Protein Structure Prediction

BCH394P/364C Systems Biology/Bioinformatics
April 1, 2025
Daryl Barth

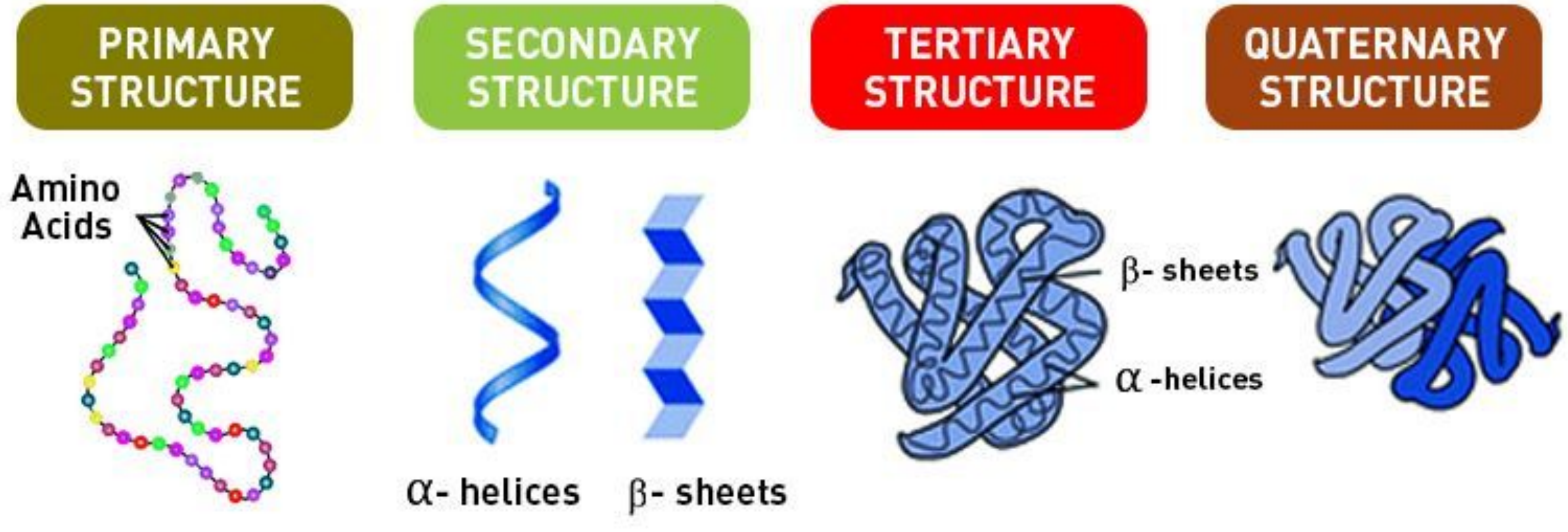
Today’s Goals

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

Why?

https://www.ted.com/talks/david_baker_5_challenges_we_could_solve_by_designing_new_proteins

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?

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

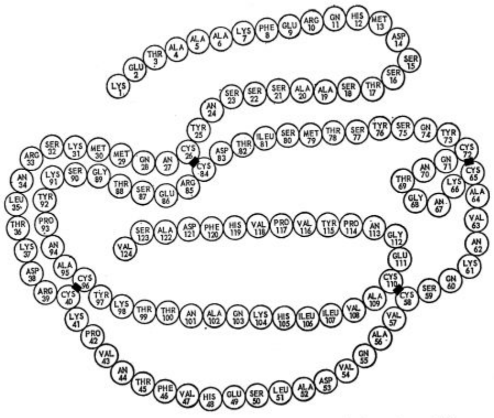
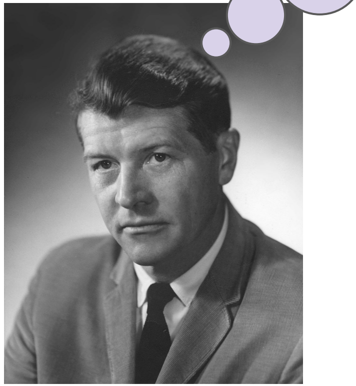
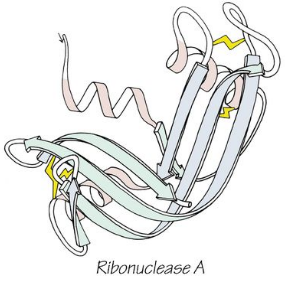
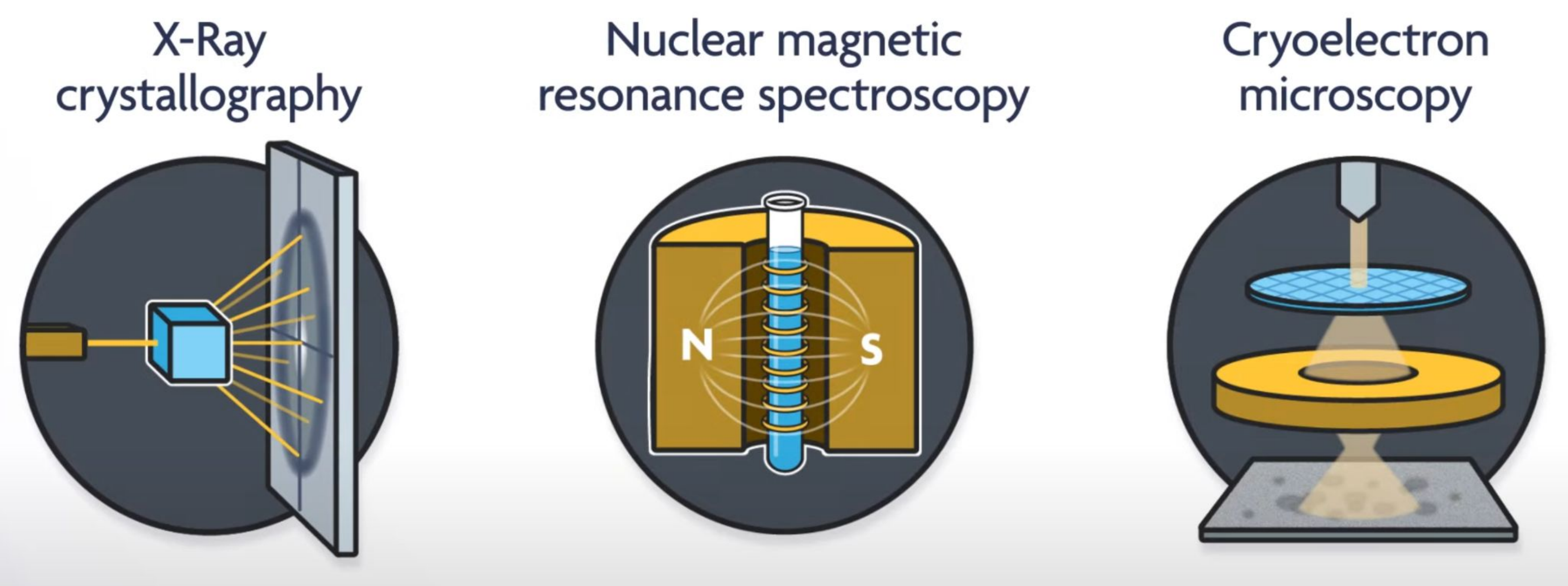


Fig. 1. The amino acid sequence of bovine pancreatic ribonuclease (*50*).



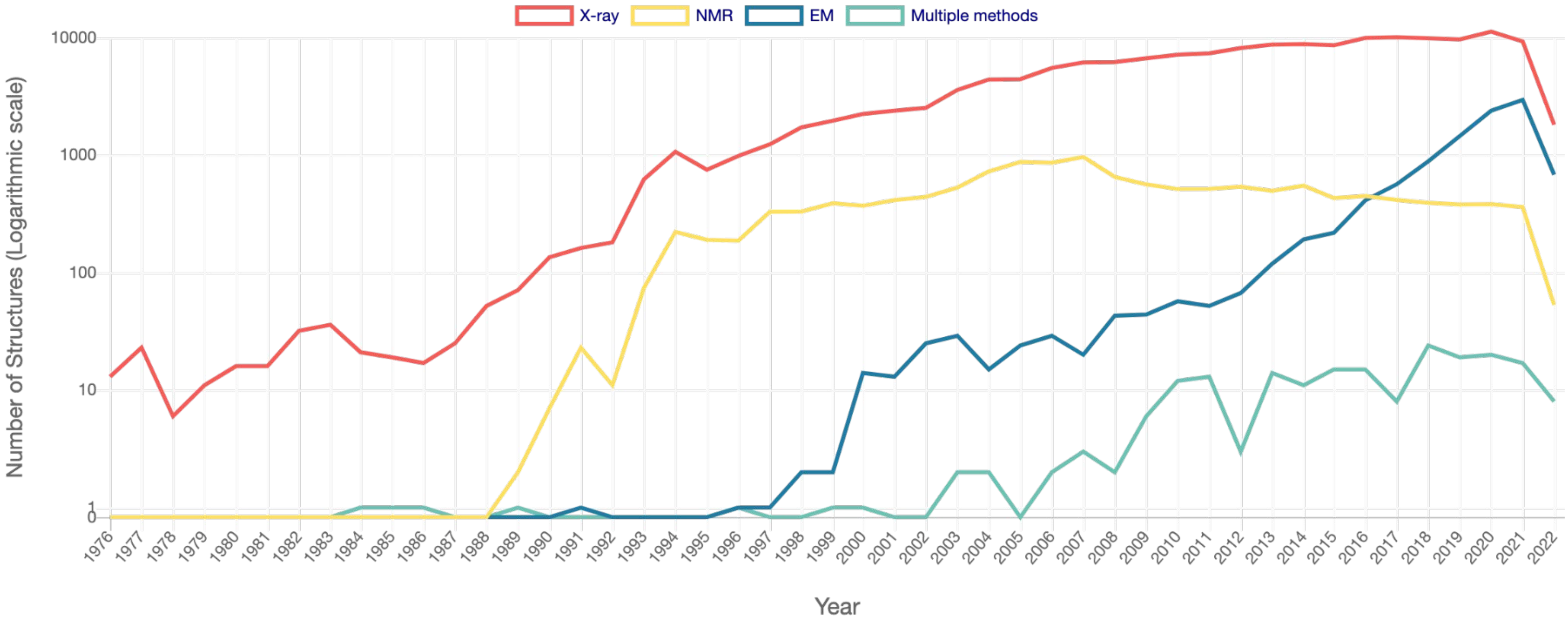


Experimental Methods for Determining Protein Structure

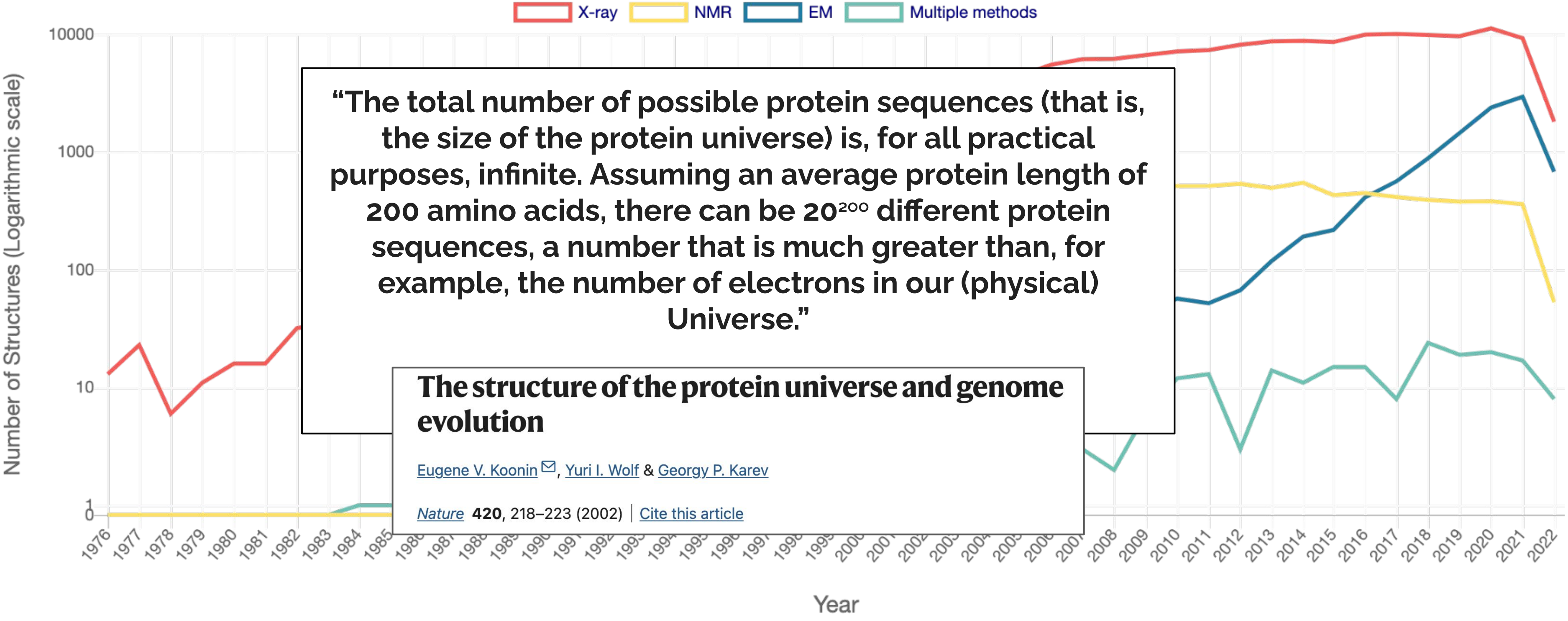


What is AlphaFold? New England Journal of Medicine:
https://www.youtube.com/watch?v=7q8Uw3rmXyE

How many structures have been determined?



How many structures have been determined?



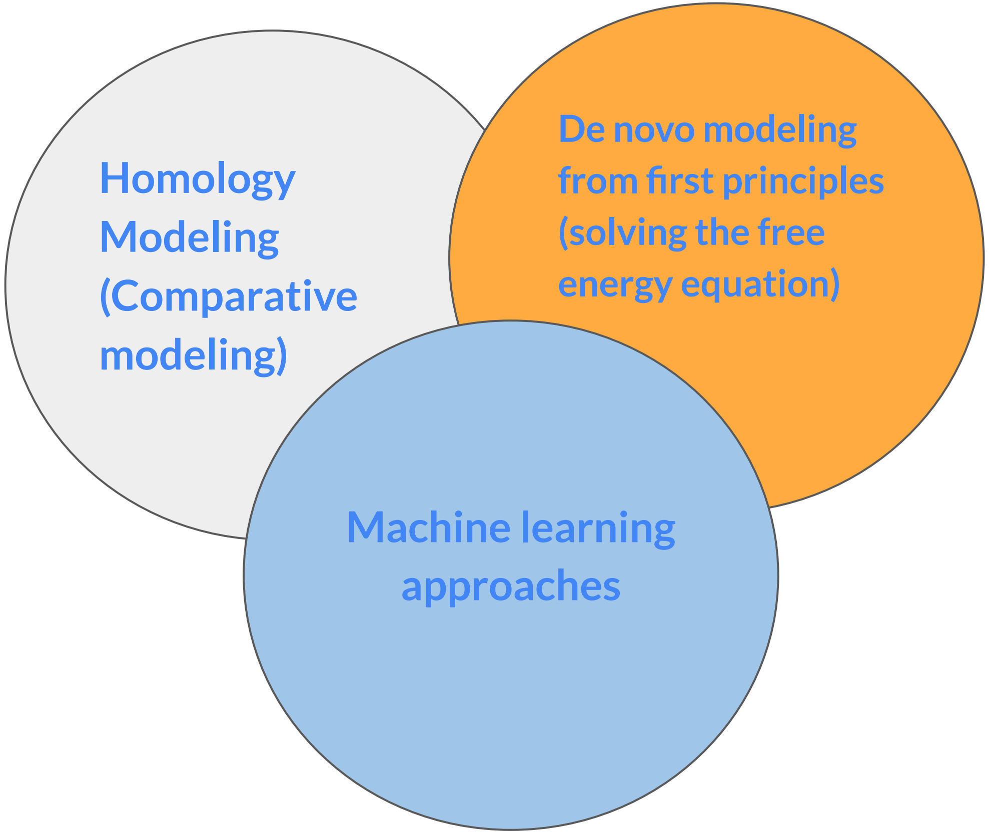
“The total number of possible protein sequences (that is, the size of the protein universe) is, for all practical purposes, infinite. Assuming an average protein length of 200 amino acids, there can be 20^{200} different protein sequences, a number that is much greater than, for example, the number of electrons in our (physical) Universe.”

The structure of the protein universe and genome evolution

Eugene V. Koonin, Yuri I. Wolf & Georgy P. Karev

Nature **420**, 218–223 (2002) | Cite this article

Computational Methods



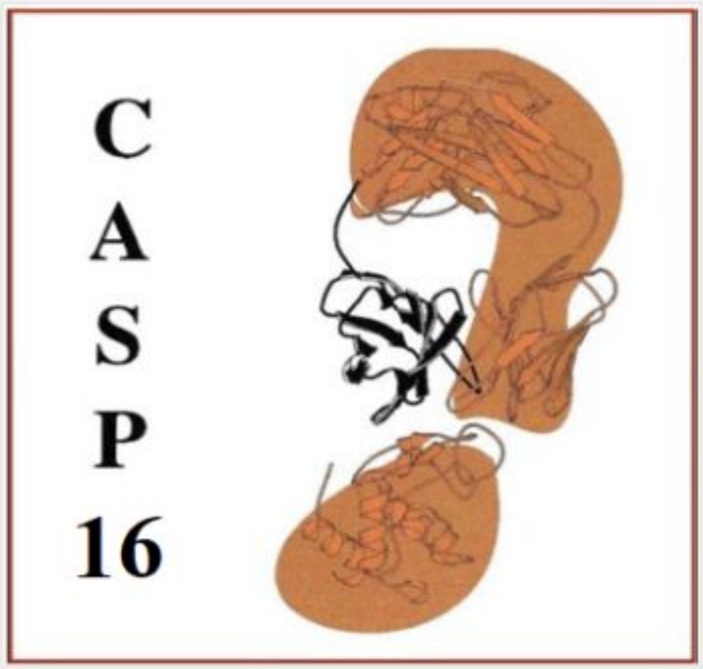
Homology Modeling (Comparative modeling)

De novo modeling from first principles (solving the free energy equation)

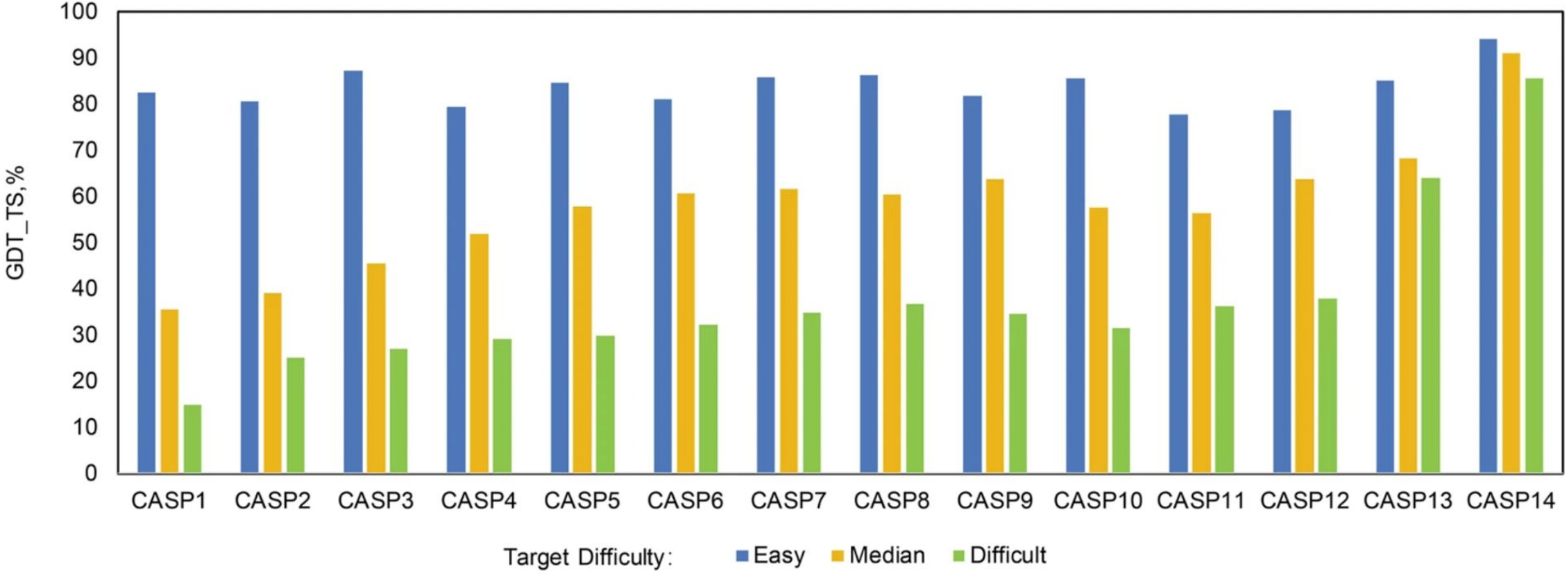
Machine learning approaches

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

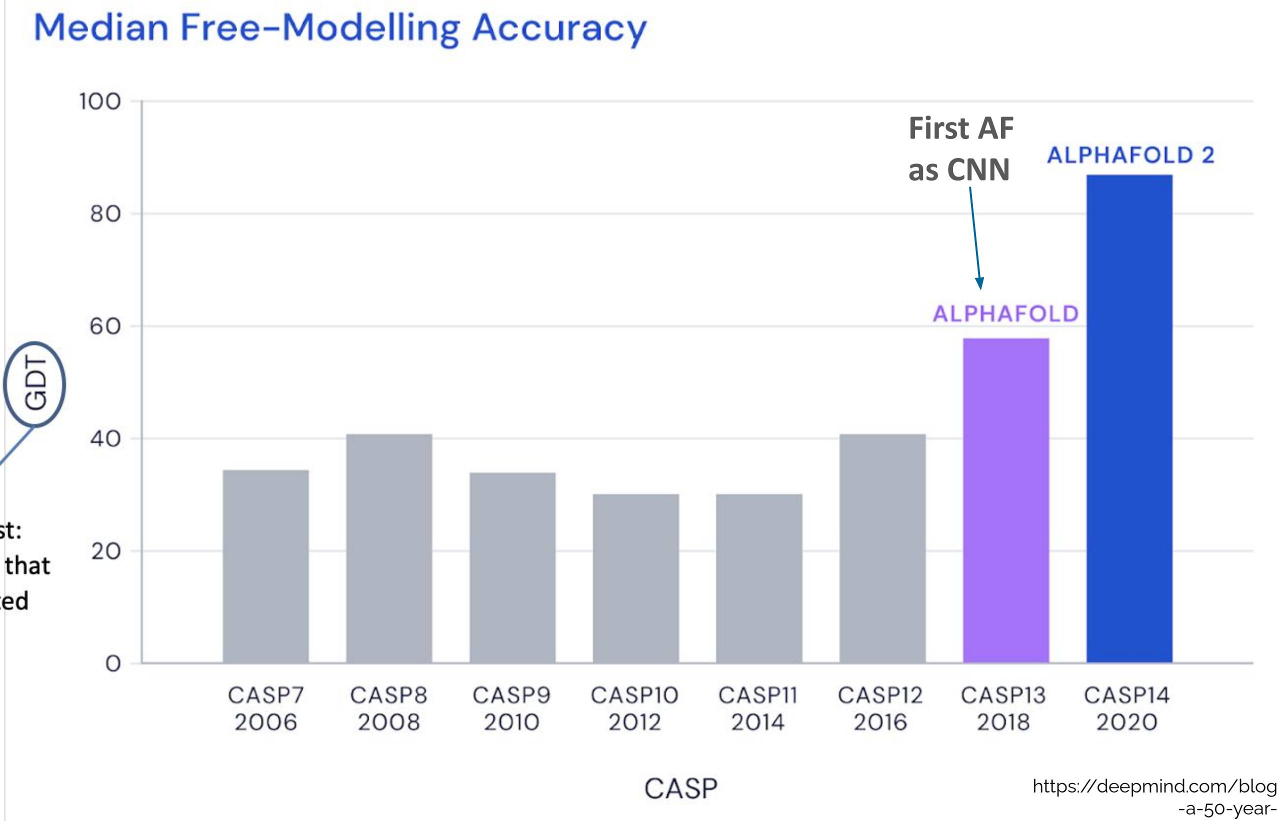


CASP Competition drives protein structure prediction



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

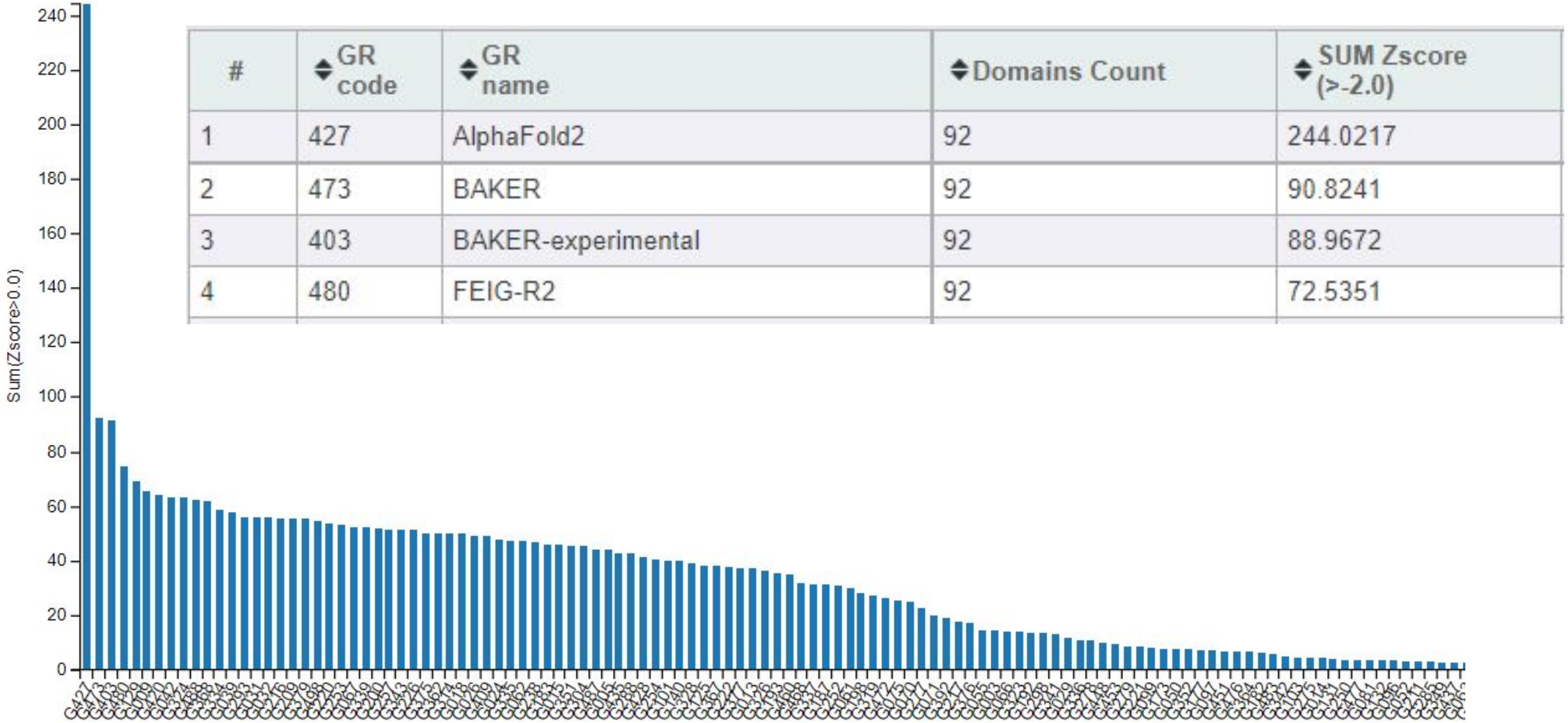
AF2 compared to CASP winners over the years



Global distance test: fraction of protein that is correctly predicted

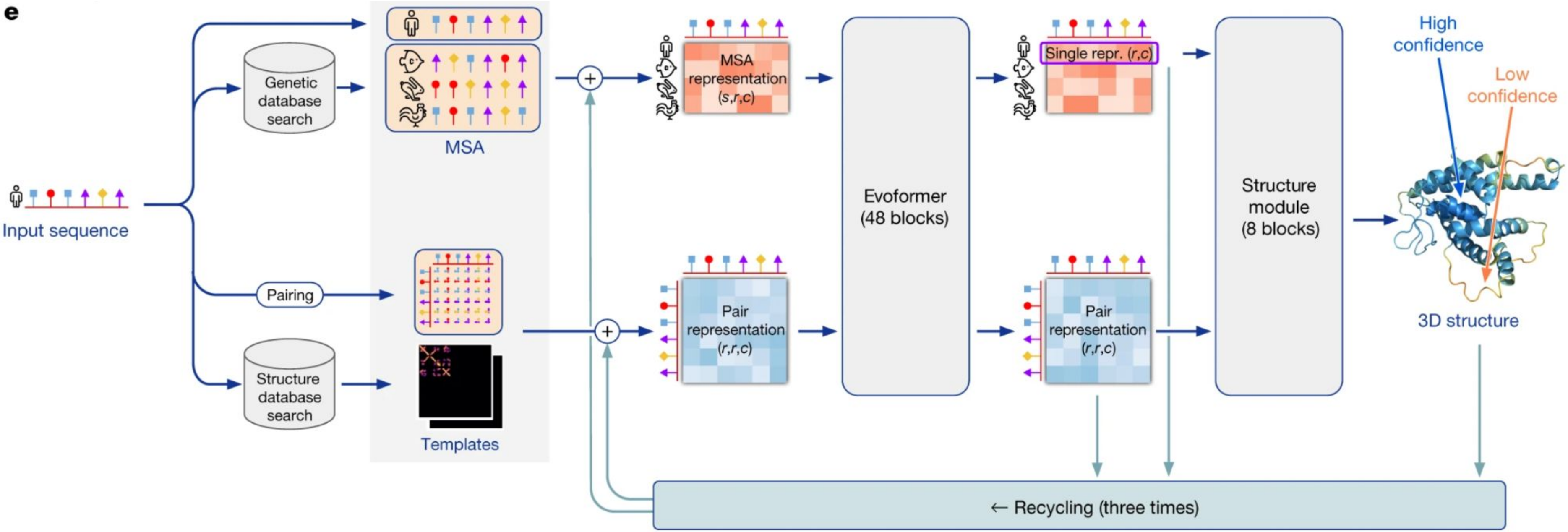
https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology

CASP14 (2020) Results: Entrance of AlphaFold2



#	⬍GR code	⬍GR name	⬍Domains Count	⬍SUM Zscore (>-2.0)
1	427	AlphaFold2	92	244.0217
2	473	BAKER	92	90.8241
3	403	BAKER-experimental	92	88.9672
4	480	FEIG-R2	92	72.5351

AlphaFold2: the dawn of a new age

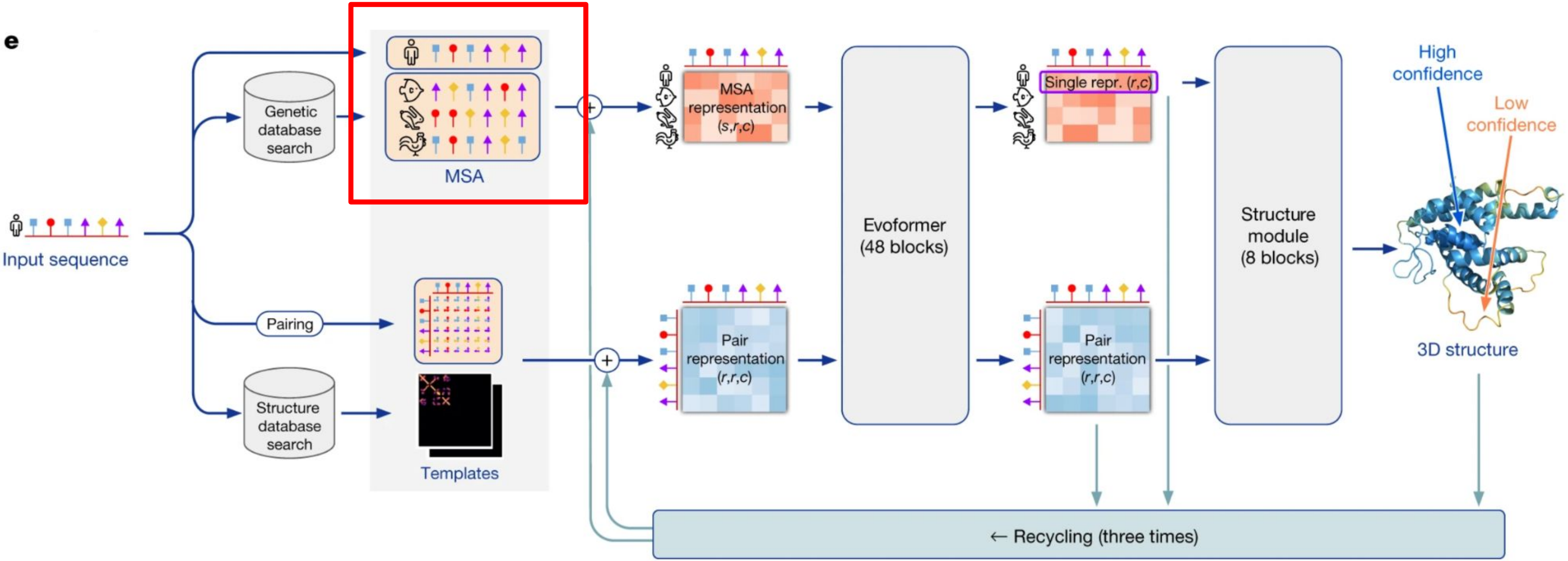

e
Input sequence
Genetic database search
MSA
Pairing
Structure database search
Templates
MSA representation (s,r,c)
Pair representation (r,r,c)
Evoformer (48 blocks)
Single repr. (r,c)
Pair representation (r,r,c)
Structure module (8 blocks)
High confidence
Low confidence
3D structure
← Recycling (three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

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

Start Demo!

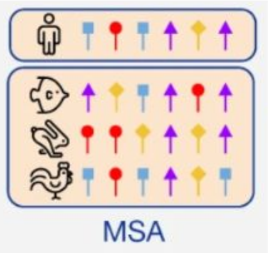
- ColabFold Server
- >2E8I_1|Chain A|6-aminohexanoate-dimer hydrolase|Flavobacterium sp. (37931)
- MNARSTGQHPARYPGAAAGEPTLDSWQEPPHNRWAFAHLGEMVPSAAVSRRPVNAPGHALARLGAIAAQLPDLEQRLEQTYTDAFLVLRGTEVVAEYYRAGFAPDDRHLLMSVSKSLCGTVVGALVDEGRIDPAQPVTEYVPELAGSVYDGPSVLQVLDMQISIDYNEDYVDPASEVQTHDRSAGWRTRRHGDPADTYEFLTTLRGDGSTGEFQYCSANTDVLAWIVERVTGLRYVEALSTYLWAKLDADRDATITVDTTGFGFAHGGVSCTARDLARVGRMMLDGGVAPGGRVVSEDWVRRVLAGGSHEAMTDKGFTNTFPDGSYTRQWWCTGNERGNVSGIGIHGQNLWLDPLTDSVIVKLSSWPDPDTEHWHRLQNGILLDVSRALDAV
- 392 Amino Acids, Nylon Hydrolase

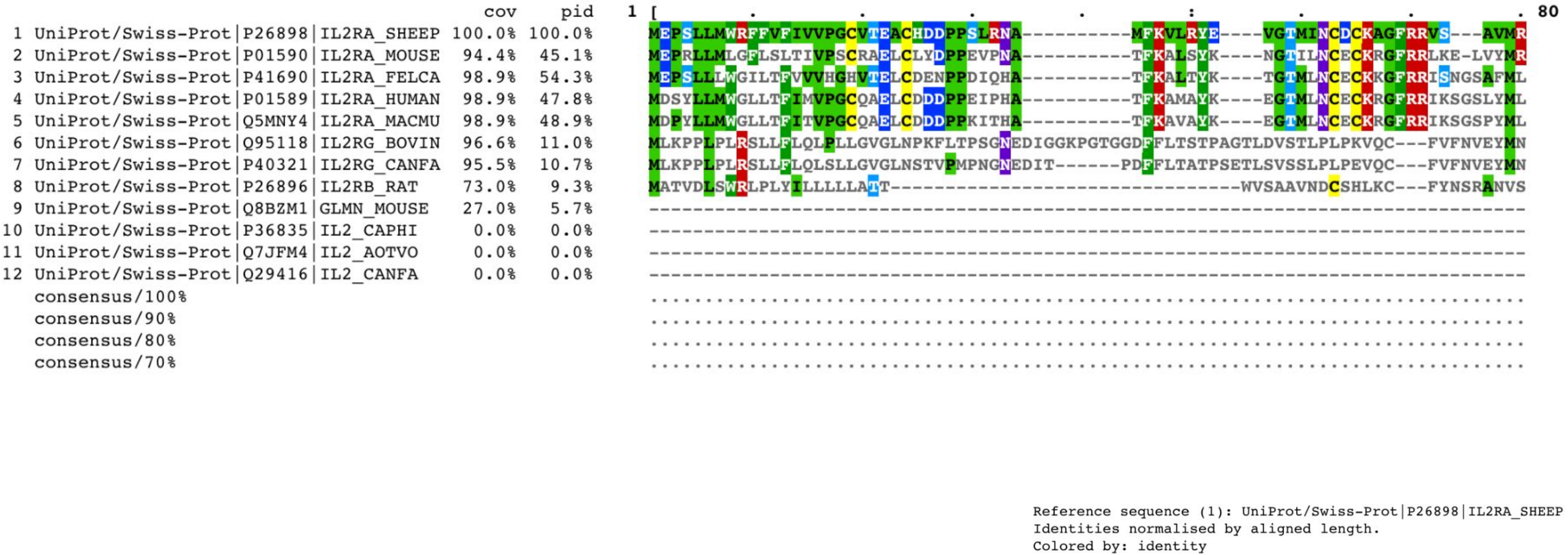
What is an MSA?


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)

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

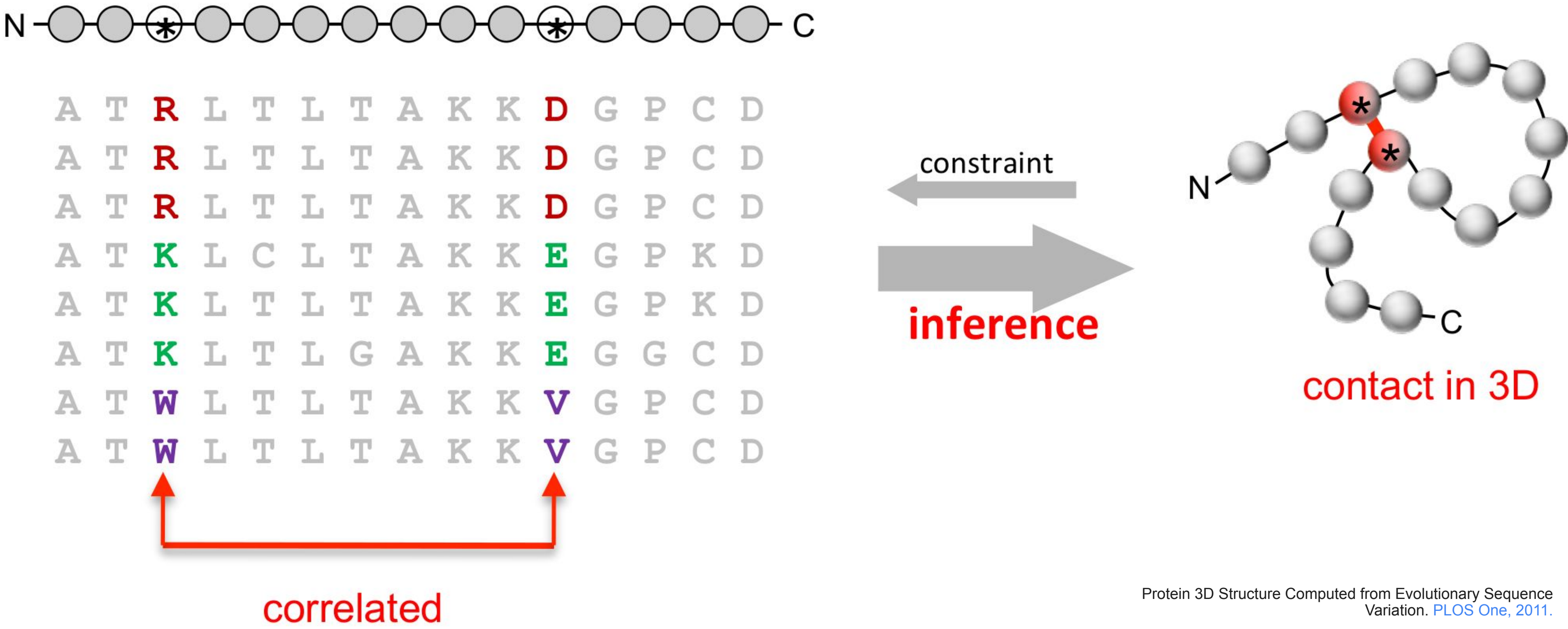




```
                                            cov    pid
                                                      1 [        .         .         .         .         :         .         .         . 80
 1 UniProt/Swiss-Prot|P26898|IL2RA_SHEEP 100.0% 100.0%  MEPSLLMWRFFVFIVVPGCVTEACHDDPPSLRNA----------MFKVLRYE----VGTMINCDCKAGFRRVS---AVMR
 2 UniProt/Swiss-Prot|P01590|IL2RA_MOUSE  94.4%  45.1%  MEPRLLMLGFLSLTIVPSCRAELCLYDPPEVPNA----------TFKALSYK----NGTILNCECKRGFRRLKE-LVYMR
 3 UniProt/Swiss-Prot|P41690|IL2RA_FELCA  98.9%  54.3%  MEPSLLLWGILTFVVVHGHVTELCDENPPDIQHA----------TFKALTYK----TGTMLNCECKKGFRRISNGSAFML
 4 UniProt/Swiss-Prot|P01589|IL2RA_HUMAN  98.9%  47.8%  MDSYLLMWGLLTFIMVPGCQAELCDDDPPEIPHA----------TFKAMAYK----EGTMLNCECKRGFRRIKSGSLYML
 5 UniProt/Swiss-Prot|Q5MNY4|IL2RA_MACMU  98.9%  48.9%  MDPYLLMWGLLTFITVPGCQAELCDDDPPKITHA----------TFKAVAYK----EGTMLNCECKRGFRRIKSGSPYML
 6 UniProt/Swiss-Prot|Q95118|IL2RG_BOVIN  96.6%  11.0%  MLKPPLPLRSLLFLQLPLLGVGLNPKFLTPSGNEDIGGKPGTGGDFFLTSTPAGTLDVSTLPLPKVQC---FVFNVEYMN
 7 UniProt/Swiss-Prot|P40321|IL2RG_CANFA  95.5%  10.7%  MLKPPLPLRSLLFLQLSLLGVGLNSTVPMPNGNEDIT------PDFFLTATPSETLSVSSLPLPEVQC---FVFNVEYMN
 8 UniProt/Swiss-Prot|P26896|IL2RB_RAT    73.0%   9.3%  MATVDLSWRLPLYILLLLLATT--------------------------------WVSAAVNDCSHLKC---FYNSRANVS
 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
```

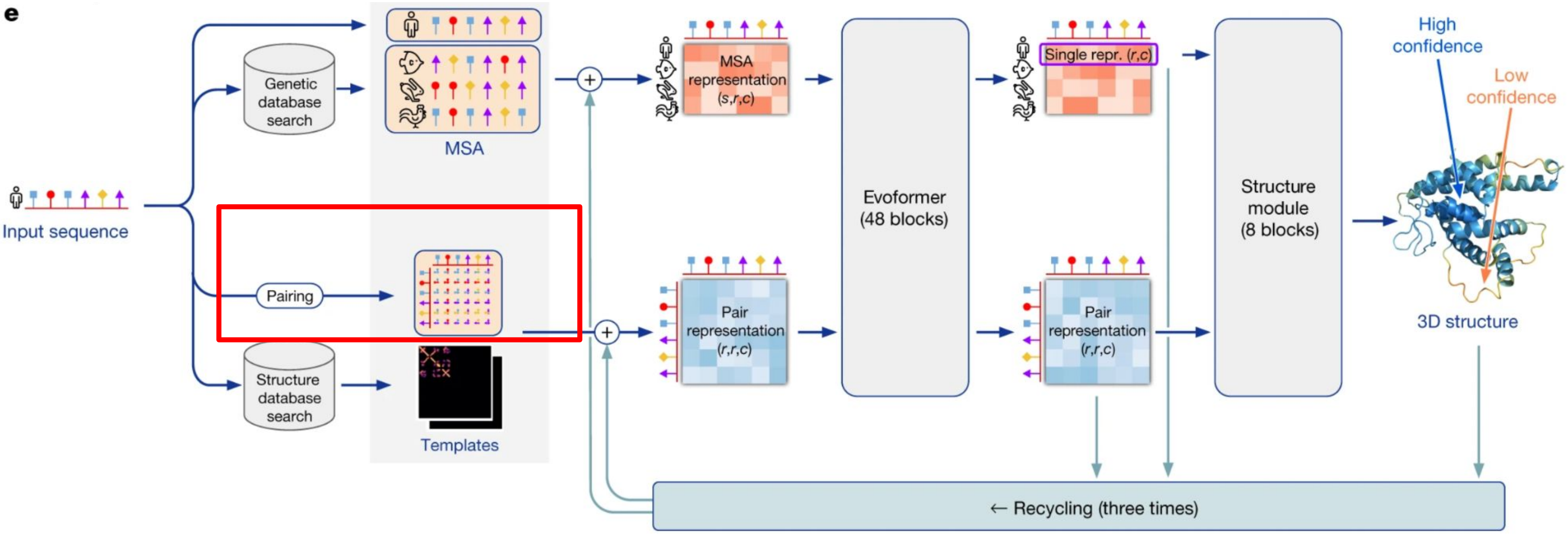
Direct-coupling analysis of residue coevolution captures native contacts across many protein families



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

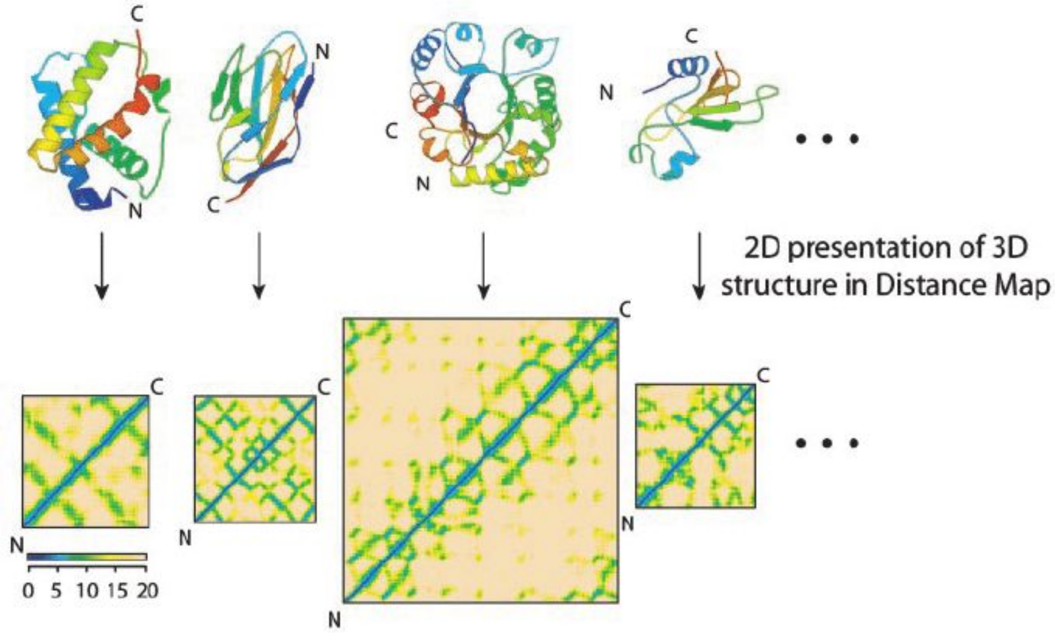
Protein 3D Structure Computed from Evolutionary Sequence Variation. PLOS One, 2011.

Initializing the Pair Representation


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)

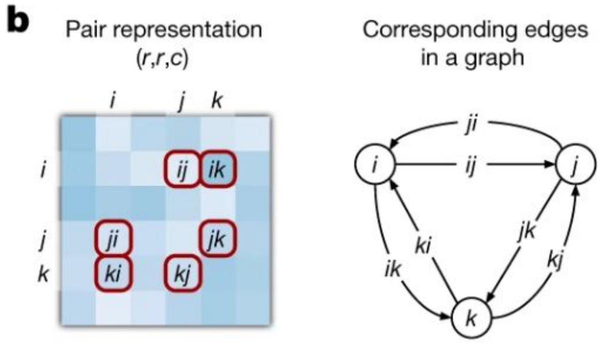
Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589 (2021). https://doi.org/10.1038/s41586-021-03819-2

What is a pair representation?



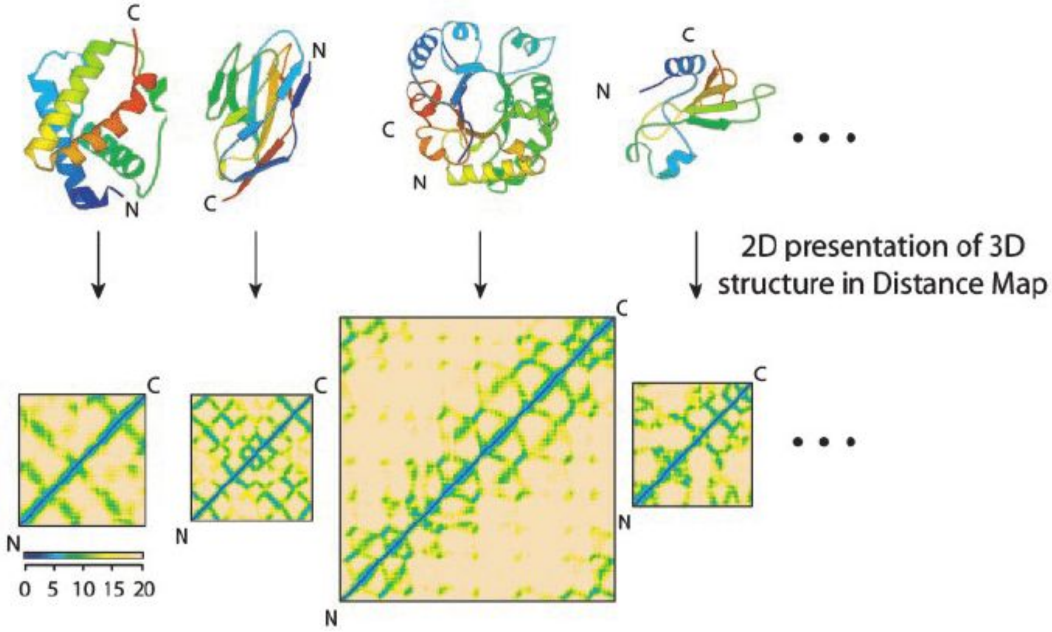
Distogram is used to map 2D pairwise distances

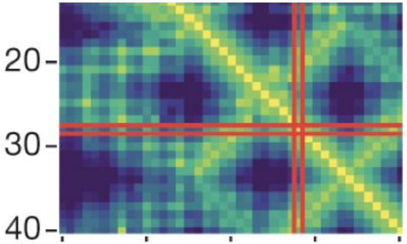
Distogram are independent of translations and rotations, so no need to align structures (much faster)



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

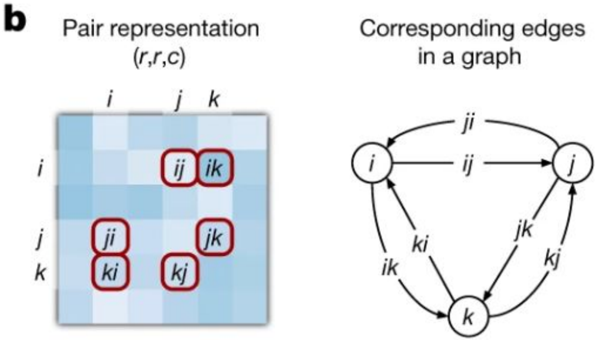
What is a pair representation?





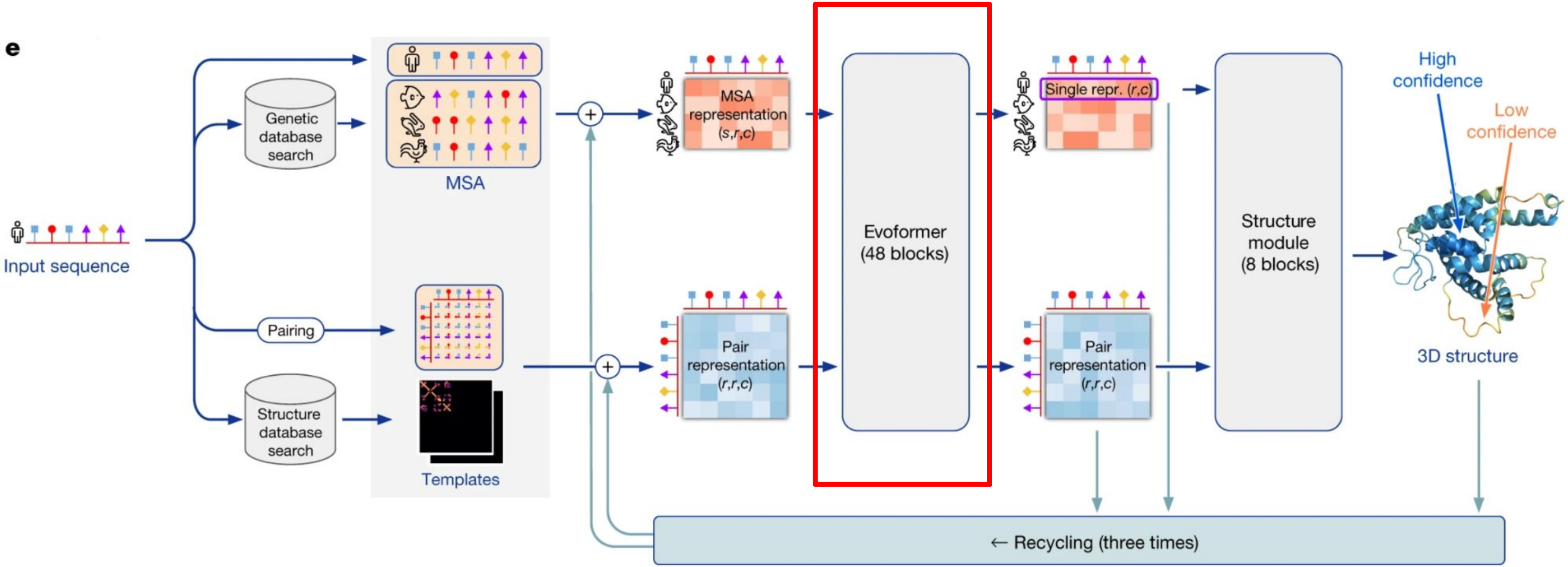
Distogram is used to map 2D pairwise distances

Distogram are independent of translations and rotations, so no need to align structures (much faster)

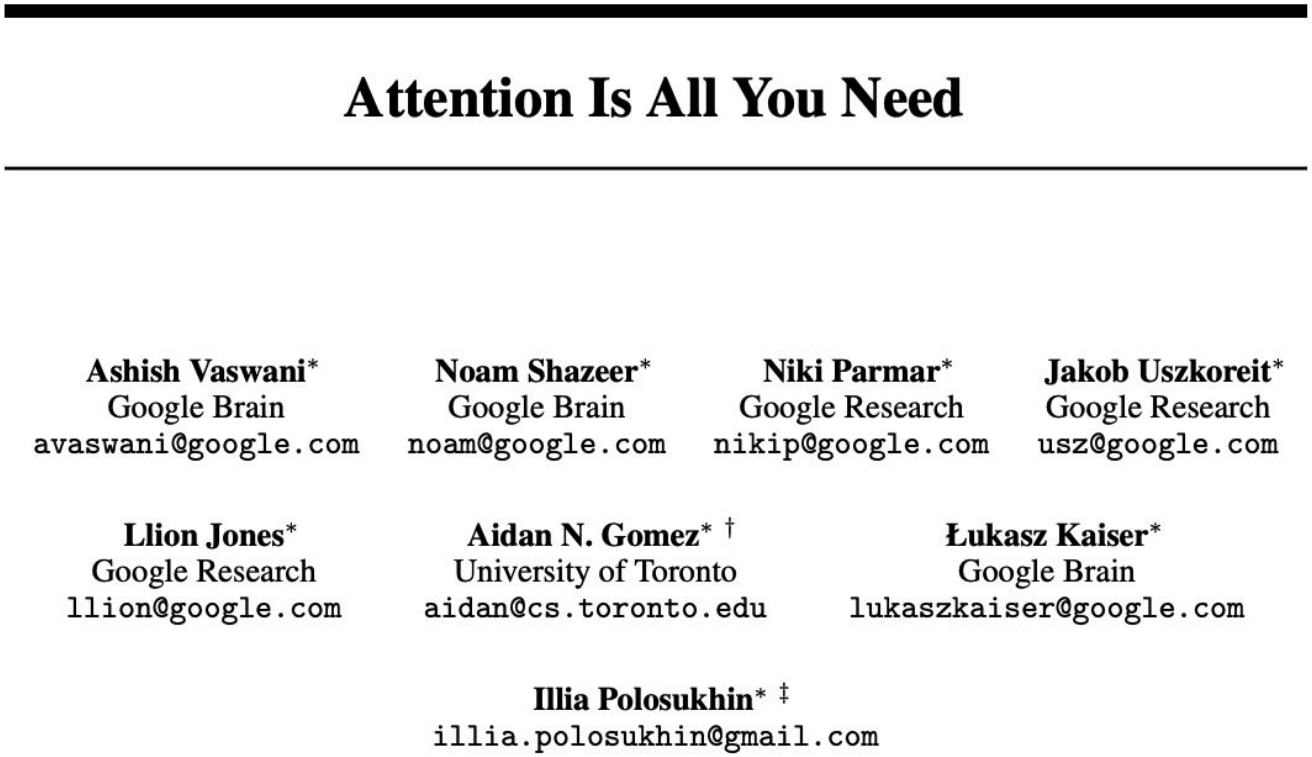


Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., ... & Hassabis, D. (2020). Nature.

Transformer-like module updates representations


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)

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

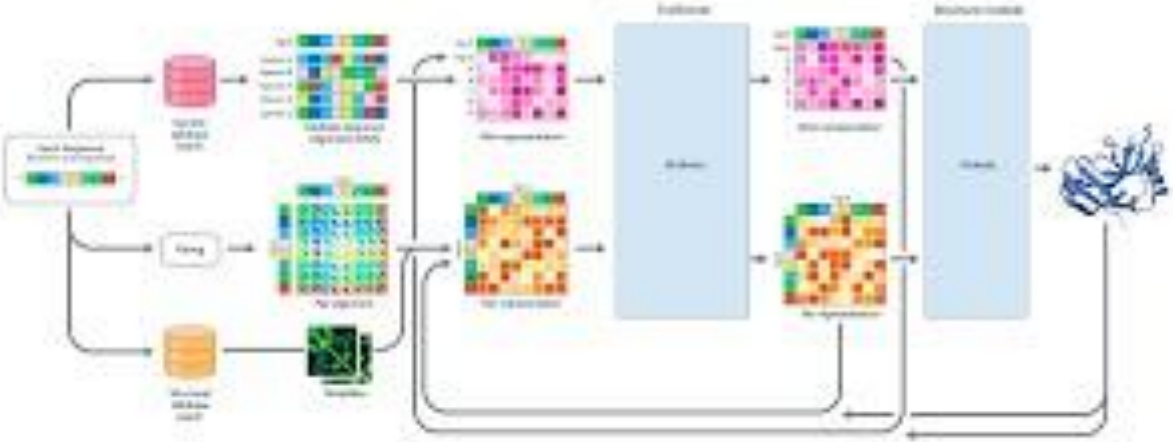
Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser*
Google Brain
lukaszkaiser@google.com

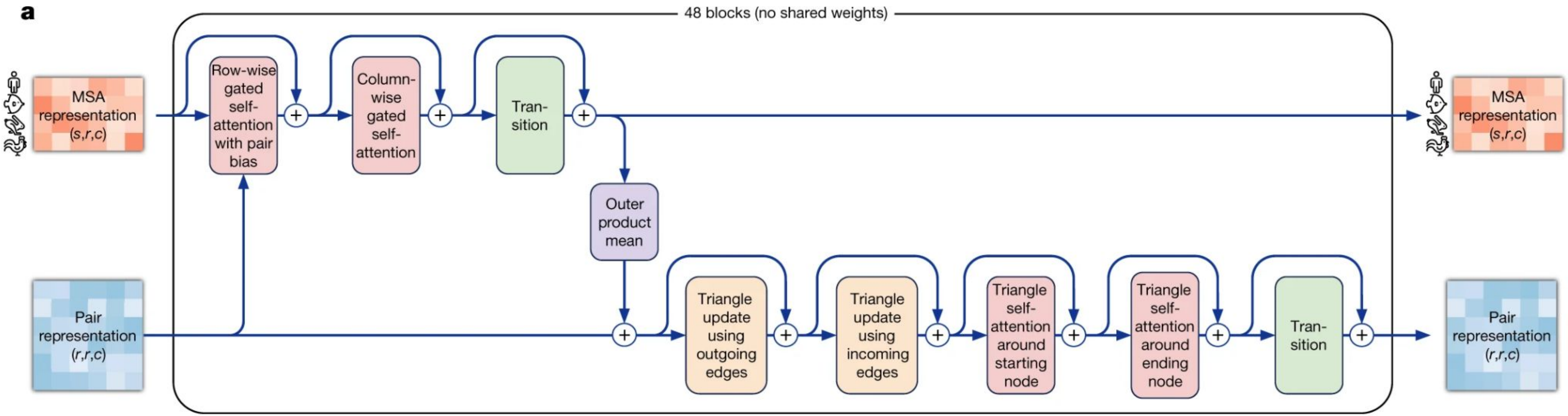
Illia Polosukhin* ‡
illia.polosukhin@gmail.com

https://arxiv.org/abs/1706.03762

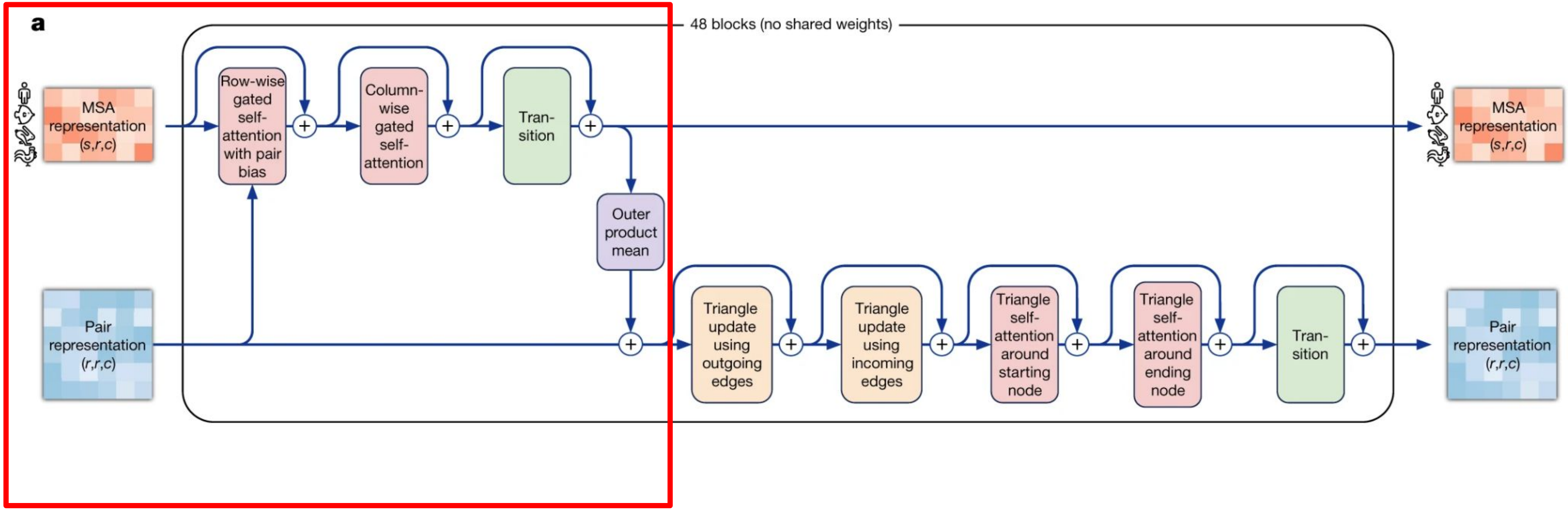
What Is AlphaFold?



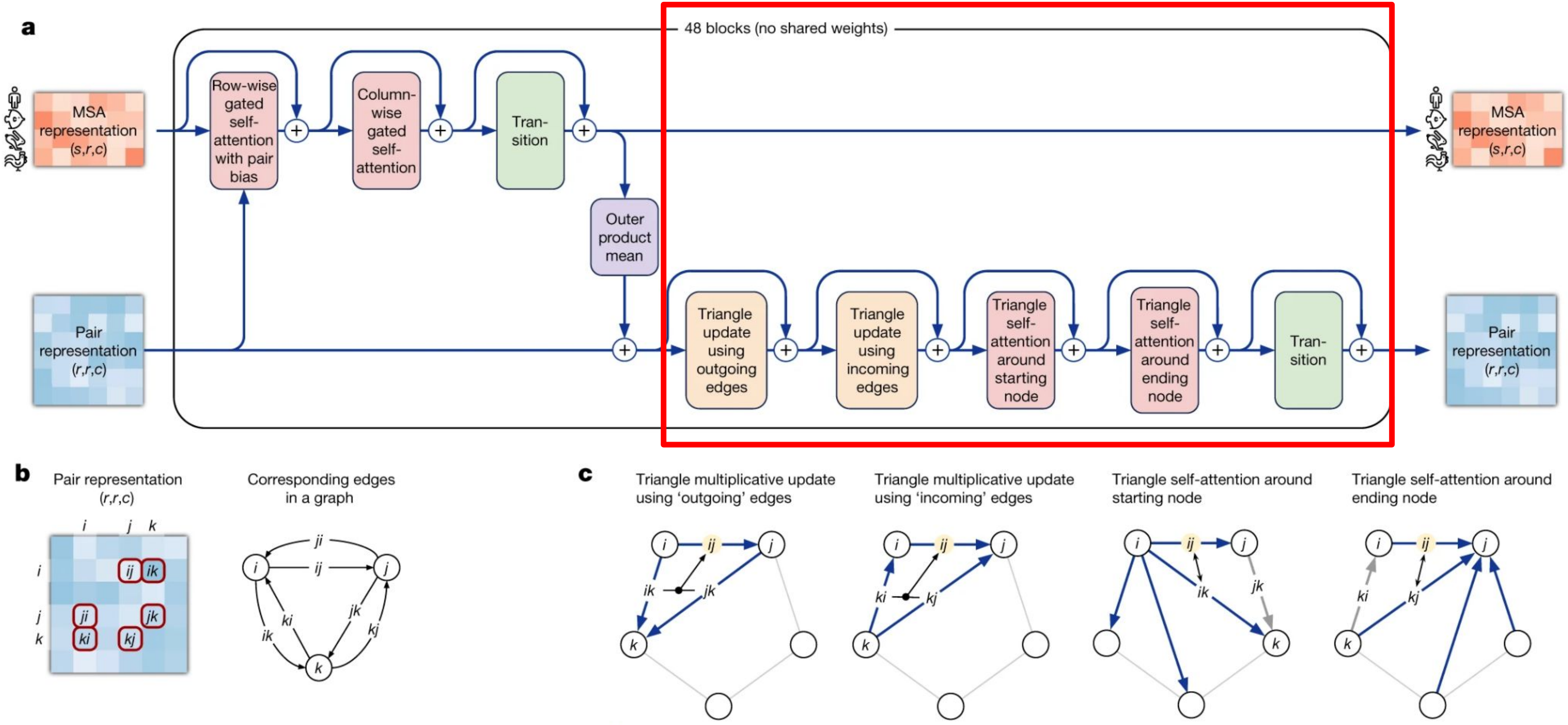
The Evoformer: 48 blocks of back and forth between evolutionary and spatial reasoning.


a
48 blocks (no shared weights)
MSA representation (s,r,c)
Row-wise gated self-attention with pair bias
+
Column-wise gated self-attention
+
Tran-sition
+
MSA representation (s,r,c)
Outer product mean
Pair representation (r,r,c)
+
Triangle update using outgoing edges
+
Triangle update using incoming edges
+
Triangle self-attention around starting node
+
Triangle self-attention around ending node
+
Tran-sition
+
Pair representation (r,r,c)

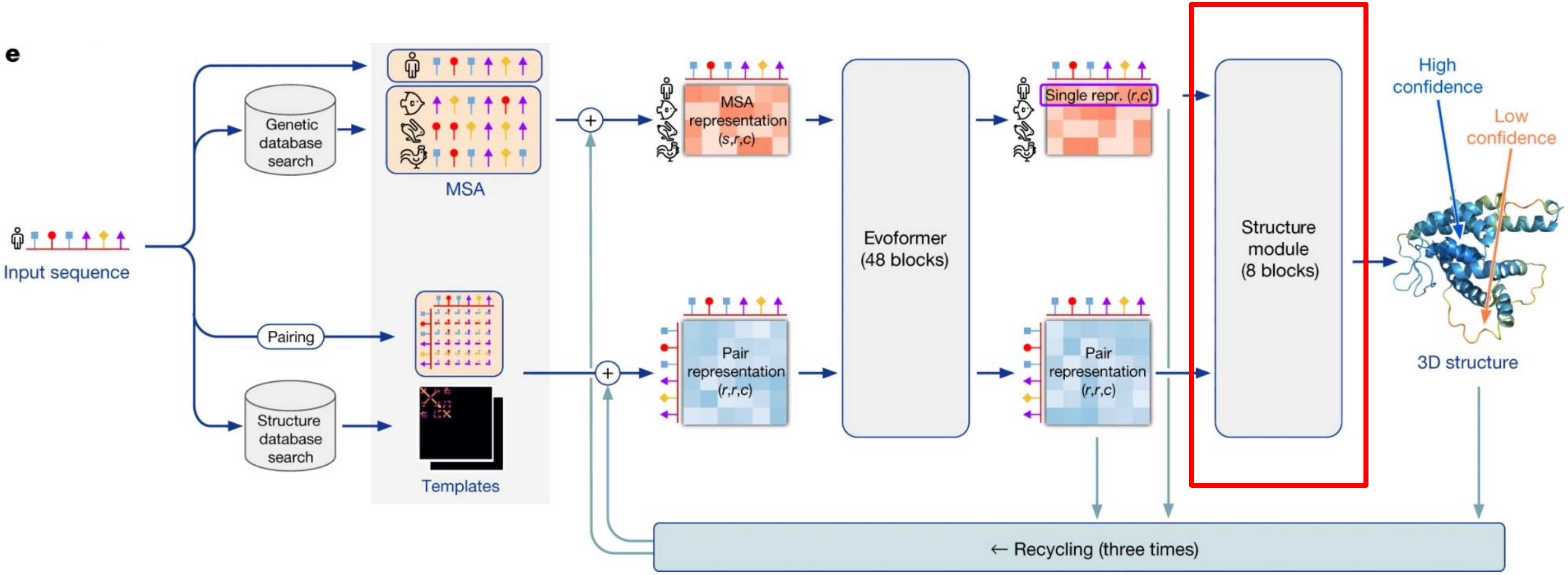
The Evoformer: 48 blocks of back and forth between evolutionary and spatial reasoning.


a
48 blocks (no shared weights)
MSA representation (s,r,c)
Pair representation (r,r,c)
Row-wise gated self-attention with pair bias
Column-wise gated self-attention
Tran-sition
Outer product mean
Triangle update using outgoing edges
Triangle update using incoming edges
Triangle self-attention around starting node
Triangle self-attention around ending node
Tran-sition
MSA representation (s,r,c)
Pair representation (r,r,c)

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

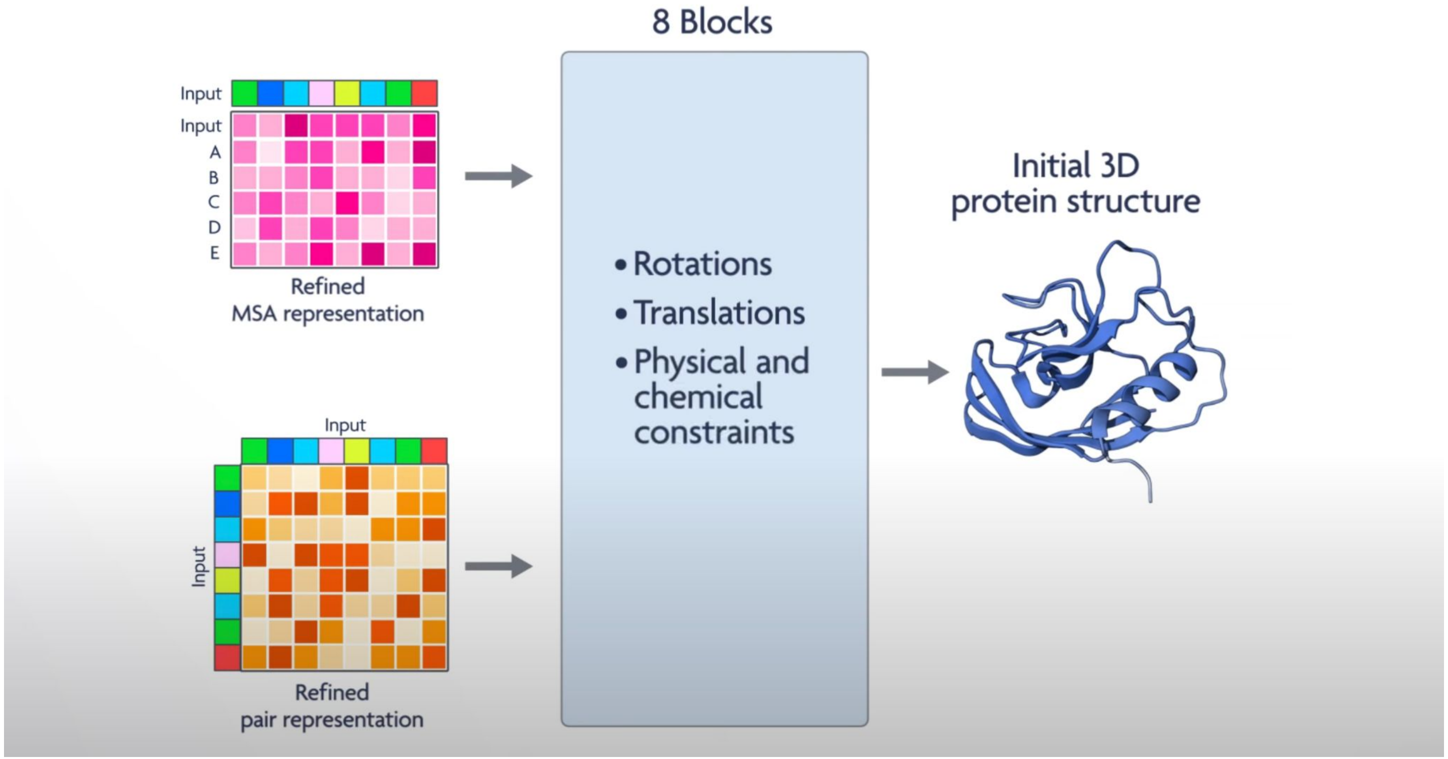

a
48 blocks (no shared weights)
MSA representation (s,r,c)
Row-wise gated self-attention with pair bias
Column-wise gated self-attention
Tran-sition
Outer product mean
Pair representation (r,r,c)
Triangle update using outgoing edges
Triangle update using incoming edges
Triangle self-attention around starting node
Triangle self-attention around ending node
Tran-sition
MSA representation (s,r,c)
Pair representation (r,r,c)
b
Pair representation (r,r,c)
i j k
i
j
k
ij ik ji jk ki kj
Corresponding edges in a graph
ji ij jk kj ik ki
i j k
c
Triangle multiplicative update using 'outgoing' edges
Triangle multiplicative update using 'incoming' edges
Triangle self-attention around starting node
Triangle self-attention around ending node
i ij j ik jk k
i ij j ki kj k
i ij j ik jk k
i ij j ki kj k

Structure module converts representation to structure


e
Input sequence
Genetic database search
Pairing
Structure database search
MSA
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)

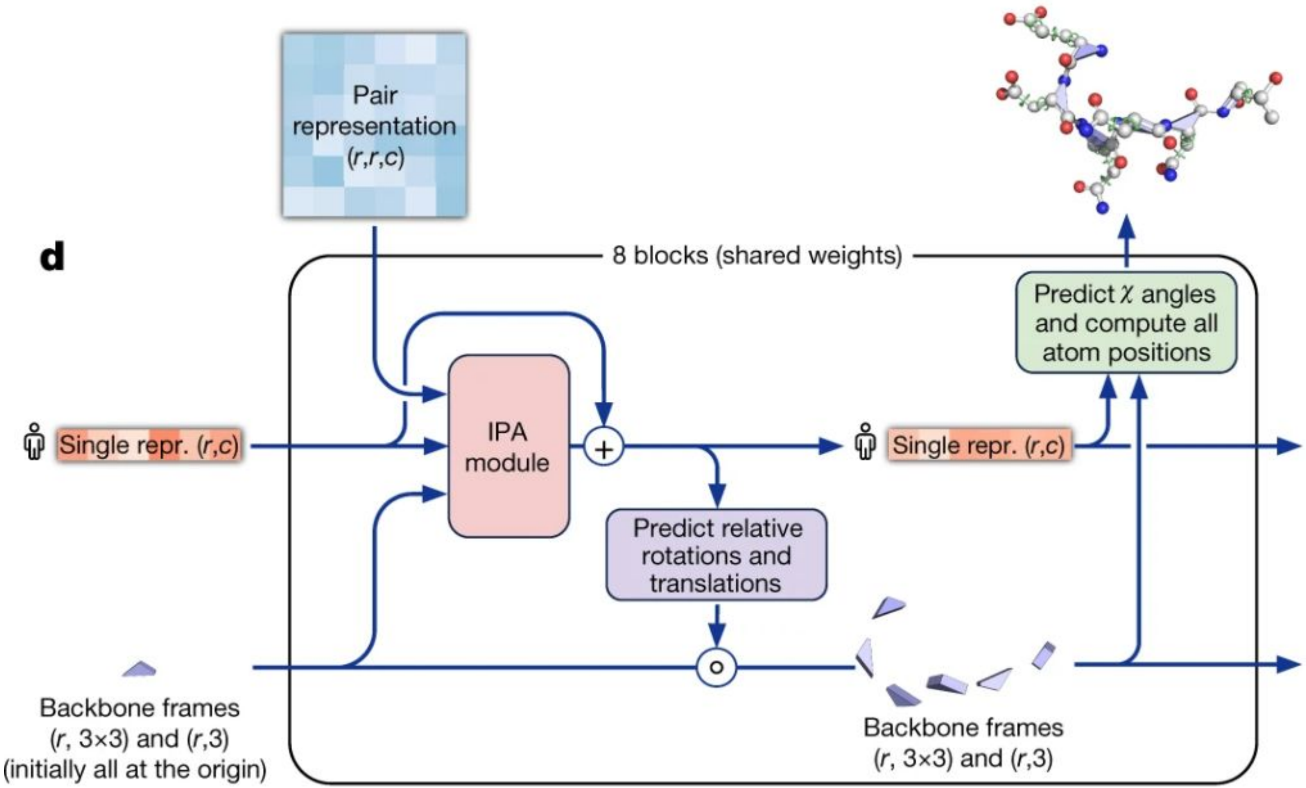
Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589 (2021). https://doi.org/10.1038/s41586-021-03819-2

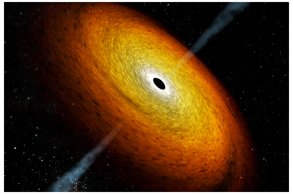
Structure module converts representation to structure



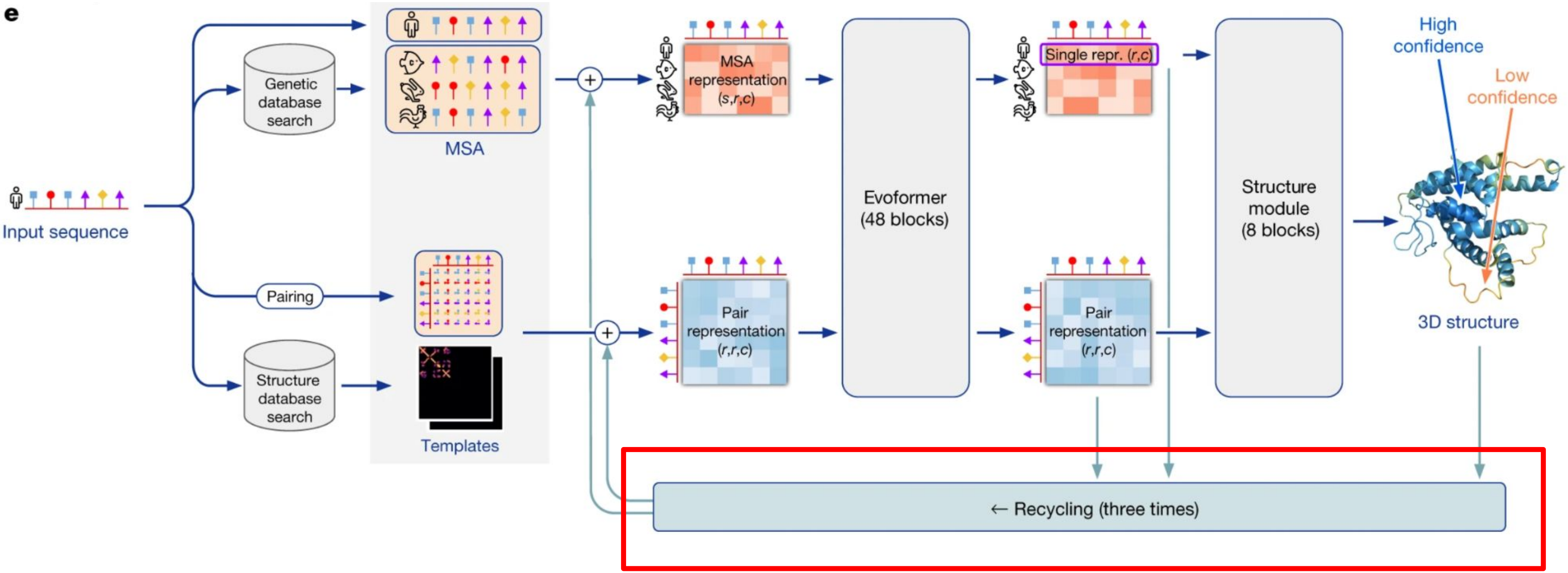
- Rotations
- Translations
- Physical and chemical constraints

The Structure Module: iteratively updating a residue gas


Pair representation (r,r,c)
d
8 blocks (shared weights)
Predict χ angles and compute all atom positions
Single repr. (r,c)
IPA module
+
Single repr. (r,c)
Predict relative rotations and translations
o
Backbone frames (r, 3×3) and (r,3) (initially all at the origin)
Backbone frames (r, 3×3) and (r,3)

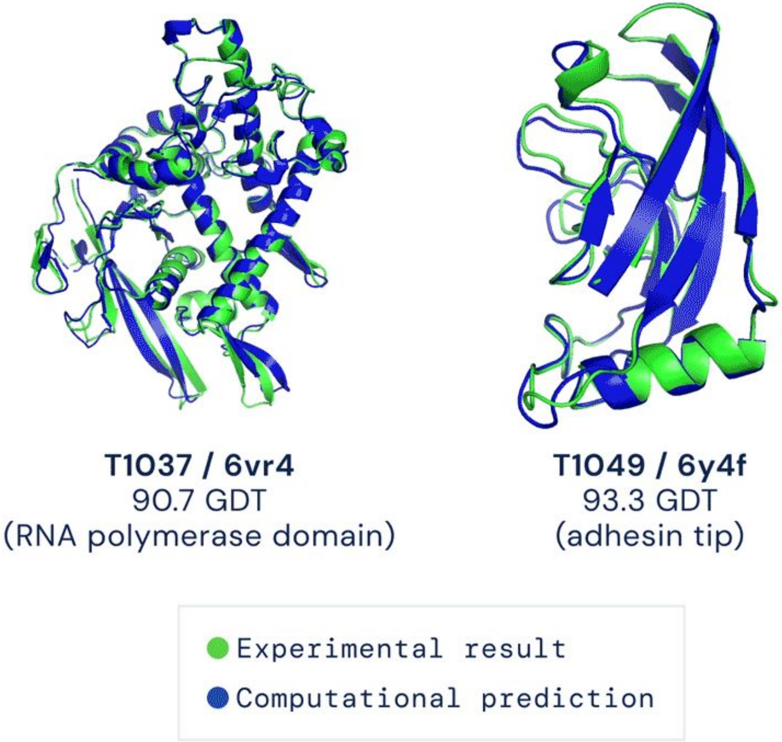


Recycling iteratively refines structure


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)

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!



T1037 / 6vr4
90.7 GDT
(RNA polymerase domain)

T1049 / 6y4f
93.3 GDT
(adhesin tip)

Check in on Demo

The 3Ps: pLDDT, PAE, and pTM

- LDDT, AE, TM require ground truth. What can we do?

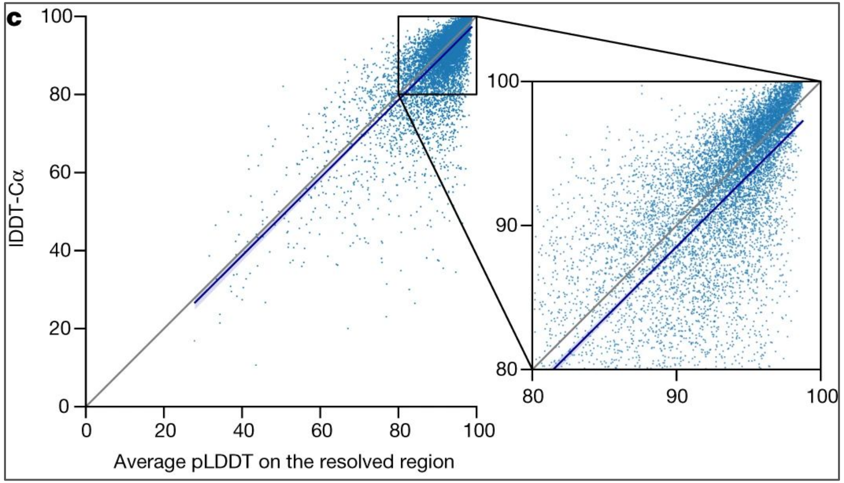
- pLDDT: predicted **L**ocal **D**istance **D**ifference **T**est

- 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

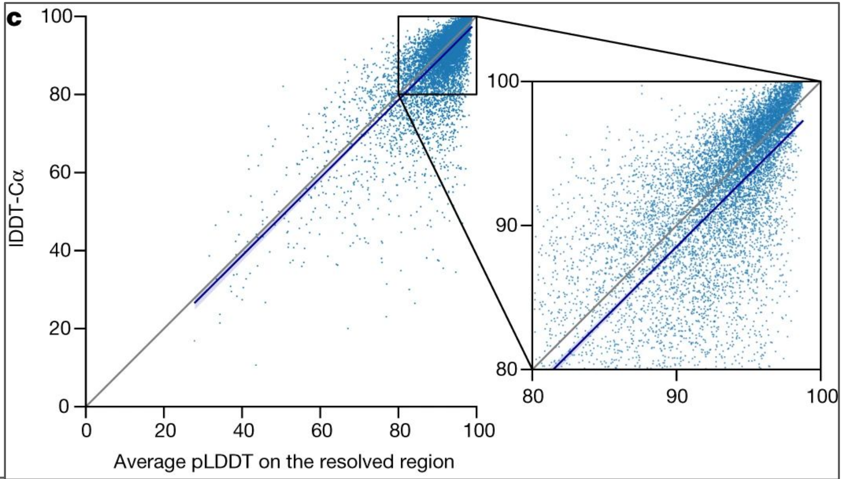
- LDDT, AE, TM require ground truth. What can we do?
- pLDDT: predicted **L**ocal **D**istance **D**ifference **T**est
- PAE: **P**redicted **A**ligned **E**rror
- pTM: predicted **T**emplate **M**odeling score
  - Global comparison of similarity between two structures
  - Measure of 0 to 1

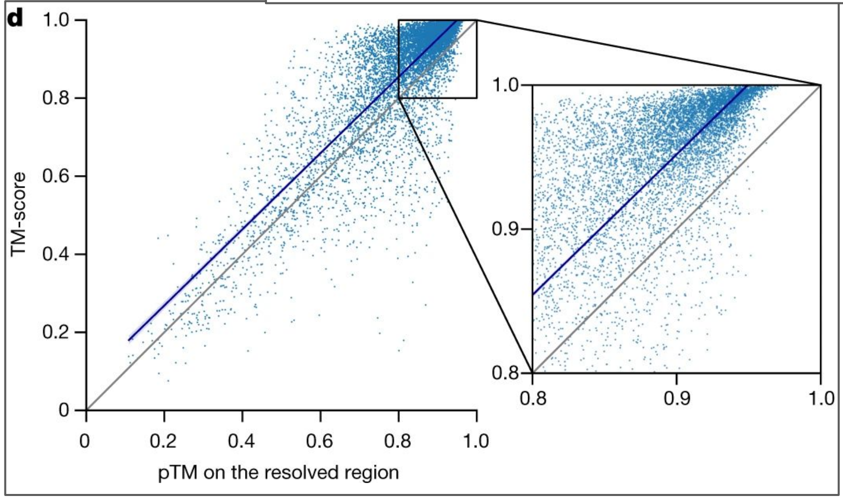


Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589 (2021). https://doi.org/10.1038/s41586-021-03819-2

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

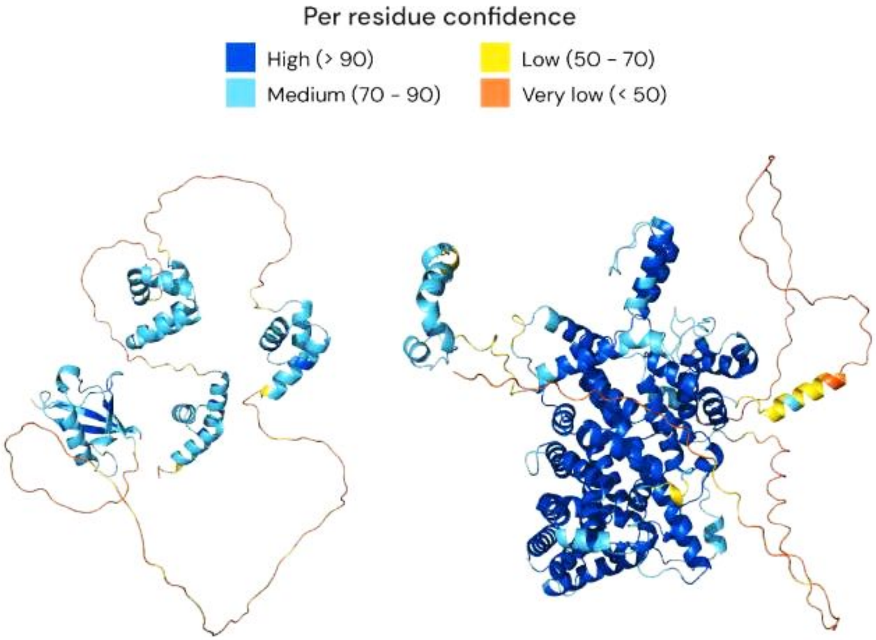




Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589 (2021). https://doi.org/10.1038/s41586-021-03819-2

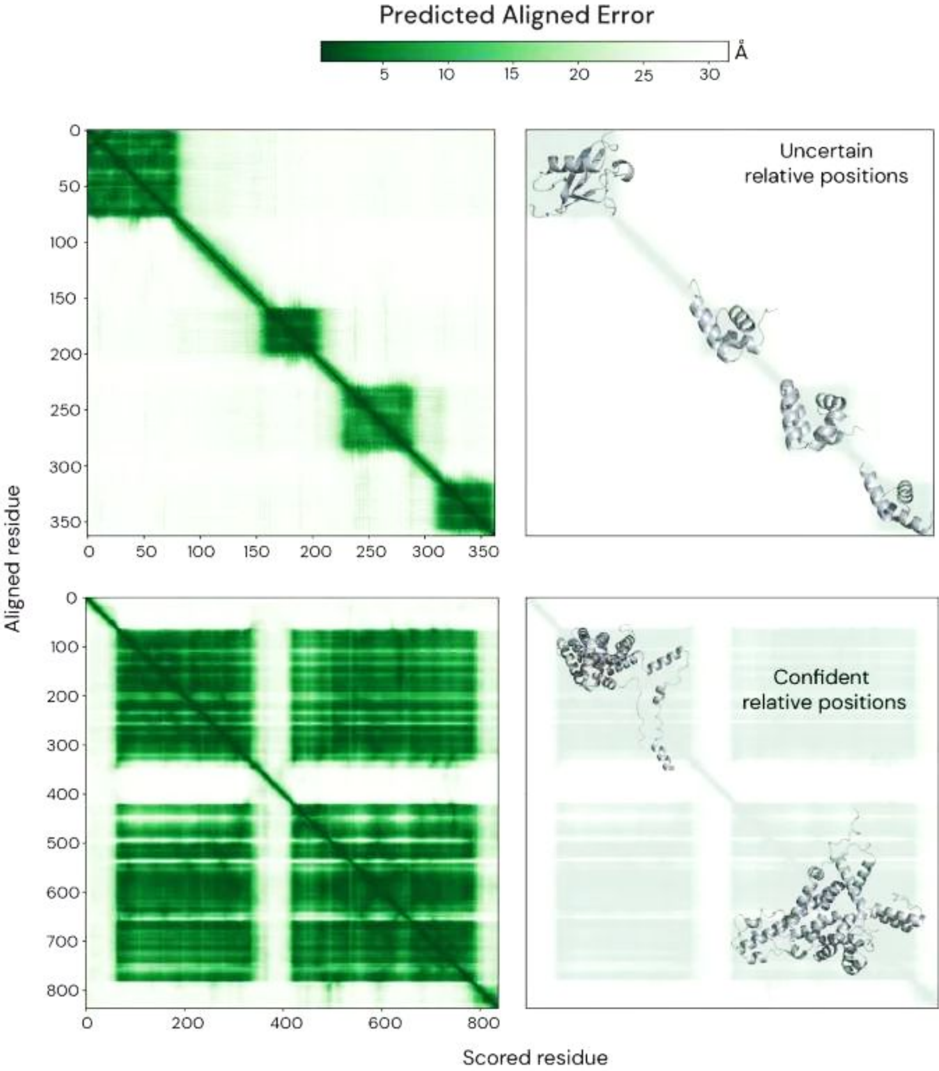
Predicted Local Distance Difference Test

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



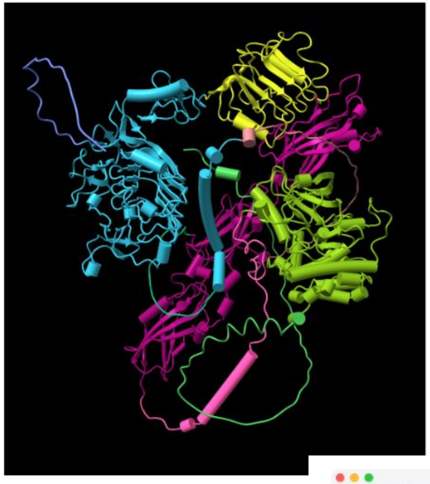
Predicted Aligned Error (PAE)

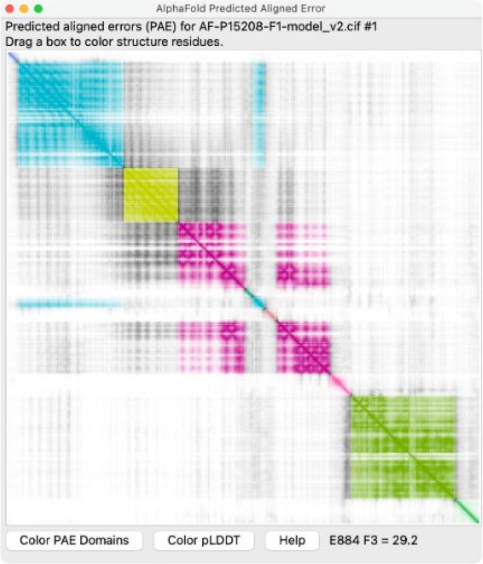
- Prediction of position error at residue x if the predicted and the true structures were aligned on y
- PAE aims to measure confidence in the relative positions of pairs of residues
- Use where pairwise confidence is relevant – interpreting domain distances in a multi domain protein
- Suppose residue 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





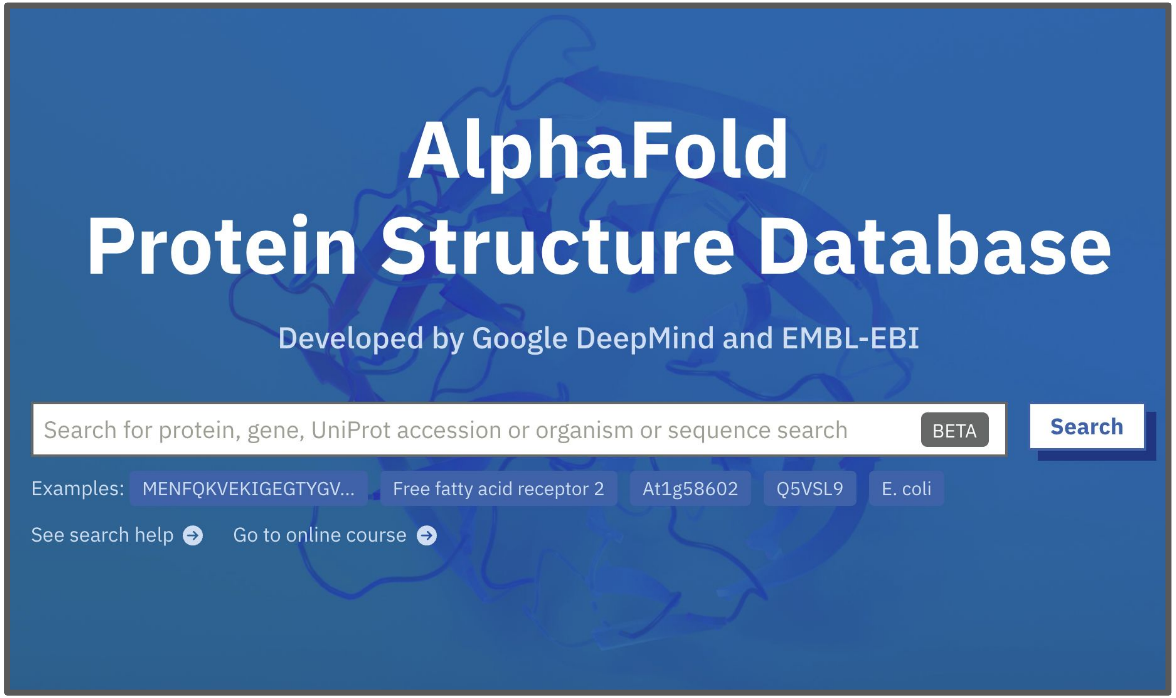
AlphaFold Predicted Aligned Error

Predicted aligned errors (PAE) for AF-P15208-F1-model_v2.cif #1

Drag a box to color structure residues.

E884 F3 = 29.2

214,683,839 Predicted Protein Structures on AFDB!



AlphaFold Protein Structure Database

Developed by Google DeepMind and EMBL-EBI

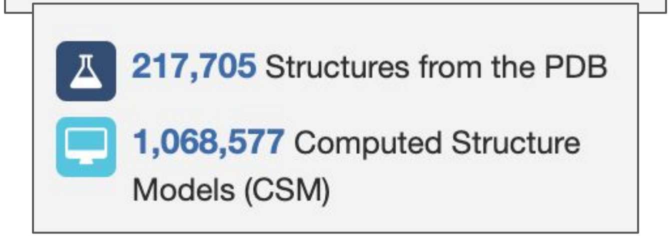
Search for protein, gene, UniProt accession or organism or sequence search
BETA
Search
Examples: MENFQKVEKIGEGTYGV... Free fatty acid receptor 2 At1g58602 Q5VSL9 E. coli
See search help
Go to online course

PROTEIN DATA BANK

Years

RCSB PDB

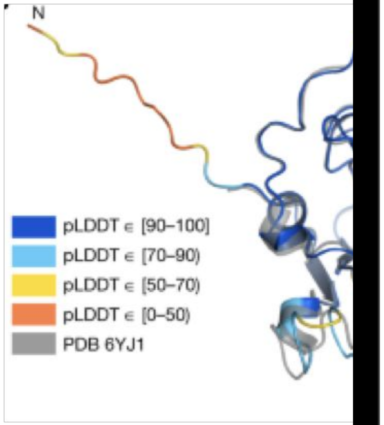
PROTEIN DATA BANK

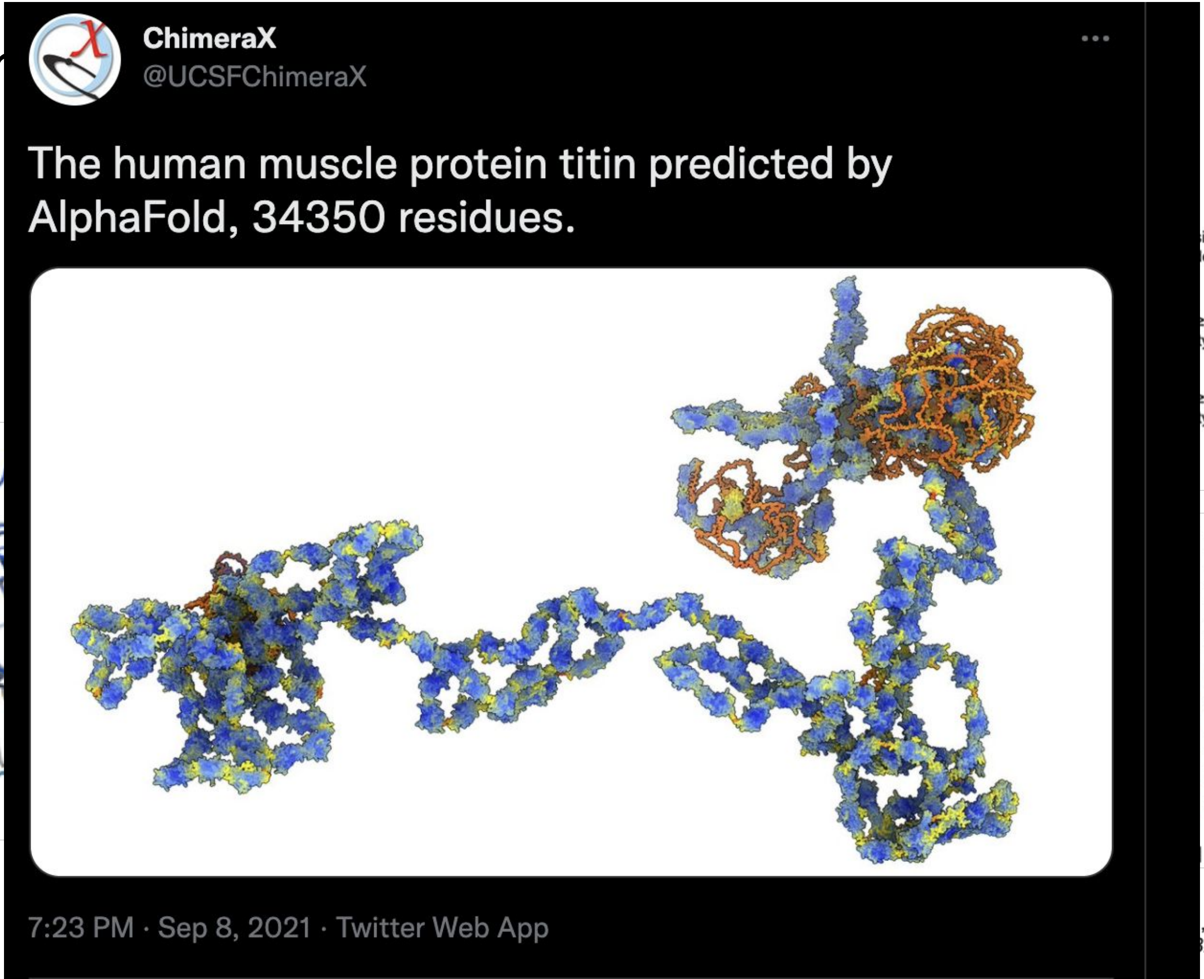


217,705 Structures from the PDB

1,068,577 Computed Structure Models (CSM)

Proteor

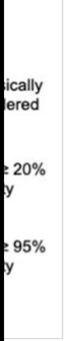




ChimeraX
@UCSFChimeraX

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

7:23 PM · Sep 8, 2021 · Twitter Web App

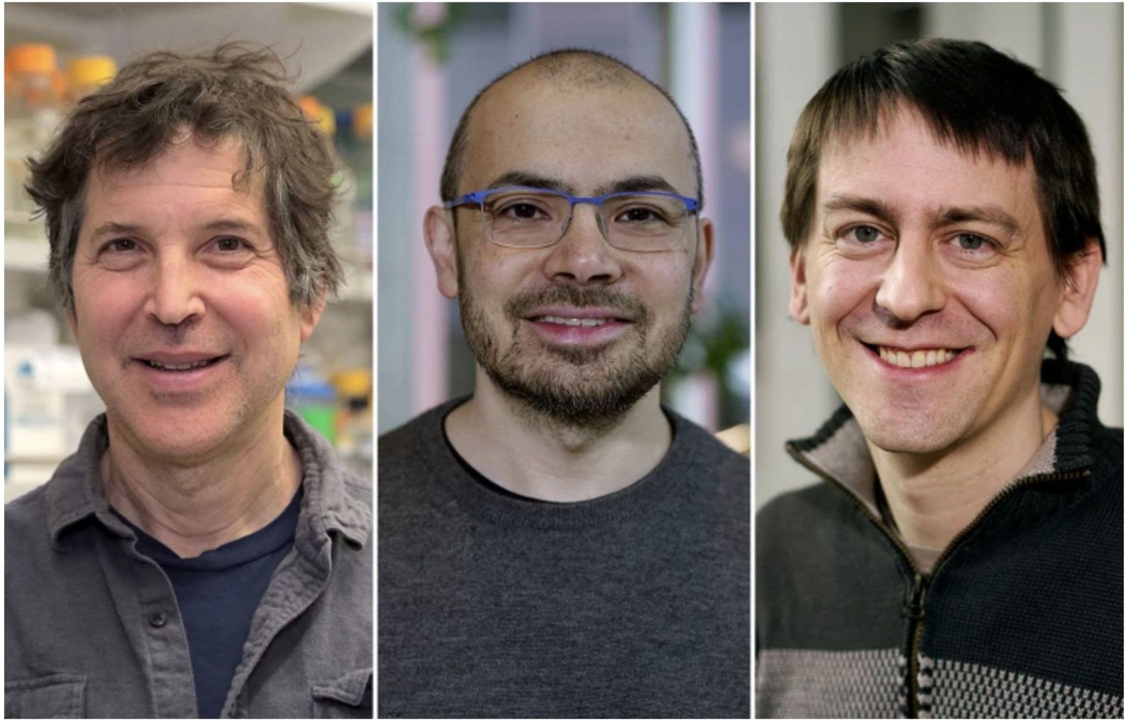


Dark proteome
Intrinsically disordered
PFAM – No PDB / AF
Gained AF – 70 > pLDDT ≥ 50
Gained AF – 90 > pLDDT ≥ 70
Gained AF – pLDDT ≥ 90
PDB 20% to 50% - pLDDT < 90
PDB 50% to 95%
PDB ≥ 95%

Tunyasuvunakool, K., Adler, J., Wu,

, Valentini, S., & Valencia, A.
iology.

2024 Nobel Prize in Chemistry



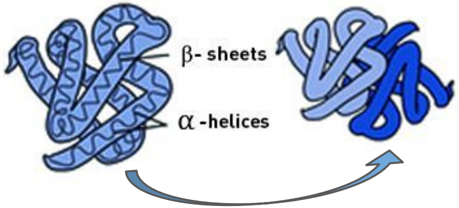
David Baker, Demis Hassabis and John Jumper (left to right) won the chemistry Nobel for developing computational tools that can predict and design protein structures. Credit: BBVA Foundation

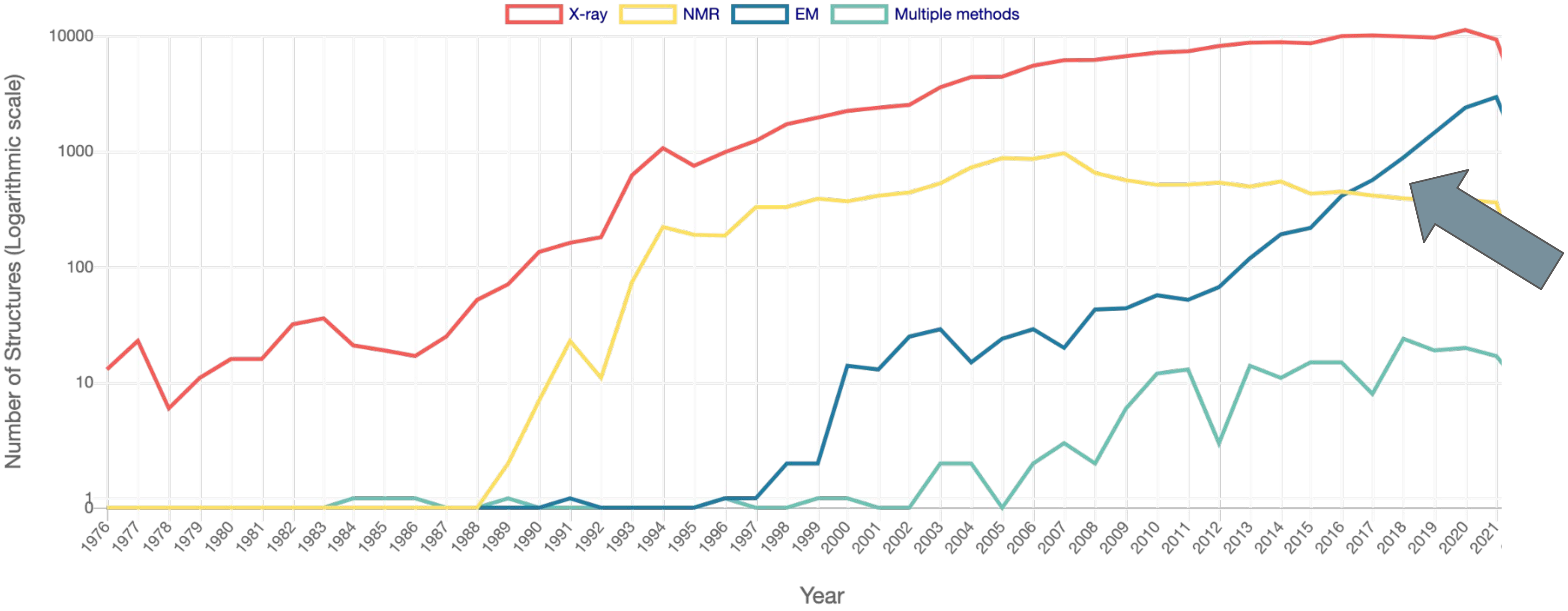
https://www-nature-com.ezproxy.lib.utexas.edu/articles/d41586-024-03214-7

AlphaFold Multimer!

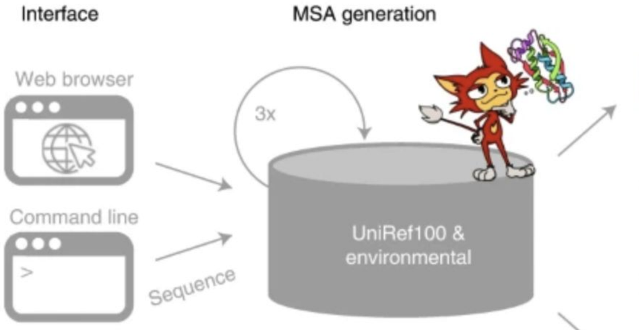
TERTIARY STRUCTURE

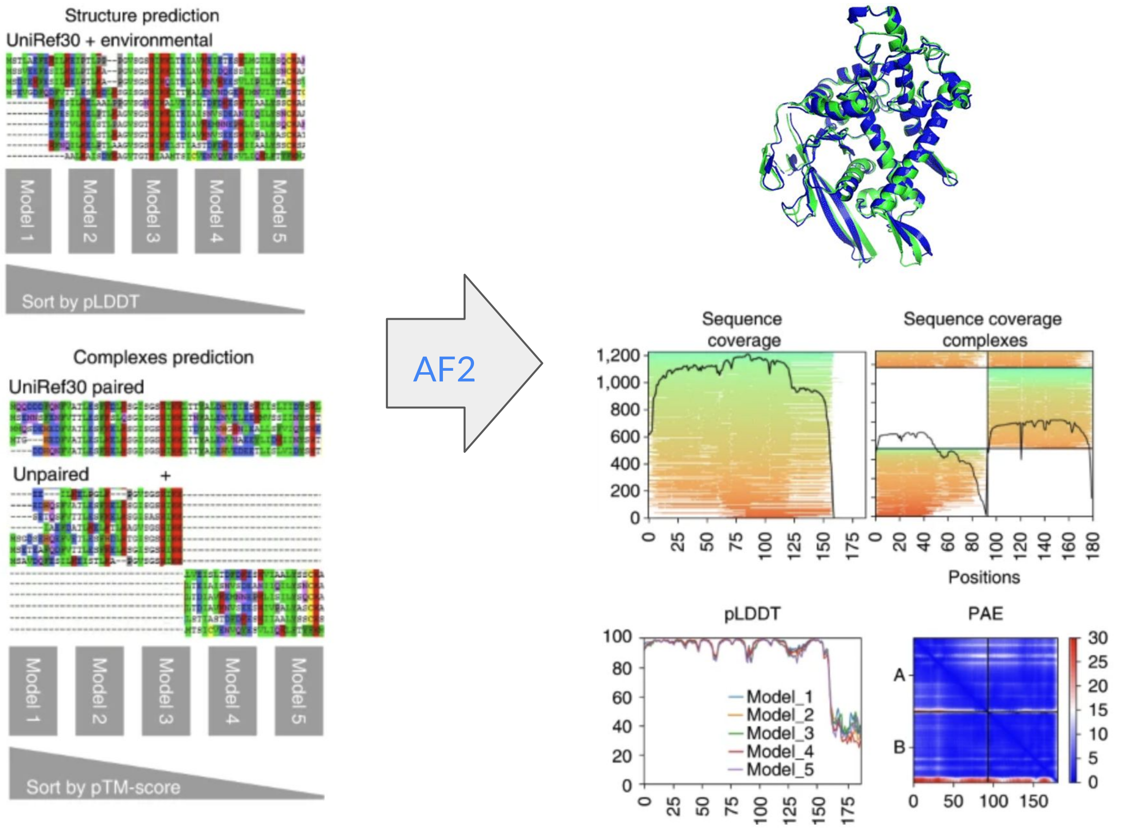
QUATERNARY STRUCTURE





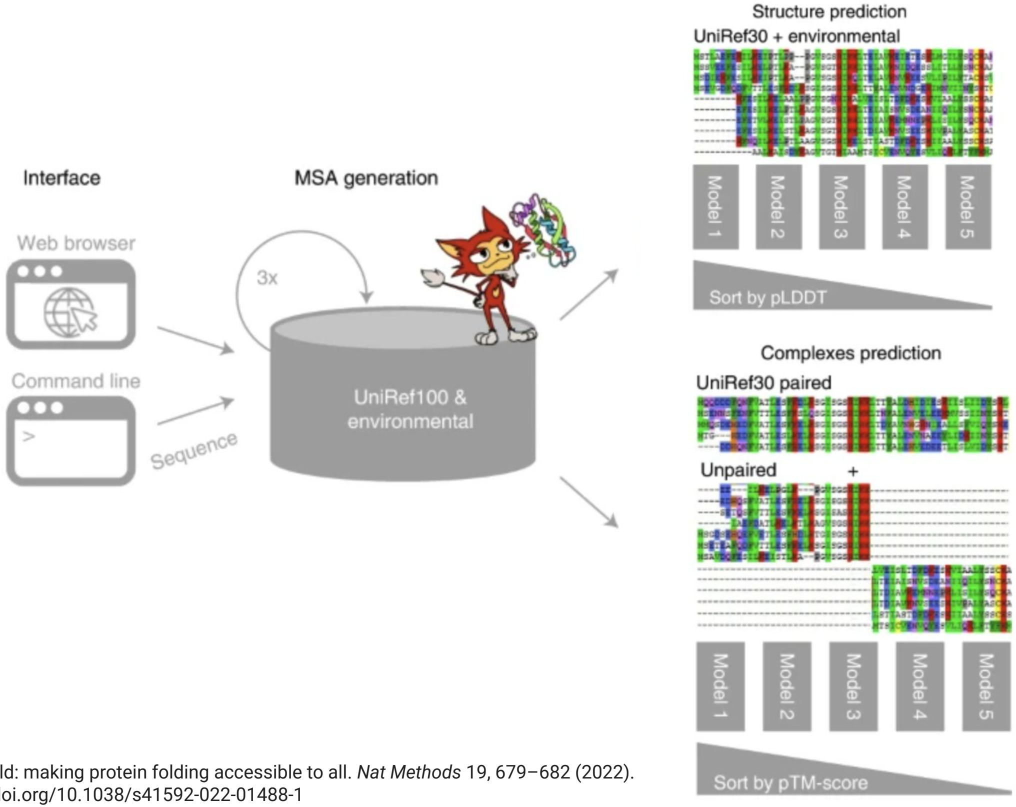
ColabFold speeds up AlphaFold by 40-60 fold!

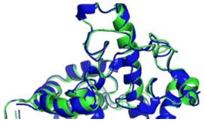




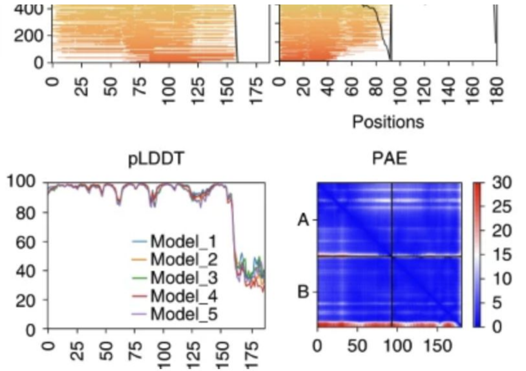
ColabFold: making protein folding accessible to all. *Nat Methods* 19, 679–682 (2022). https://doi.org/10.1038/s41592-022-01488-1

ColabFold speeds up AlphaFold by 40-60 fold!



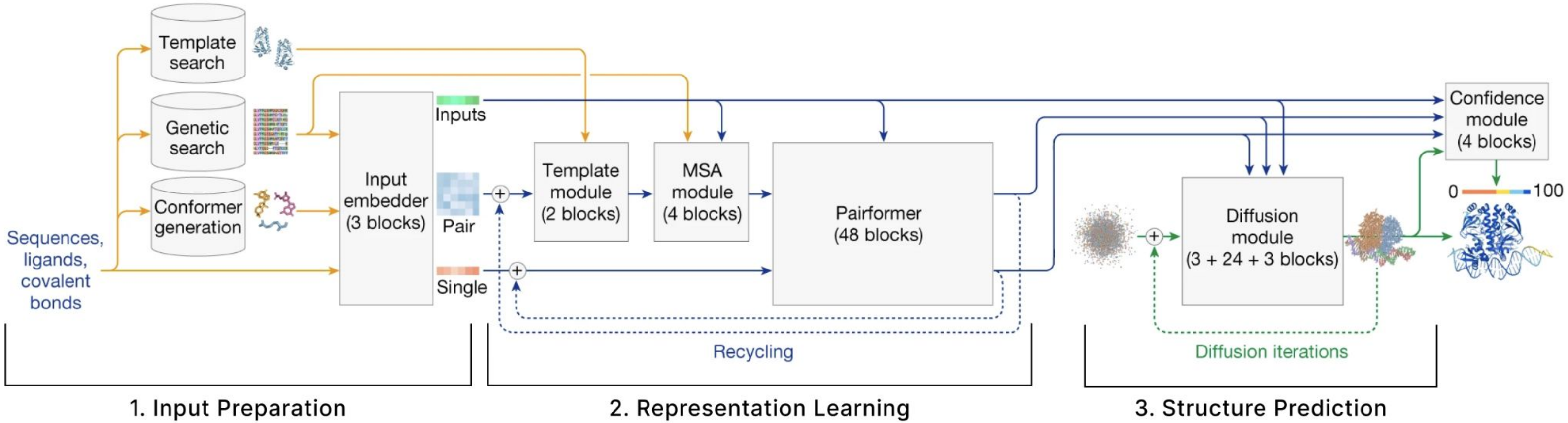


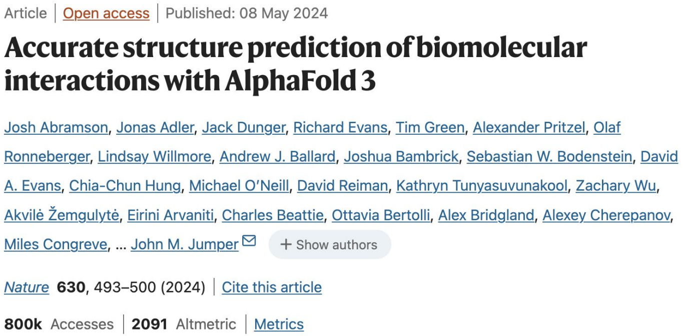
“ColabFold’s 40–60-fold faster search and optimized model utilization enables prediction of close to 1,000 structures per day on a server with one graphics processing unit.”



ColabFold: making protein folding accessible to all. *Nat Methods* 19, 679–682 (2022). https://doi.org/10.1038/s41592-022-01488-1

A new era: All Atom Models (AF3)



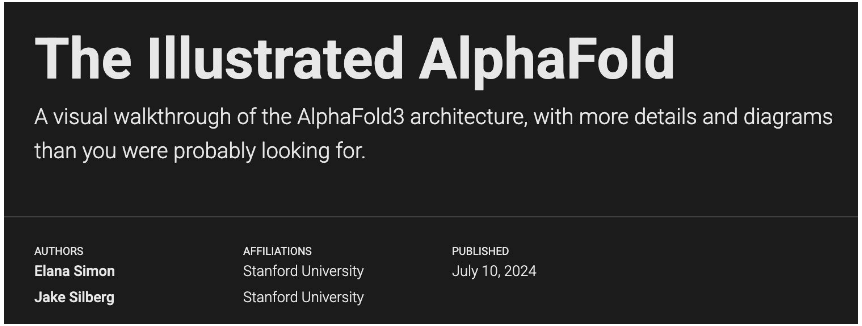

Article | Open access | Published: 08 May 2024

Accurate structure prediction of biomolecular interactions with AlphaFold 3

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J. Ballard, Joshua Bambrick, Sebastian W. Bodenstein, David A. Evans, Chia-Chun Hung, Michael O'Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvilė Žemgulytė, Eirini Arvaniti, Charles Beattie, Ottavia Bertolli, Alex Bridgland, Alexey Cherepanov, Miles Congreve, ... John M. Jumper + Show authors

Nature **630**, 493–500 (2024) | Cite this article

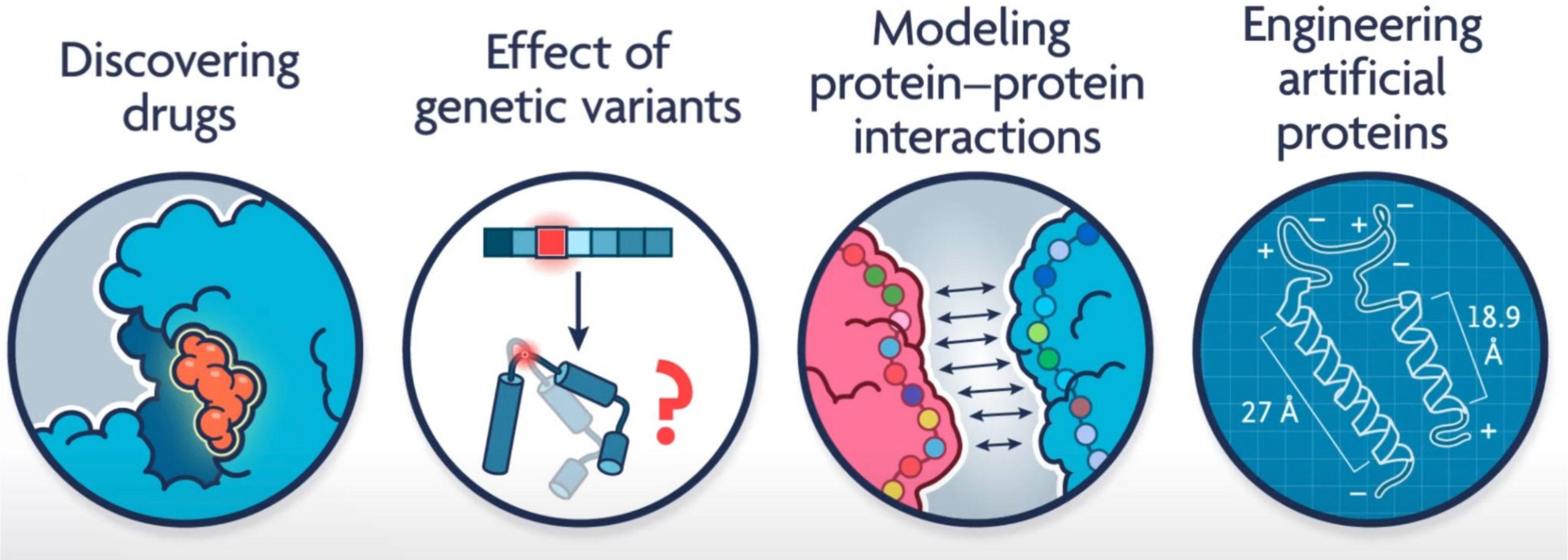
800k Accesses | **2091** Altmetric | Metrics


The Illustrated AlphaFold

A visual walkthrough of the AlphaFold3 architecture, with more details and diagrams than you were probably looking for.

AUTHORS	AFFILIATIONS	PUBLISHED
Elana Simon	Stanford University	July 10, 2024
Jake Silberg	Stanford University	

Applications and Frontiers!


Discovering drugs
Effect of genetic variants
Modeling protein–protein interactions
Engineering artificial proteins
18.9 Å
27 Å

Has protein structure prediction been solved?

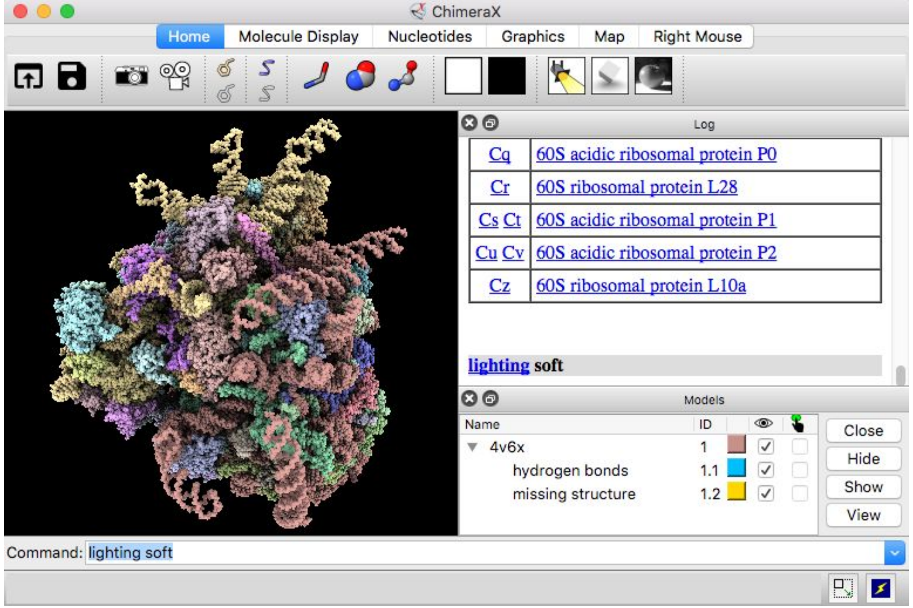
- **Sort of.**
- **The Protein Folding Problem (de novo)**
  - What is the folding code?
  - **What is the folding mechanism?**
  - Can we predict a native protein structure from its primary, amino acid sequence?
    - **No for a sequence in isolation…**
    - **Yes when informed by like sequences and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

Resources & Useful Links

- Prediction Servers/Colabs:
  - AF3 Server
  - 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
  - The Illustrated AlphaFold
  - 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
  - AlphaFold2 & Equivariance

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



Cq	60S acidic ribosomal protein P0
Cr	60S ribosomal protein L28
Cs Ct	60S acidic ribosomal protein P1
Cu Cv	60S acidic ribosomal protein P2
Cz	60S ribosomal protein L10a

```
lighting soft
```

Name	ID		👁	
▼ 4v6x	1	🟫	☑	□
hydrogen bonds	1.1	🟦	☑	□
missing structure	1.2	🟨	☑	□

```
Command: lighting soft
```

Thank you! Questions?

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

https://www.biomlsociety.org

We meet every other Thursday from 11am to noon, in MBB 3.204

TACOS and **COFFEE** provided!

Watch the Commander Complex assemble!