Sargeant, Rosalia G. Schneider, Andrew W. Senior, John Jumper, Demis Hassabis, Pushmeet Kohli^{*}, Žiga Avsec^{*}

PROTEOME

PROTEIN
LANDSCAPE

PROTEOME
ATLAS

PROTEIN
UNIVERSE



McDonald et al. *Benchmarking AlphaMissense Pathogenicity Prediction Against Cystic Fibrosis Variants*, <https://www.biorxiv.org/content/10.1101/2023.10.05.561147v1>



Questions and Tutorial Time..



[linkedin.com/in/eli-f-mcdonald/](https://www.linkedin.com/in/eli-f-mcdonald/)



[@EliFritzMcD](https://twitter.com/EliFritzMcD)



[researchgate.net/profile/Eli-Mcdonald](https://www.researchgate.net/profile/Eli-Mcdonald)



eli.f.mcdonald@vanderbilt.edu

