

Protein Structure Prediction

BCH394P/364C Systems Biology/Bioinformatics
April 2, 2024
Daryl Barth

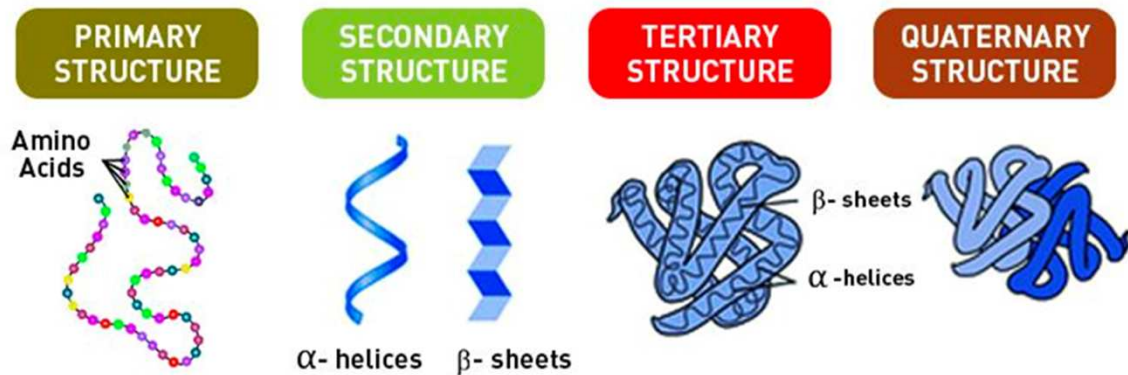
Today's Goals



- Motivation and bit of history on protein structure prediction
- CASP
- AlphaFold
- Metrics you need to know
- ColabFold
- Demo!

Why?

The four levels of protein structure



Is an amino acid sequence all you need?

- Anfinsen's dogma:
- The Protein Folding Problem
 - What is the folding code?
 - What is the folding mechanism?
 - Can we predict a native protein structure from its primary, amino acid sequence?

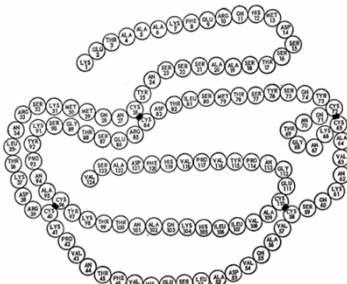
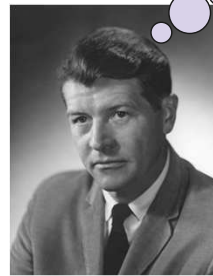
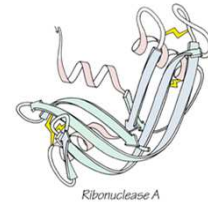


Fig. 1. The amino acid sequence of bovine pancreatic ribonuclease (50).

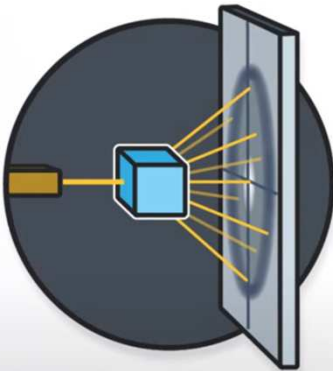


A protein's native structure stands for a free energy minimum determined by its amino acid sequence...

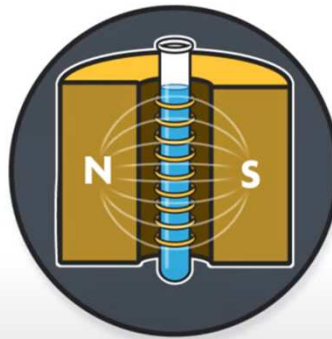


Experimental Methods for Determining Protein Structure

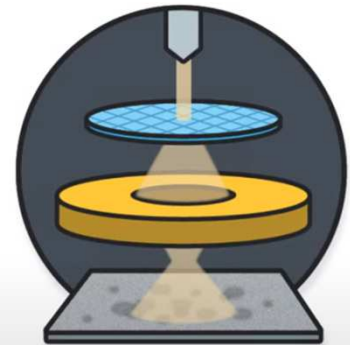
X-Ray crystallography



Nuclear magnetic resonance spectroscopy

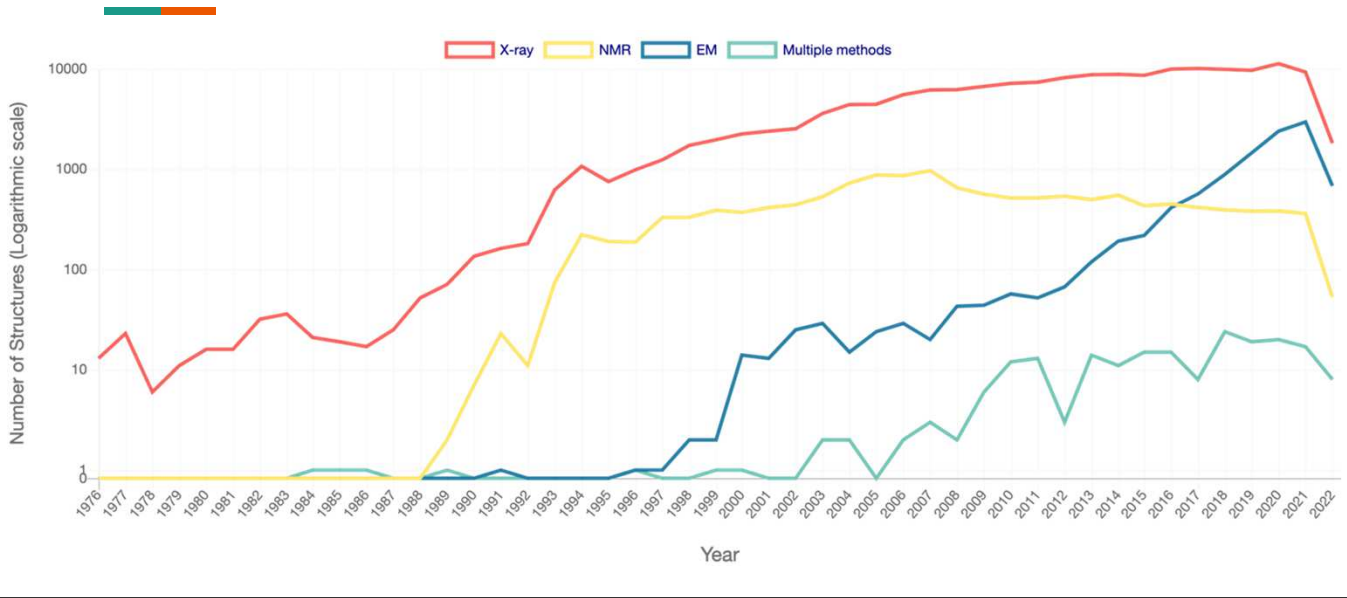


Cryoelectron microscopy

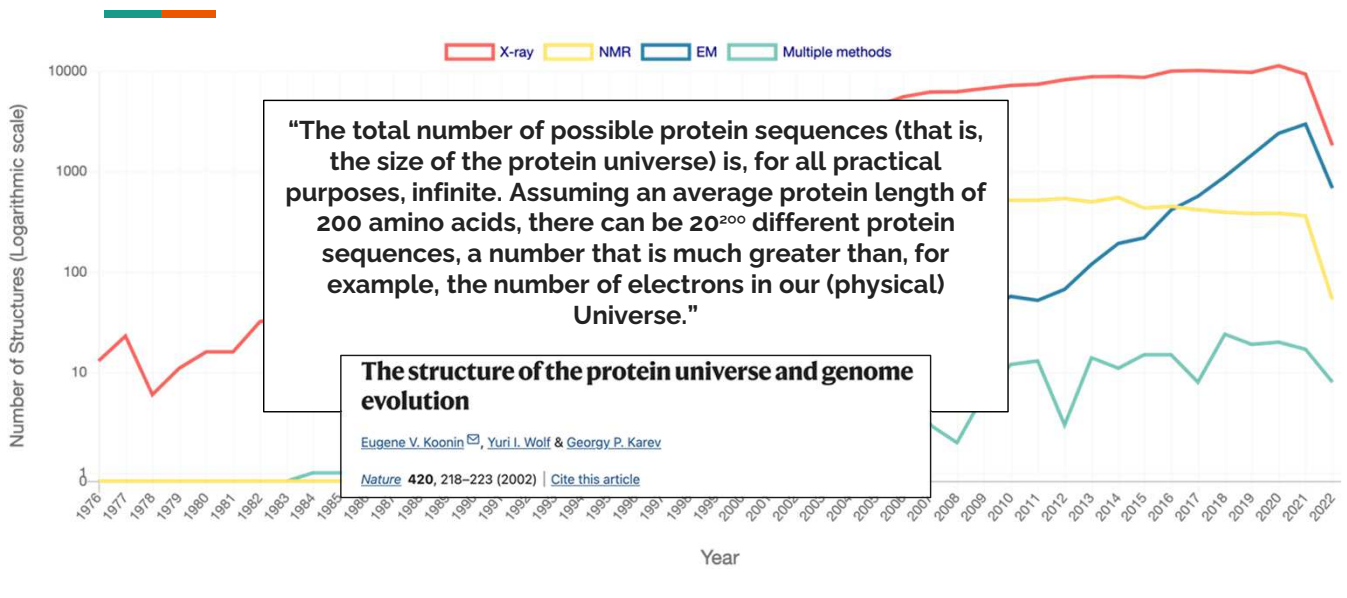


What is AlphaFold? New England Journal of Medicine:
<https://www.youtube.com/watch?v=7q8Uw3rmXyE>

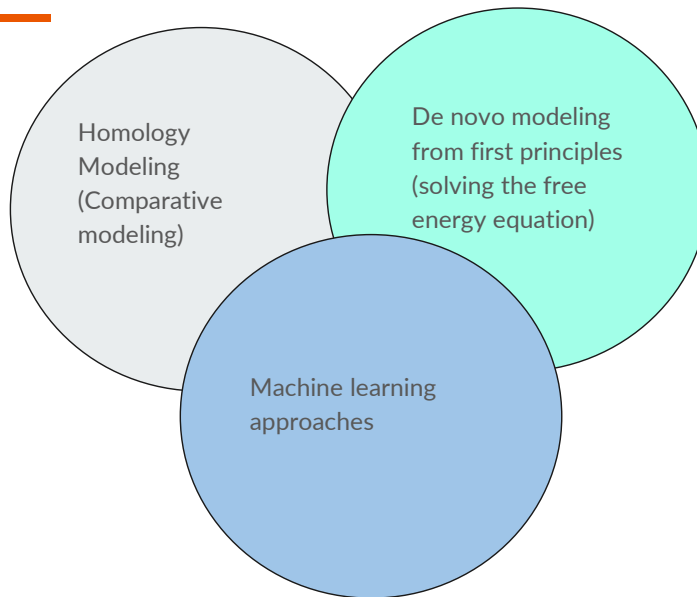
How many structures have been determined?



How many structures have been determined?

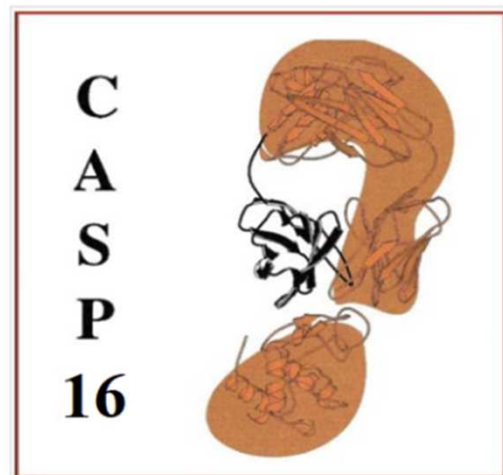


Computational Methods

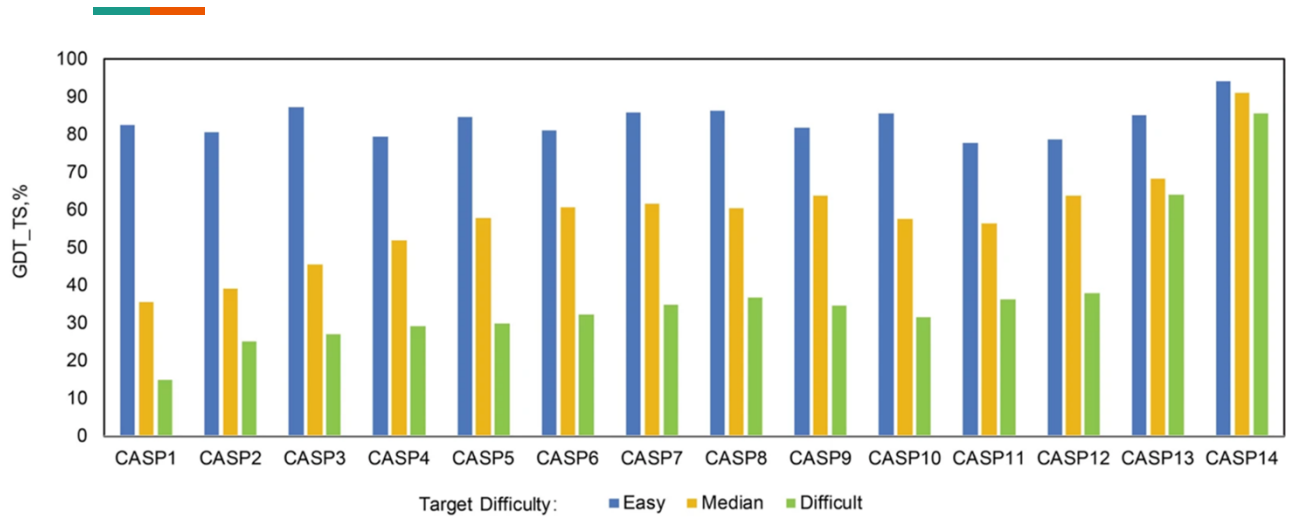


CASP Competition drives protein structure prediction

- Critical Assessment of Structure Prediction (CASP)
- Founded in 1994
- 'Olympics' of protein structure prediction
- Supported by Lawrence Livermore National Labs, DOE, etc.
- Check it out:
<https://www.predictioncenter.org>

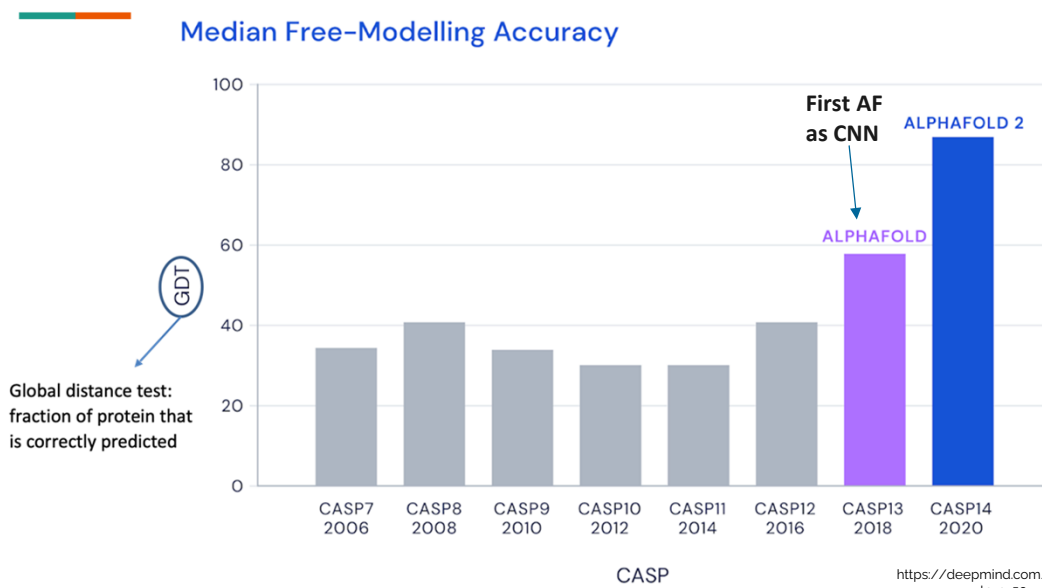


CASP Competition drives protein structure prediction



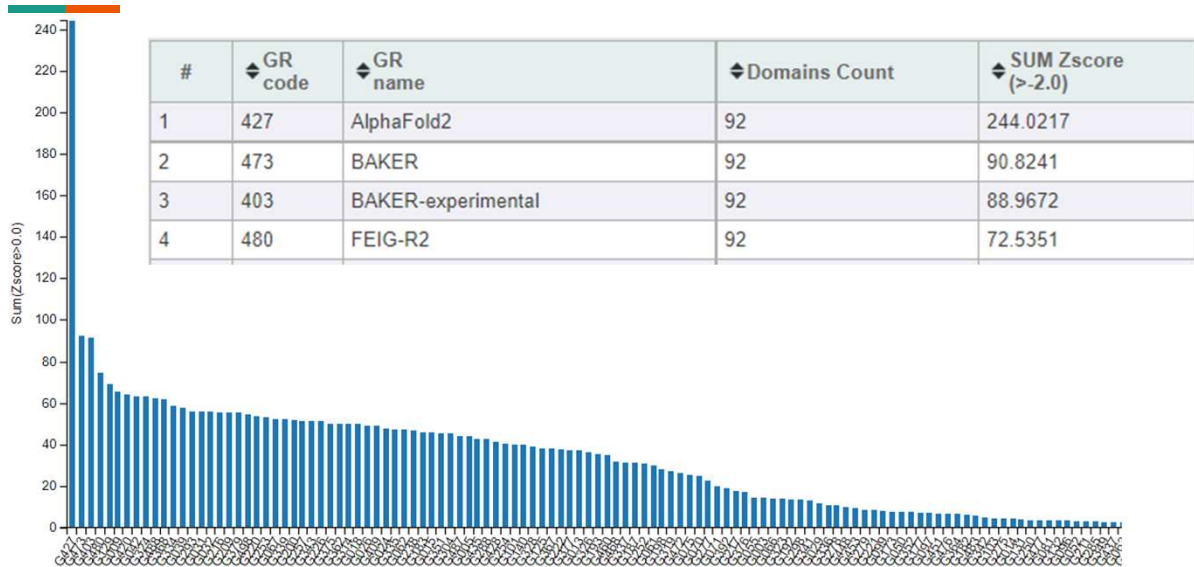
AlphaFold2 and its applications in the fields of biology and medicine. *Sig Transduct Target Ther* 8, 115 (2023). <https://doi.org/10.1038/s41392-023-01381-z>

AF2 compared to CASP winners over the years

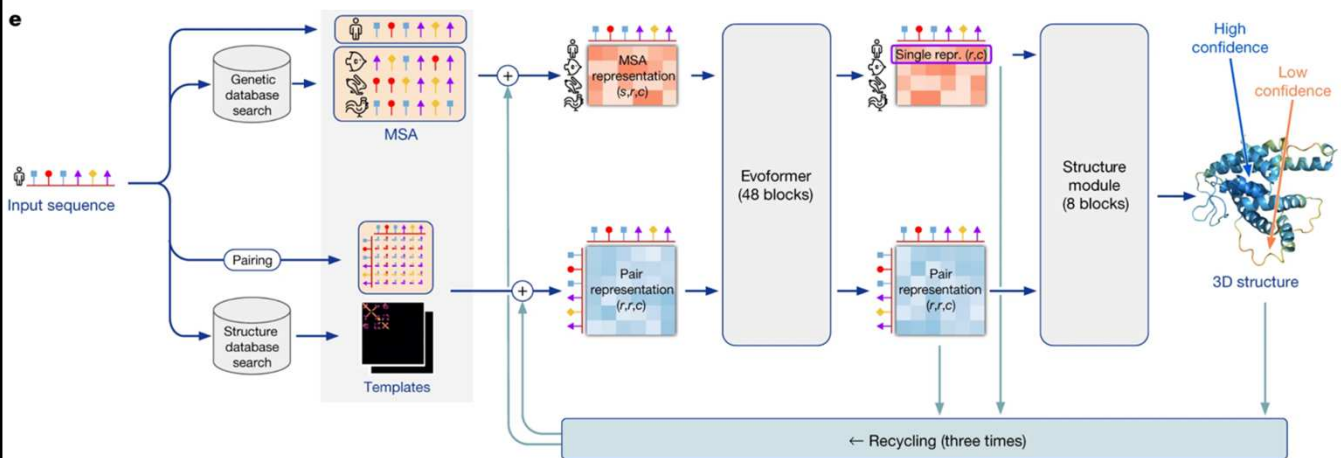


<https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology>

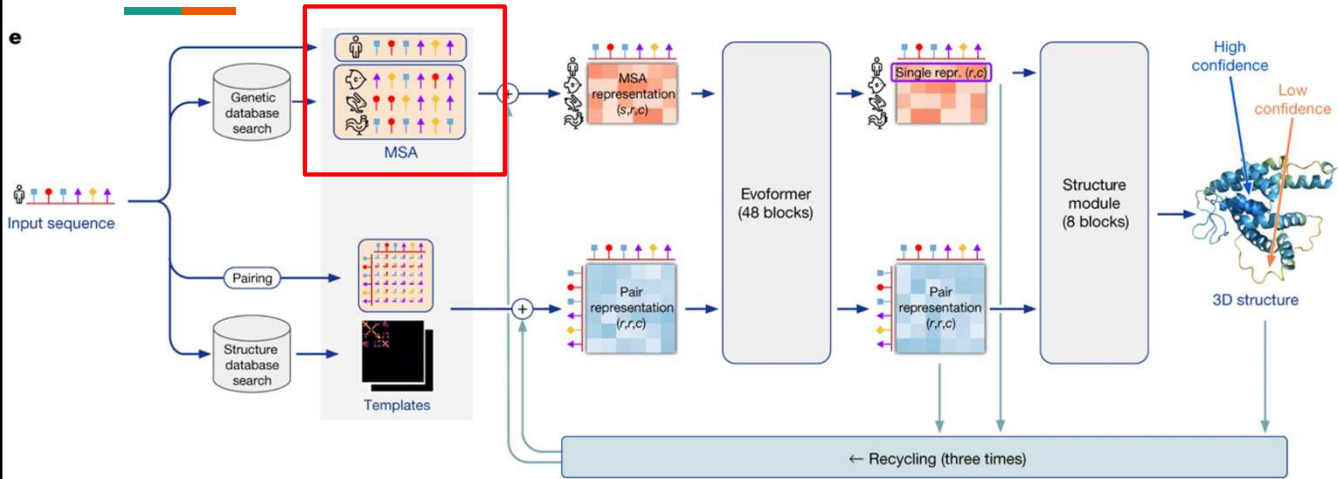
CASP14 (2020) Results: Entrance of AlphaFold2



AlphaFold2: the dawn of a new age

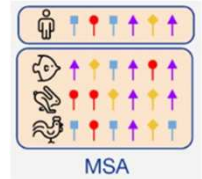


What is an MSA?



Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589 (2021). <https://doi.org/10.1038/s41586-021-03819-2>

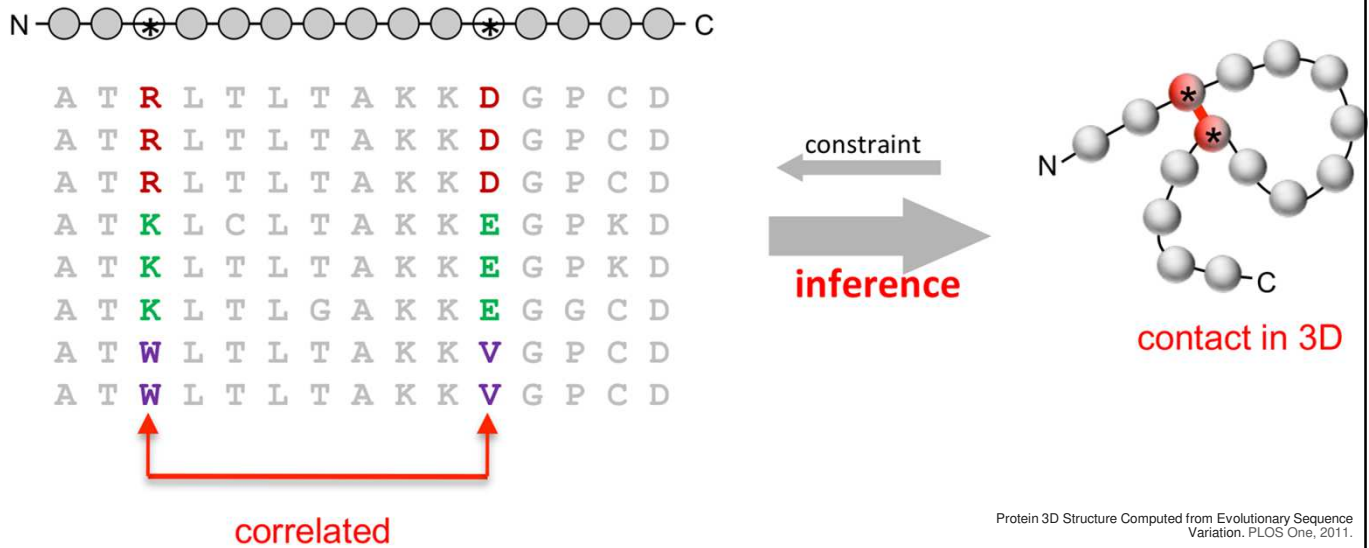
MSA: Multiple Sequence Alignment



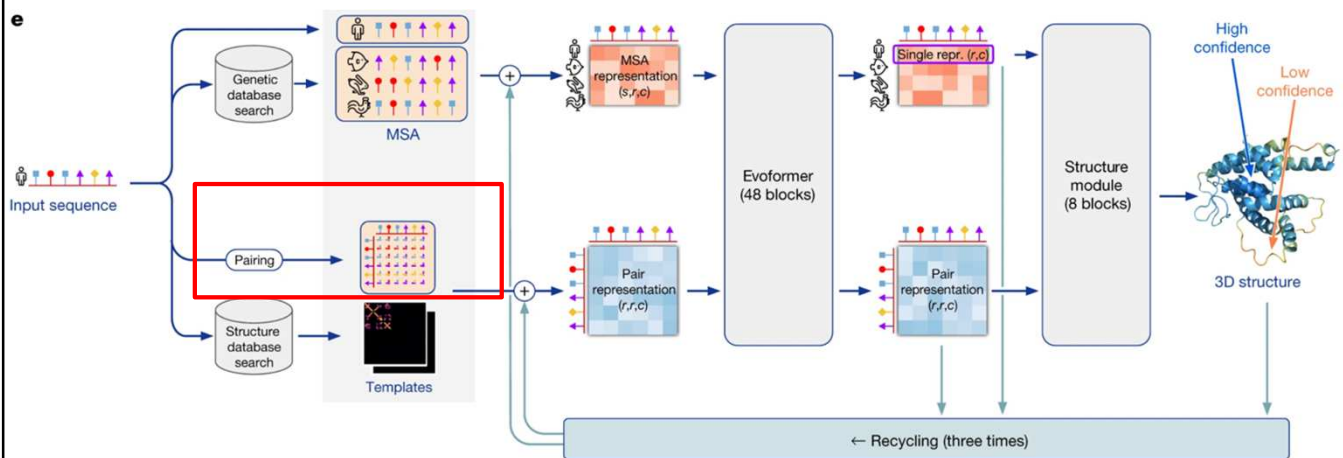
			cov	pid	1	80
1	UniProt/Swiss-Prot	P26898	IL2RA_SHEEP	100.0%	100.0%	E-SLLWRFFVEIVVPGC-TE-CHDDPFSRN-----FKVLRVE-VCTMINCDCKAGFRRS-AYVIR
2	UniProt/Swiss-Prot	P01590	IL2RA_MOUSE	94.4%	45.1%	E-RLNLIGLISLTIVSCRMELCLYDPREVFNA-----TFKALSKK-NCITLNCECKRGRFRRLKE-LVYIR
3	UniProt/Swiss-Prot	P41690	IL2RA_FELCA	98.9%	54.3%	E-SLLLIGLIRFVVVHGHVTELCDENPPDIQHA-----TFKALTYK-TCITLNCECKRGRFRRLSNGS-FML
4	UniProt/Swiss-Prot	P01589	IL2RA_HUMAN	98.9%	47.8%	NDSYLLNIGLLTFTIVPGCQBELCDDDPPEIPHA-----TFKAMAWK-ECTLNCECKRGRFRRIKSGSLYML
5	UniProt/Swiss-Prot	Q5MNY4	IL2RA_MACMU	98.9%	48.9%	NDEYLLNIGLLTFTIVPGCQBELCDDDPPEKITHA-----TFKAVAWK-ECTLNCECKRGRFRRIKSGSPYML
6	UniProt/Swiss-Prot	Q95118	IL2RG_BOVIN	96.6%	11.0%	MLKPPPLRSLLELQLSLLGVGLNPKFLTPSGNEDIGGKPGTGGD-FLTSTPAGTLDVSTLPLPKVQC--FVFNVEYMN
7	UniProt/Swiss-Prot	P40321	IL2RG_CANFA	95.5%	10.7%	MLKPPPLRSLLELQLSLLGVGLNSTVEMPNGNEDIT----PD-FLTATPSETLSVSSLPLPEVQC--FVFNVEYMN
8	UniProt/Swiss-Prot	P26896	IL2RB_RAT	73.0%	9.3%	NATVDSNRPLPLYILLLLLAIT-----PDS-FLTATPSETLSVSSLPLPEVQC--FVFNVEYMN
9	UniProt/Swiss-Prot	Q8BZM1	GLMN_MOUSE	27.0%	5.7%	-----
10	UniProt/Swiss-Prot	P36835	IL2_CAPHI	0.0%	0.0%	-----
11	UniProt/Swiss-Prot	Q7JFM4	IL2_AOTVO	0.0%	0.0%	-----
12	UniProt/Swiss-Prot	Q29416	IL2_CANFA	0.0%	0.0%	-----
	consensus/100%					
	consensus/90%					
	consensus/80%					
	consensus/70%					

Reference sequence (1): UniProt/Swiss-Prot|P26898|IL2RA_SHEEP
Identities normalised by aligned length.
Colored by: identity

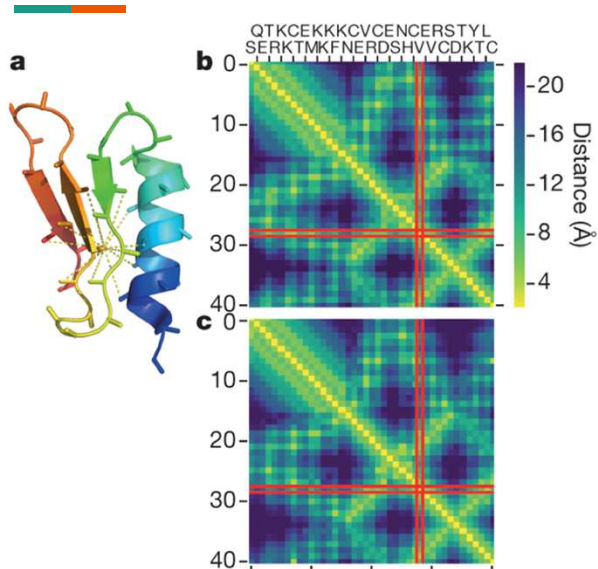
Direct-coupling analysis of residue coevolution captures native contacts across many protein families



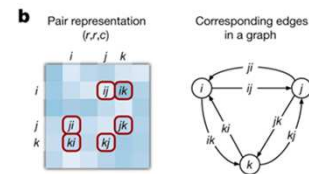
Initializing the Pair Representation



What is a pair representation?

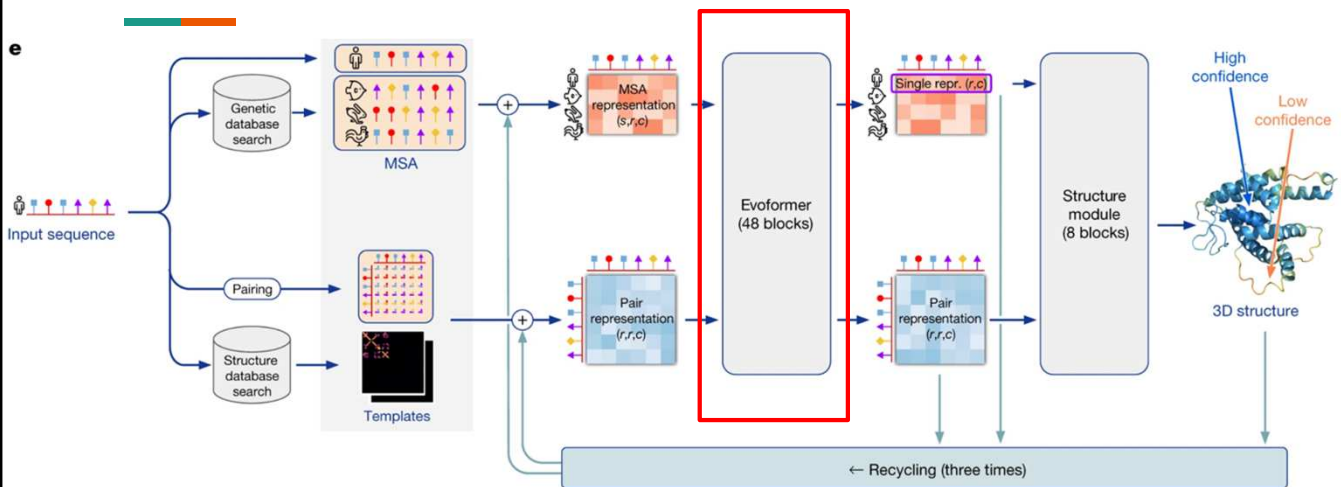


- Distogram is used to map 2D pairwise distances
- Distograms are independent of translations and rotations, so no need to align structures (much faster)



Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., ... & Hassabis, D. (2020). Nature.

Transformer-like module updates representations



Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589 (2021). <https://doi.org/10.1038/s41586-021-03819-2>

Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

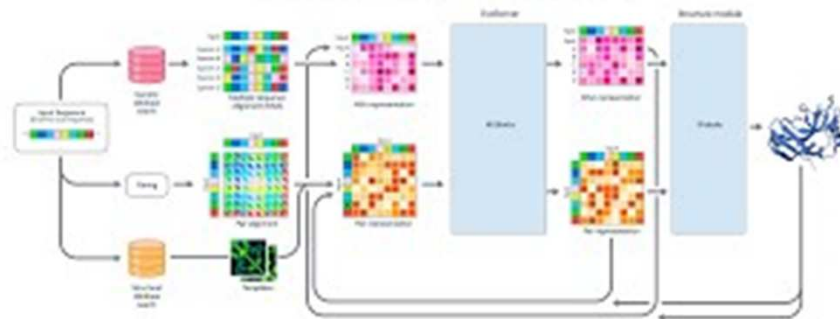
Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser*
Google Brain
lukaszkaizer@google.com

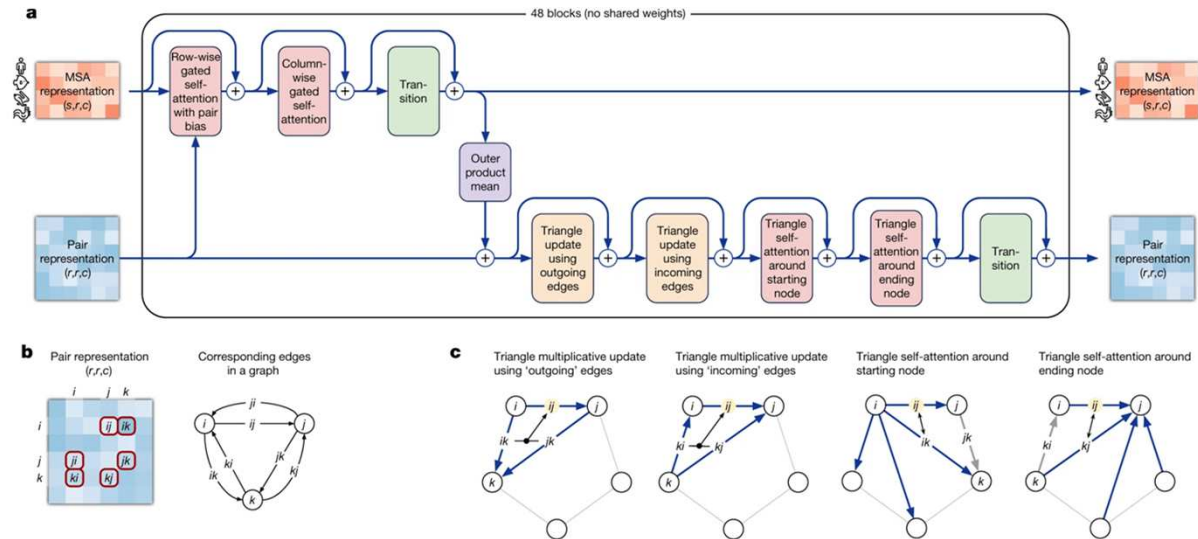
Illia Polosukhin* †
illia.polosukhin@gmail.com

<https://arxiv.org/abs/1706.03762>

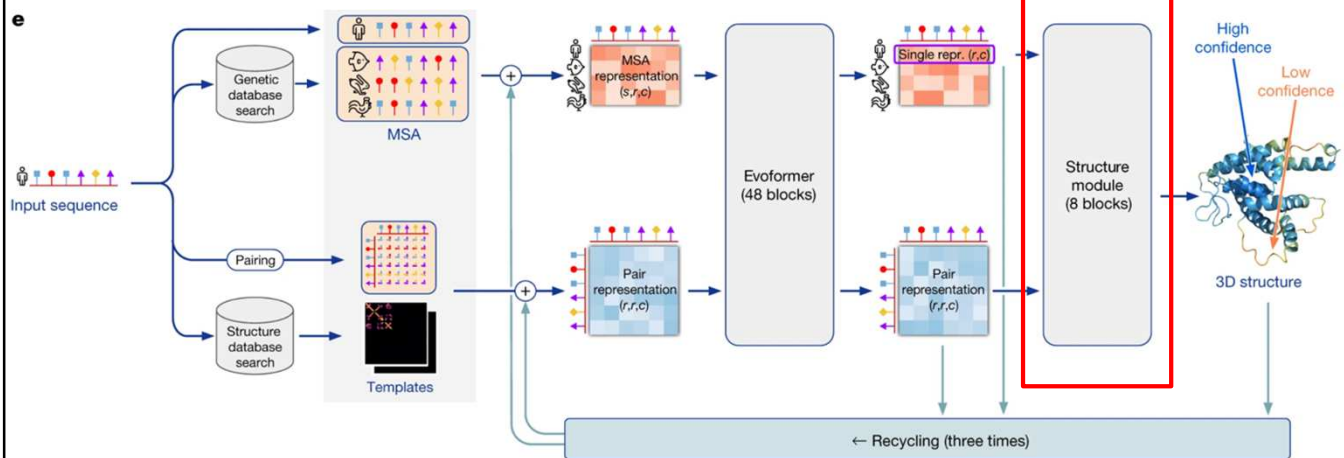
What Is AlphaFold?



The Evoformer: 48 blocks of back and forth between evolutionary and spatial reasoning.

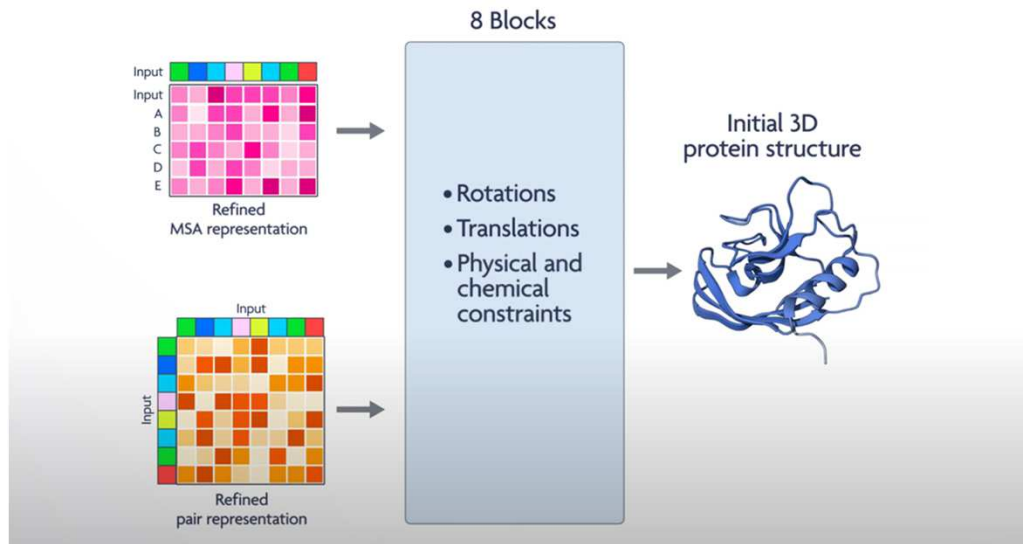


Structure module converts representation to structure

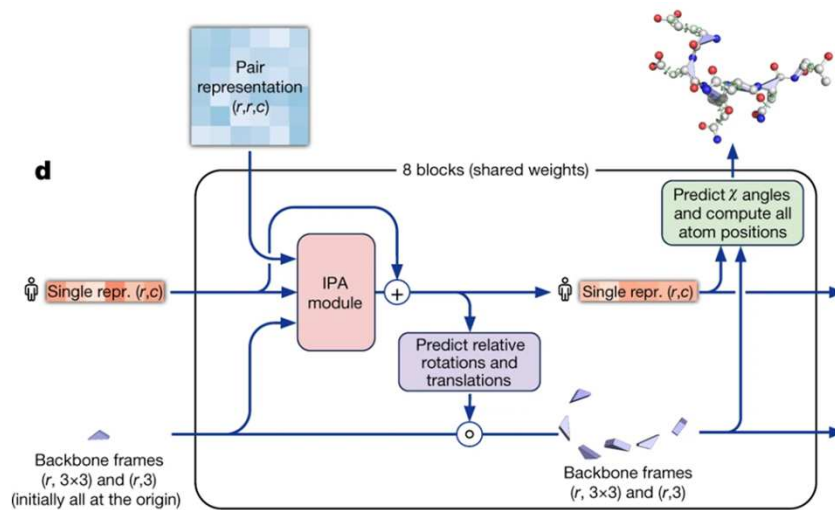


Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589 (2021). <https://doi.org/10.1038/s41586-021-03819-2>

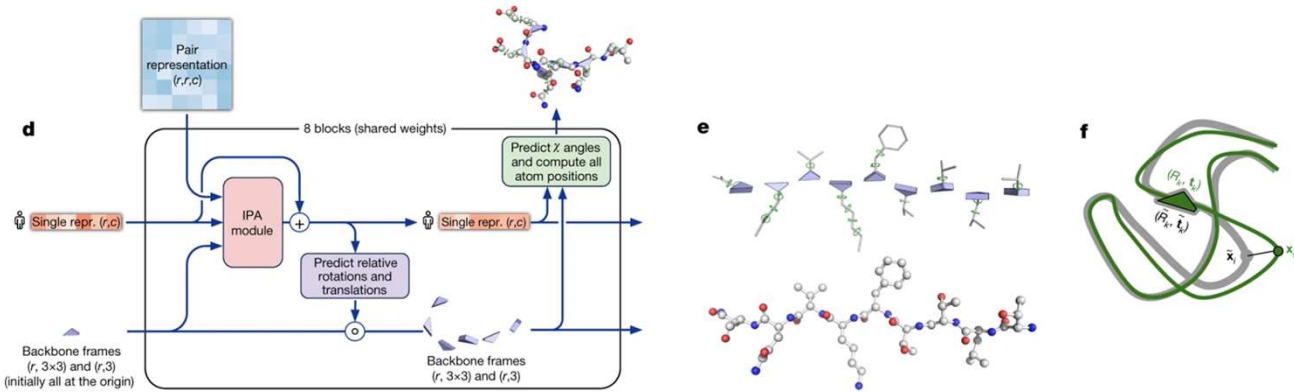
Structure module converts representation to structure



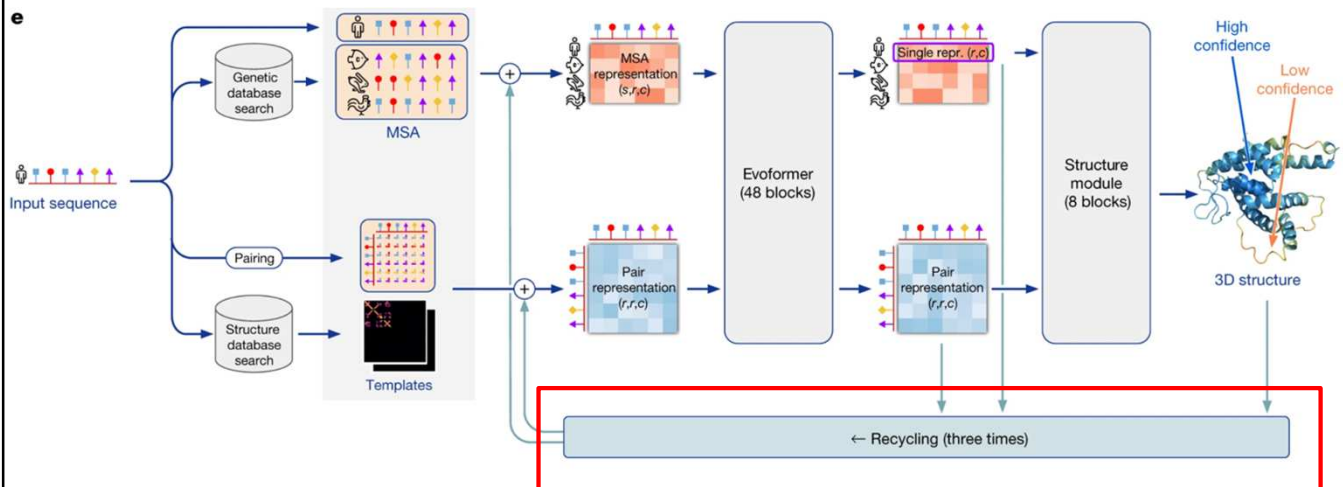
Structure Module



Structure Module

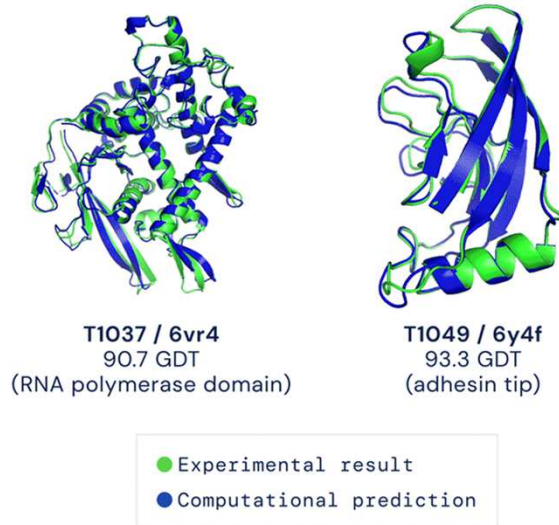


Recycling iteratively refines structure



Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589 (2021). <https://doi.org/10.1038/s41586-021-03819-2>

How do we know it works? – ask the model!

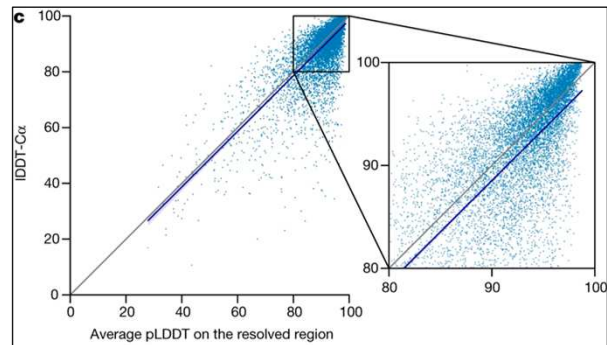


The 3Ps: pLDDT, PAE, and pTM

- LDDT, AE, TM require ground truth. What can we do?
- pLDDT: predicted **L**ocal **D**istance **D**ifference **T**est
- PAE: **P**redicted **A**ligned **E**rror
- pTM: predicted **T**emplate **M**odeling score
 - Global comparison of similarity between two structures
 - Measure of 0 to 1

The 3Ps: pLDDT, PAE, and pTM

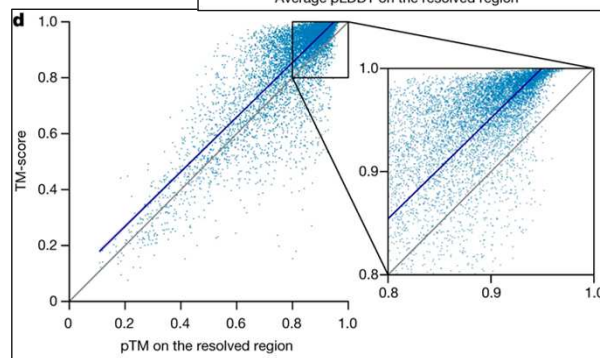
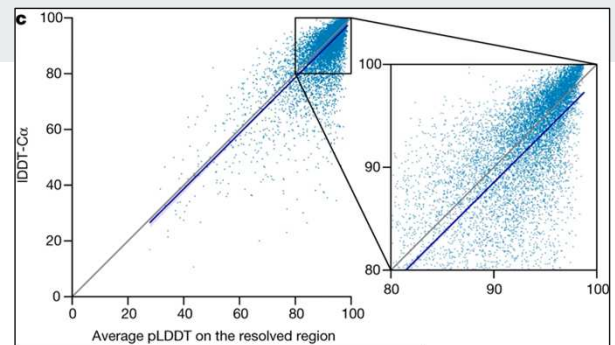
- LDDT, AE, TM require ground truth. What can we do?
- pLDDT: predicted **L**ocal **D**istance **D**ifference **T**est
- PAE: **P**redicted **A**ligned **E**rror
- pTM: predicted **T**emplate **M**odeling score
 - Global comparison of similarity between two structures
 - Measure of 0 to 1



Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589 (2021). <https://doi.org/10.1038/s41586-021-03819-2>

The 3Ps: pLDDT, PAE, and pTM

- LDDT, AE, TM require ground truth. What can we do?
- pLDDT: predicted **L**ocal **D**istance **D**ifference **T**est
- PAE: **P**redicted **A**ligned **E**rror
- pTM: predicted **T**emplate **M**odeling score
 - Global comparison of similarity between two structures
 - Measure of 0 to 1

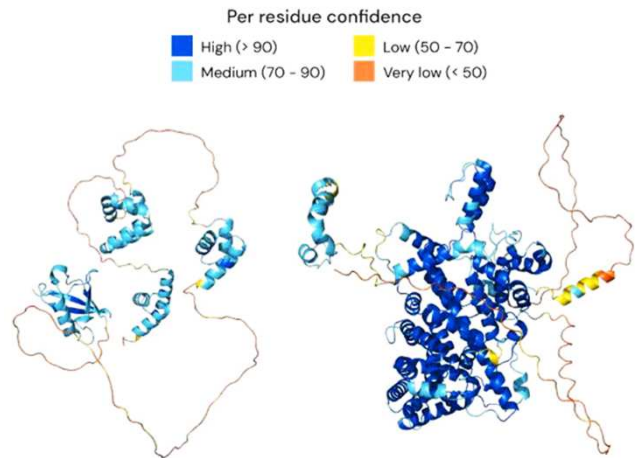


Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589 (2021). <https://doi.org/10.1038/s41586-021-03819-2>

Predicted Local Distance Difference Test



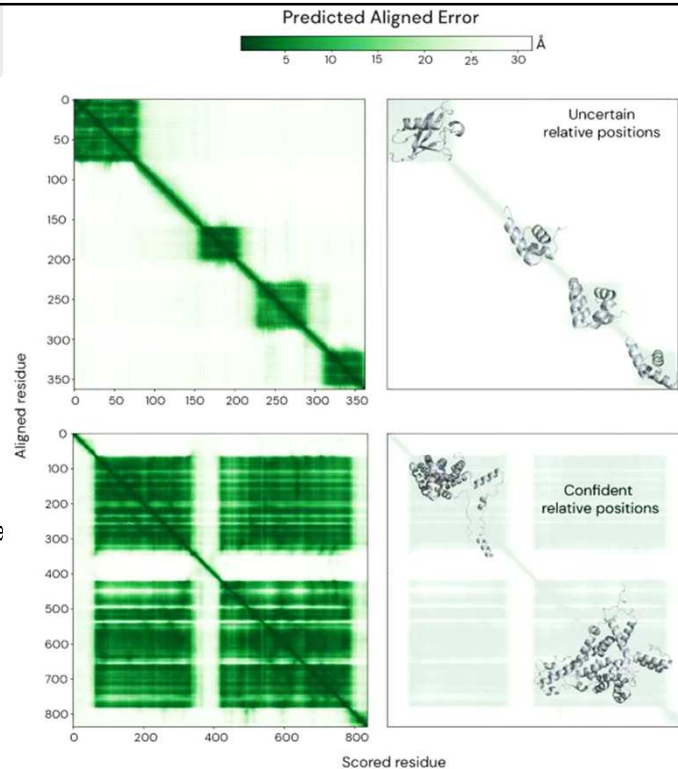
- AlphaFold's per-residue prediction of its IDDT-Ca score
- Low IDDT commonly associated with disorder
- High pLDDT on each domain doesn't imply confidence in relative positions!



Predicted Aligned Error (PAE)

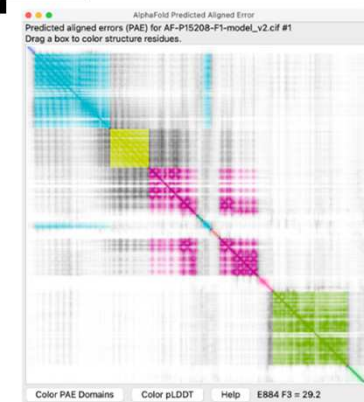
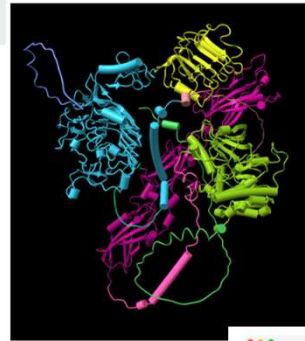


- Prediction of position error at residue x if the predicted and the true structures were aligned on y
- PAE aims to measure confidence in the relative positions of pairs of residues
- Use where pairwise confidence is relevant – interpreting domain distances in a multi domain protein
- Suppose residue y were aligned to the true structure and we measured the position error at residue x . The color at (x,y) is AF's prediction of that error



Predicted Aligned Error (PAE)

- Prediction of position error at residue x if the predicted and the true structures were aligned on y
- PAE aims to measure confidence in the relative positions of pairs of residues
- Use where pairwise confidence is relevant – interpreting domain distances in a multi domain protein
- Suppose residue y were aligned to the true structure and we measured the position error at residue x. The color at (x,y) is AF's prediction of that error





214,683,839 Predicted Protein Structures on AFDB!



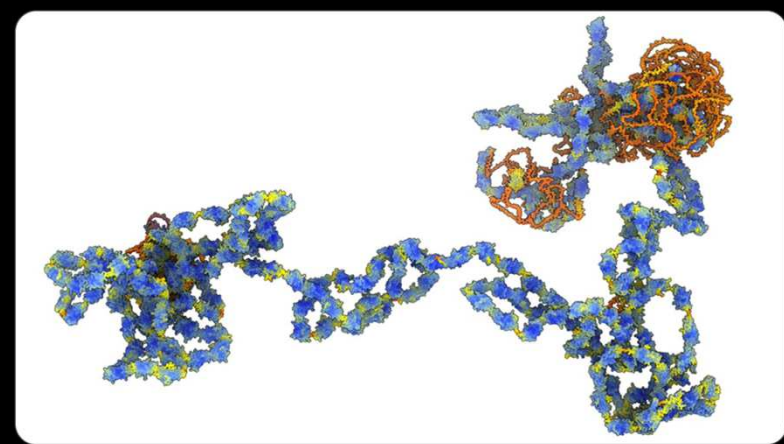
 **217,705** Structures from the PDB
 **1,068,577** Computed Structure Models (CSM)

Proteome



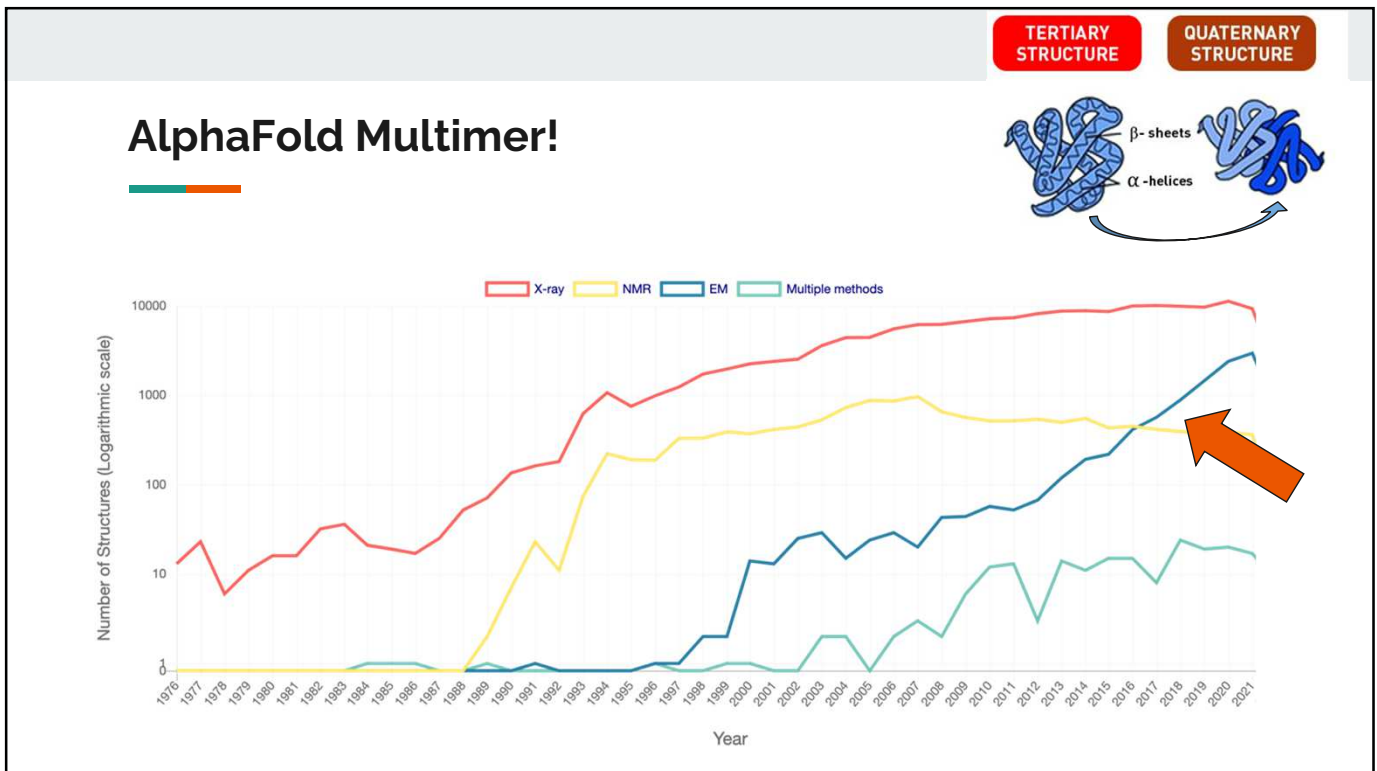


The human muscle protein titin predicted by AlphaFold, 34350 residues.

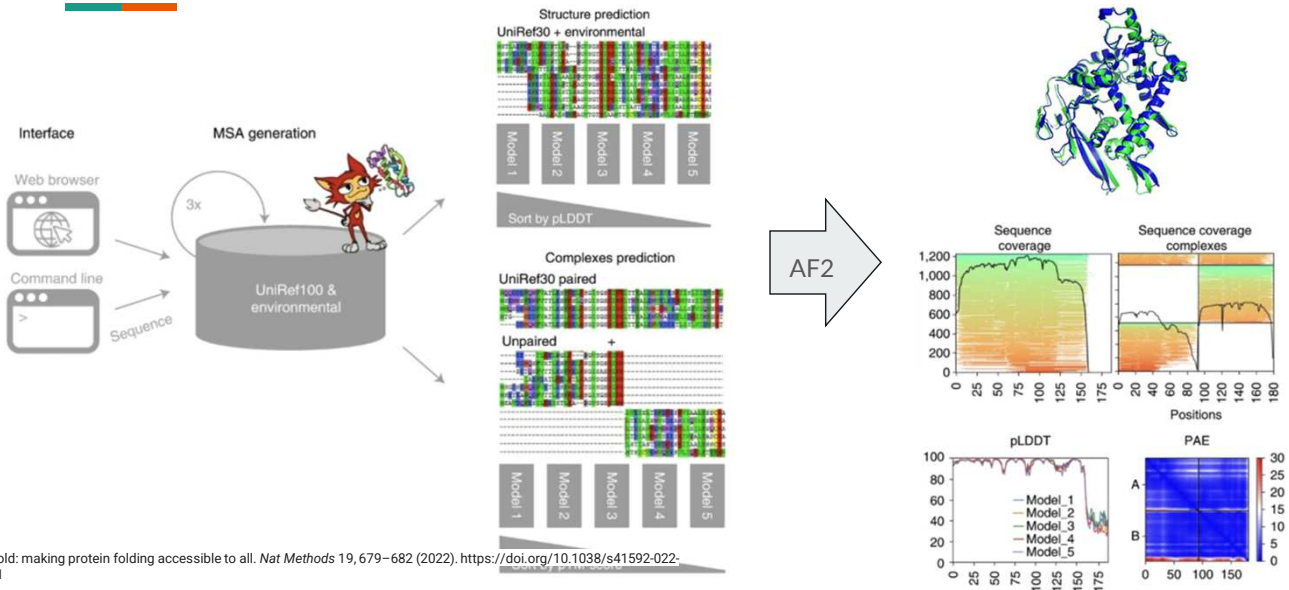


Dark proteome
Intrinsically disordered
PFAM – No PDB / AF
Gained AF – 70 > pLDDT ≥ 50
Gained AF – 90 > pLDDT ≥ 70
Gained AF – pLDDT ≥ 90
PDB 20% to 50% - pLDDT < 90
PDB 50% to 95%
PDB ≥ 95%

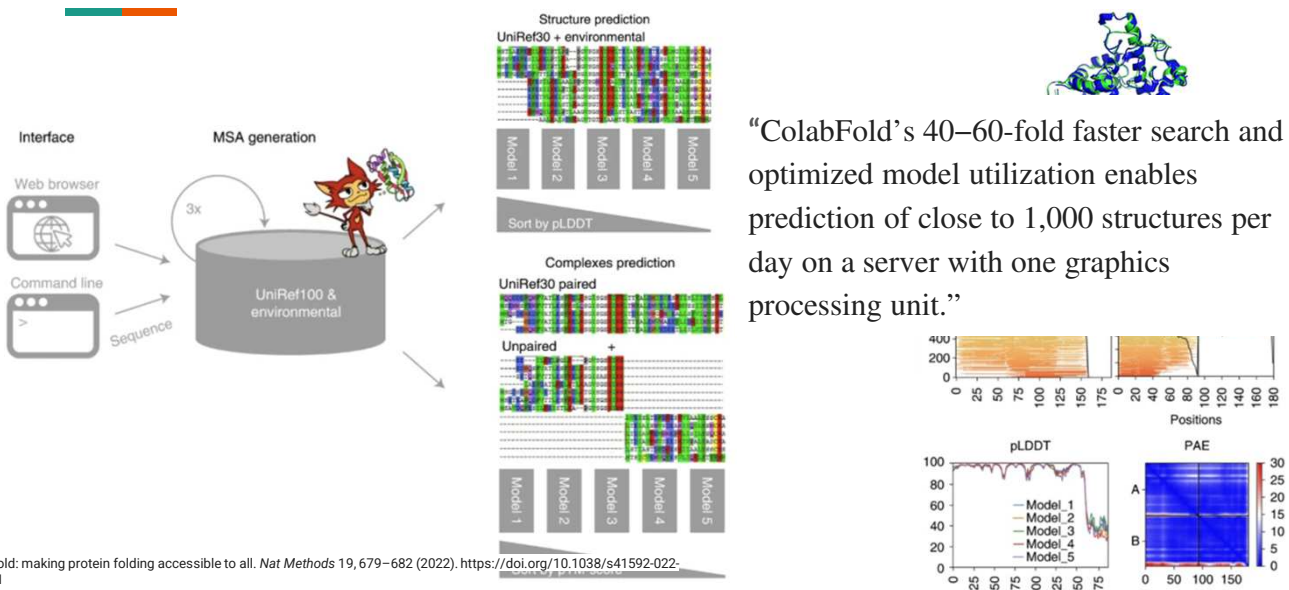
Tunyasuvunakool, K., Adler, J., Wu, ...
7:23 PM · Sep 8, 2021 · Twitter Web App
Valentini, S., & Valencia, A. ...
Biology.



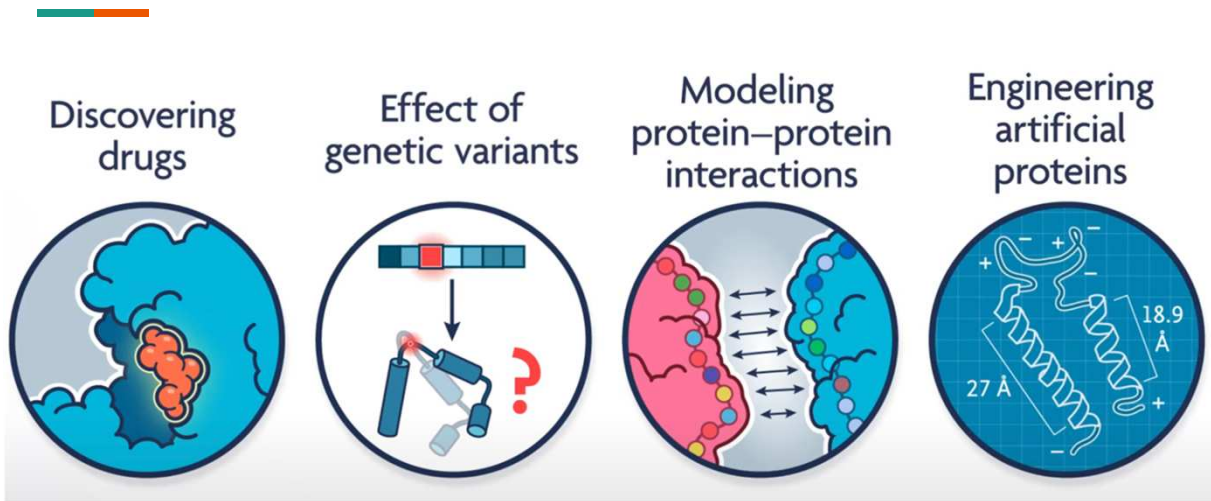
ColabFold speeds up AlphaFold by 40-60 fold!



ColabFold speeds up AlphaFold by 40-60 fold!



Applications and Frontiers!



Has protein structure prediction been solved?

- Sort of.
- The Protein Folding Problem (de novo)
 - What is the folding code?
 - **What is the folding mechanism?**
 - Can we predict a native protein structure from its primary, amino acid sequence?
 - **No for a sequence in isolation...**
 - **Yes when informed by like sequences and their structures**

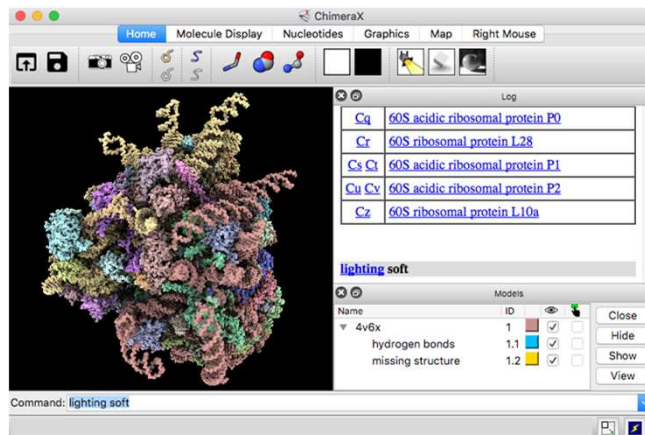
Let's give it a try!

- Google Colab Notebooks:
 - [AlphaFold](#)
 - [ColabFold](#) (AF2 w/MMSeqs2)
 - Maintained list of Google Colabs: <https://github.com/sokrypton/ColabFold>

- AlphaFold Resources:
 - Github page: <https://github.com/google-deepmind/alphafold>
 - AFDB Protein Structure Database and links: <https://alphafold.com>
 - [Lovely & more in depth lecture from John Jumper at Vanderbilt University](#)
 - Running AlphaFold on the BRCF Servers:
 - [AMD GPUs](#)
 - [NVIDIA GPUs](#)
 - [Running AlphaFold on TACC](#)

ChimeraX

- Download ChimeraX: <https://www.cgl.ucsf.edu/chimerax/download.html>
- Quick Start: <https://www.cgl.ucsf.edu/chimerax/docs/quickstart/index.html>
- Very comprehensive user guide: <https://www.cgl.ucsf.edu/chimerax/docs/user/index.html>



Thank you! Questions?



Looking for people to work with/learn more about Machine Learning applied to Biology?

<https://www.biomsociety.org>

We meet every other Thursday from 9:30am to 10:30am, in MBB 3.204

TACOS and **COFFEE** provided :)

Watch the Commander Complex assemble!

