DeepMind

Highly accurate protein structure prediction with AlphaFold

Michael Figurnov
Staff Research Scientist, DeepMind

John Jumper[1*†], Richard Evans[1*], Alexander Pritzel[1*], Tim Green[1*], Michael Figurnov[1*], Kathryn Tunyasuvunakool[1*], Olaf Ronneberger[1*], Russ Bates[1*], Augustin Žídek[1*], Alex Bridgland[1*], Clemens Meyer[1*], Simon A A Kohl[1*], Anna Potapenko[1*], Andrew J Ballard[1*], Andrew Cowie[1*], Bernardino Romera-Paredes[1*], Stanislav Nikolov[1*], Rishub Jain[1*], Jonas Adler[1], Trevor Back[1], Stig Petersen[1], David Reiman[1], Martin Steinegger[2], Michalina Pacholska[1], David Silver[1], Oriol Vinyals[1], Andrew W Senior[1], Koray Kavukcuoglu[1], Pushmeet Kohli[1], Demis Hassabis[1*†]

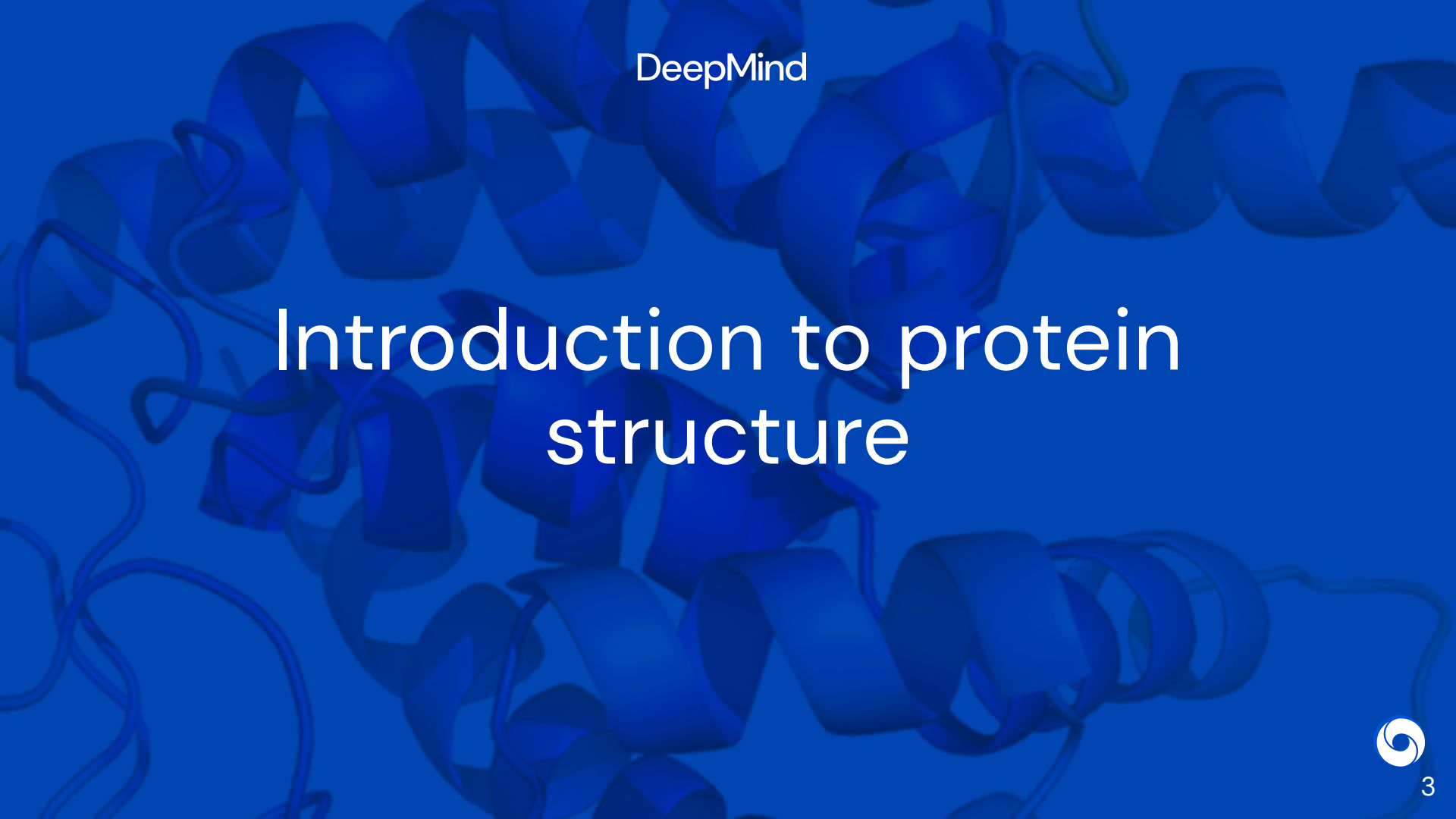
[1]DeepMind, London, UK, [2]Seoul National University, South Korea
* Equal contribution
† Corresponding authors: John Jumper (jumper@deepmind.com), Demis Hassabis (dhcontact@deepmind.com)

Outline

1. Introduction to protein structure
2. How AlphaFold works
3. How AlphaFold understands proteins
4. Results
5. What's next?

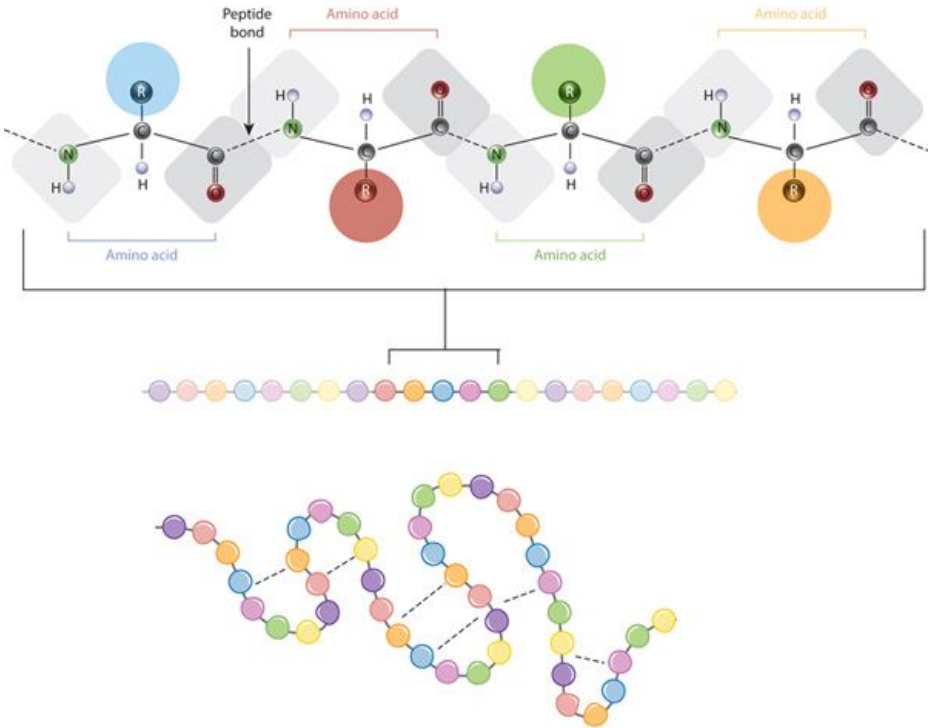
DeepMind

Introduction to protein structure



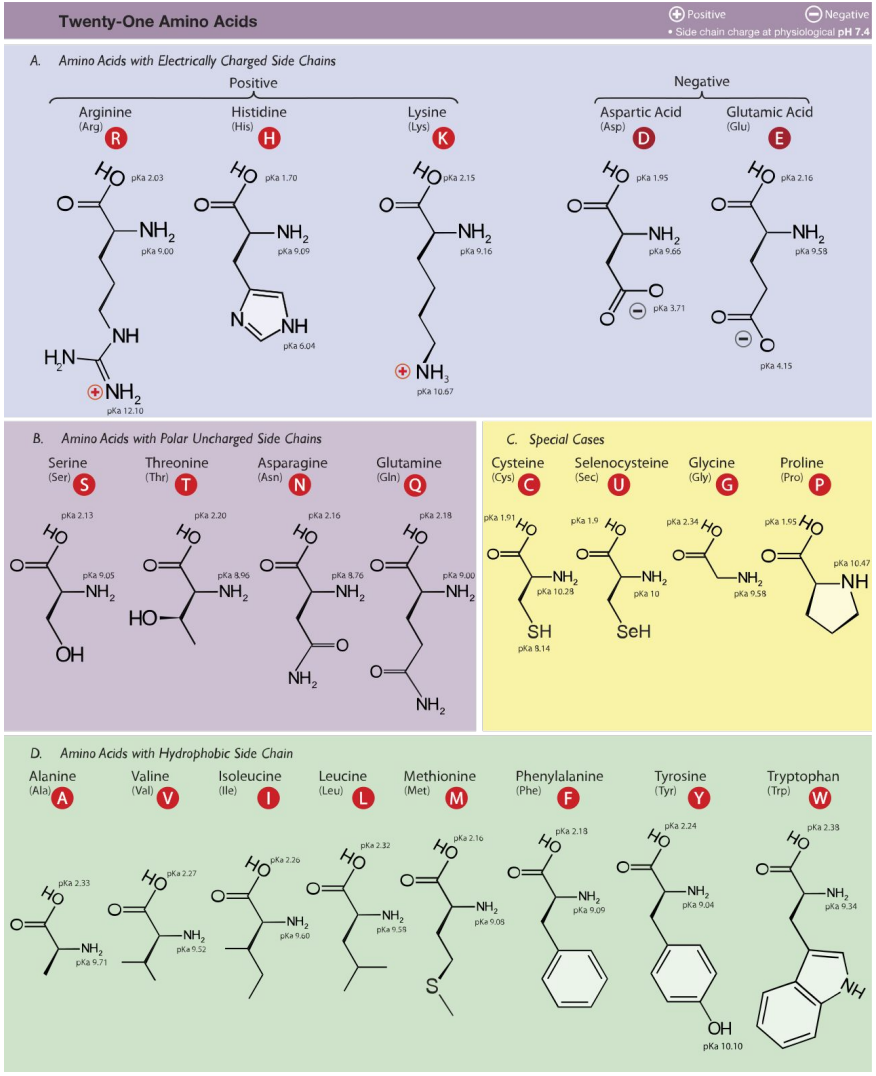
What is a protein?

- A long molecule (chain) of amino acids
- Computer scientist view: Strings over an alphabet of 20 letters (called “residues” or “amino acids”)
- Each letter designates a specific collection of atoms and their associated bonds
- DNA sequences directly encode the protein sequence



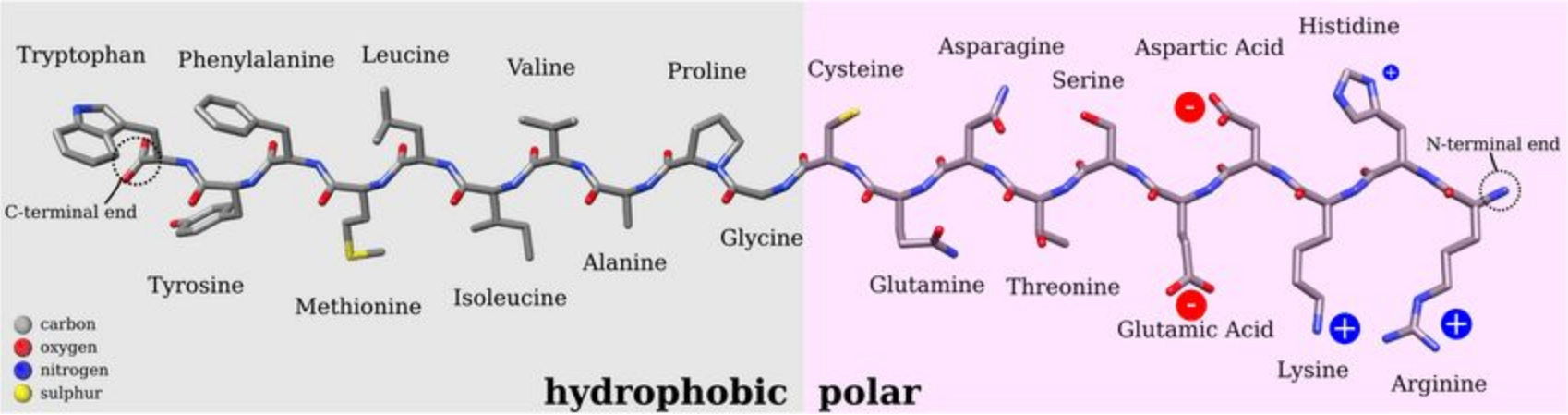
https://www.nature.com/scitable/content/ne0000/ne0000/ne0000/ne0000/14711381/U2CP4-1_PeptideBondSidechain_ksm.jpg

Amino acids



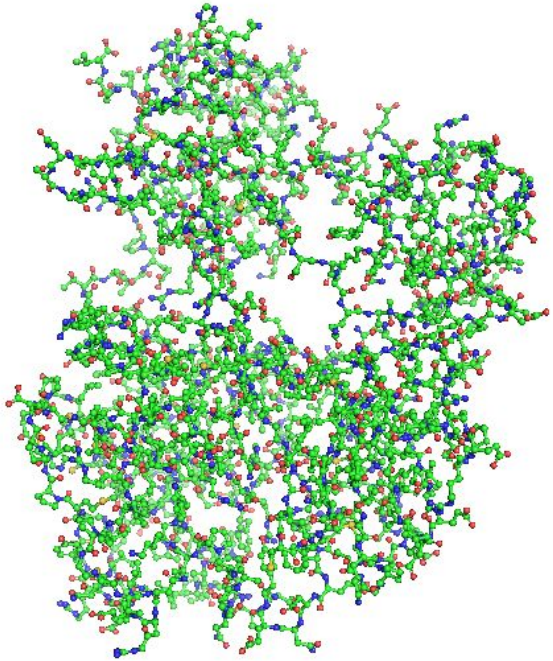
https://en.wikipedia.org/wiki/Proteinogenic_amino_acid#/media/File:Molecular_structures_of_the_21_proteinogenic_amino_acids.svg

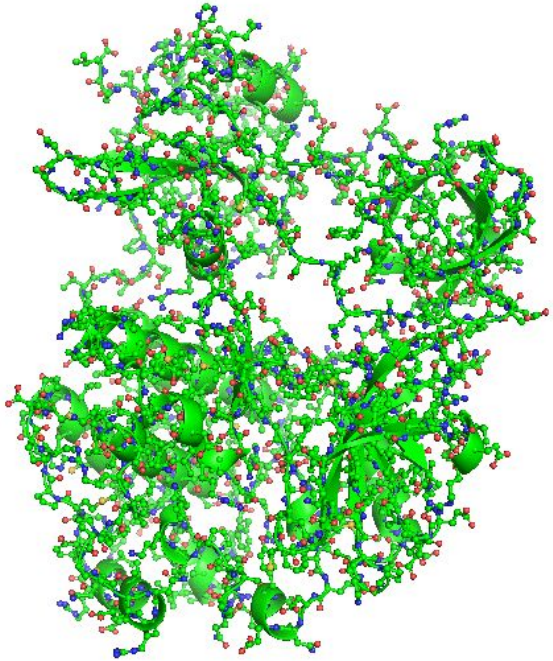
Amino acids



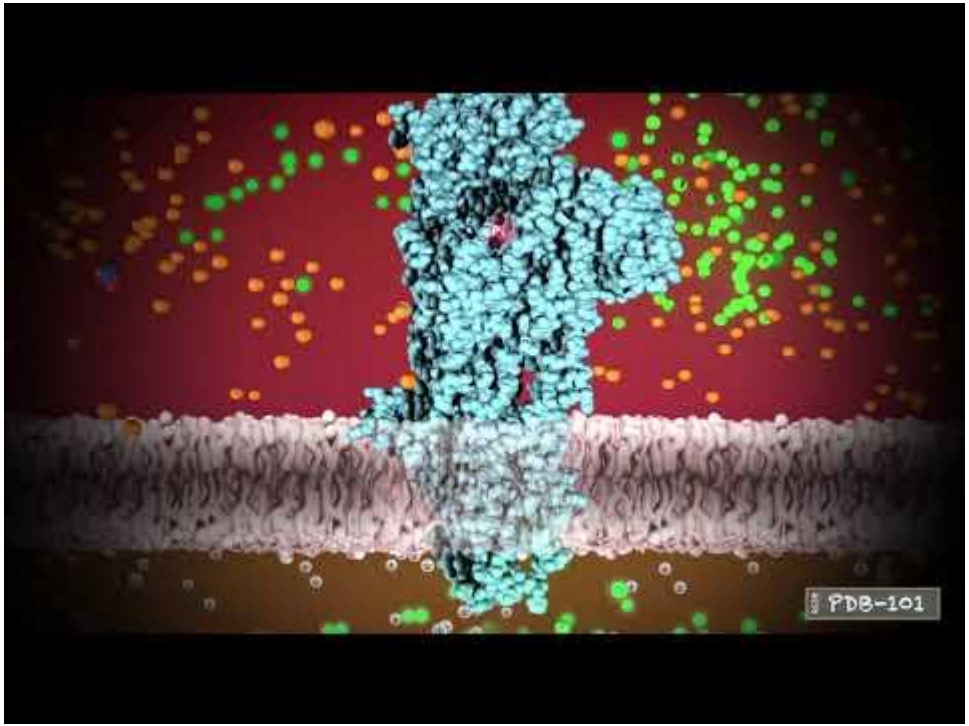
https://www.researchgate.net/profile/Tobias_Sikosek/publication/264933592/figure/fig1/AS:668940372045831@1536499221408/The-20-different-amino-acids-occuring-in-proteins-The-image-shows-a-protein-chain-in.ppm

Protein structure (3-D coordinates)





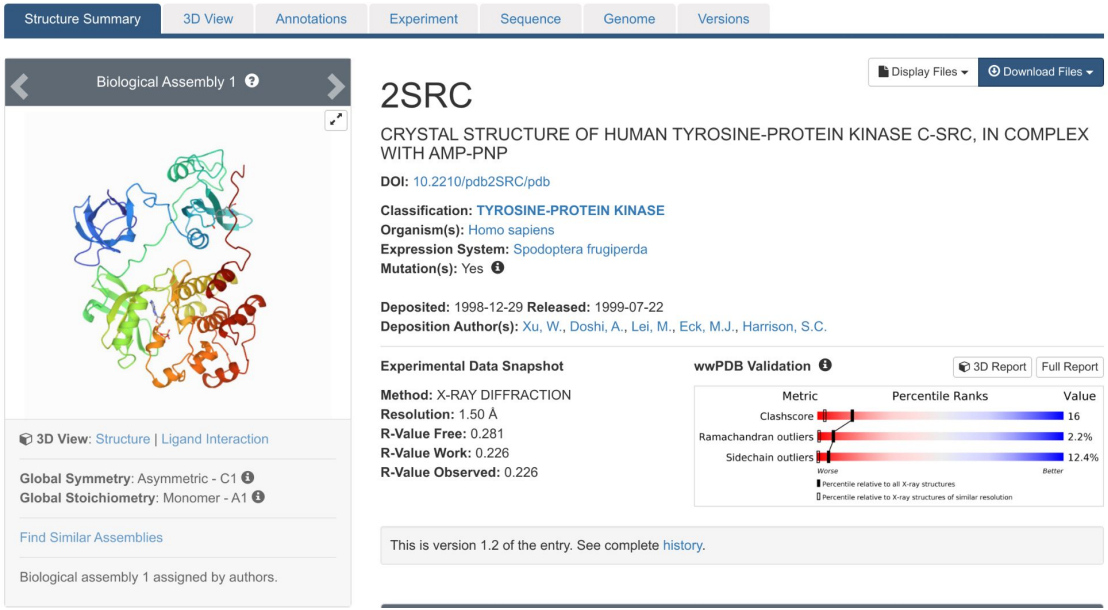
Proteins are the machines of the cell



Source: PDB-101

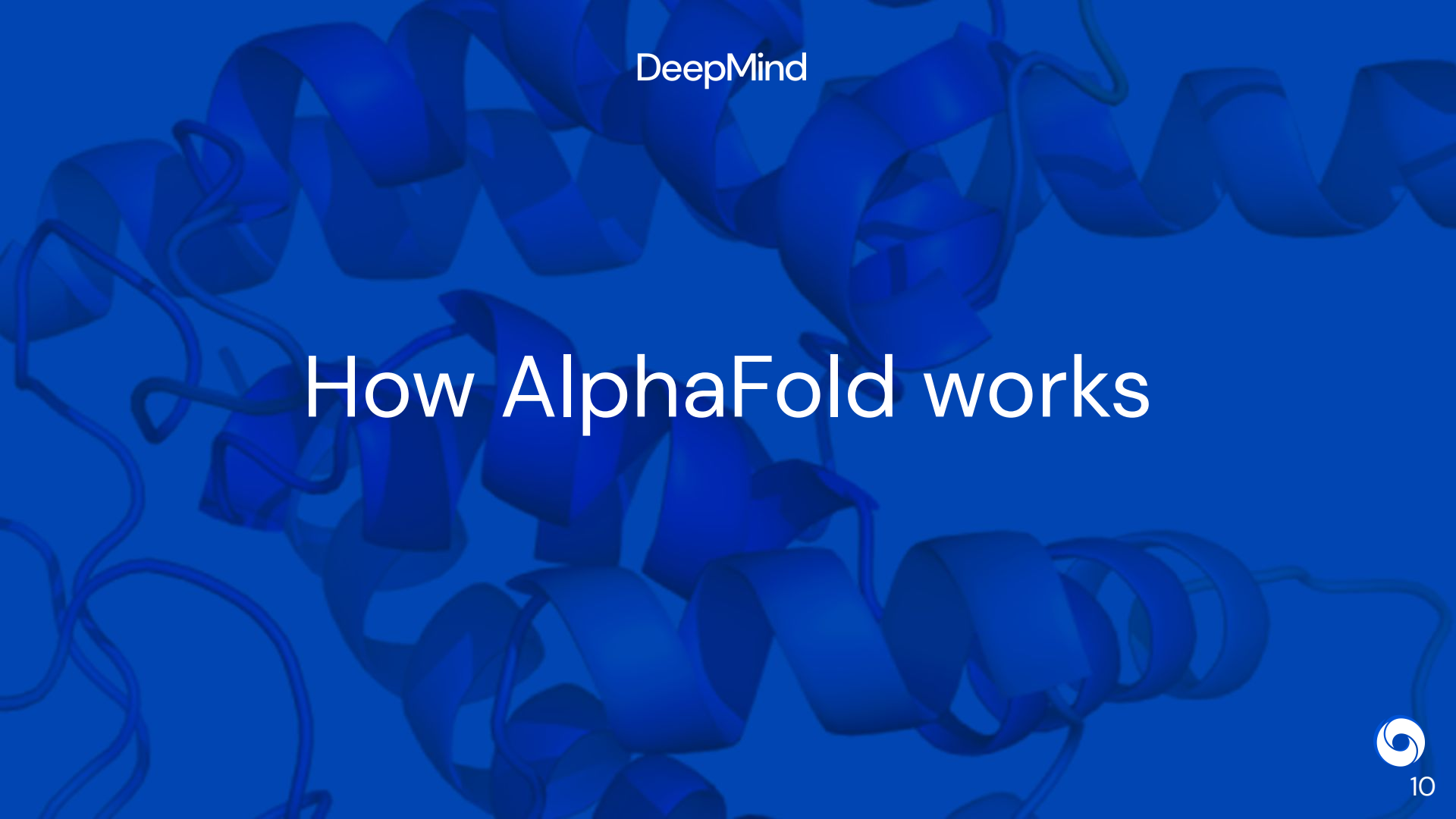
Experimental protein structure determination

- Hard to determine experimentally – we know only **~200k structures** and a single structure can take years for a PhD student
- All published structures are contributed to the Protein Data Bank



DeepMind

How AlphaFold works



AlphaFold methods paper

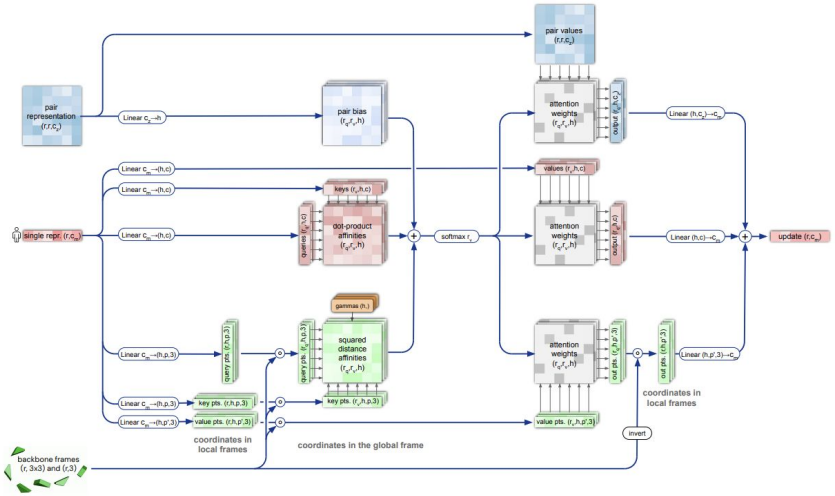
Highly accurate protein structure prediction with AlphaFold (Jumper et al., 2021, Nature)

Comprehensive description of AlphaFold

- 60 pages of supplementary materials
- 32 algorithms

Alongside it we **open sourced the code**, which enables the community to build on our work

- Code: github.com/deepmind/alphafold
- Colab: dpmd.ai/alphafold-colab



Algorithm 31 Generic recycling training procedure

def RecyclingTraining($\{\text{inputs}_c\}$, N_{cycle}) :

$N' = \text{uniform}(1, N_{\text{cycle}})$ *# shared value across the batch*

outputs $= \mathbf{0}$

 # Recycling iterations

for all $c \in [1, \ldots, N']$ **do** *# shared weights*

  outputs $\leftarrow$ stopgrad(outputs) *# no gradients between iterations*

  outputs $\leftarrow$ Model(inputs_c, outputs)

end for

return loss(outputs) *# only the final iteration's outputs are used*

Determining Structure from Evolution - Intuition

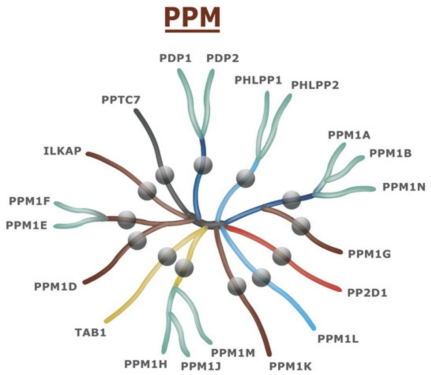
Evolutionary
Structure Prediction
[Bayesian]

Sequence

Physics
[causal]

Structure

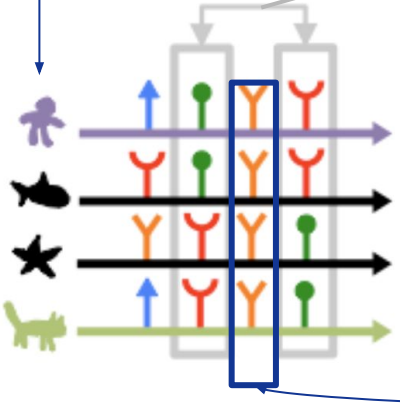
Function
[causal]



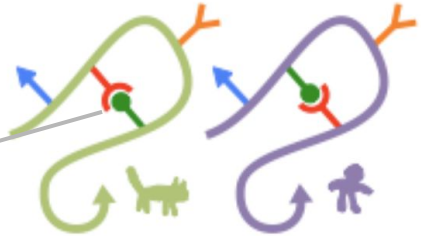
Evolutionary History

Determining Structure from Evolution - Intuition

Search genetics database



Multiple Sequence Alignment (MSA)



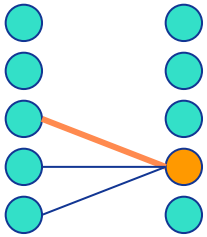
Co-evolution: residues in contact tend to mutate together.

(mutation of a single residue breaks the contact and the organism with the mutated protein does not survive)

Evolution conserves some properties like hydrophobic/hydrophilic amino acids on the ”inside”/“outside”

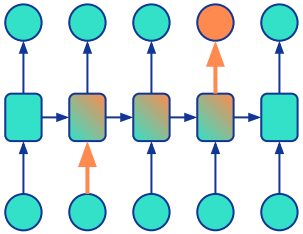
Coevolution cartoon by Sergey Ovchinnikov (https://jgi.doe.gov/seeking-structure-metagenome-sequences/cartoon-coevolution-sergey-o/)

Inductive Bias for Deep Learning Models



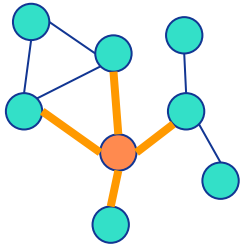
Convolutional Networks (e.g. computer vision)

- data in regular grid
- information flow to local neighbours
- AlphaFold 1



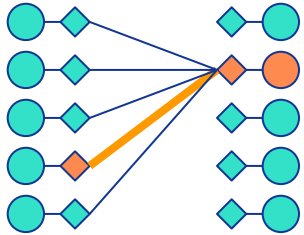
Recurrent Networks (e.g. language)

- data in ordered sequence
- information flow sequentially



Graph Networks (e.g. recommender systems or molecules)

- data in fixed graph structure
- information flow along fixed edges

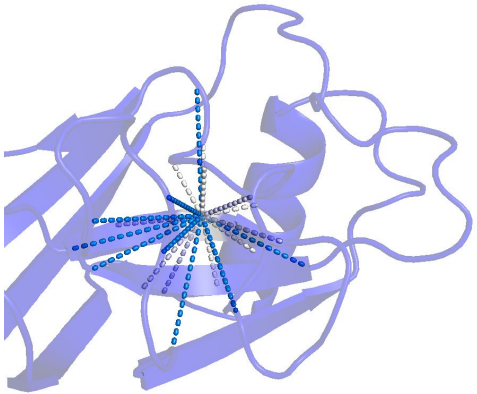


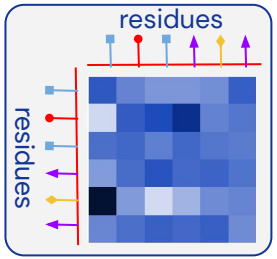
Attention Module (e.g. language)

- data in unordered set
- information flow dynamically controlled by the network (via keys and queries)

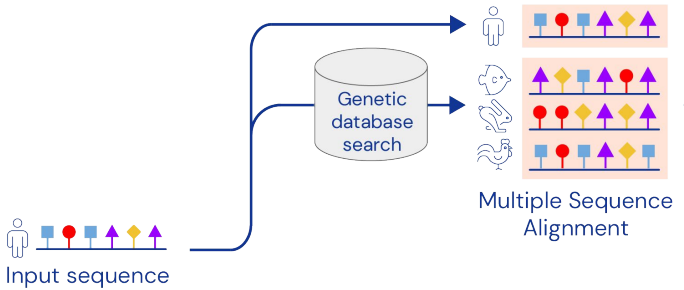
Putting our protein knowledge into the model

- Physical and geometric insights are built into the network structure, not just a process around it
- End-to-end system directly producing a structure instead of inter-residue distances
- Inductive biases reflect our knowledge of protein physics and geometry
  - The positions of residues in the sequence are de-emphasized
  - Instead residues that are close in the folded protein need to communicate
  - The network iteratively learns a graph of which residues are close, while reasoning over this implicit graph as it is being built



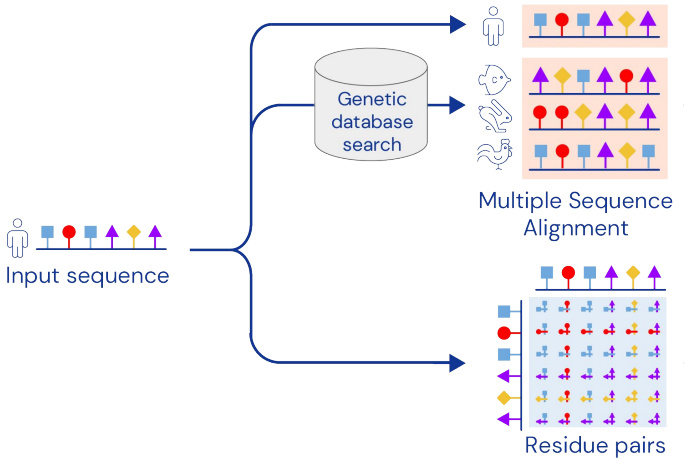


Inputs



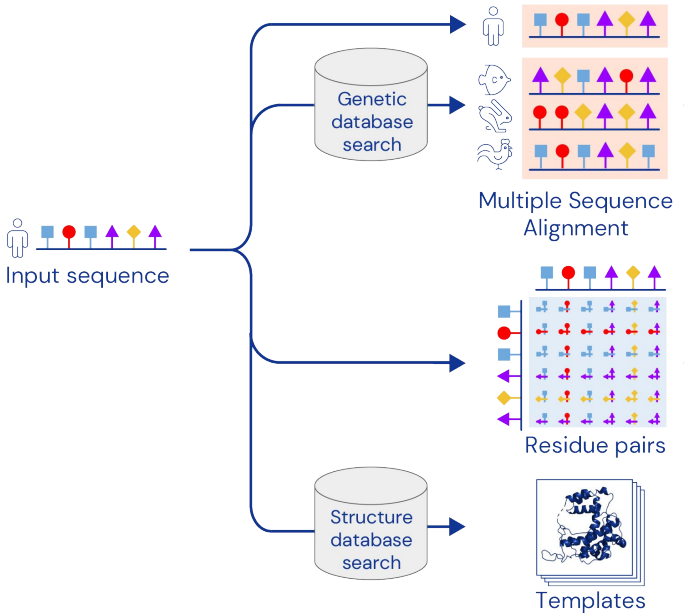
A key AlphaFold input is the MSA, containing sequences evolutionarily related to the target. Related sequences are found using standard tools and public databases.

Inputs



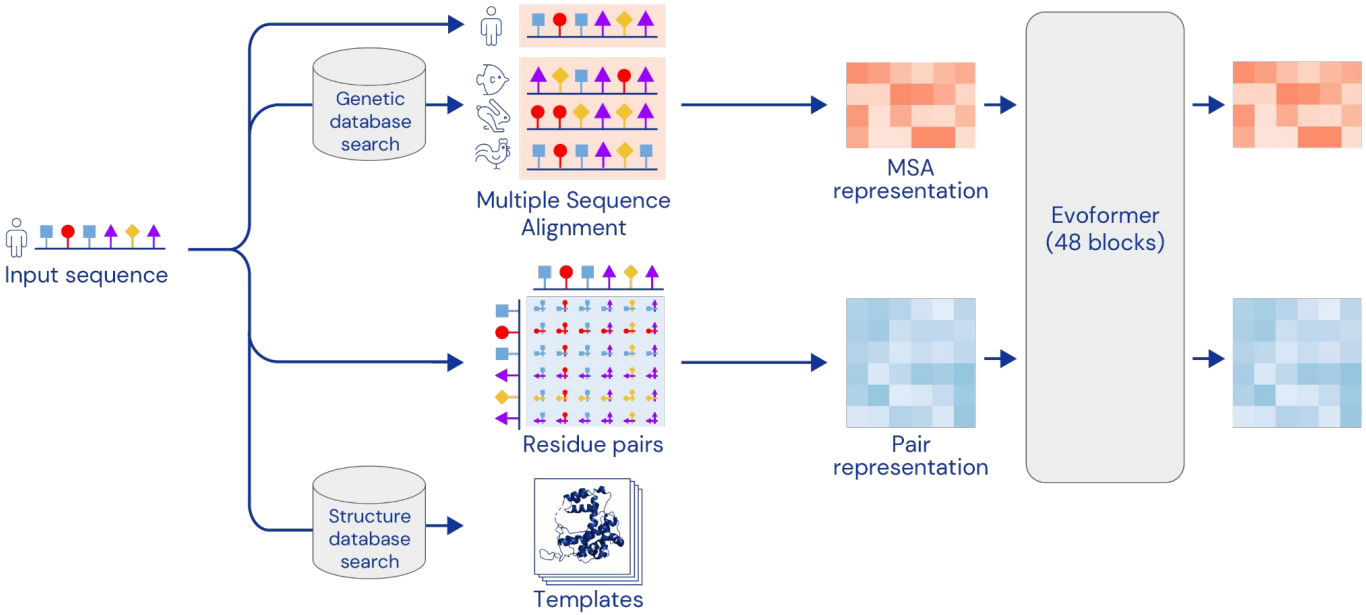
The input sequence is used to create an array of representations representing all residue pairs.

Inputs



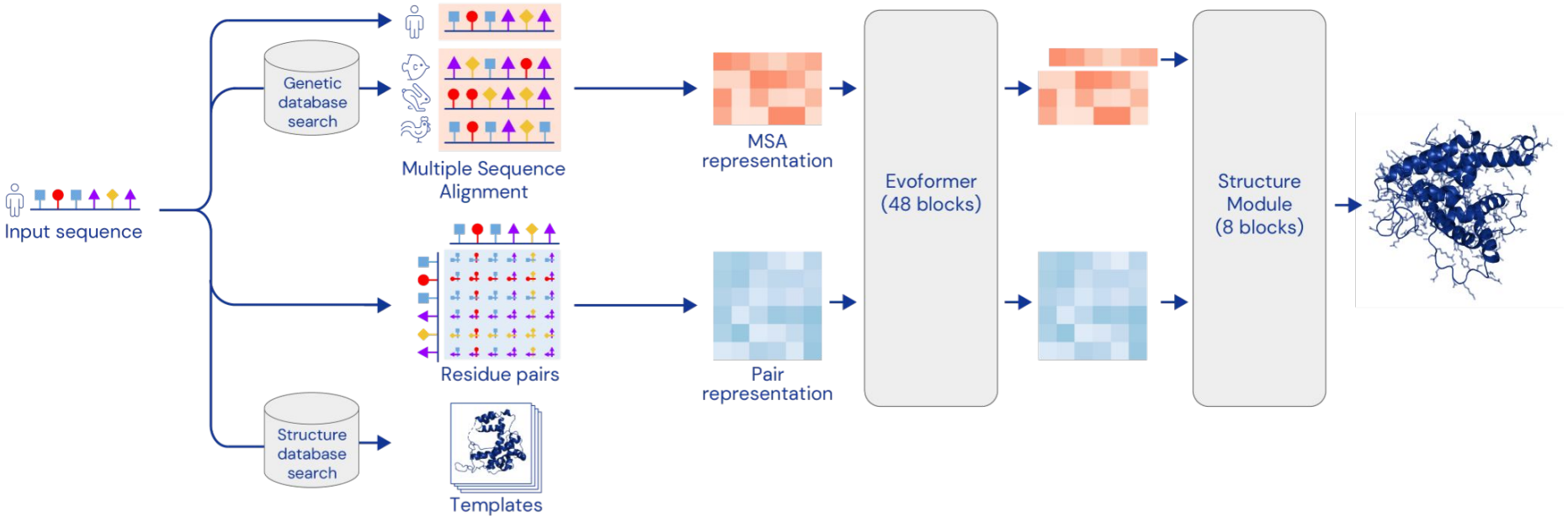
AlphaFold can also use template structures from the PDB, found using standard tools. However, it often produces accurate predictions without a template.

Network



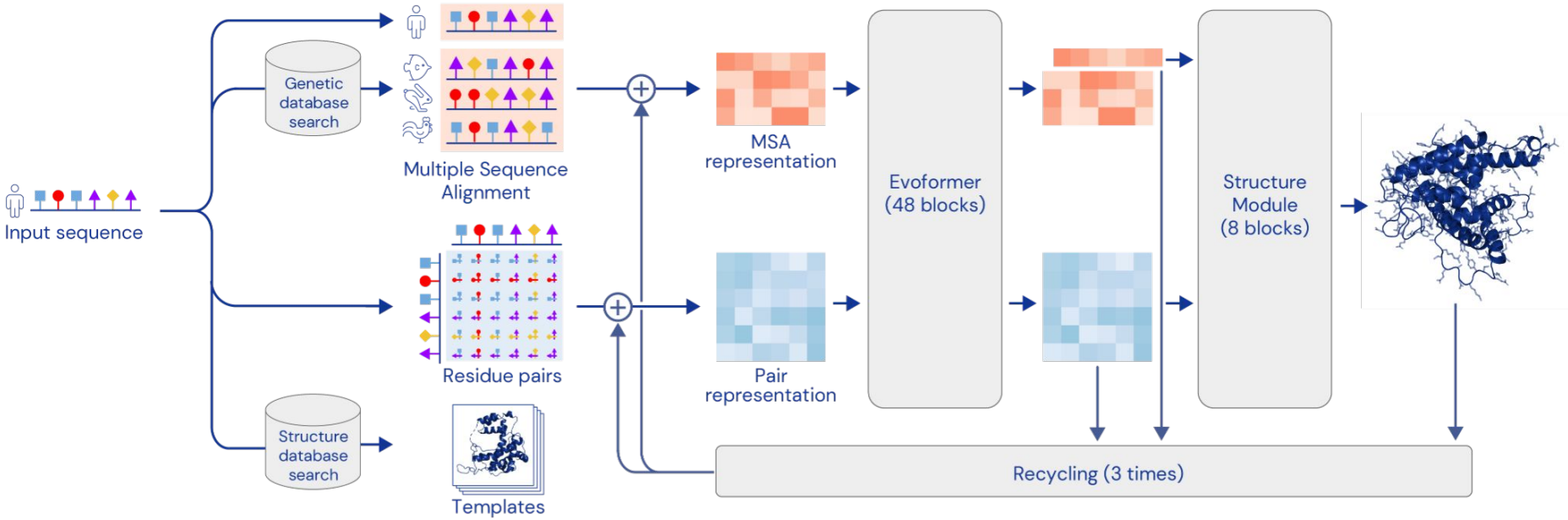
The Evoformer blocks extract information about the relationship between residues.
The MSA representation can update the pair representation and vice versa.

Network



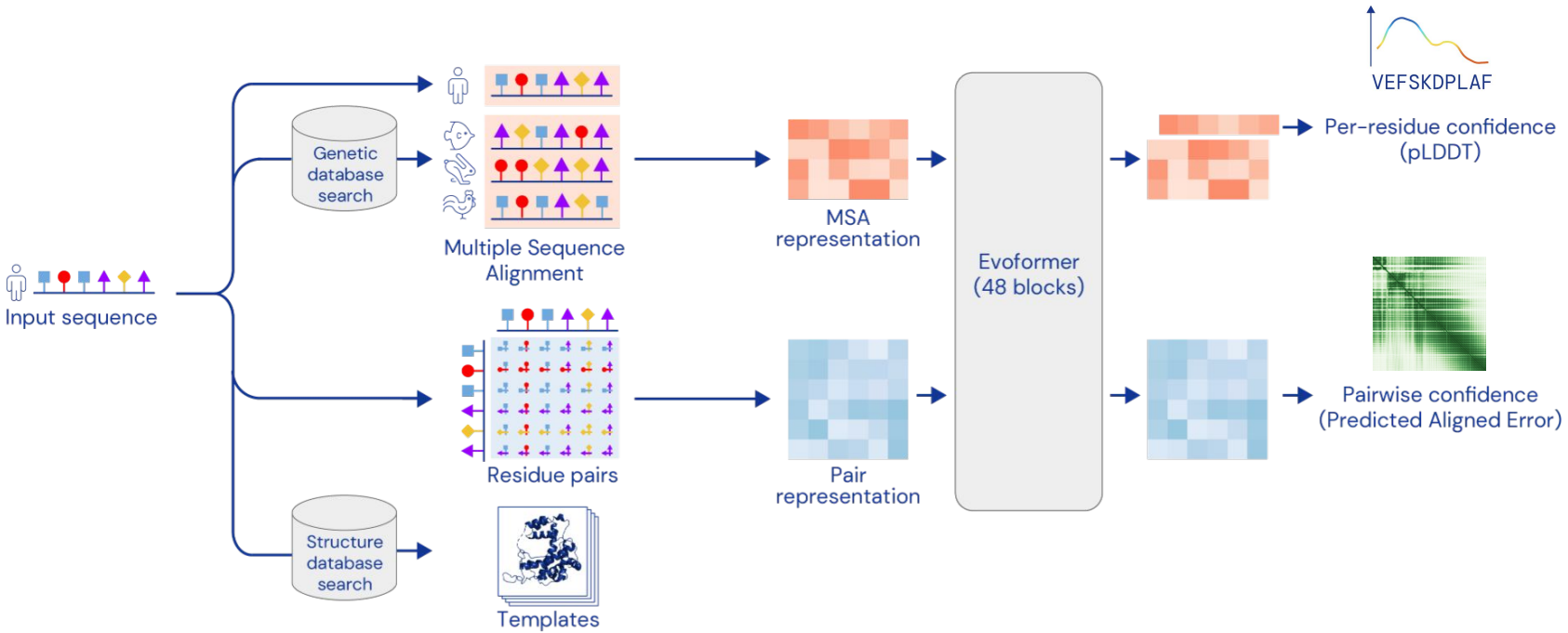
The Structure Module predicts a rotation + translation to place each residue.
A small network predicts side chain chi angles.

Network



Feeding certain outputs back through the network again improves performance

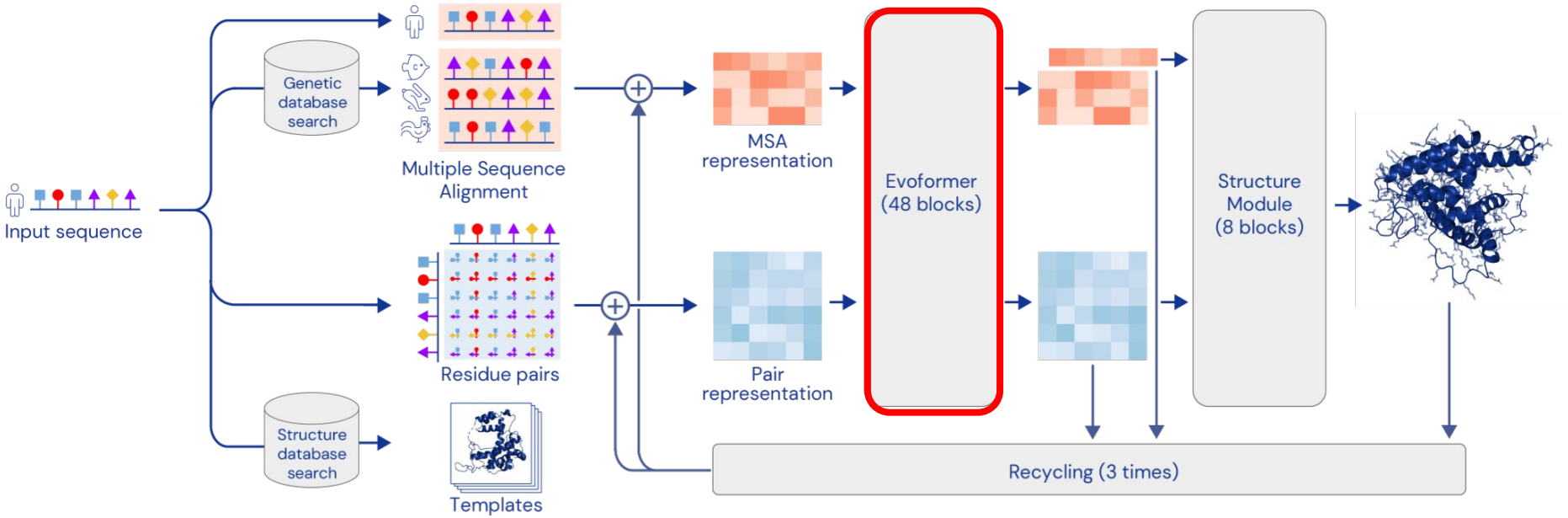
Confidence estimates



AlphaFold is trained to predict its confidence via simple supervision on the training set

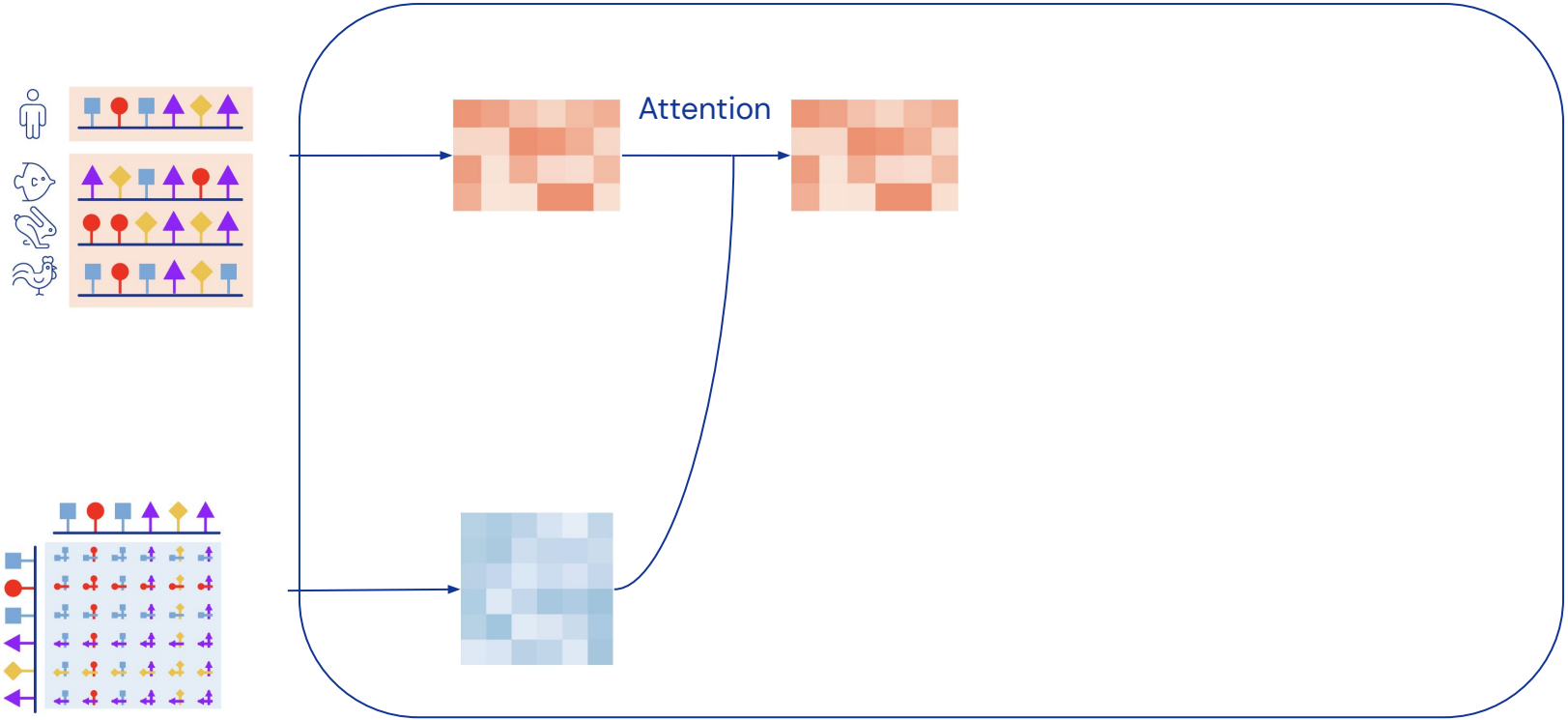
We use these confidences for ranking of predictions

Network

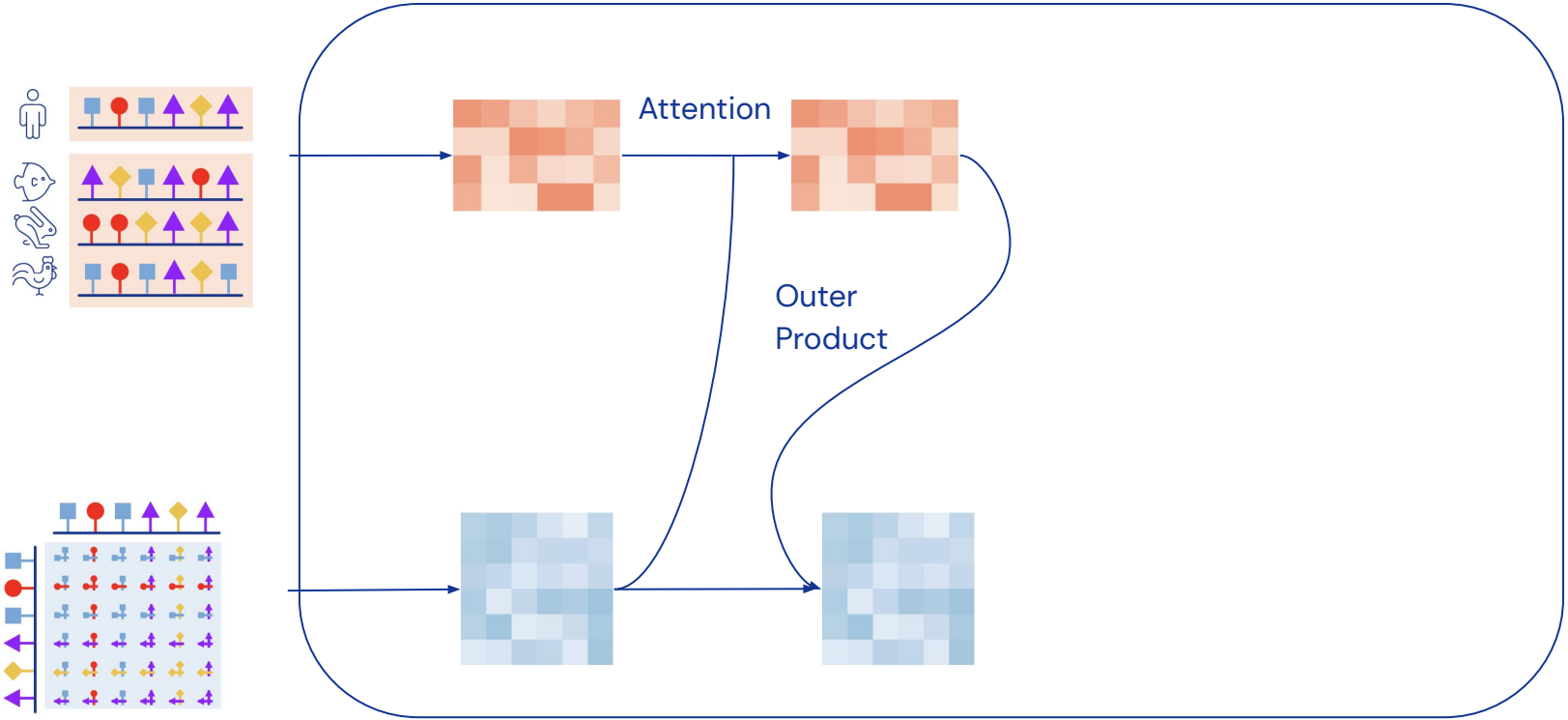

Input sequence
Genetic database search
Multiple Sequence Alignment
Residue pairs
Structure database search
Templates
MSA representation
Pair representation
Evoformer (48 blocks)
Structure Module (8 blocks)
Recycling (3 times)

Evoformer

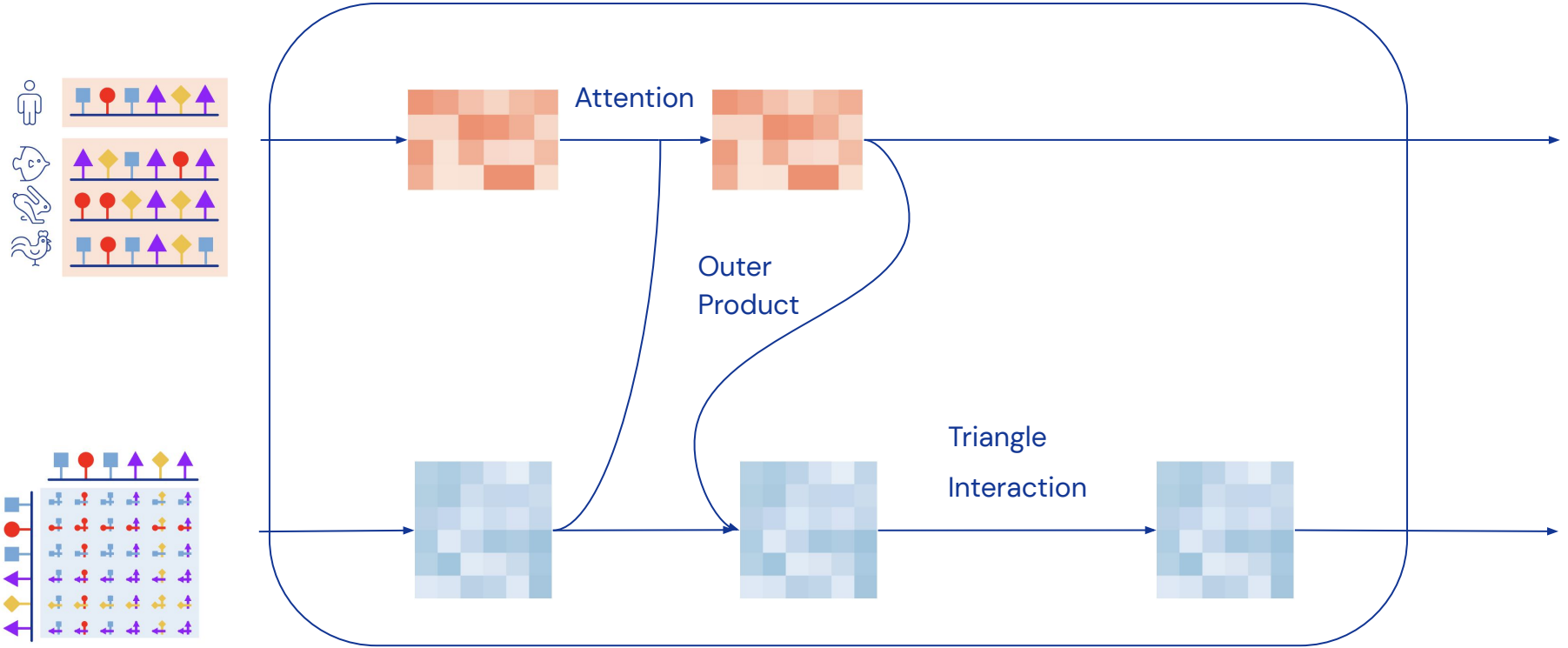
Evoformer



Evoformer

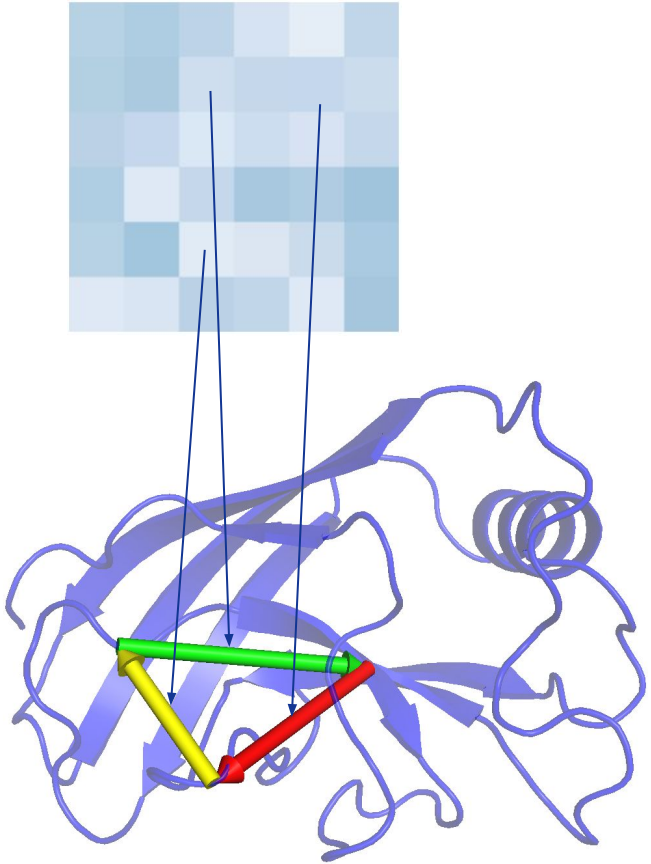


Evoformer



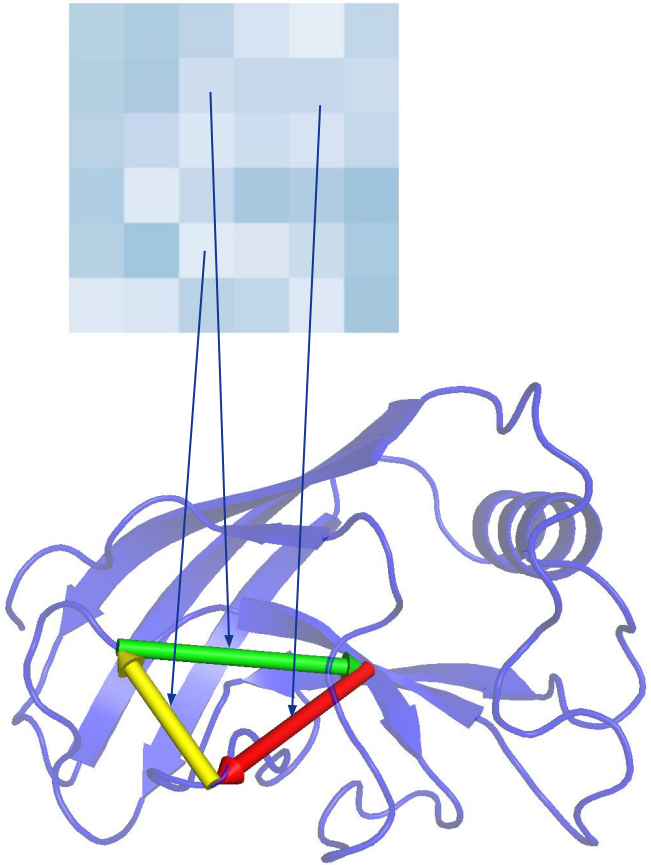
Triangle Interaction

- **Take 3 points A, B, C**
  - If Distance AB and distance BC known strong constraint on AC (triangle inequality)
  - Evolution & Sequence gives information about relations between residues
- **Pair Embedding encodes relations**
  - Update for pair AC should depend on BC, AB

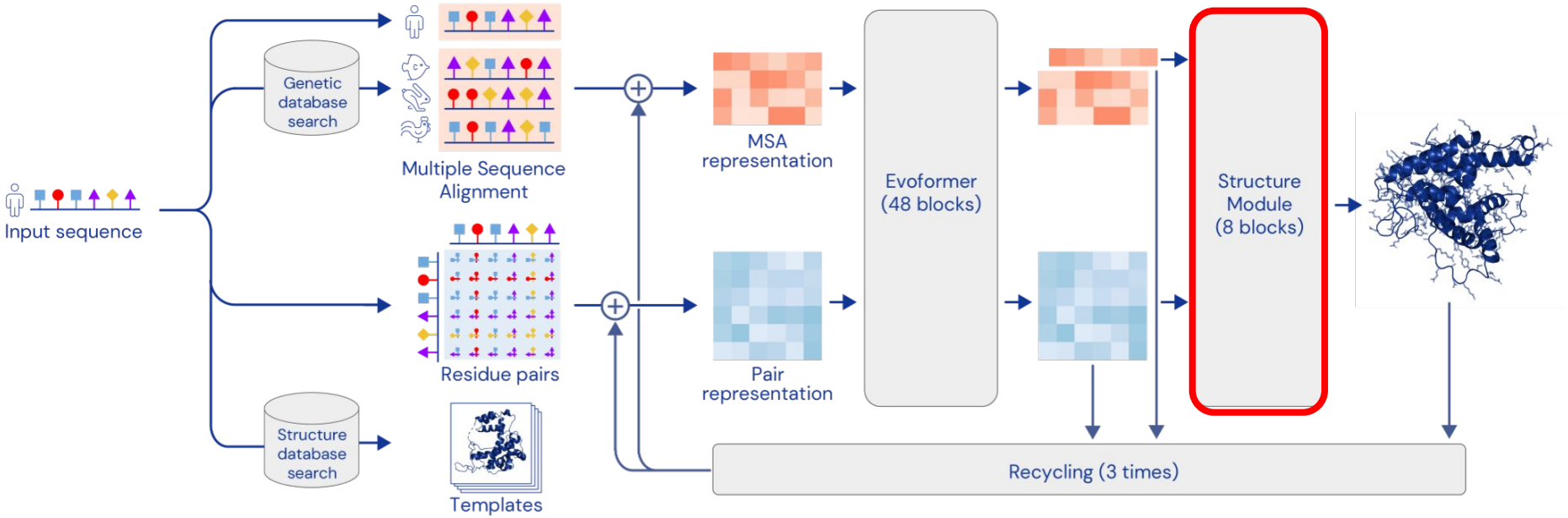


Graph Inference

- **Graph**
  - Edges = Pairs of residues
  - We don't know the graph, need to infer it
  - Triplet relation in this language:
    length 3 cycles in graph
  - Update  based on all cycles involving edge
- **Transitivity inductive bias**
  - This encodes transitivity bias of relations
    (E.g. triangle inequality, Loop Closure)



Network


Input sequence
Genetic database search
Multiple Sequence Alignment
Residue pairs
Structure database search
Templates
MSA representation
Pair representation
Evoformer (48 blocks)
Structure Module (8 blocks)
Recycling (3 times)

Structure module

- **End-to-end folding** instead of gradient descent
- Protein backbone = gas of 3-D rigid bodies (chain is learned!)

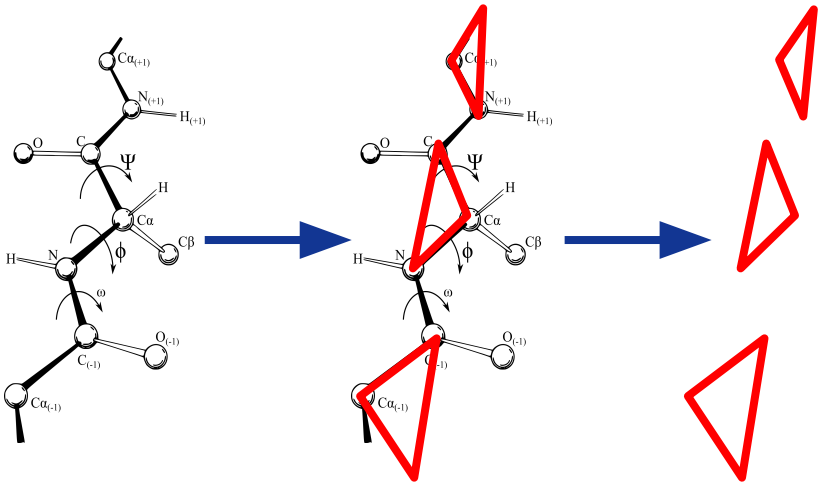
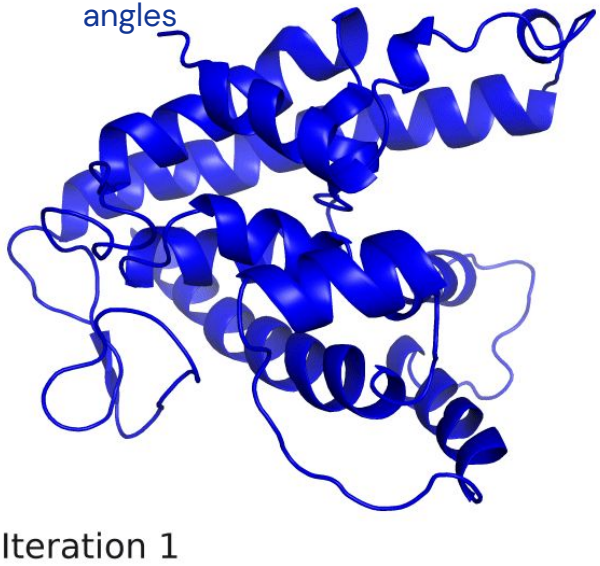


Image: Dcrjsr, vectorised Adam Rędzikowski (CC BY 3.0, Wikipedia)

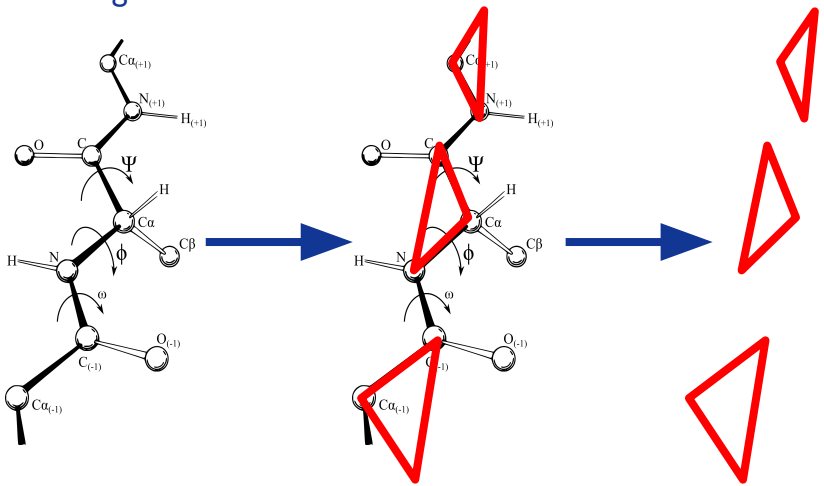
- **3-D equivariant transformer architecture** updates the rigid bodies / backbone (Invariant Point Attention)
  - Also builds the side chains from torsion angles



Target: T1041

Local Frames

- Protein backbone = gas of 3-D rigid bodies (chain is learned!)
- Rigid body orientation defines local coordinate frame along chain



- Coordinates measured in local Frame are **invariant to choice of global frame**

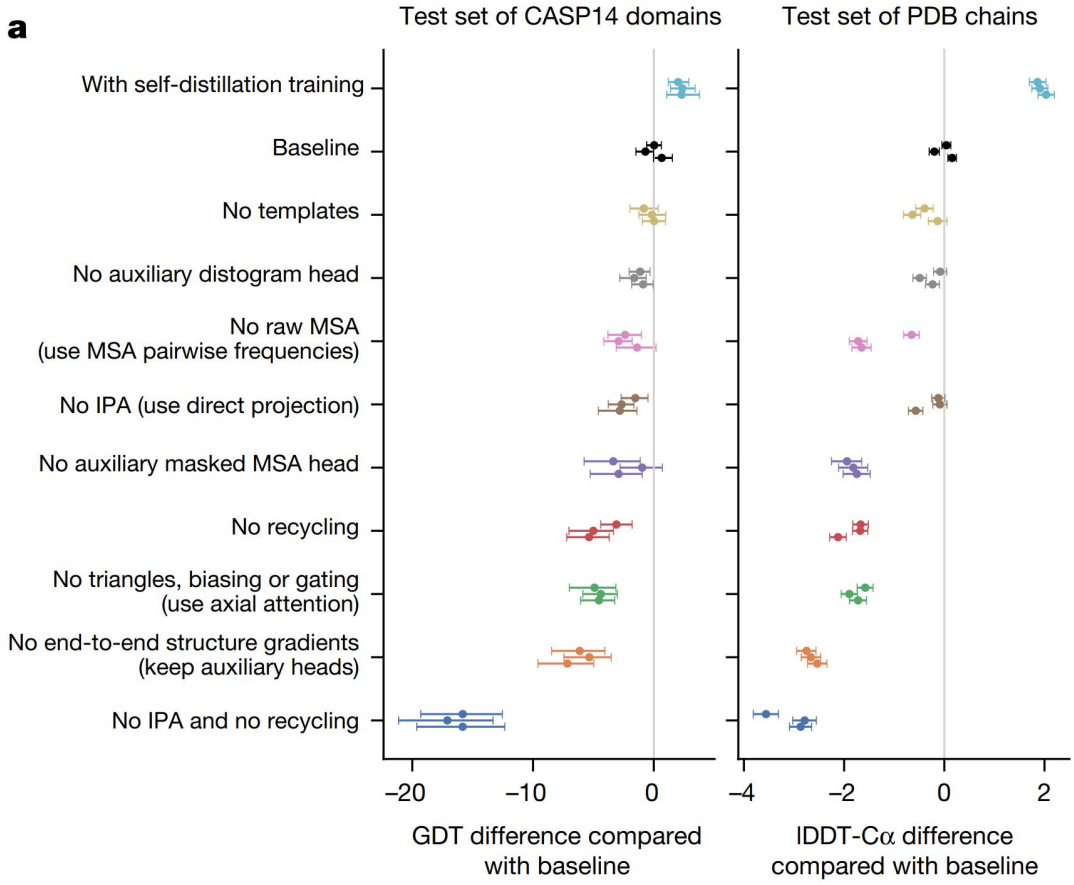
Image: Dcrjsr, vectorised Adam Rędzikowski (CC BY 3.0, Wikipedia)

Which parts mattered? All of it

No single improvement is dominant

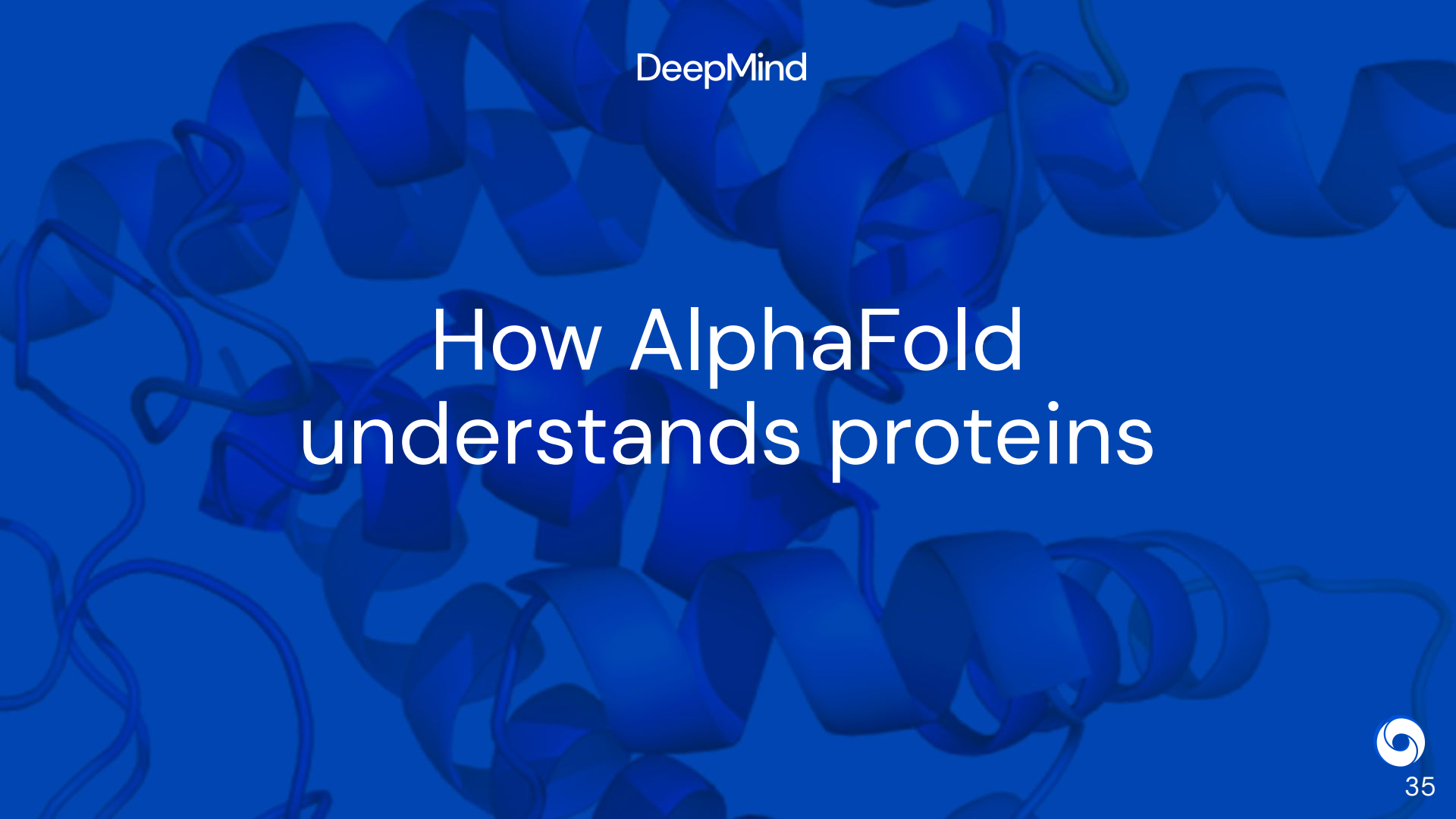
More important was the methodology of building the protein intuition into the model

Multiple ablations suggest strong interactions between many of the components



DeepMind

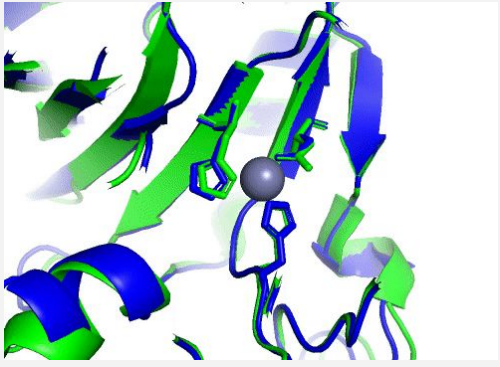
How AlphaFold understands proteins



Biological context

- Computational structure prediction is typically underspecified
  - Oligomeric state, ligands, DNA-binding, experimental conditions, multiple conformations etc.
- Our network is tolerant to missing context
- AlphaFold is just as good at membrane proteins or novel folds as more typical PDB structures

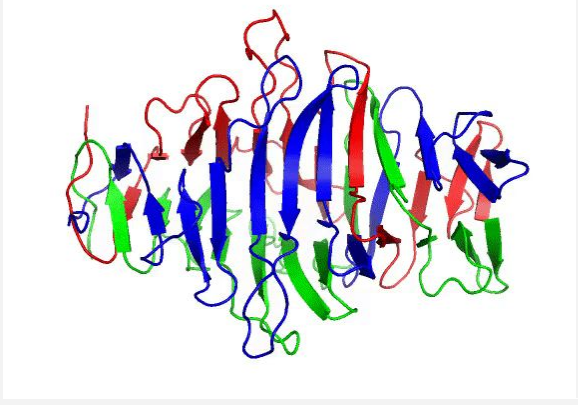
T1056 (zinc binding)



AlphaFold / Experiment

TBM-hard, 98.2 GDT

T1080 (trimer)

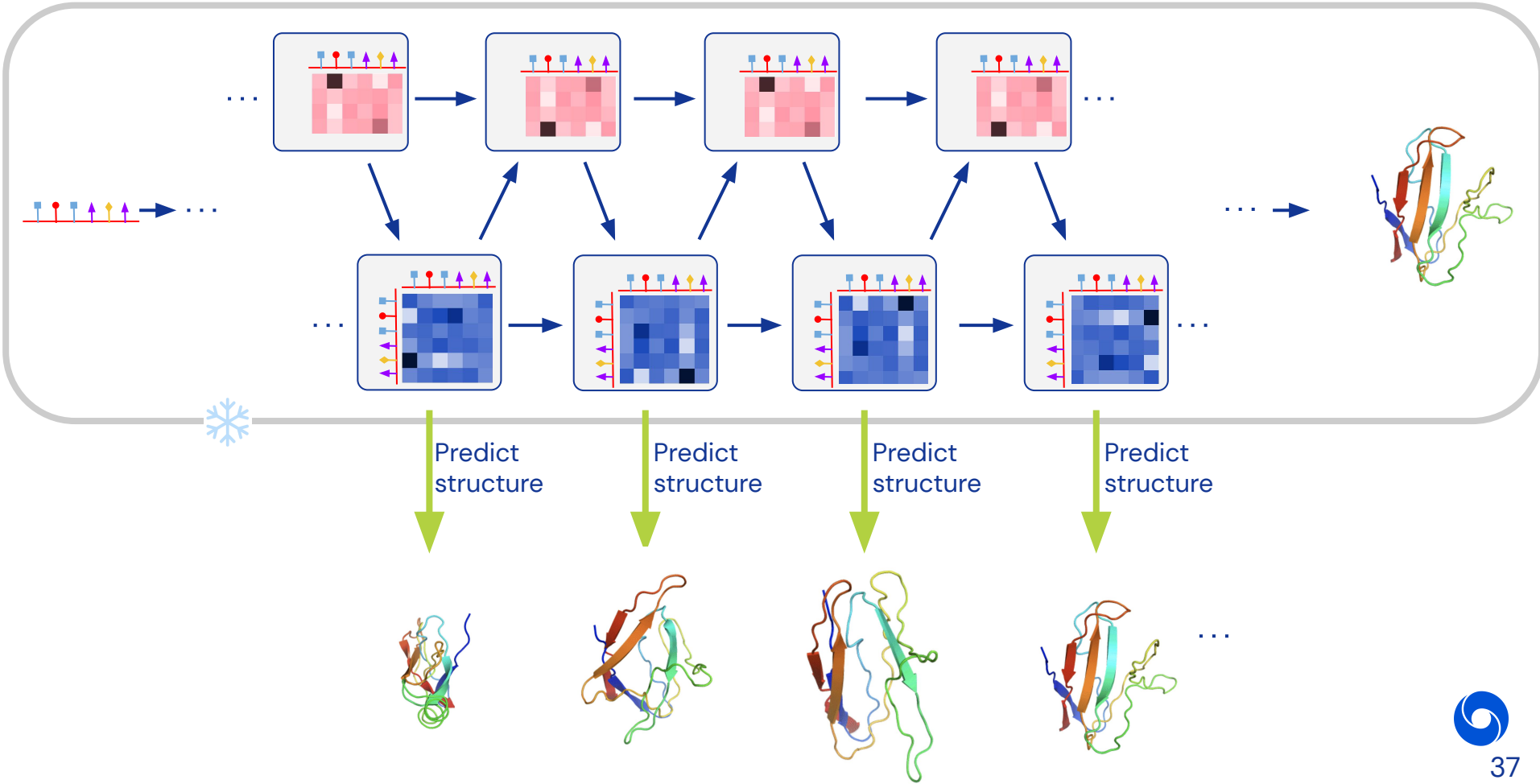


AlphaFold (monomer prediction x3)

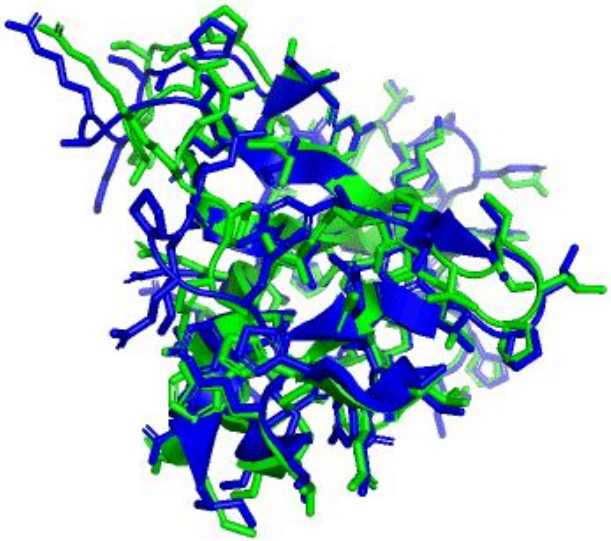
FM/TBM, 85.9 GDT

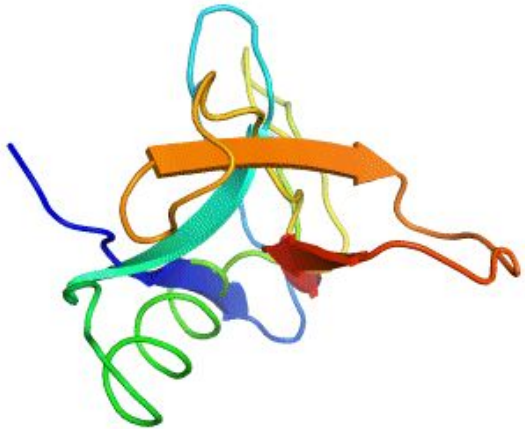
Experimental structure

Interrogating the Network



Model interpretability - ORF8 - Sars-Cov2

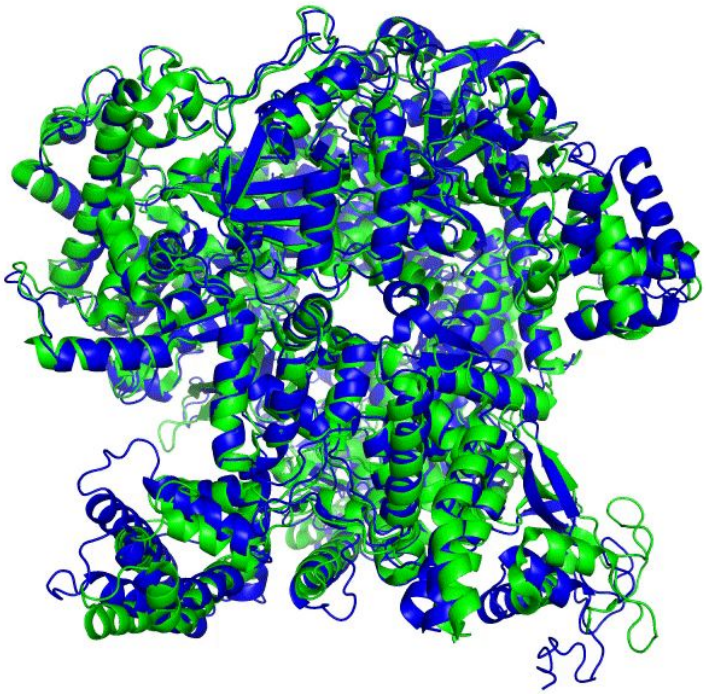




1

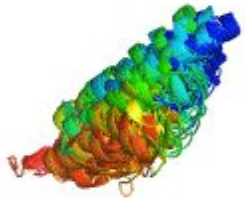
7JTL: Flower, T.G., et al. (2020) Structure of SARS-CoV-2 ORF8, a rapidly evolving coronavirus protein implicated in immune evasion. Biorxiv.

Model interpretability - T1044



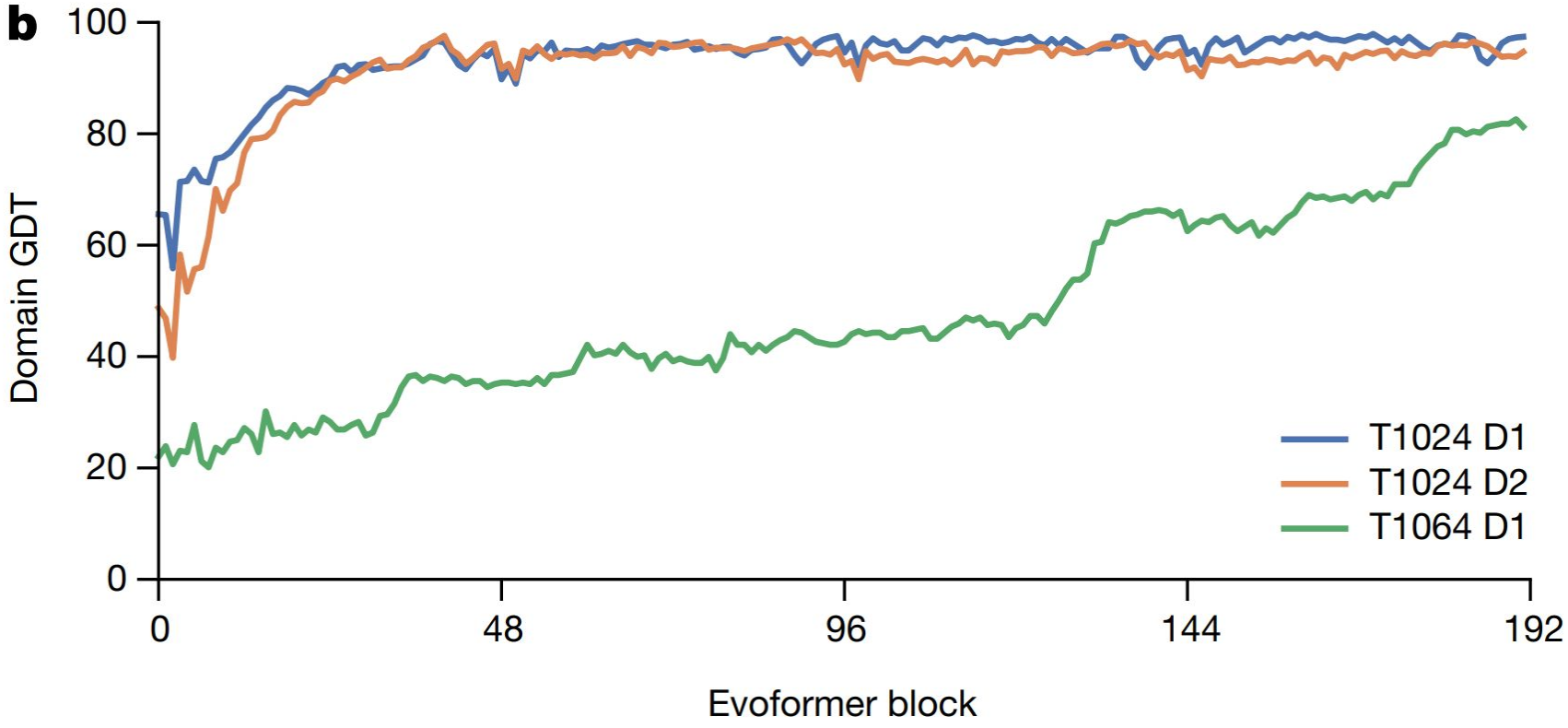
6VR4: Drobysheva, A.V., et al. Structure and function of virion RNA polymerase of a crAss-like phage. Nature (2020). (CASP14 target T1044)

T1044



1

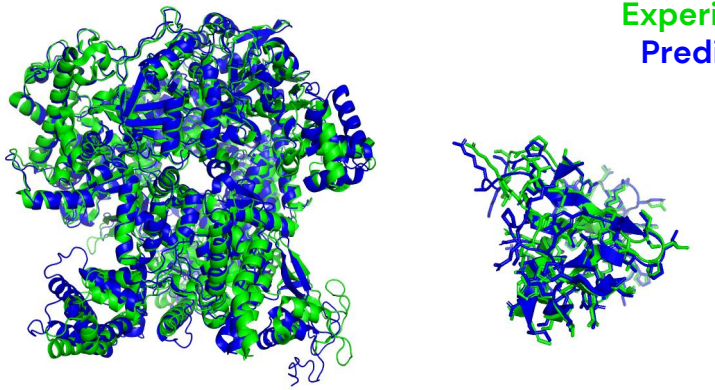
Model interpretability - Role of depth

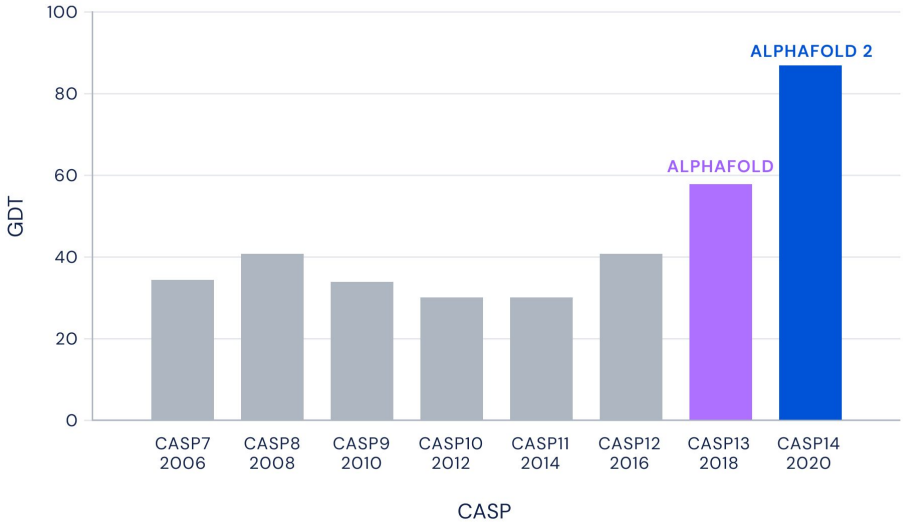


DeepMind

Results

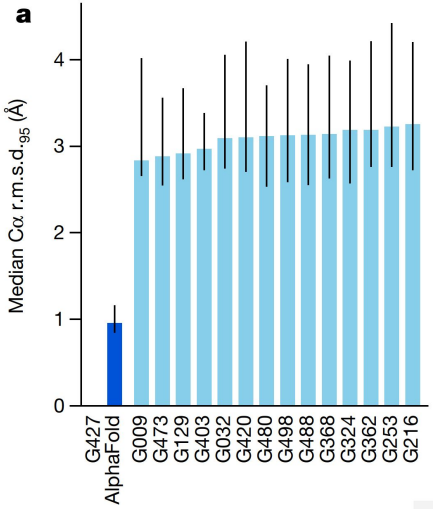
AlphaFold at CASP





CASP has been essential in both showing what works and aligning the community on common definition of the problem

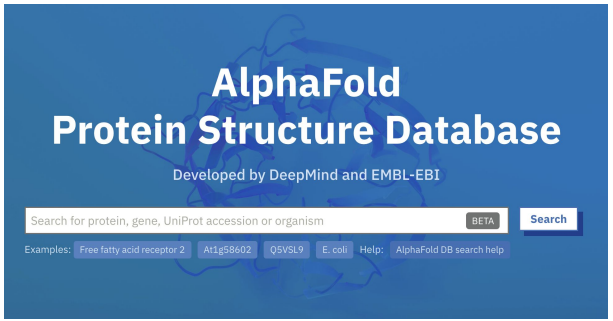
This wouldn't have been possible without the CASP organizers or the cooperation of the experimental community

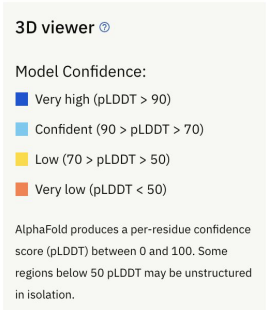


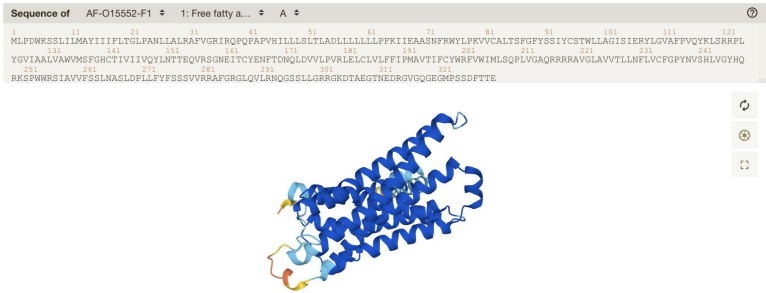
AlphaFold Protein Structure Database

Website developed and hosted by EMBL-EBI: alphafold.ebi.ac.uk

Contains pre-run predictions for **21 model organisms**, with plans to expand to **Uniref90** (~100M structures)



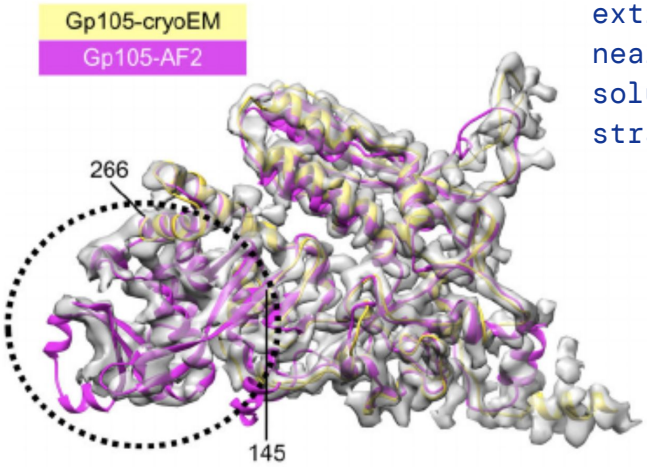




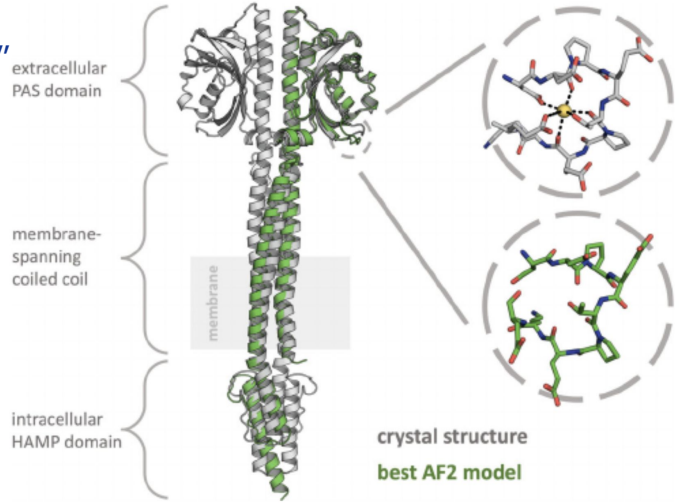
EMBL DeepMind

AlphaFold as an aid to experimental structure determination

"Thus, in this case the availability of high-quality structure predictions has reduced an extremely challenging, near-intractable structure solution into a straightforward task."



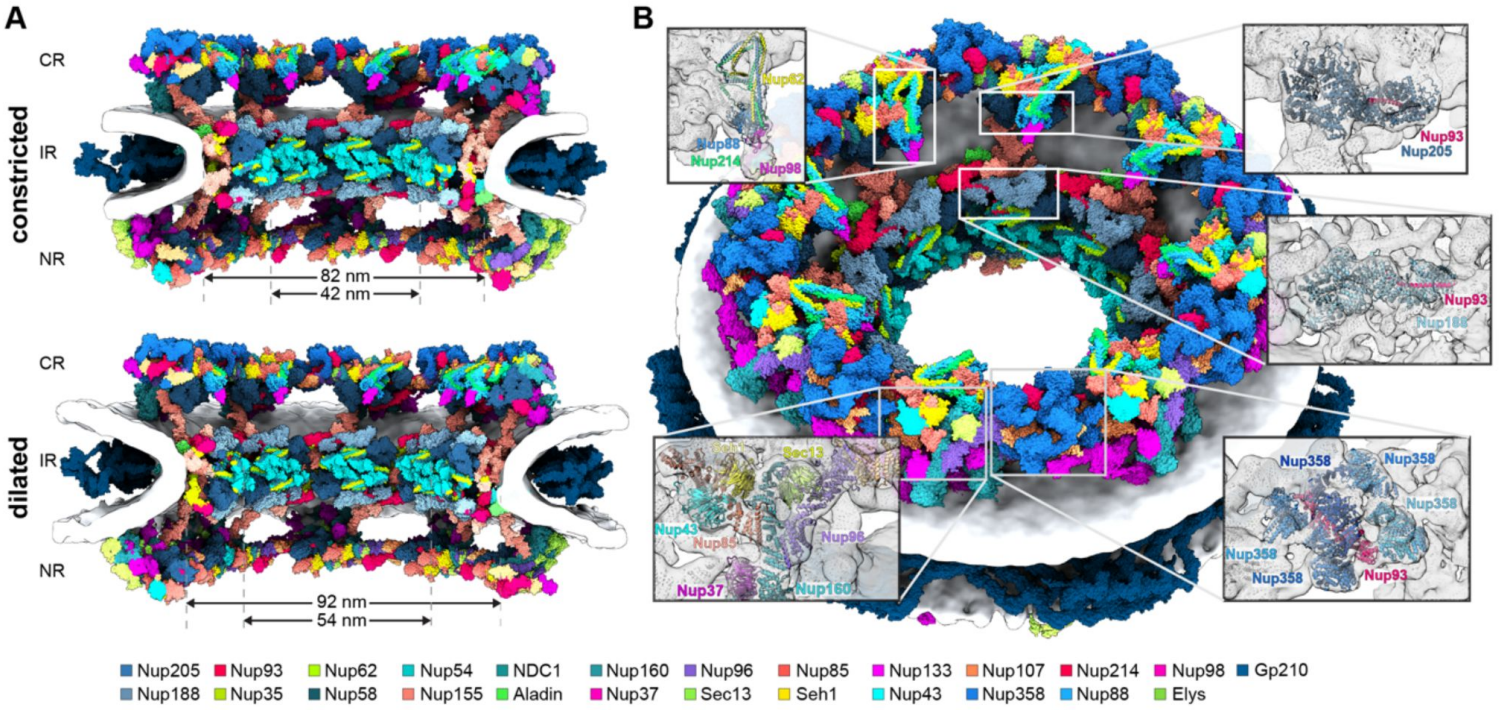
AlphaFold models of AR9 nvRNAP proteins fit the cryo-EM density nearly perfectly



Crystal structure of dimeric Af1503

Figures taken from: Kryshtafovych, Andriy, et al. Computational models in the service of X-ray and cryo-EM structure determination. Proteins: Structure, Function, and Bioinformatics (2021).

Integrative modelling to combine AlphaFold with experimental data for large complexes



Mosalaganti et al. *Artificial intelligence reveals nuclear pore complexity. Biorxiv 2021.*

DeepMind

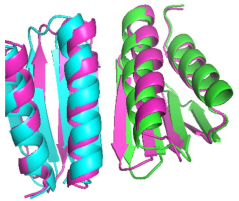
What's next?

Hacking AlphaFold for multimers

Minkyung Baek
@minkbaek

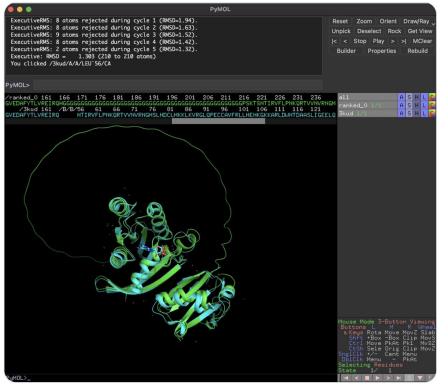
Adding a big enough number for "residue_index" feature is enough to model hetero-complex using AlphaFold (green&cyan: crystal structure / magenta: predicted model w/ residue_index modification).
#AlphaFold #alphafold2

```
to residue index
residue_index']

 in each chain

 += 200

dex'] = idx_res
```



Yoshitaka Moriwaki
@Ag_smith

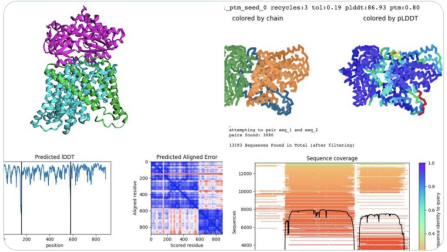
AlphaFold2 can also predict heterocomplexes. All you have to do is input the two sequences you want to predict and connect them with a long linker.



Sergey Ovchinnikov @sokrypton · 8 Aug

ColabFold AlphaFold2_advanced, now supports higher-order homo/hetero-complexes (or however number you can fit into memory) 🤣

WARNING: #alphafold was only trained and validated on single chains (monomers). Modeling complexes is an unintended use case, very experimental!



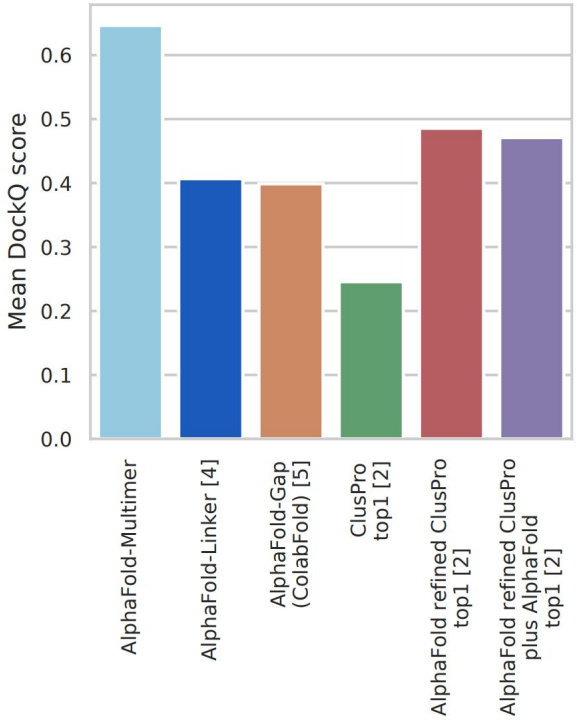
4 12 63

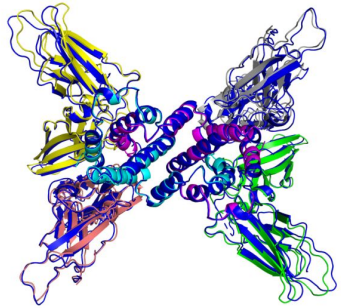
"Even using the default settings, it is clear that **AF2 is superior to all other docking methods**, including other Fold and Dock methods, methods based on shape complementarity and template-based docking"

Bryant, Pozzati, and Elofsson. Improved prediction of protein-protein interactions using AlphaFold2 and extended multiple-sequence alignments, biorxiv 2021. Emphasis ours.

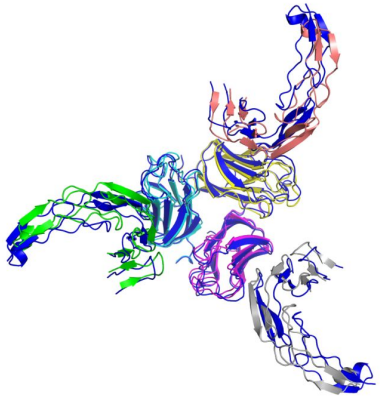
Training AlphaFold to predict multimers (AlphaFold-Multimer)

Adapting the inputs, loss function, and training of AlphaFold to handle multimers and then training the model from scratch

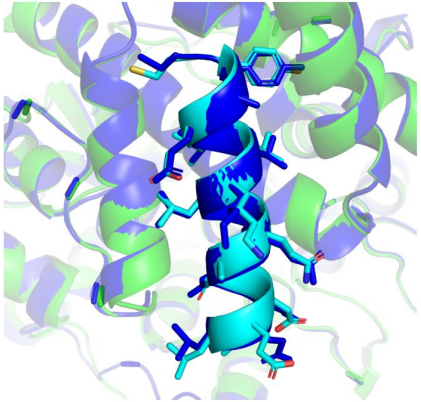




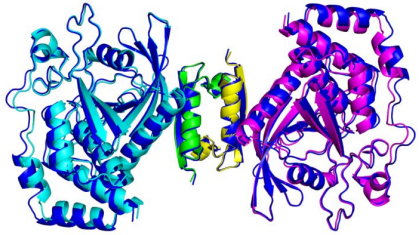
(a) A2B2C2 heteromer
TM-score = 98.0, N_{res} = 1,246, PDB ID = 6E3K



(b) A3B3 heteromer
TM-score = 89.3, N_{res} = 795, PDB ID = 7KHD



(c) Protein-peptide complex
TM-score = 96.0, DockQ = 0.948,
N_{res} = 385, PDB ID = 6JMT



(d) A2B2 heteromer
TM-score = 98.3, N_{res} = 716, PDB ID = 6IWD

Three major challenges

Extend AlphaFold to other major areas of structure prediction

- Increased multimer accuracy and completing the human structural interaction map
- Structure of nucleic acid and ligand interactions
- Conformational diversity

Structural metaproteome

- We will expand Alphafold database to cover UniRef (~100 million structures)
- Need new tools to make these data useful

Prediction of mutational effects

- Changes in structure and stability
- Binding affinity

Acknowledgments

Thank you to everyone who made AlphaFold possible!

AlphaFold 2 Methods

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger (Seoul NU), Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, Demis Hassabis

Human Proteome

Kathryn Tunyasuvunakool, Jonas Adler, Zachary Wu, Tim Green, Michal Zielinski, Augustin Žídek, Alex Bridgland, Andrew Cowie, Clemens Meyer, Agata Laydon, Sameer Velankar (EBI), Gerard J. Kleywegt (EBI), Alex Bateman (EBI), Richard Evans, Alexander Pritzel, Michael Figurnov, Olaf Ronneberger, Russ Bates, Simon Kohl, Anna Potapenko, Andrew J. Ballard, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Ellen Clancy, David Reiman, Stig Petersen, Andrew W. Senior, Koray Kavukcuoglu, Ewan Birney (EBI), Pushmeet Kohli, John Jumper, Demis Hassabis

The wider team at DeepMind

DeepMind

Our wonderful collaborators at EMBL-EBI

EMBL

The CASP community

CASP14

PDB & the experimental biology community

50 YEARS OF Protein Data Bank