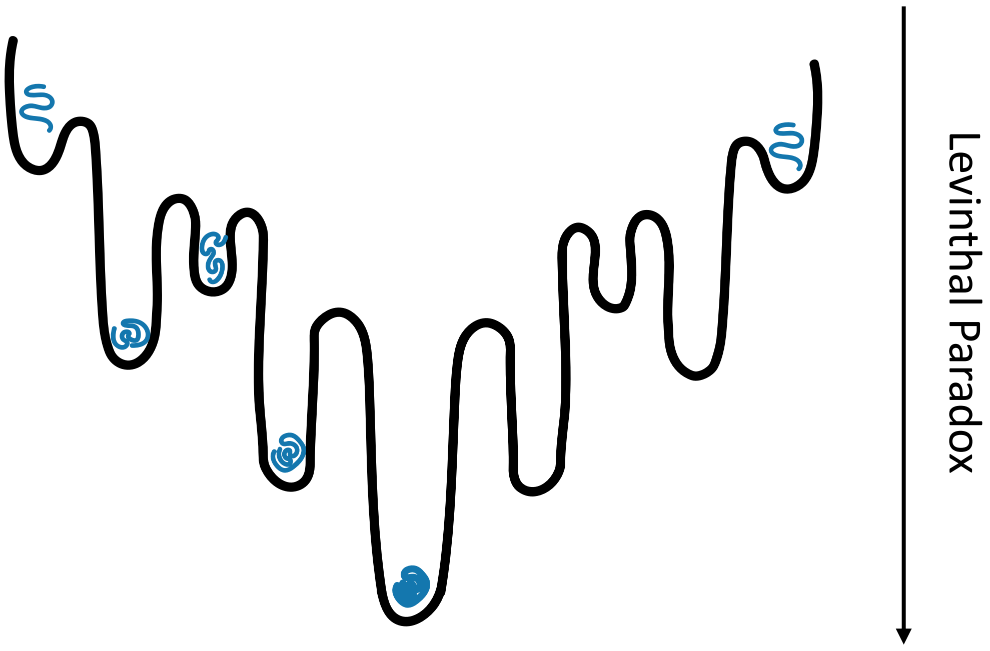
ML Based Structural Modeling and Prediction with AlphaFold2 and RoseTTAFold

Meilerlab

VANDERBILT
UNIVERSITY

Rosetta Workshop July 2024

Protein folding problem

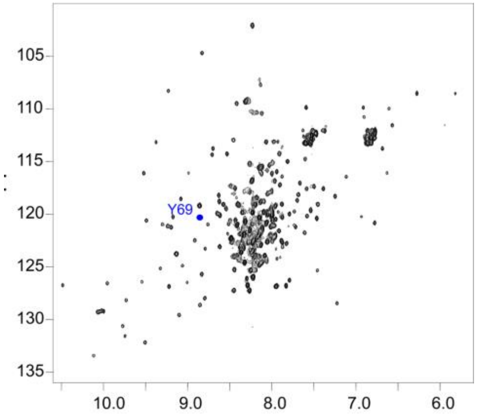


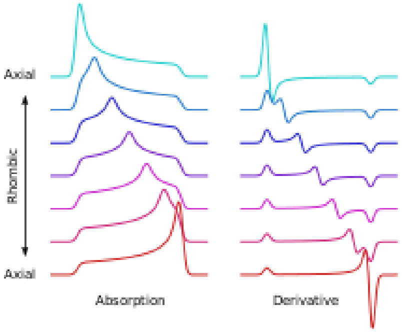
Thermodynamic Hypothesis

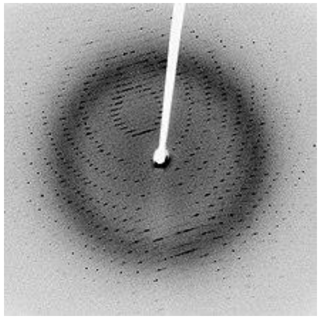
- Consider a 100 residue protein w/ 2 DOF/residue
- # of possible conformations $= 2^{100} \sim 10^{30}$
- vibration rate ~ $k_v = 10^{13}$ conformations/sec
- Time to sample all confs:
- $T = j^N / Nk_v$
  $= 10^{30} / (100 * 10^{13})$
  $= 10^{17}$ sec $= 10^{10}$ years
- Age of universe $= 13.4 *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







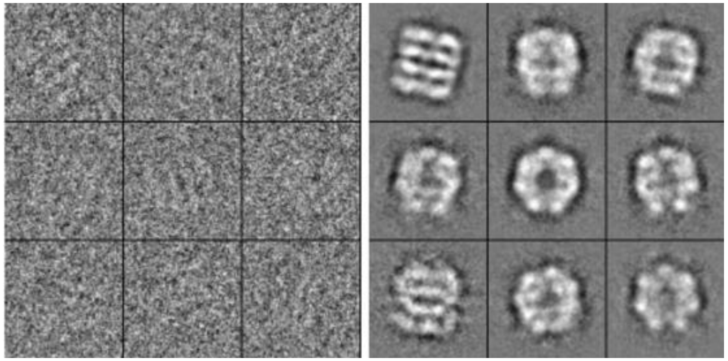
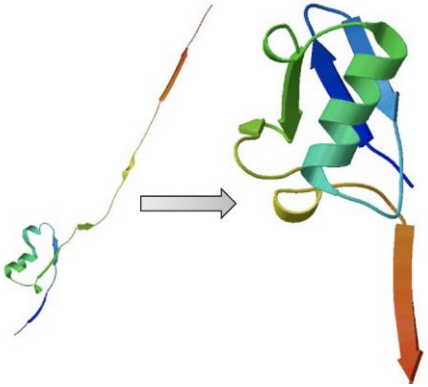


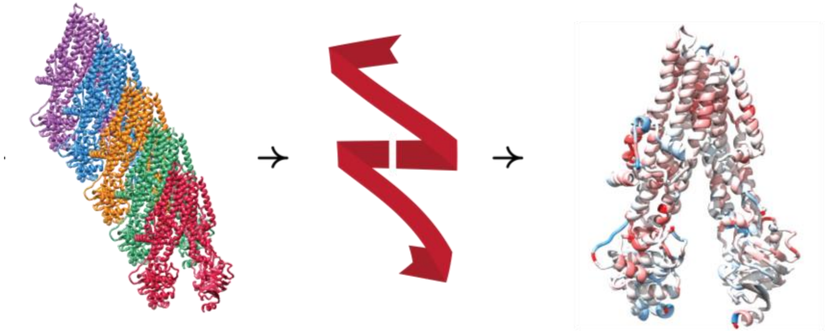
image sources: Wikipedia, jiang.bio.purdue.edu/

Computational structure approaches

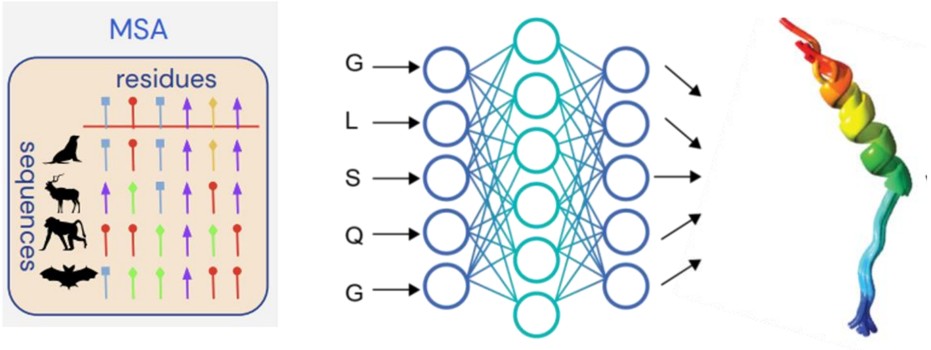
- De novo folding
  - sequence information
  - fragment libraries
  - physics/statistics based



- Homology modeling
  - structural information
  - templates + fragment libraries

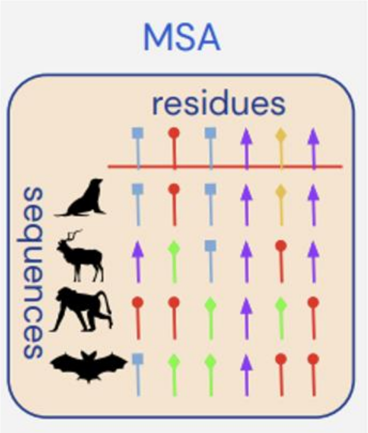


- Deep learning methods
  - evolutionary information
  - predict residue distances
  - structural training data



Deep learning methods

- MSA-based methods
  - use MSAs to infer information about residue pairs
  - predicts 3D structures based on pairs
  - Often includes template information
  - AlphaFold2, RoseTTAFold, ColabFold



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

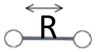
What is AlphaFold2 (AF2)

- 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, but can use template information

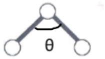
- Generated models are “relaxed” further using a classical mechanical force field and analyzed for accuracy

Covalent Interactions



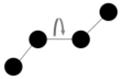
$E_{\text{Bond}} = \frac{1}{2} k_B (R - R_0)^2$

1) bond length



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

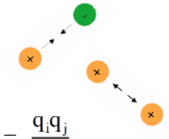
2) bond angle



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

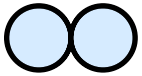
3) torsion angle

Non-covalent Interactions

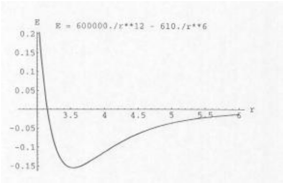


$E = \frac{q_i q_j}{\varepsilon r_{ij}}$

1) electrostatic



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



2) Van der Waals

AF2 Inputs

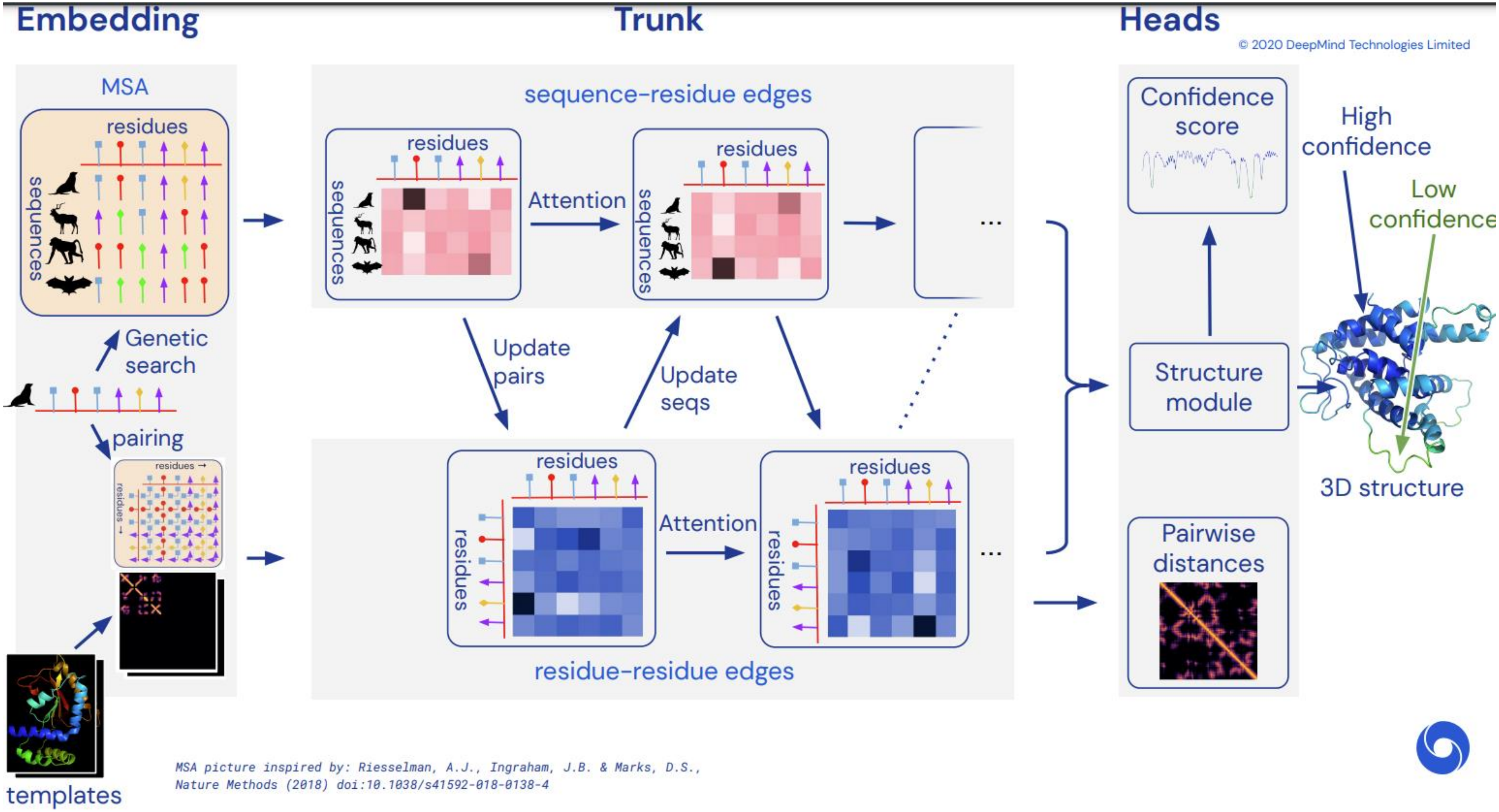
Sequence Databases

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

Structural Databases

- PDB (training)
- PDB70 clustering (hhsearch)

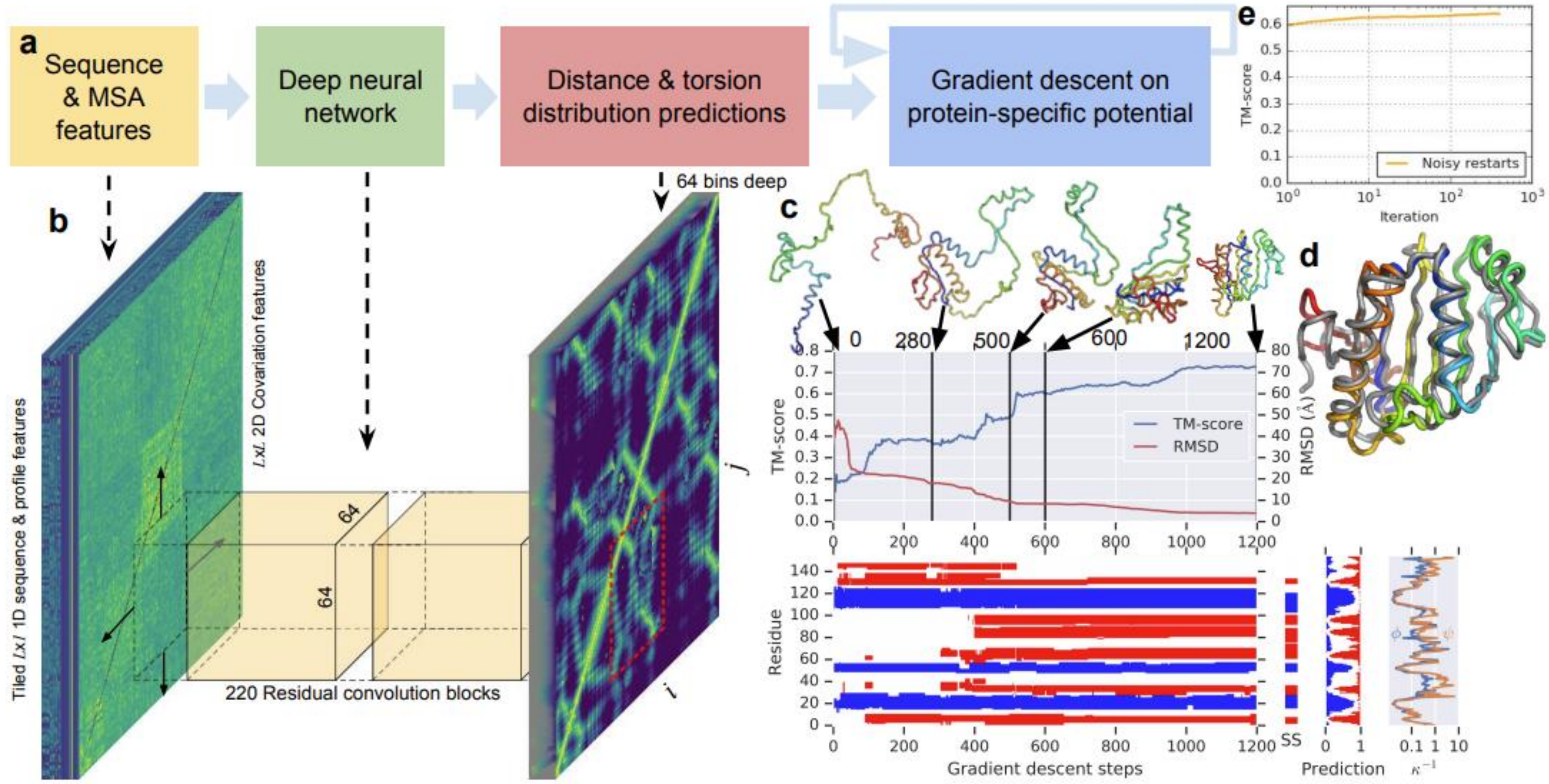
AF2 general idea



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

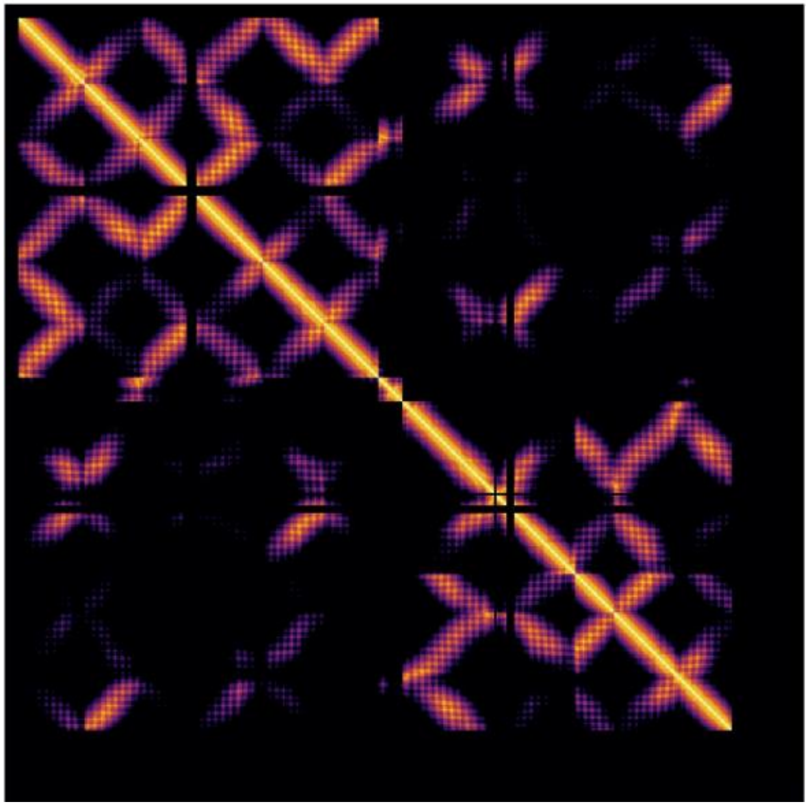

a
Sequence & MSA features
Deep neural network
Distance & torsion distribution predictions
Gradient descent on protein-specific potential
b
Tiled L×1 1D sequence & profile features
L×L 2D Covariation features
64
64
220 Residual convolution blocks
64 bins deep
j
i
c
0
280
500
600
1200
TM-score
RMSD (Å)
TM-score
RMSD
Gradient descent steps
Residue
SS
Prediction
κ⁻¹
φ
ψ
d
e
TM-score
Iteration
Noisy restarts

Adapted from Figure 2

https://discovery.ucl.ac.uk/id/eprint/10089234/1/343019_3_art_0_py4t4l_convrt.pdf

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.



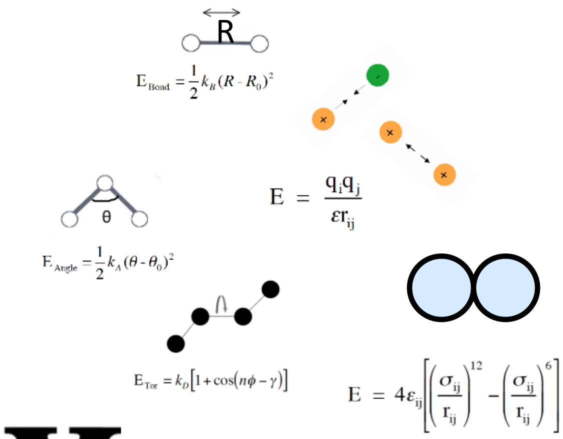
Adapted from
https://predictioncenter.org/casp15/doc/presentations/2020_12_01_TS_predictor_AlphaFold2.pdf

Relaxation

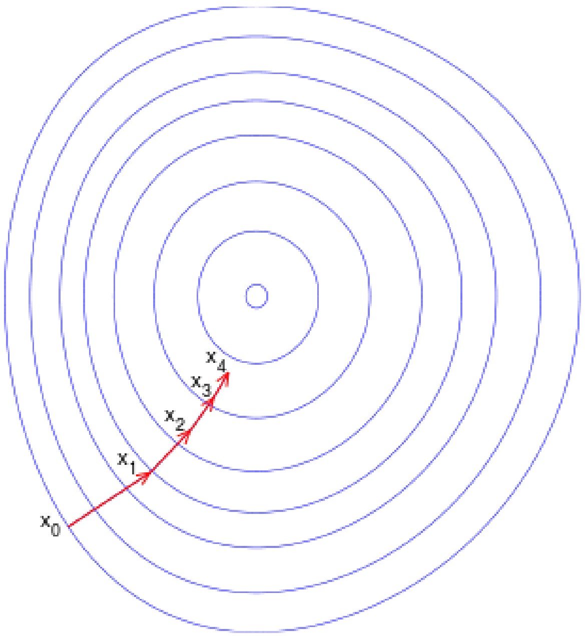
- The predicted models may not have good stereochemical properties

- A gradient-descent relaxation step w/ Amber ff99SB force field and coordinate restraints

- note: AF2 models lack hydrogens


R
$E_{Bond} = \frac{1}{2}k_B(R - R_0)^2$
$E = \frac{q_i q_j}{\varepsilon r_{ij}}$
θ
$E_{Angle} = \frac{1}{2}k_A(\theta - \theta_0)^2$
$E_{Tor} = k_D[1 + \cos(n\phi - \gamma)]$
$E = 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right]$

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



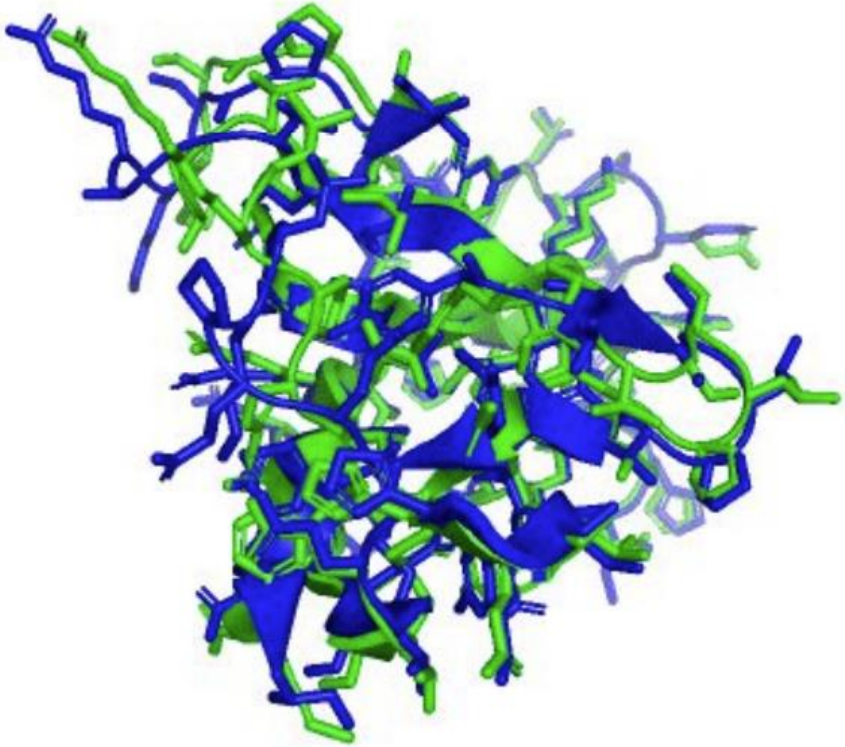
https://en.wikipedia.org/wiki/Gradient_descent

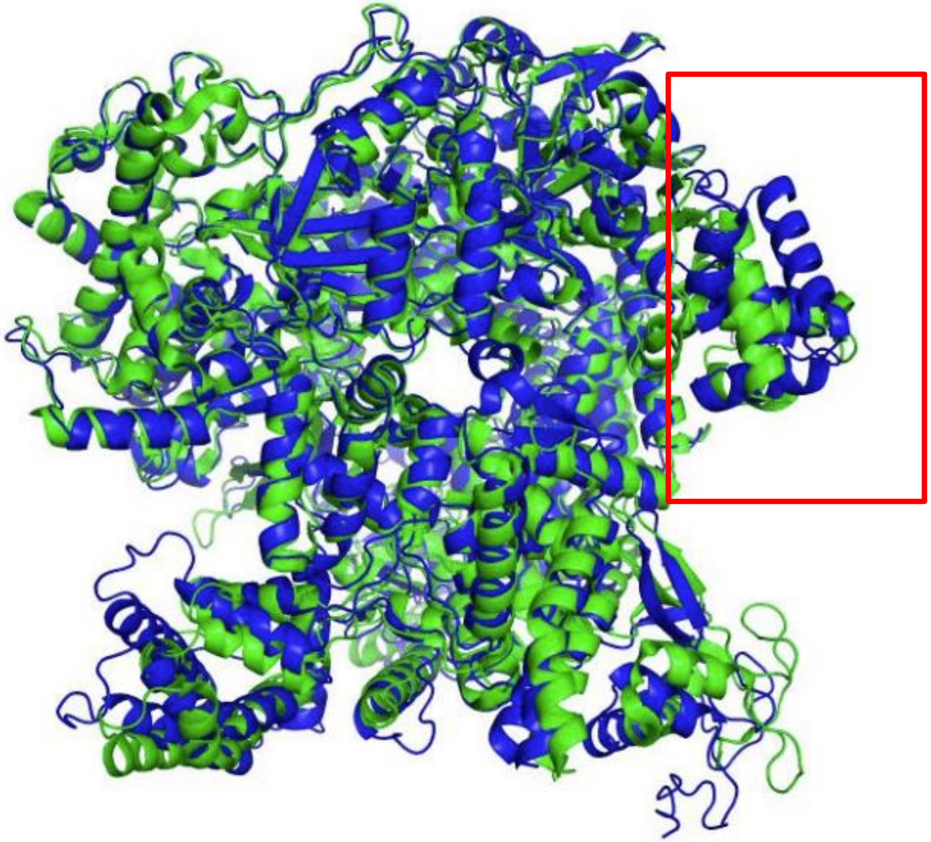
Confirmation

- pLDDT (Local Distance Difference Test) and/or lDDT-Cα 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.

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

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.

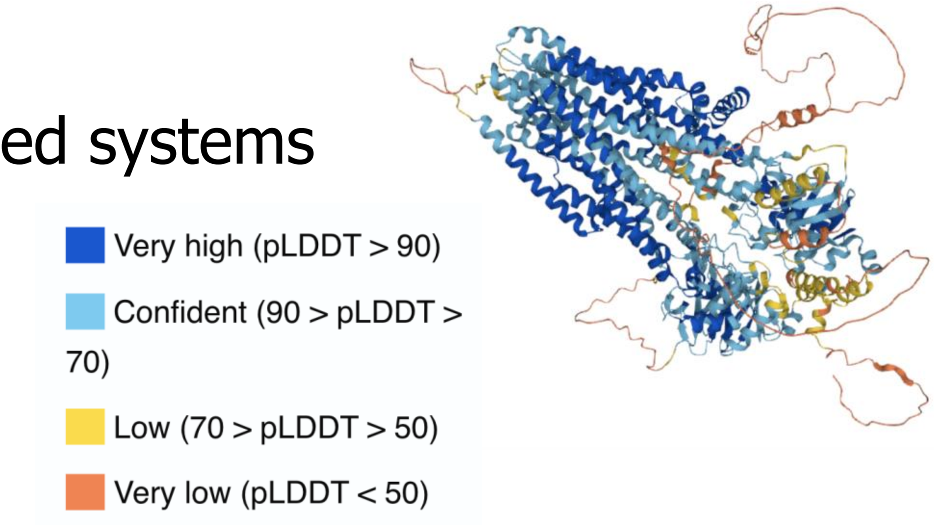
- workflow is the same except that more than one sequence are provided separated by a ” : ”

- AF2 determines the relative coordinates of the individual subunits automatically and generates a model for the complex.

Evans et al. 2022 (bioRxiv)

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



ColabFold

What is ColabFold?

- Combines MMseqs2 with AlphaFold2 or RoseTTAFold.
- 40-60 fold increase in search speed.
- There are features that user can manipulate AlphaFold2

Key Features

- Predicts up to 1,000 structures with a one GPU per day
- Free and accessible
- Available on GitHub.

https://doi.org/10.1038/s41592-022-01488-1.

How ColabFold Works

Components

- **MMseqs2 Server**: Builds MSAs, finds templates
- **Python Library**: Prepares inputs, visualizes results
- **Jupyter Notebooks**: For various use cases

Optimization

- Faster MSA generation with MMseqs2
- Possible early to stop criteria of batch predictions
- Comprehensive searches with combined databases

ColabFold Architecture

Workflow

- **Interface**: Send sequences to MMseqs2 server.
- **Search**: UniRef100 and environmental sequences.
- **Prediction**: Single structures and complexes.
- **Visualization**: MSA depth, AlphaFold2 confidence.

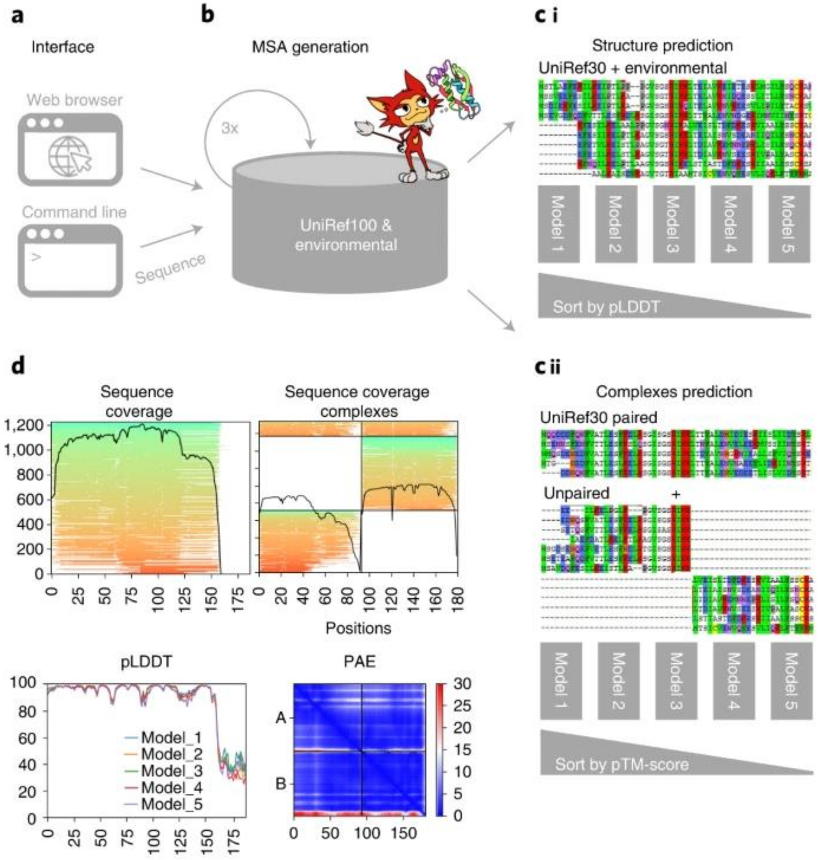
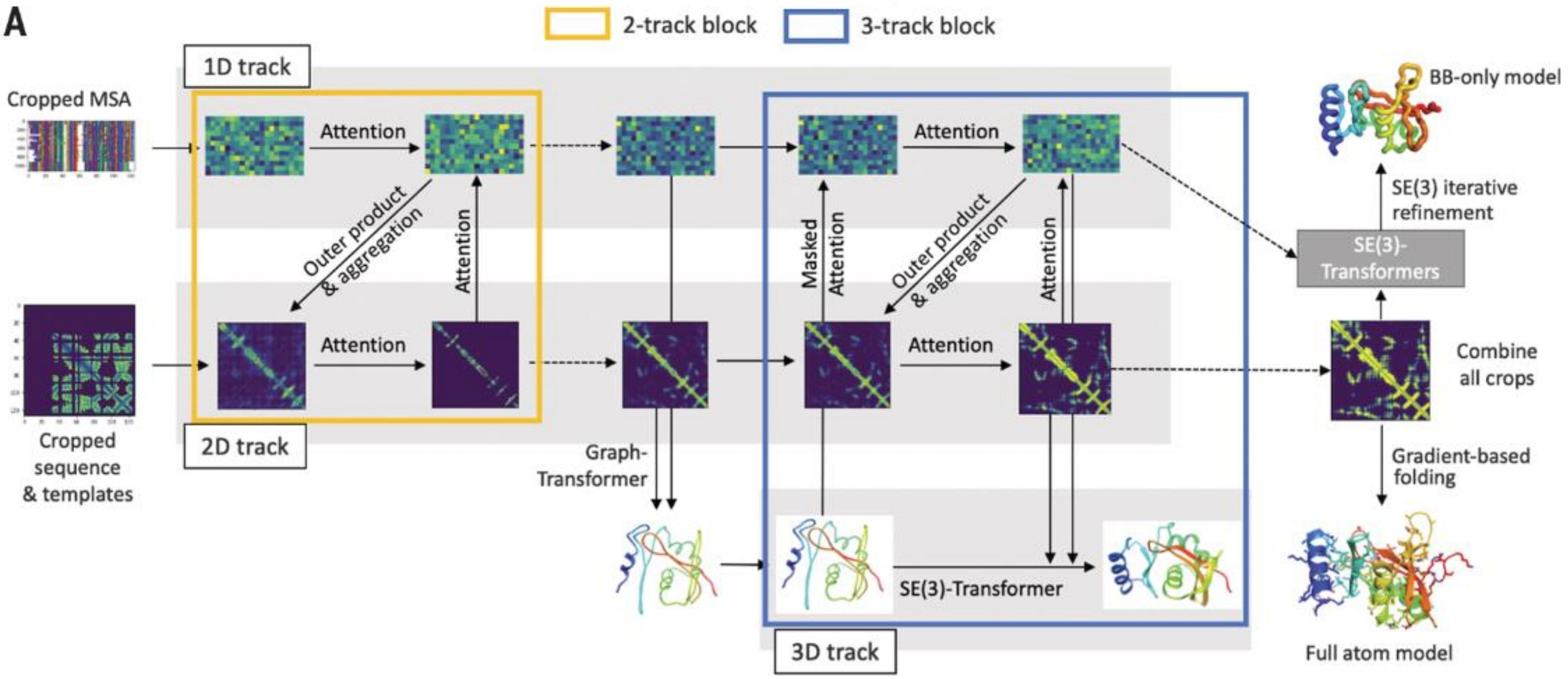


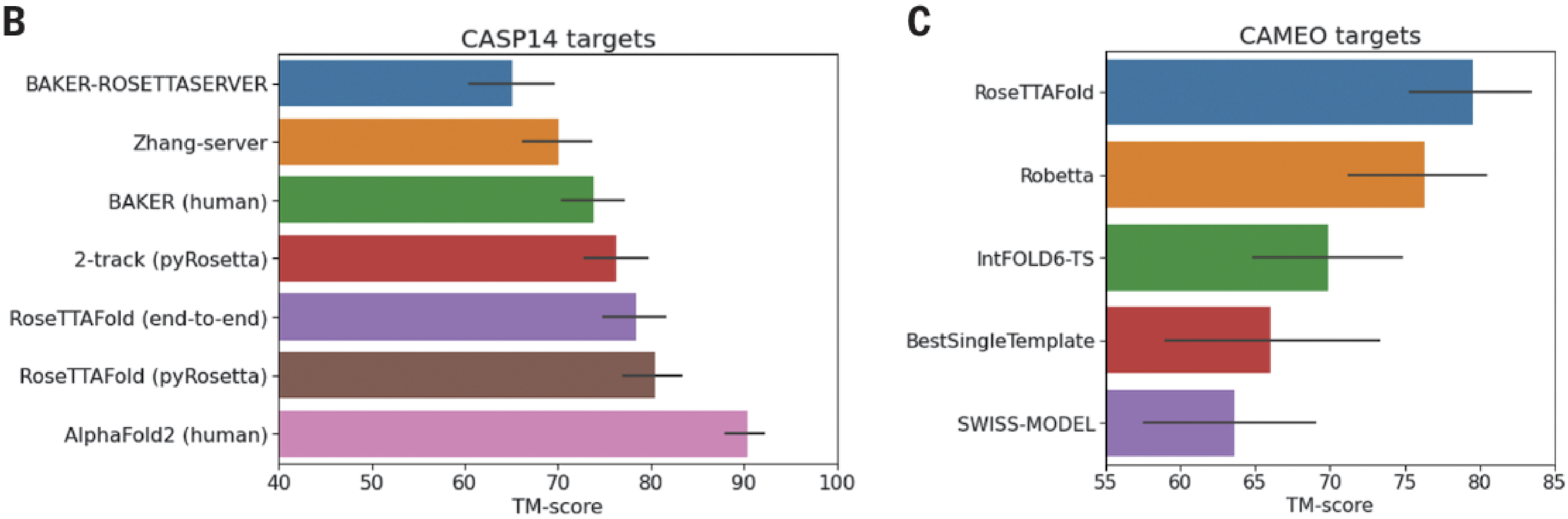
Figure 1: Schematic Diagram of ColabFold

What is RoseTTAFold?



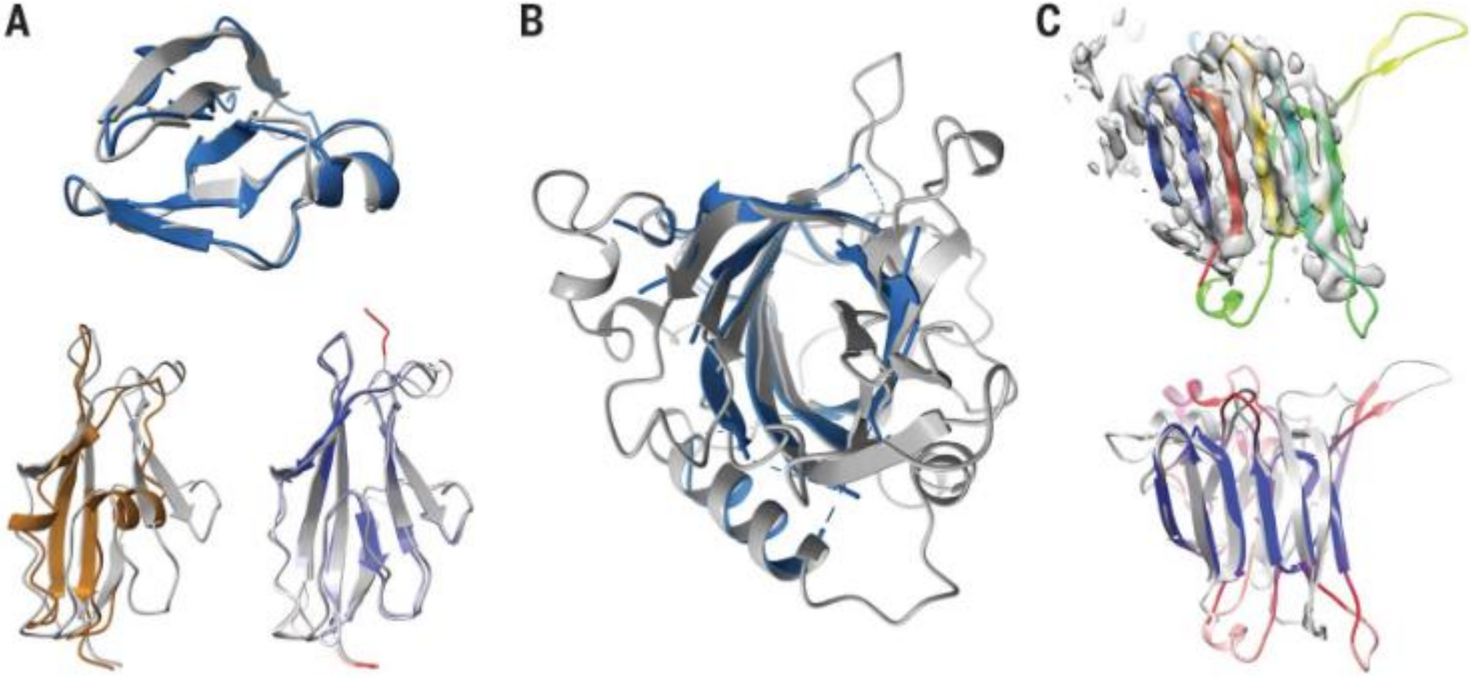
Baek et al. 2021

RoseTTAFold versus other methods



Baek et al. 2021

RoseTTAFold monomer predictions



Additional features – oligomeric structures and all atom

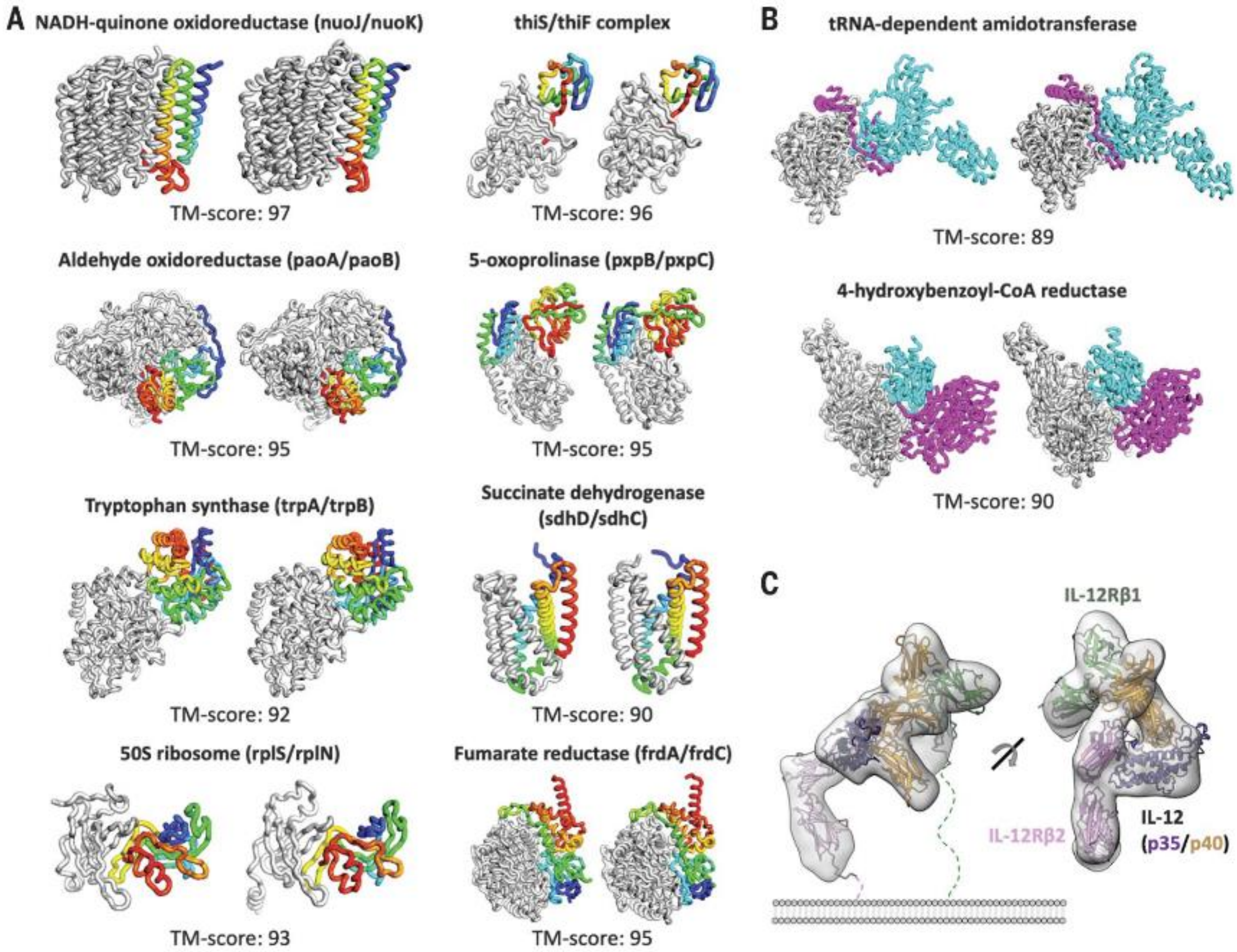
- RoseTTAFold can also model oligomeric proteins using the 3-track algorithm

- 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.

Evans et al. 2022 (bioRxiv)

RoseTTAFold complex predictions


A
NADH-quinone oxidoreductase (nuoJ/nuoK)
TM-score: 97
thiS/thiF complex
TM-score: 96
Aldehyde oxidoreductase (paoA/paoB)
TM-score: 95
5-oxoprolinase (pxpB/pxpC)
TM-score: 95
Tryptophan synthase (trpA/trpB)
TM-score: 92
Succinate dehydrogenase (sdhD/sdhC)
TM-score: 90
50S ribosome (rplS/rplN)
TM-score: 93
Fumarate reductase (frdA/frdC)
TM-score: 95
B
tRNA-dependent amidotransferase
TM-score: 89
4-hydroxybenzoyl-CoA reductase
TM-score: 90
C
IL-12Rβ1
IL-12Rβ2
IL-12
(p35/p40)

Baek et al. 2021

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*]

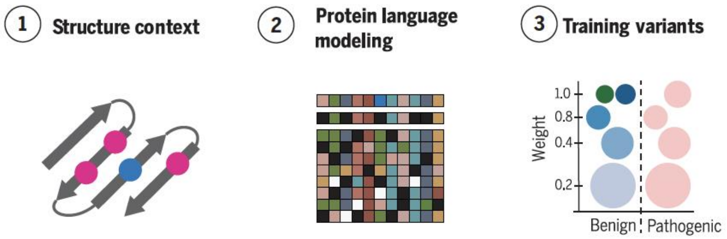
Predicting multiple conformations via sequence clustering and AlphaFold2

Received: 7 July 2023

Accepted: 3 November 2023

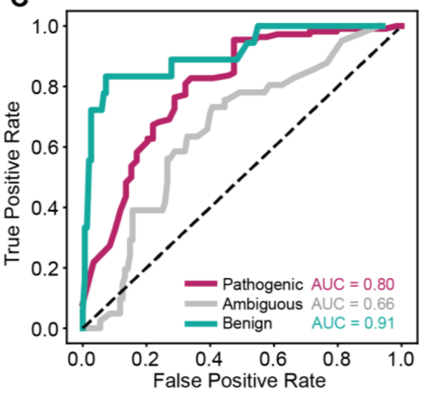
Hannah K. Wayment-Steele, Adedolapo Ojoawo, Renee Otten, Julia M. Apitz, Warintra Pitsawong, Marc Hömberger, Sergey Ovchinnikov, Lucy Colwell & Dorothee Kern

- Predict mutation effects



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*



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

- 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], Tetiana Brodiazhenko[3], Minhal Abdullah[3,4], Gabriel Studer[1,2], Gerardo Tauriello[1,2], Mehmet Akdel[5], Antonina Andreeva[6], Alex Bateman[6], Tanel Tenson[3], Vasili Hauryliuk[3,4,7,8], Torsten Schwede[1,2✉] & Joana Pereira[1,2✉]