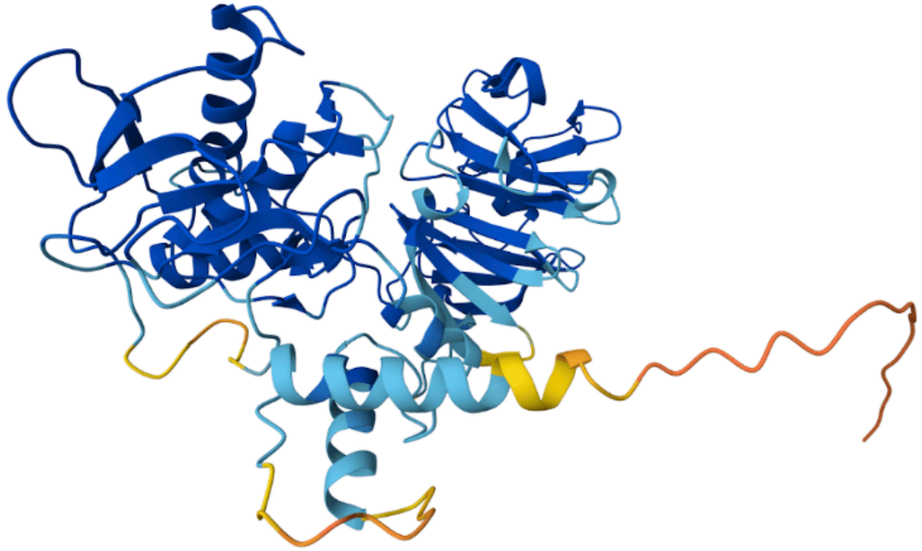
Highly accurate protein structure prediction with AlphaFold (AlphaFold2) [1]

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 A. A. 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, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli & Demis Hassabis ✉

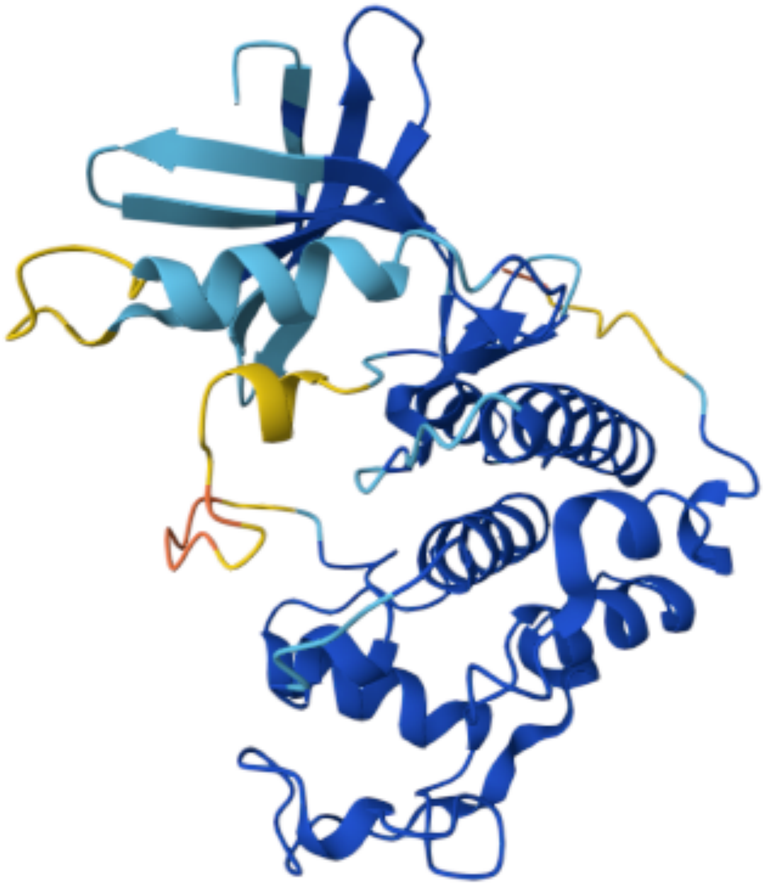
July 15, 2021



Jack Shaw

Outline

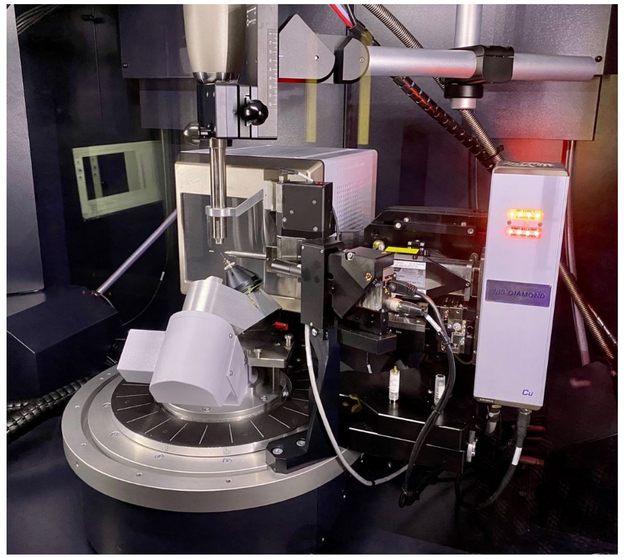
- **Introduction**
    - Motivation
    - The protein folding problem
    - Predicting structure in the light of evolution
- **Methods**
    - Model input
    - Training data
    - Architecture
    - How to interpret predictions
- **Results**
    - Recent PDB structures
    - CASP14
    - Importance of AlphaFold2 components
- **Applications**
    - Ligand binding
    - Hacking for multimeric predictions
- **Conclusion and future directions**



Introduction

Motivation

- Experimental protein structure determination
    - Extensive labor and costs
- Protein structure prediction saves time
- Enables:
    - High throughput structural bioinformatics
    - Protein design
    - etc.
- Deep learning instead of energetic simulations



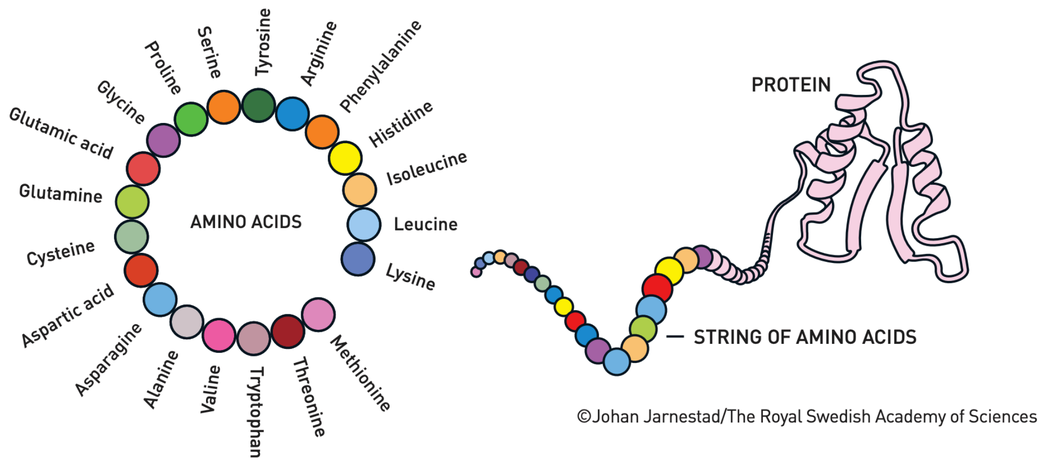
X-Ray Crystallography

Boston University

The Protein Folding Problem

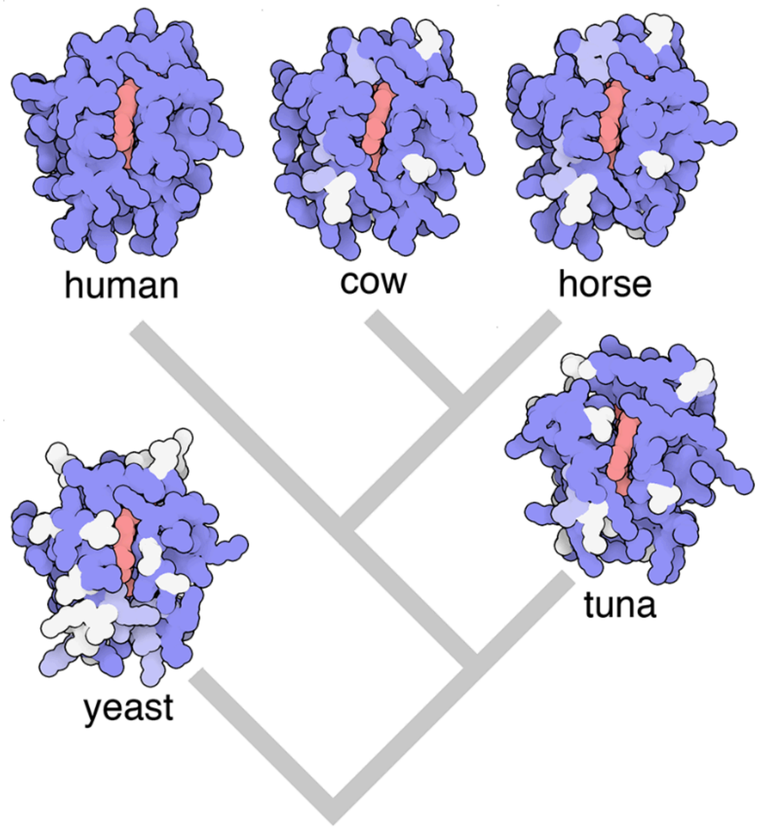
Dill et al., 2008 [2]

1. What is the folding code?
2. What is the folding mechanism?
3. Can we predict the native structure of a protein from its amino acid sequence?



Structural insights from evolution

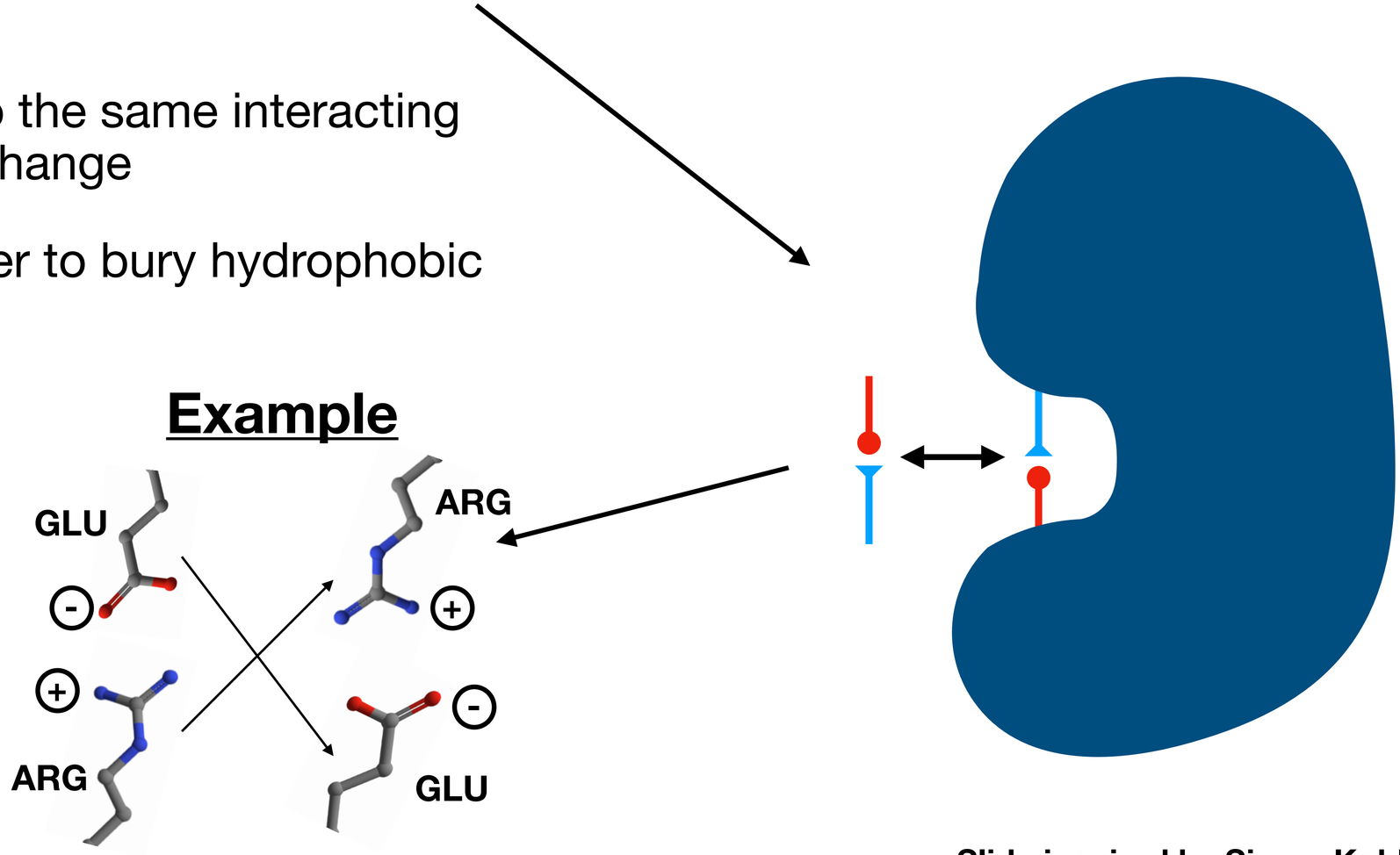
- Evolutionary history
    - Function implies structure
- Homology to solved structures
- Pairwise evolutionary correlations



Evolutionary history of cytochrome c.
Illustration by PDB-101

Pairwise evolutionary correlations indicate structure

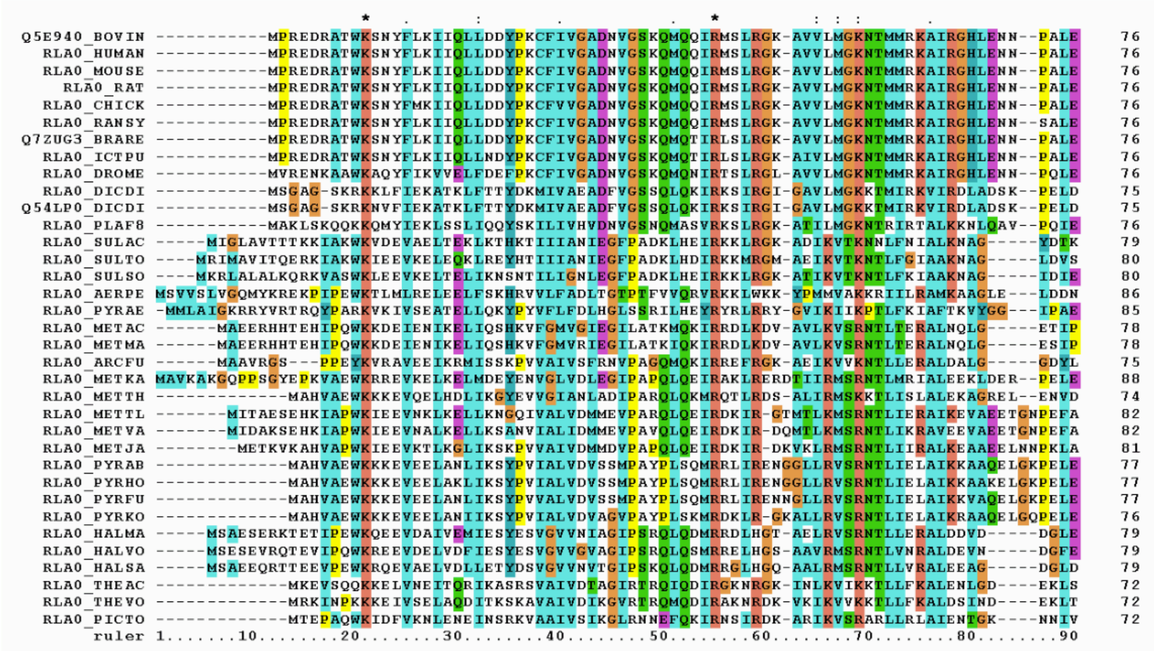
- Interacting residues must compensate for each other's changes
- Residues often bind to the same interacting molecule, which can change
- Residues work together to bury hydrophobic regions
- etc.



Slide inspired by Simon Kohl's talk [3]

Multiple sequence alignments

- Query large genetic databases
- Find homologous sequences
- Used to calculate evolutionary conservation of aa positions
- Can be used to find coevolutionary correlations



source: https://en.wikipedia.org/wiki/Multiple_sequence_alignment

How are MSAs created?

- AlphaFold uses Jackhmmr [4] and **HBBlits [5]**
  - Initial search for homologs
  - Profile HMM across homologs
    - Emission probabilities
    - Transition probabilities
  - Iterative searching with HMM profile
    - HMM:HMM pairs

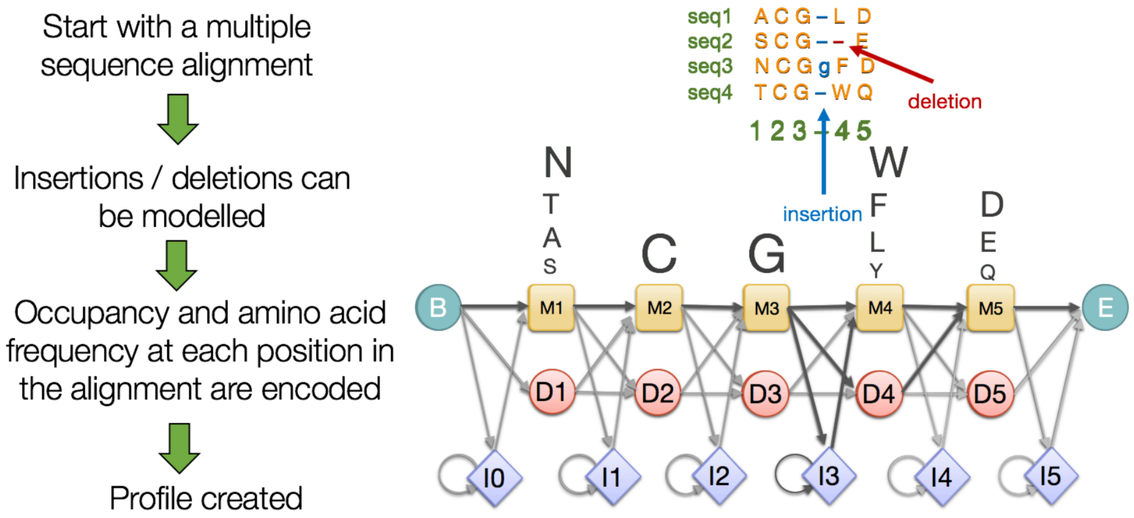


Illustration from EMBL-EBI Training

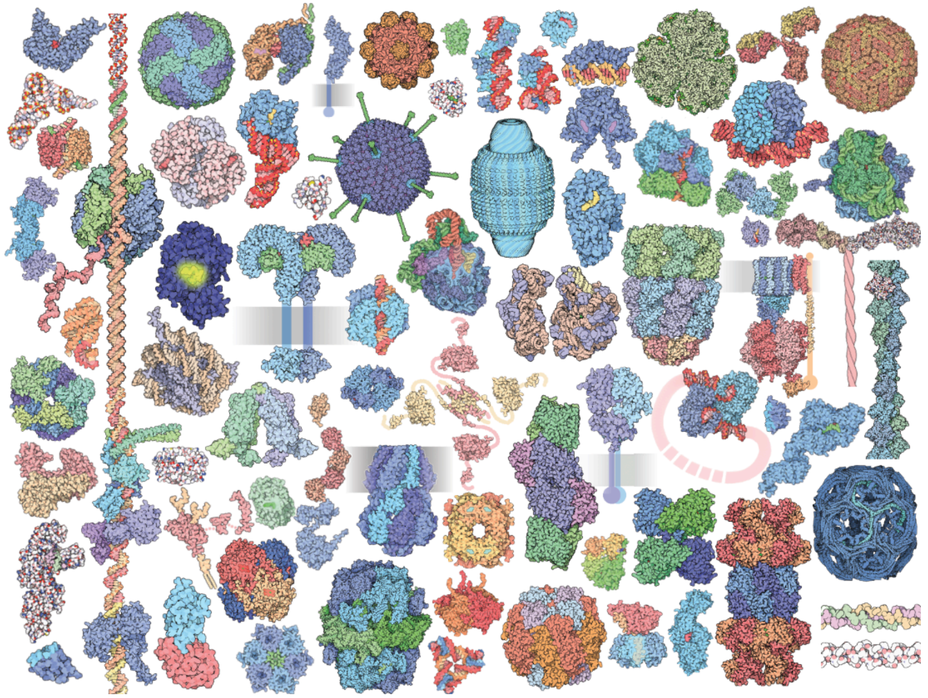
Methods

What does AlphaFold2 do?

- Goal: predict proteins' atomic coordinates given sequence input
- How: learning the mapping from MSA to structure

Training data

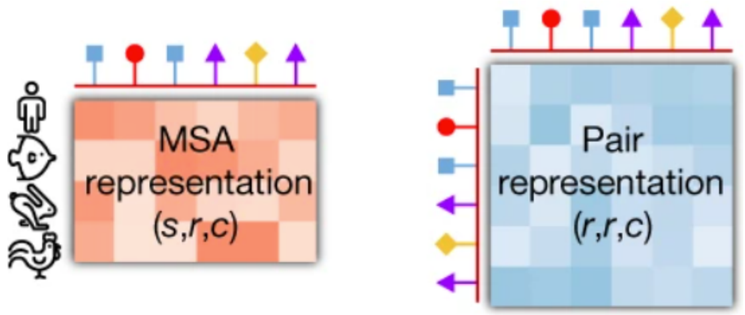
- Protein structures from the Protein Data Bank [6]
    - Structures before April 30, 2018
    - Over 150,000 known structures
- MSAs
    - Produced with Jackhmmr [4] and HHBlits [5]
    - Sequence data:
        - Big Fantastic Database
            - Uniprot (over 2 billion seqs)
            - Metaclust NR (150 million seqs)
        - UniRef
        - Uniclust30
        - MGnify

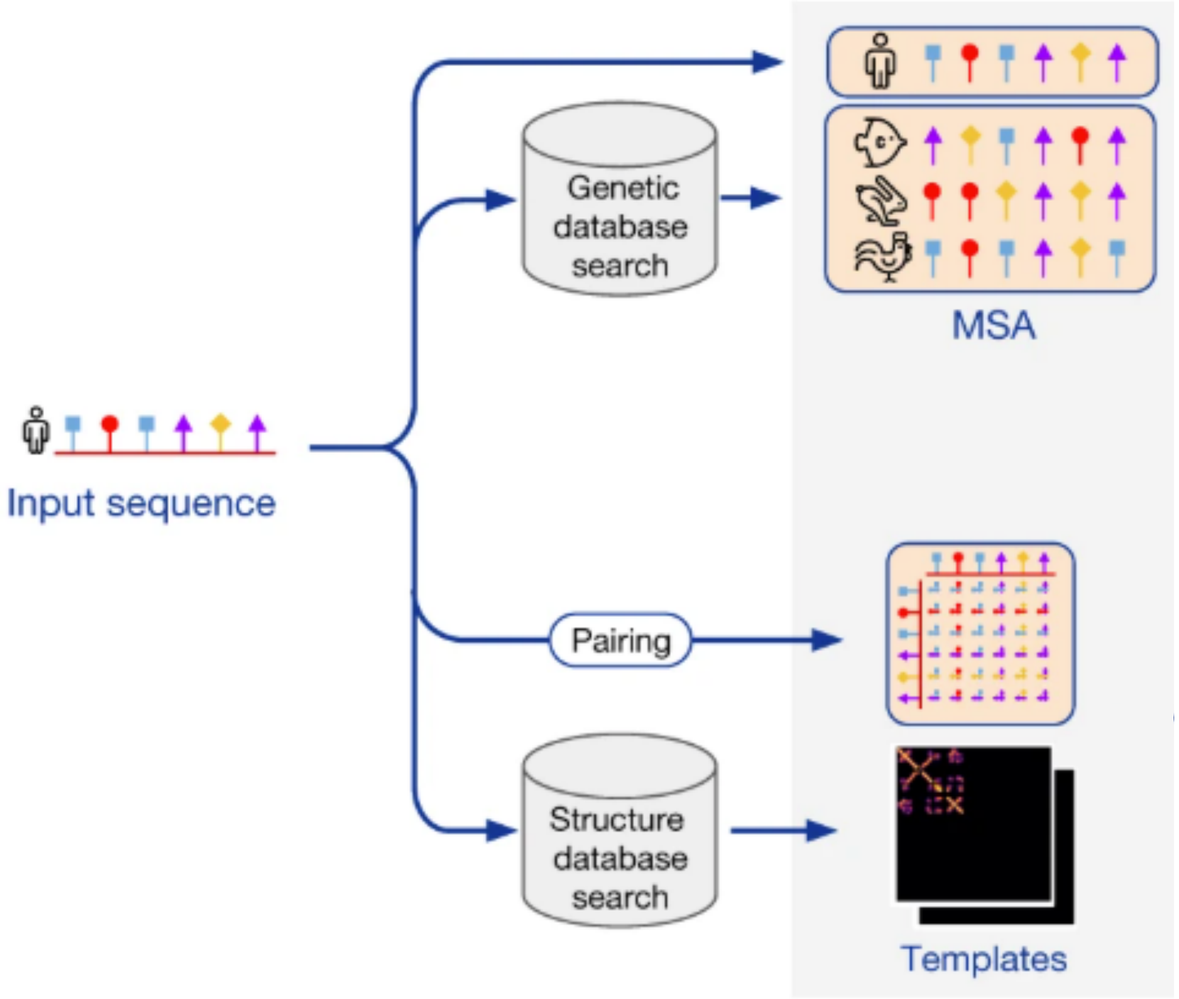


Monomers only!

Model input and processing

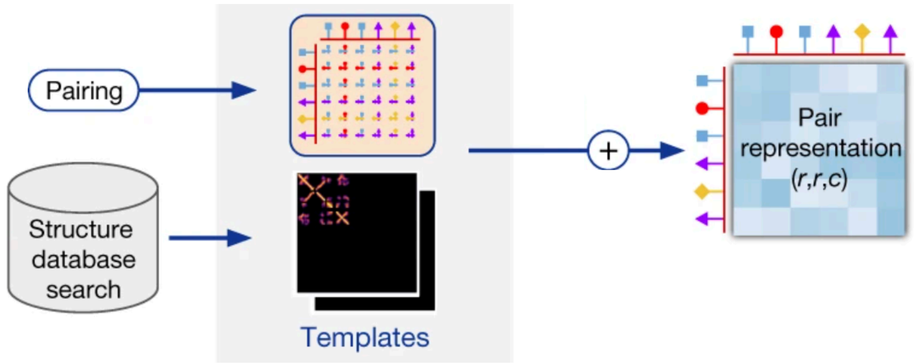
- Amino acid sequence
- MSAs with homologs
- Templates
    - Structural homology





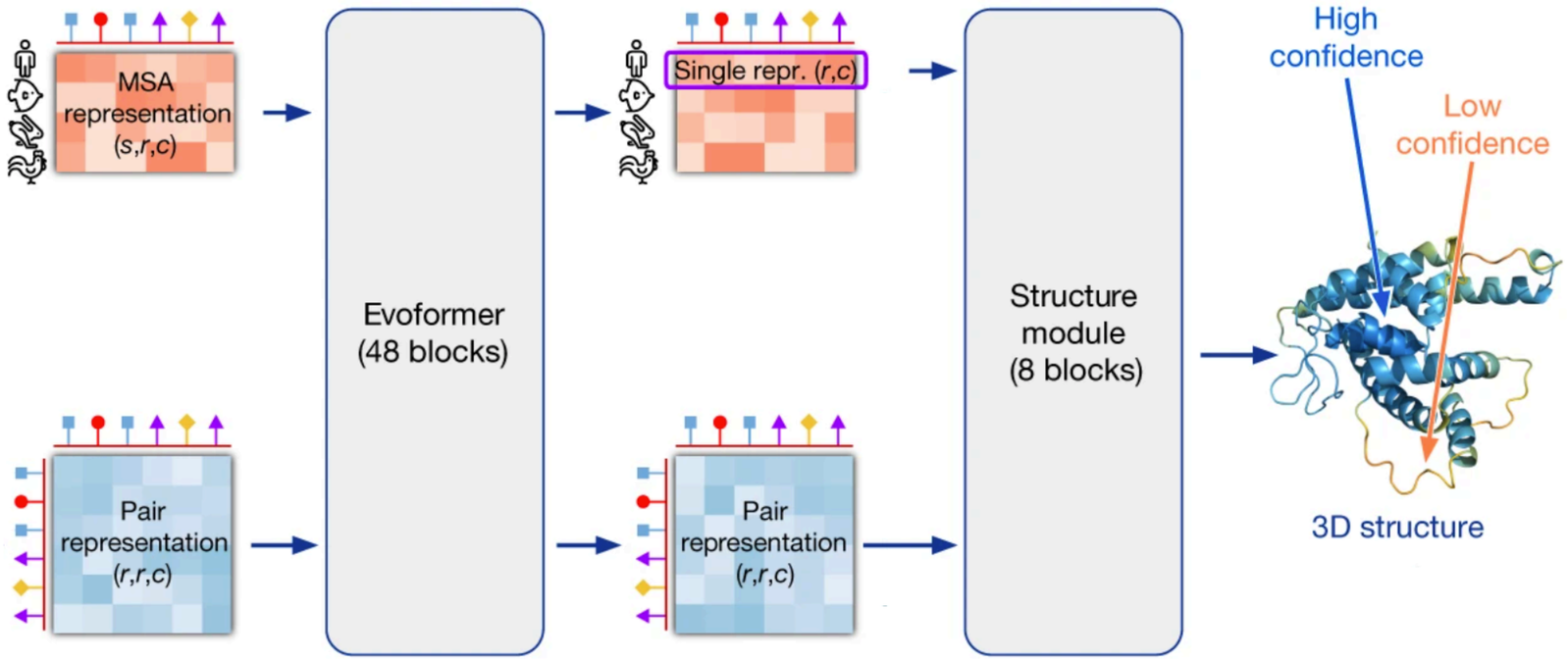
How are templates found?

- Find PDB structures whose sequences look like the input
    - Searching for structural homology
- Uses HMM profile and searches against PDB
    - Finds the best matches
    - What residues align in the structure?
- Serves as initial clue in pair representation



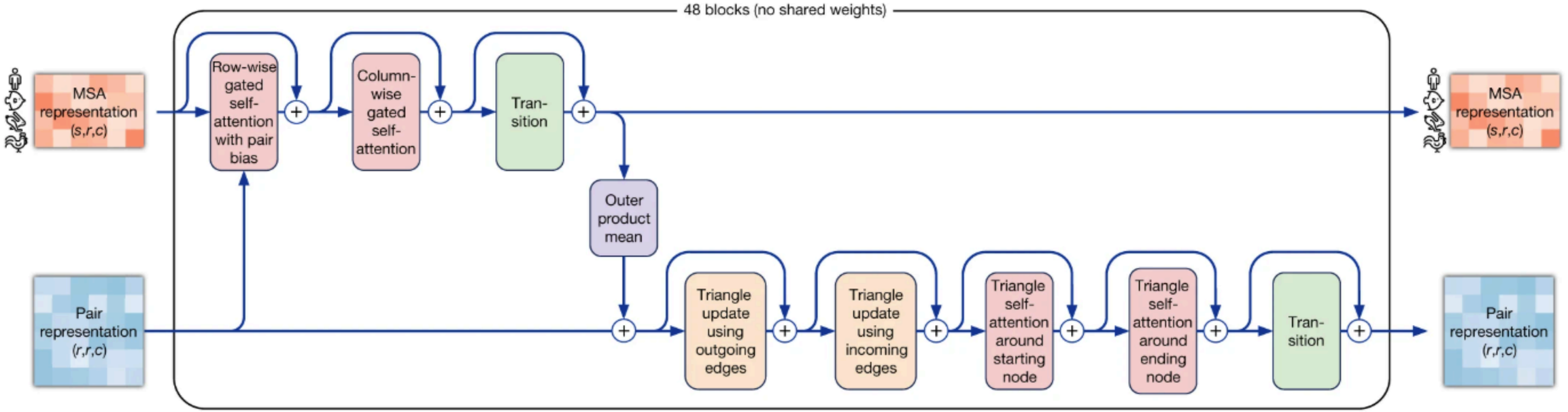
AlphaFold architecture

- Evoformer: a 48 block neural network that iteratively updates the pair and MSA representations
- Structure module: transforms the pair and MSA representations into 3d atomic coordinates



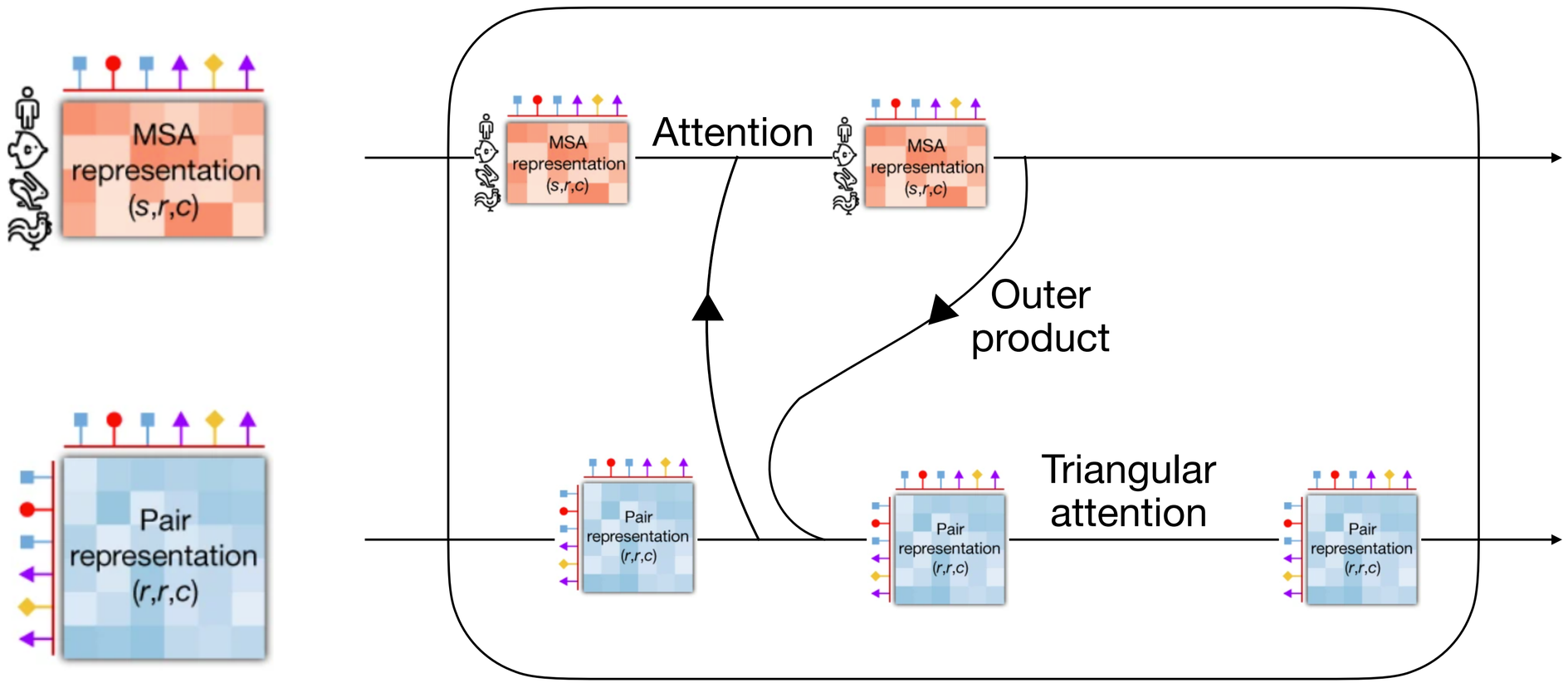
Evoformer

- Custom transformer model
- Begins with MSA and pair representations
- Generates structural insight for structure module
- Iterating process



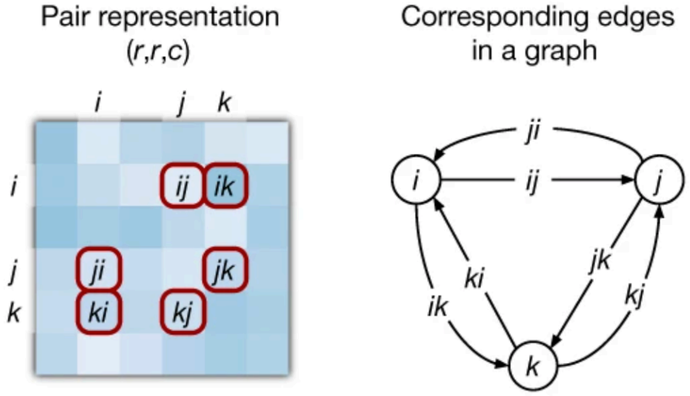
x 48

Evoformer block simplified

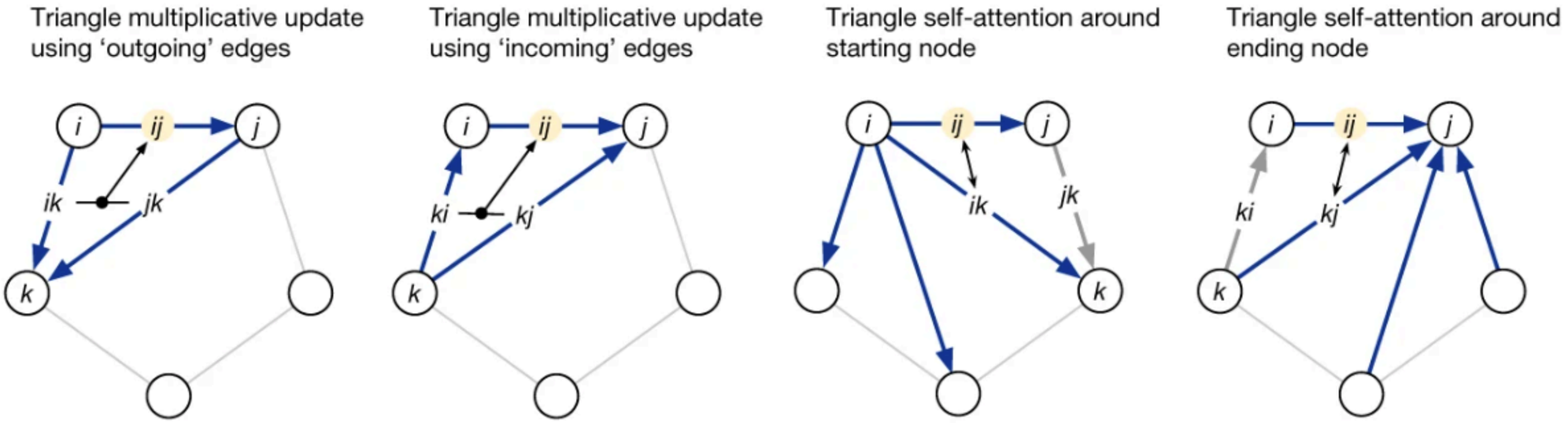


Information flow between representations

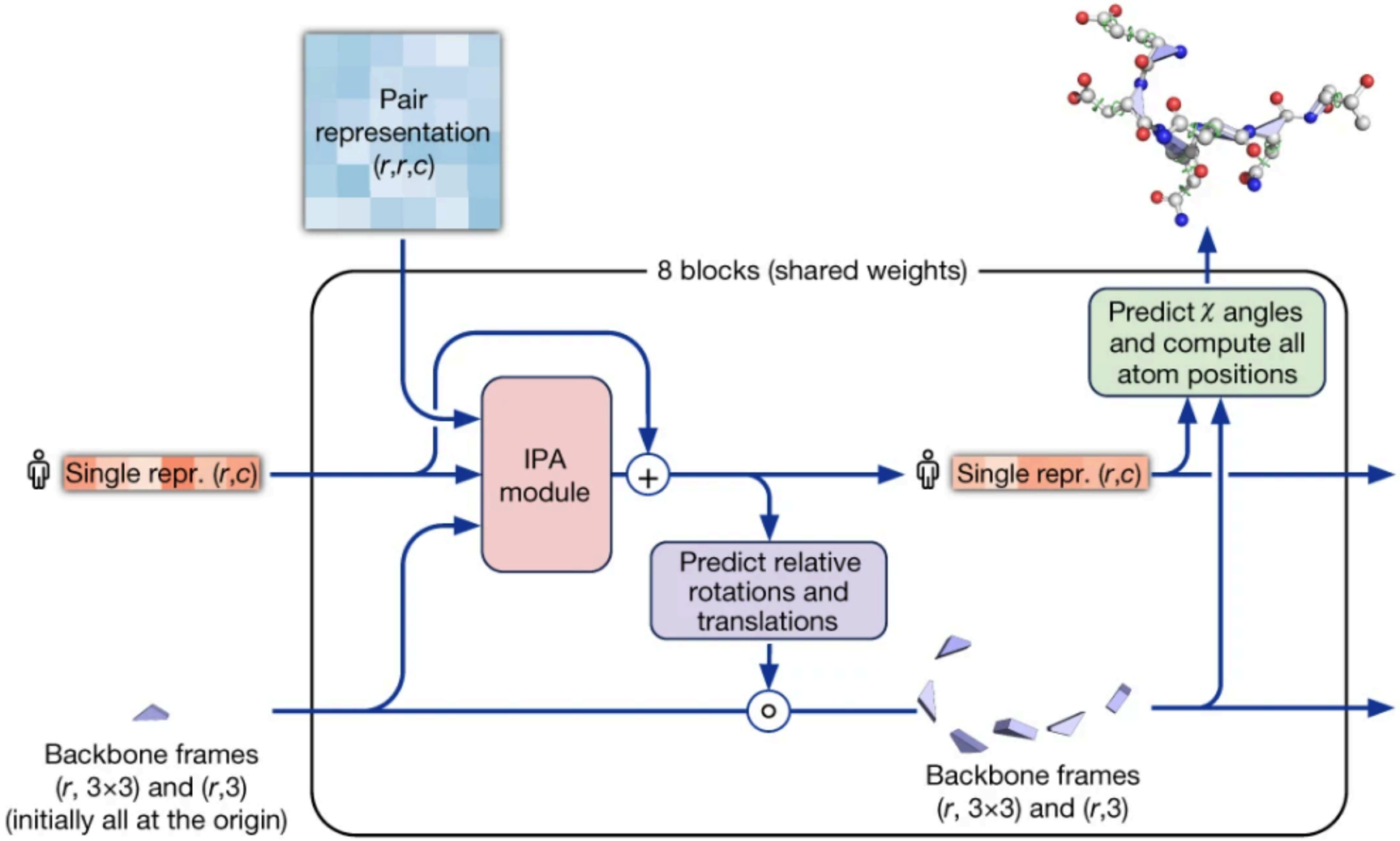
Triangular attention

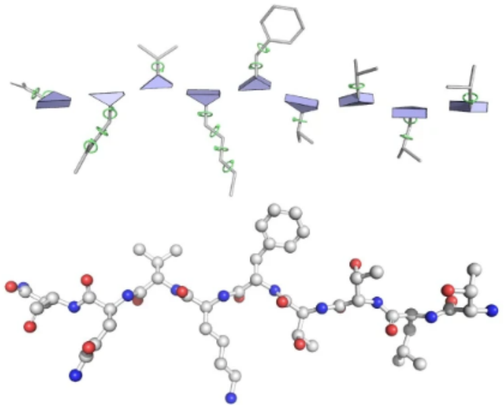


- Learns geometric constraints
- Triplets are weighted with attention
- Weights are communicated to MSA representation through Evoformer



Structure module

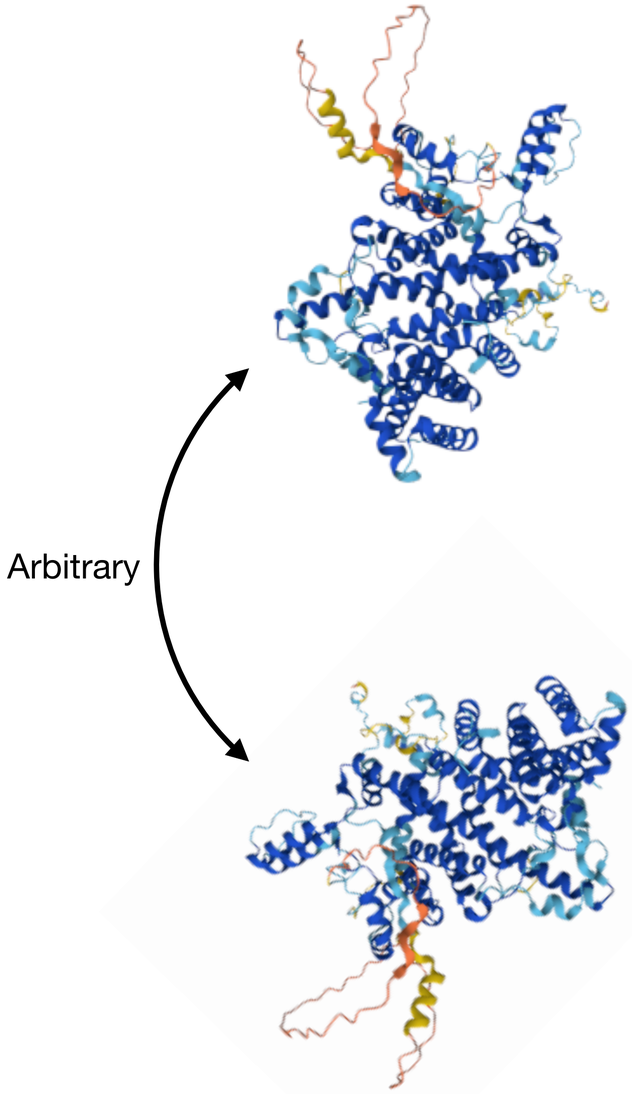




“Residue gas”

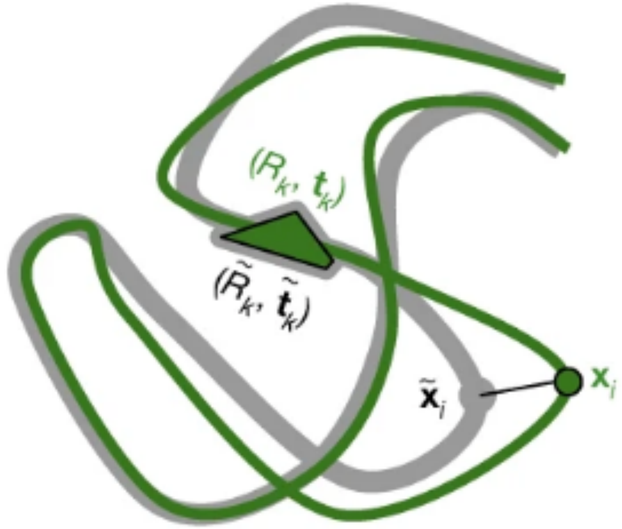
Invariant point attention module
(IPA) module

- Geometry-aware
- Updates N_{res} set of neural activations
- Equivariant update using updated activations
- Invariant to global rotations and translations



Frame Aligned Point Error (FAPE)

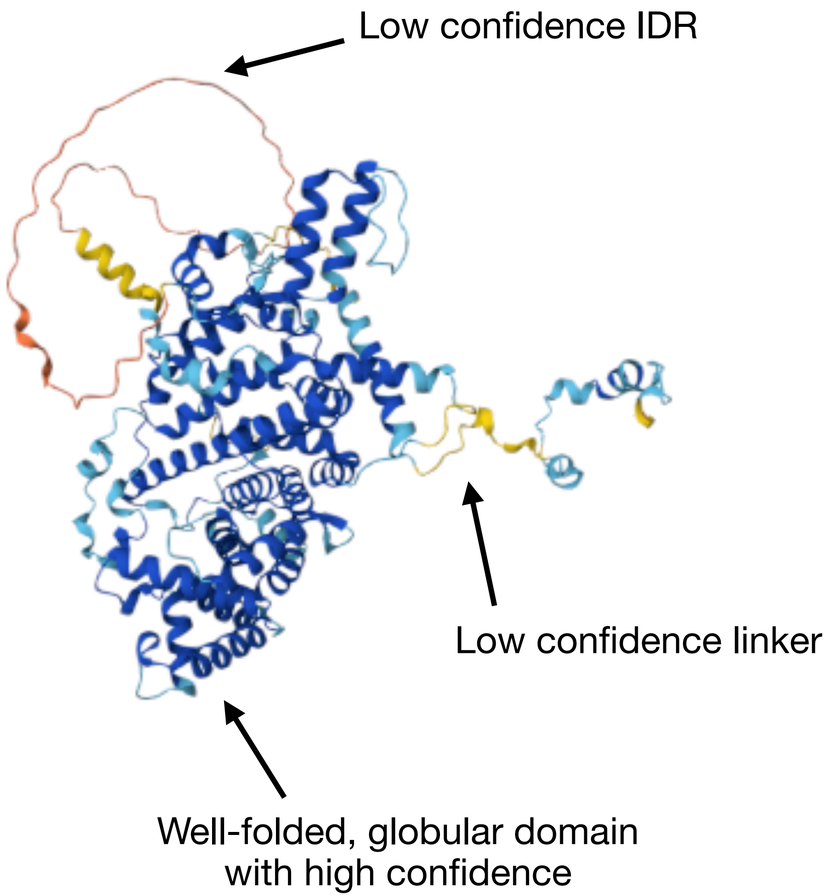
- How loss is calculated in structure module
- Compares predicted vs. experimental under many different alignments (frames)
- Error across N_{atoms} and N_{frames} creates N_{atoms} x N_{frames} distances
- Penalize with a clamped L^1 loss
- Biases residues to have correct side chains



pLDDT

predicted Local Distance Difference Test

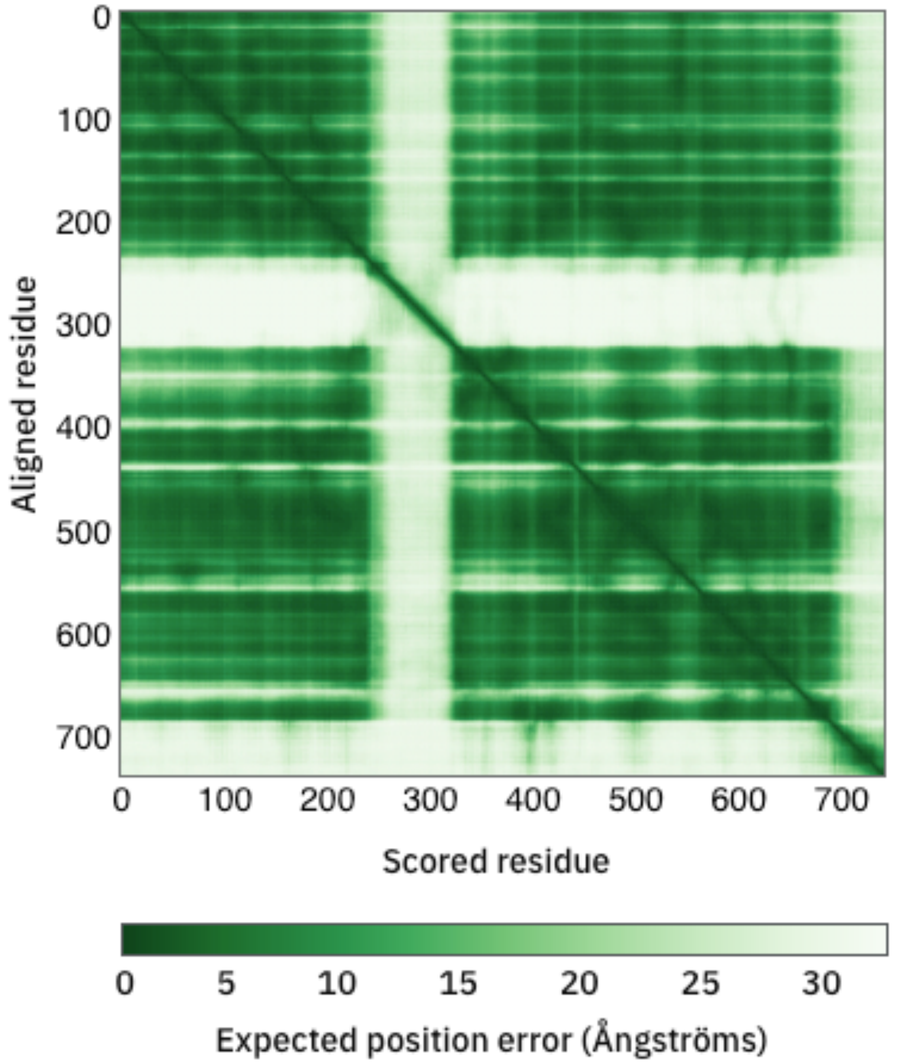
- pLDDT measures the structure prediction confidence within a domain
- pLDDT > 90 – accurate stereochemistry
- pLDDT < 50 – structure is disordered (low confidence)



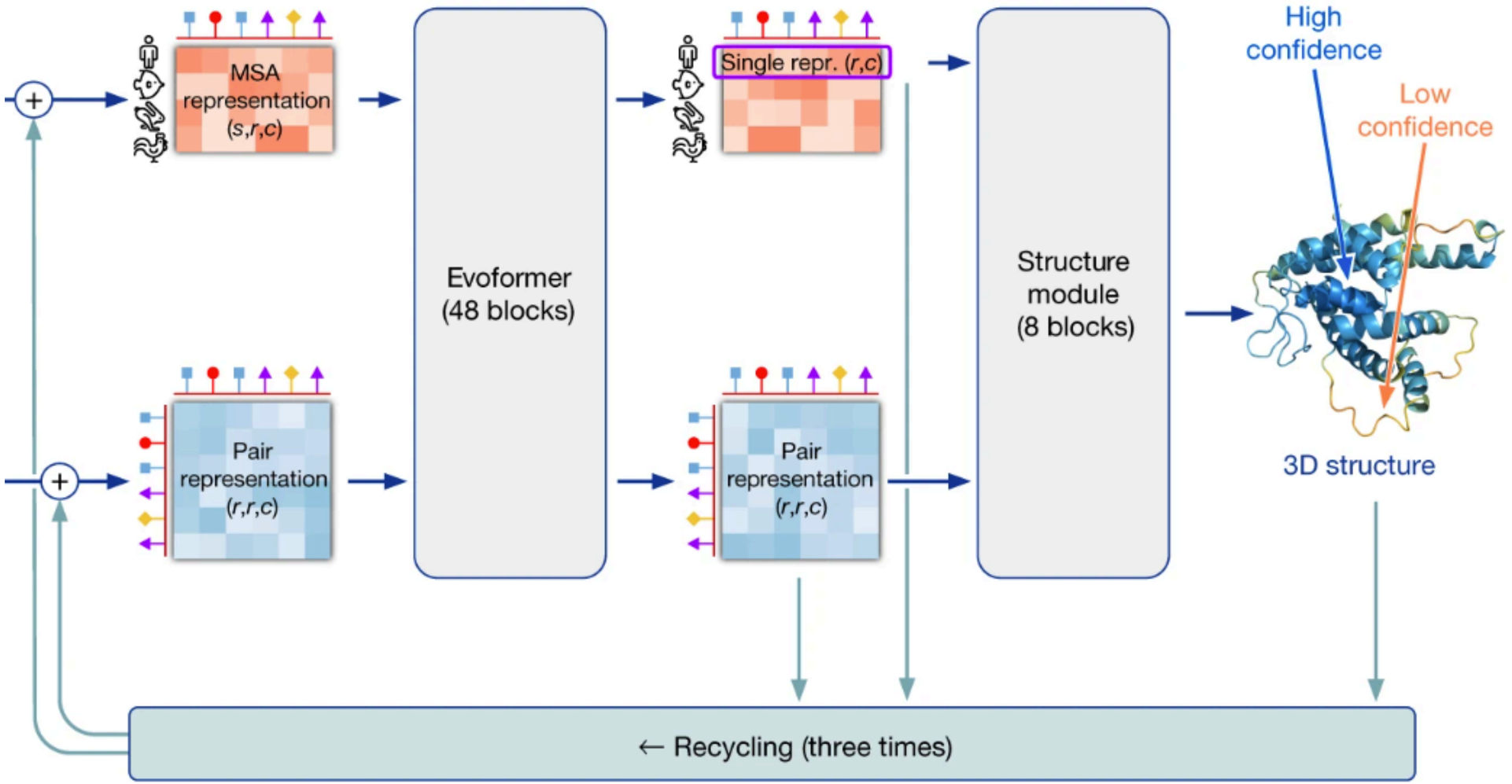
PAE

Predicted Aligned Error

- Pairwise prediction accuracy
- 0 Å - 30 Å
- Enables accurate contact mapping
    - Coevolved residue pairs

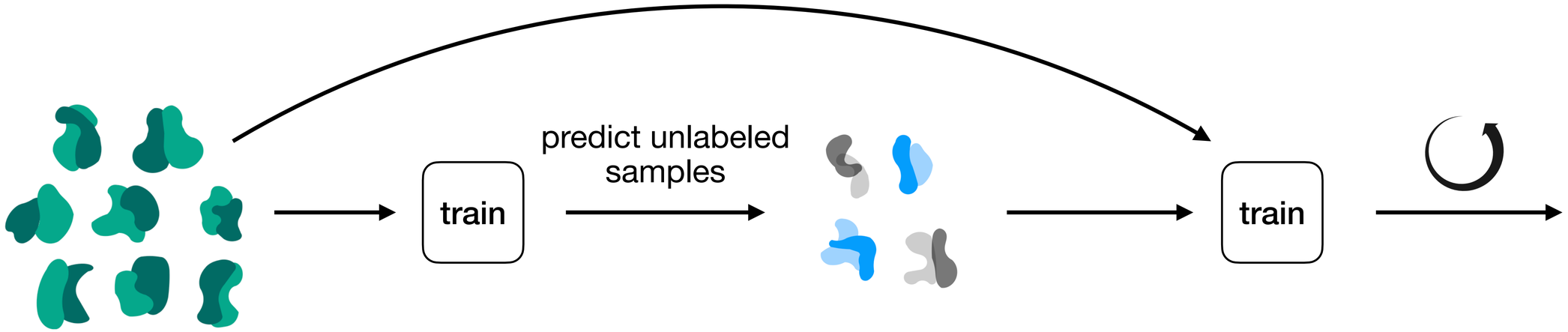


Recycling

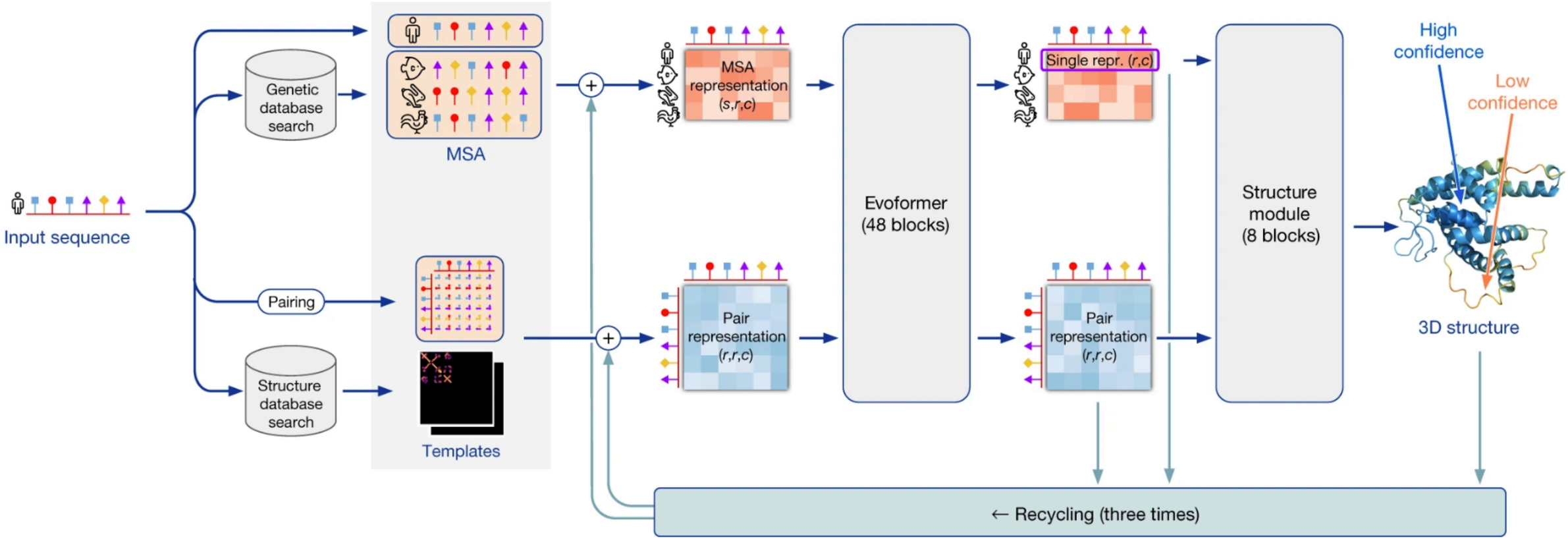

MSA representation (s,r,c)
Pair representation (r,r,c)
Evoformer (48 blocks)
Single repr. (r,c)
Pair representation (r,r,c)
Structure module (8 blocks)
High confidence
Low confidence
3D structure
← Recycling (three times)

Noisy student distillation

- AlphaFold trained multiple times
- Future models trained on accurate predictions from initial model
- Helps as a normalization step



Architecture summary

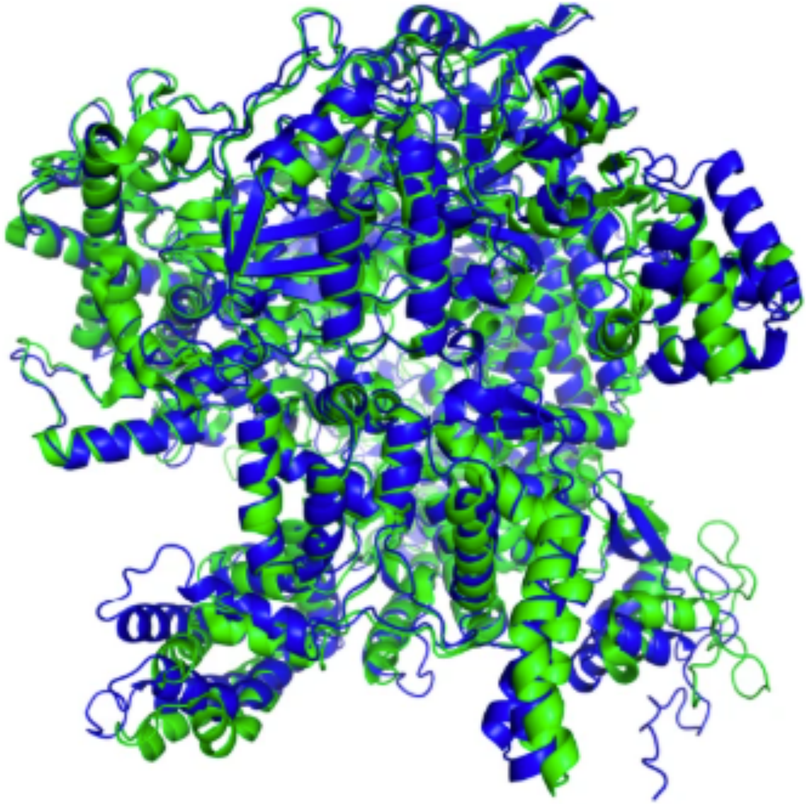


Results

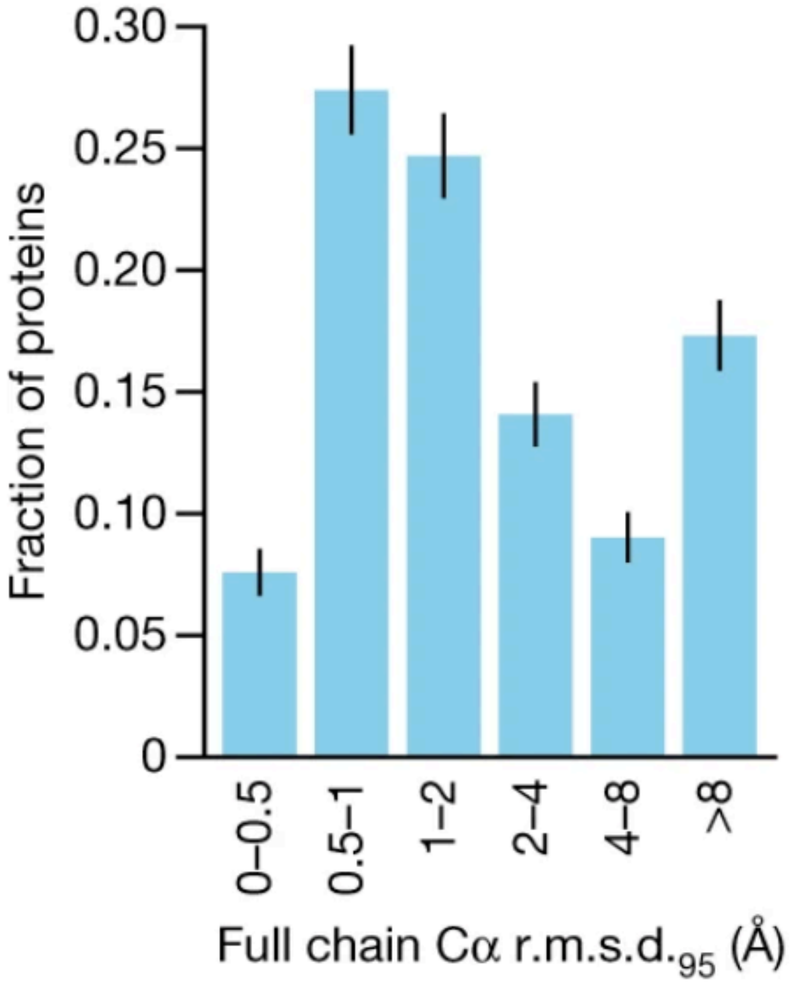
RMSD

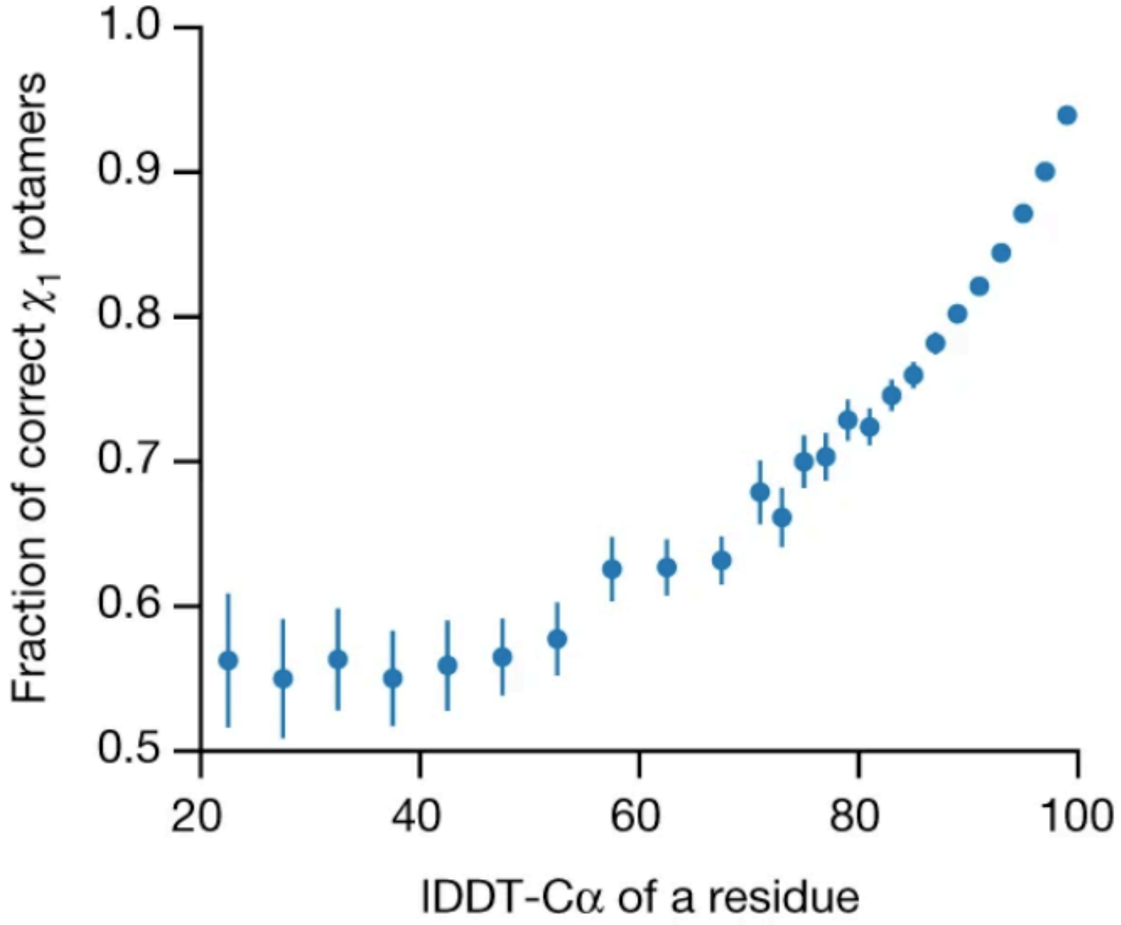
a measurement of structure prediction accuracy

- How similar is the predicted structure to the experimentally determined one?
- Calculated with alpha carbon (CA)
- Minimizes total sum of distances
- Calculated with a superimposition →



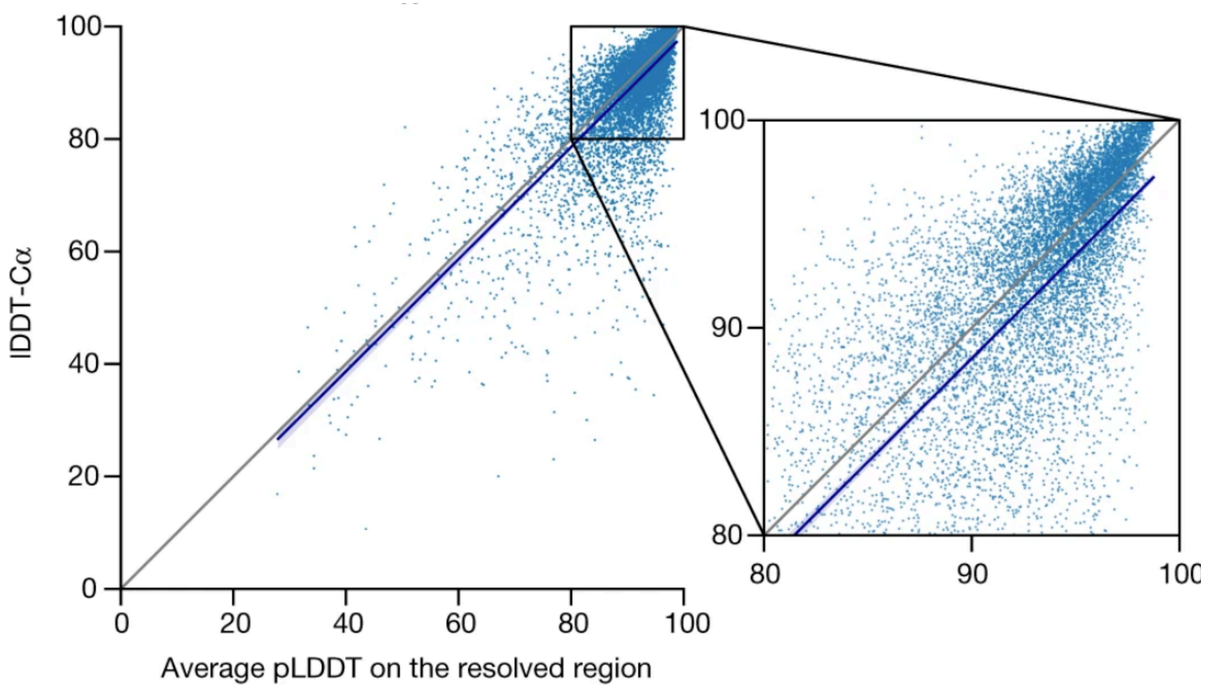
Performance on recent PDB structures

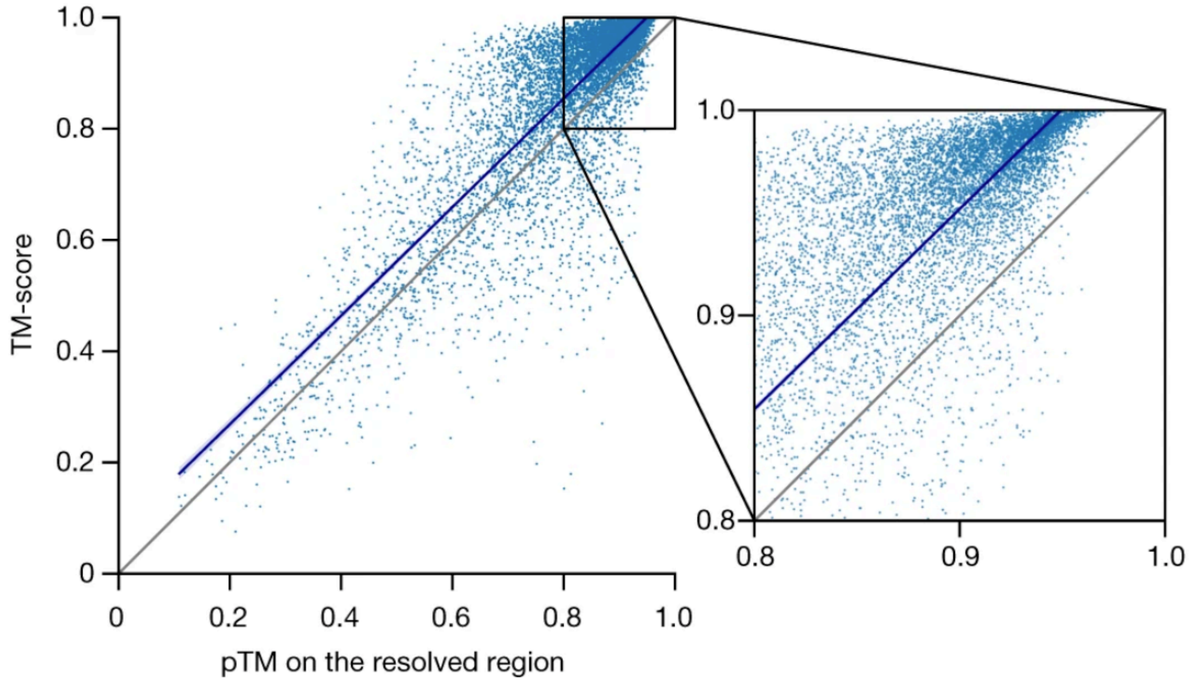




Structures deposited to PDB from April 30, 2018 to February 15, 2021

Performance on recent PDB structures



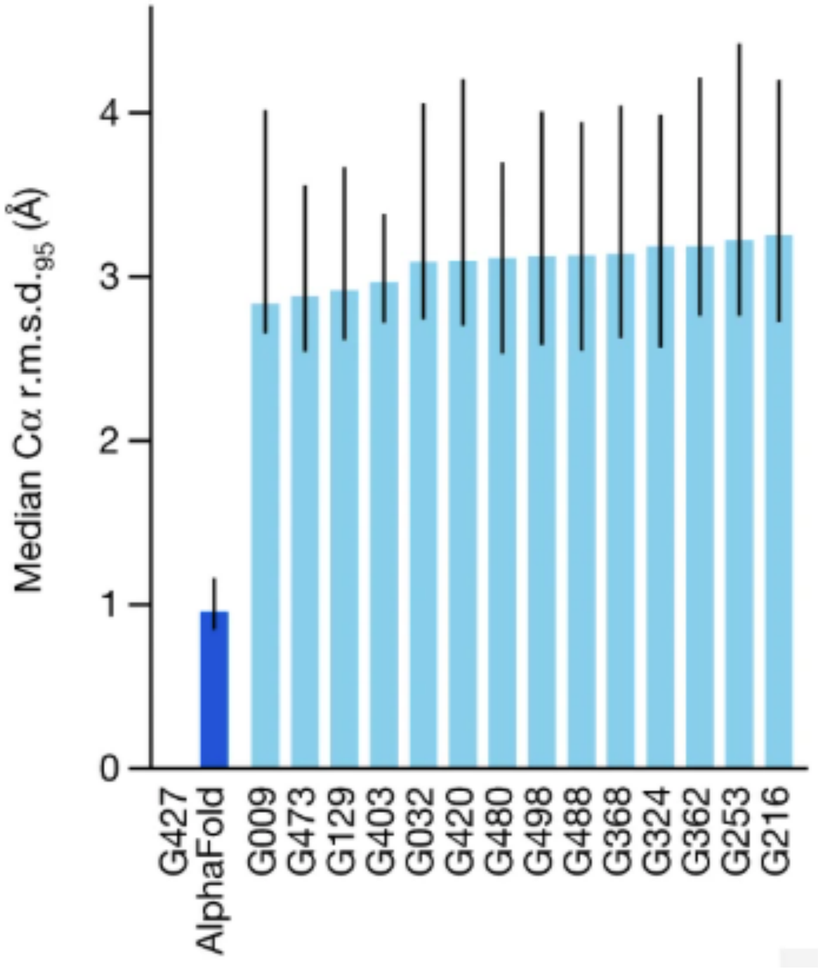


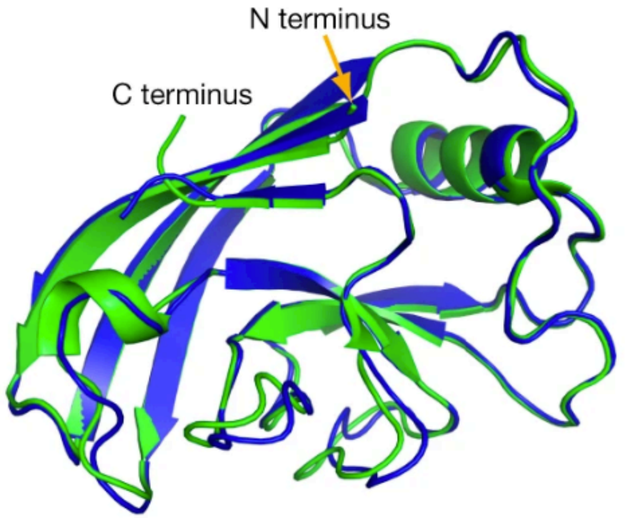
Confidence metrics indicate accuracy

Structures deposited to PDB from April 30, 2018 to February 15, 2021

CASP14

- Massive increase in performance
- Hype ⟶ **reality**
- Field of structural bioinformatics widely expanded





AlphaFold Experiment

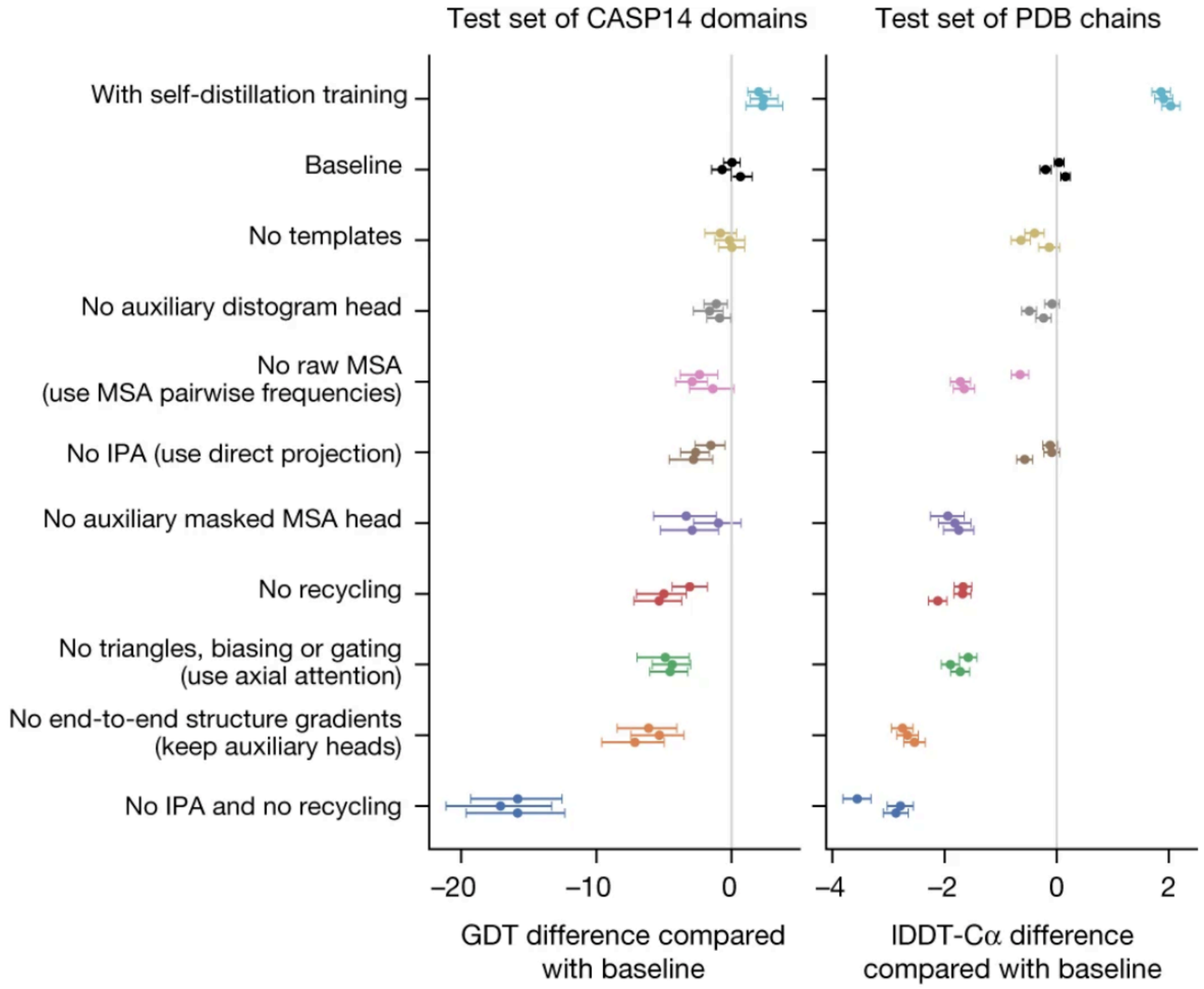
r.m.s.d.$_{95}$ = 0.8 Å; TM-score = 0.93

T049

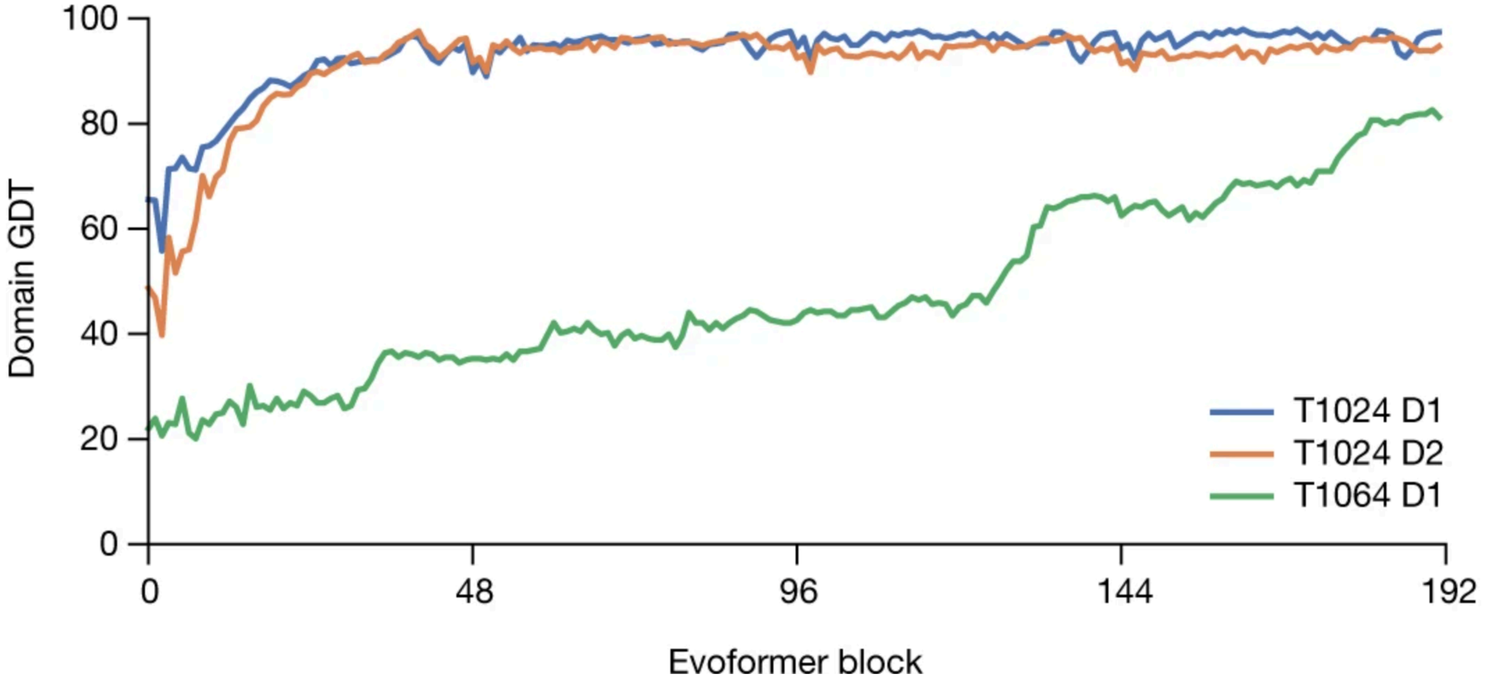
AlphaFold components

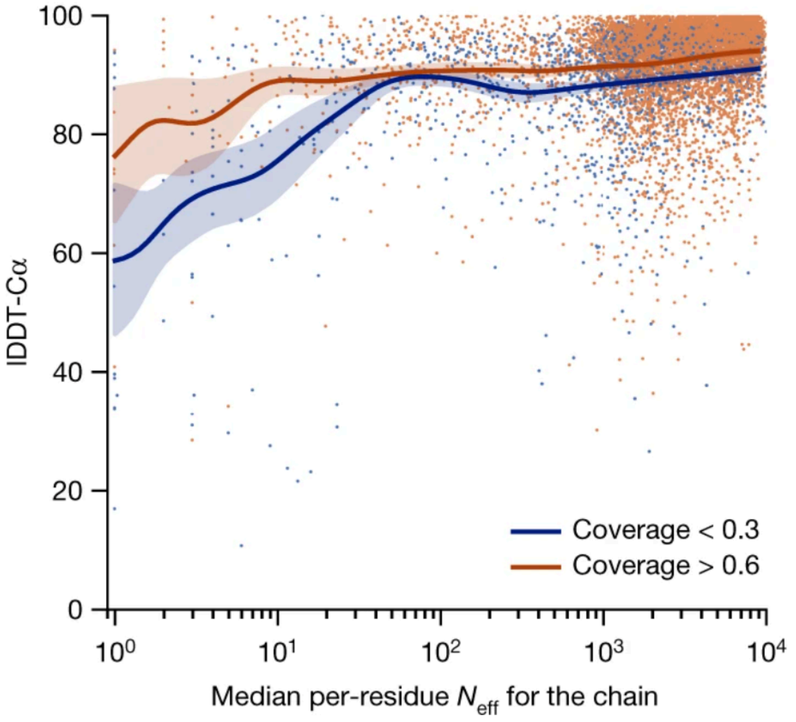
Importance evaluation

- Each component of the model is important
- Justifies complicated architecture
- Protein structure prediction is not an easy task!



How does AlphaFold predict structure?





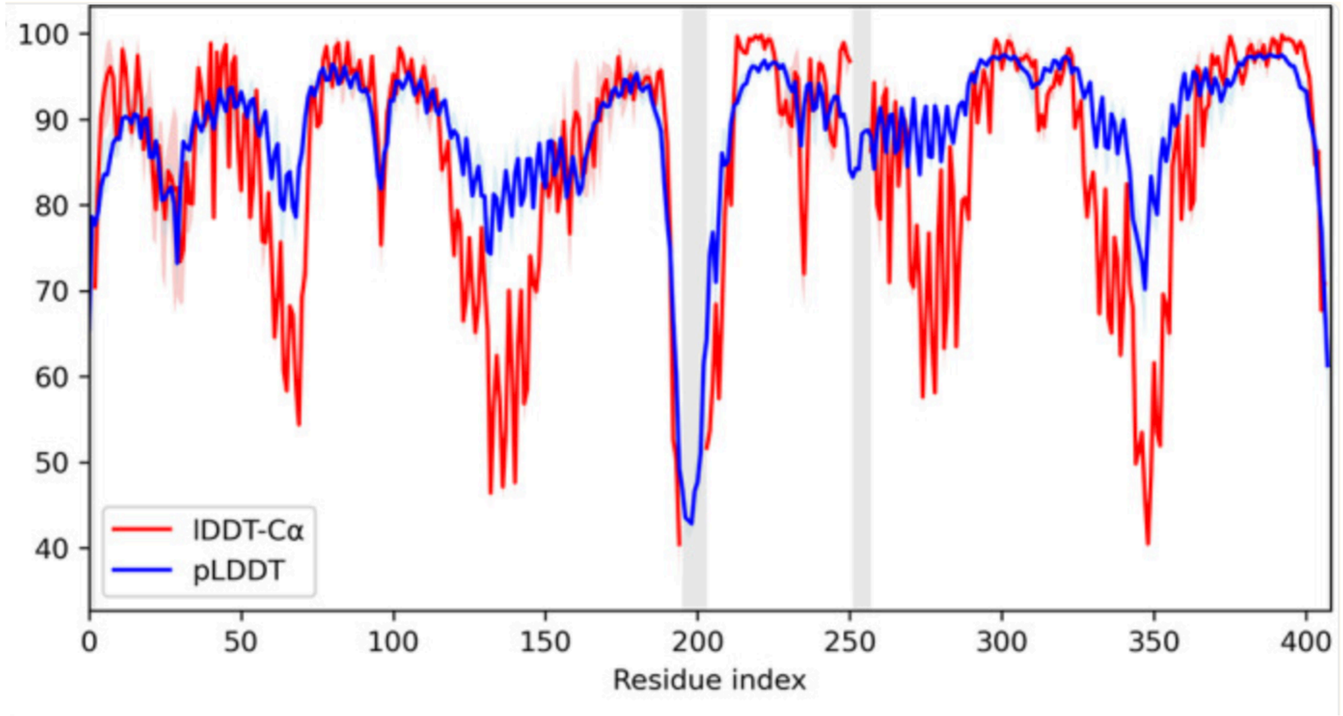
T1024

Recycling iteration 0, block 01
Secondary structure assigned from the final prediction

Single conformation only!

T1024

- Inward vs. outward conformation



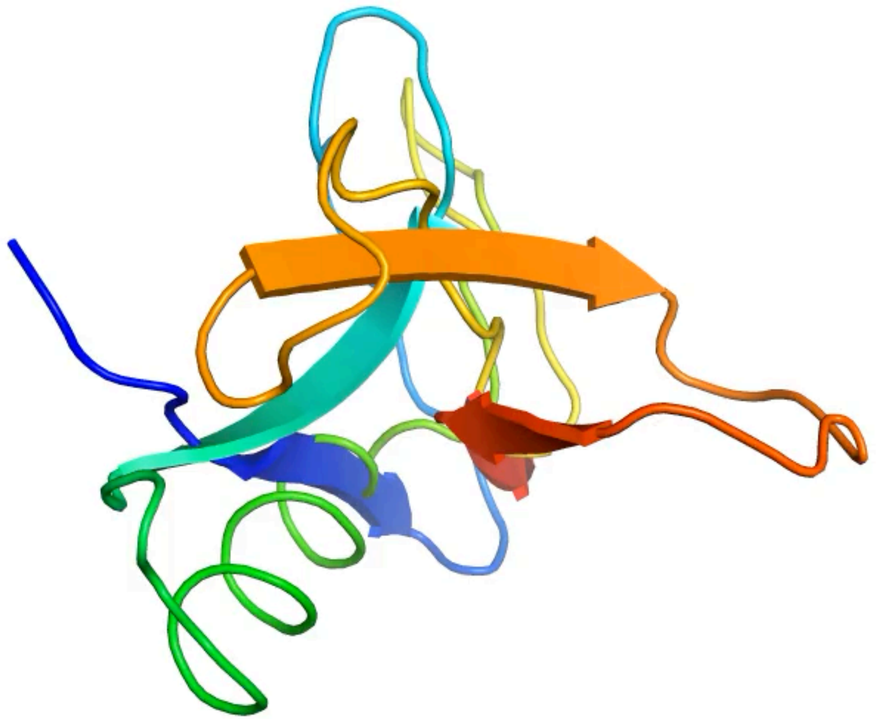
- Decrease MSA depth and pick template ⟶ successful prediction [8]

T1091

Recycling iteration 0, block 01
Secondary structure assigned from the final prediction

Iterations are important!

T1064



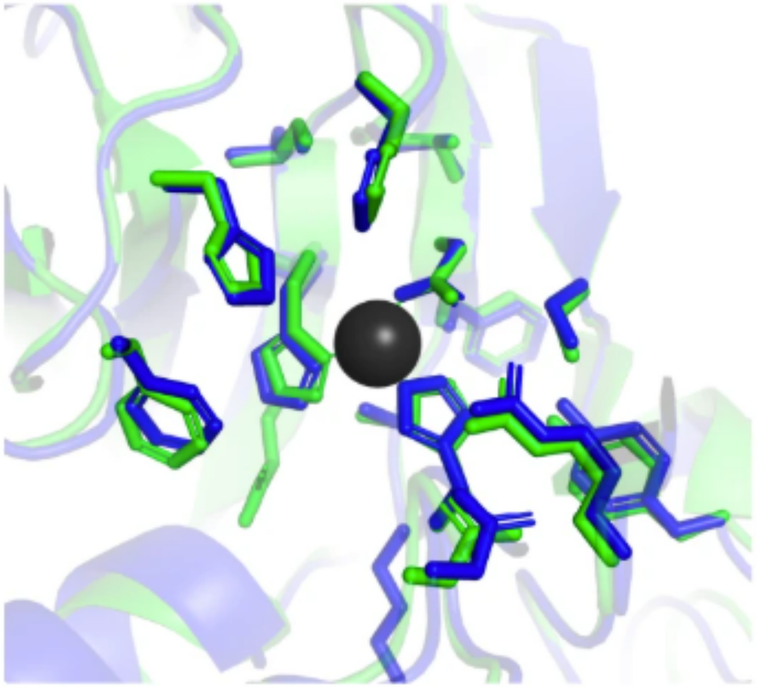
Recycling iteration 0, block 01
Secondary structure assigned from the final prediction

Applications

Imputing the presence of small molecules

- PDB contains non-polymer molecules
    - AlphaFold trains around these
- Enables discovery of novel binding pockets
- Single molecule predictions can be used for interaction mechanism analysis

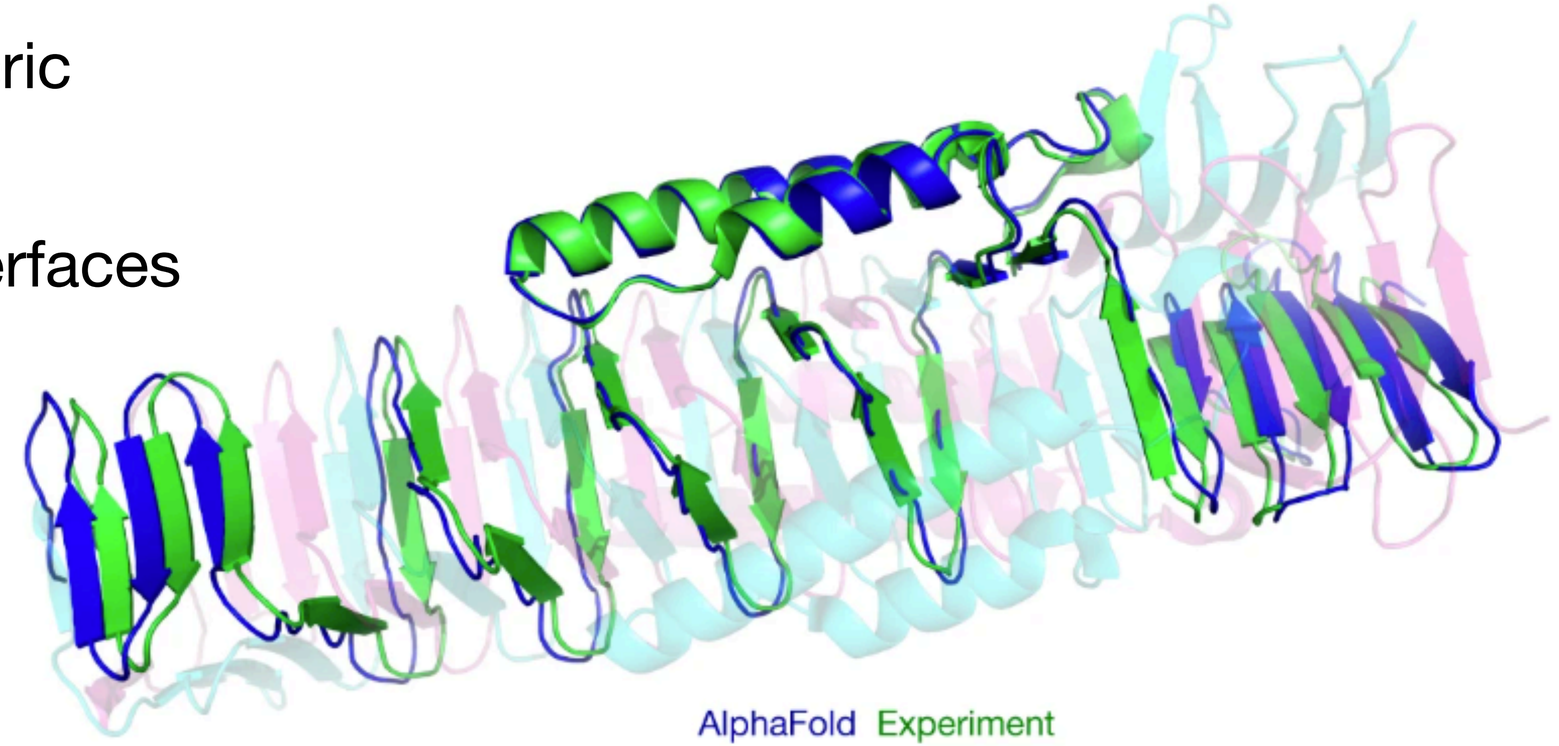
Zinc ion binding pocket



AlphaFold Experiment
r.m.s.d. = 0.59 Å within 8 Å of Zn

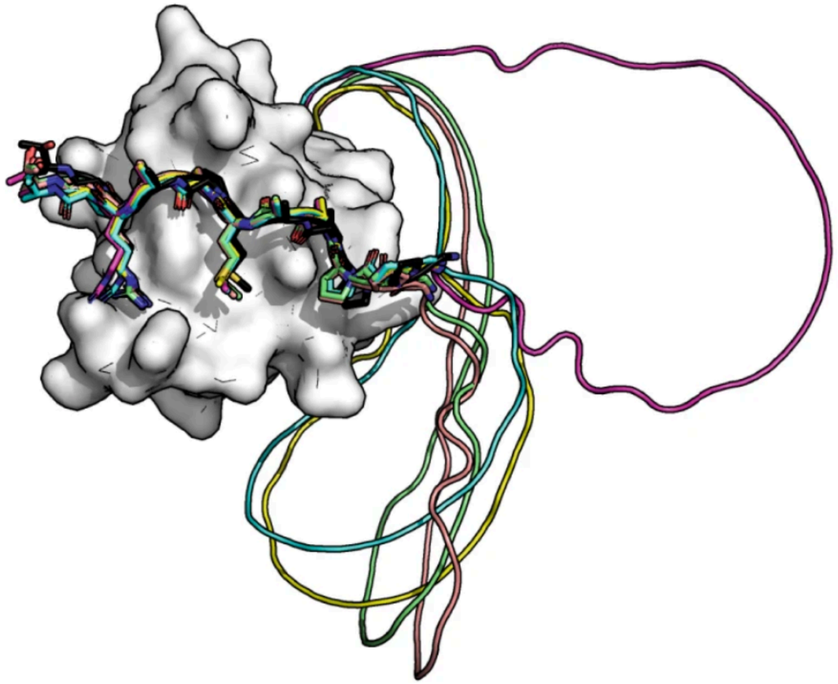
Implied multimeric conformation

- Proteins have multiple conformations
    - Depends on binding partners
- AlphaFold predicts multimeric conformations
- Enables analysis of PPI interfaces



Hacking to predict multimeric structures

- GLY: missing side chain
    - Often creates flexible segments (IDRs)
- Predict multimeric structures using a GLY linker [9]
    - AlphaFold2 is tricked into treating the input as a single protein



AlphaFold-Multimer [10] and ColabFold [11]

- Uses AlphaFold 2 architecture to predict protein complexes
    - Enables protein-protein interaction prediction
- Fuels high throughput *in silico* structural bioinformatics research
- ColabFold makes AlphaFold-Multimer easily accessible
    - Supports easy local GPU usage
    - More efficient MSAs

Conclusion and future directions

Key contributions

- The most (by far) accurate protein structure predictor yet
- Open source
    - Enables multimeric predictions and fine tuning
- Enables structural bioinformatics
    - AlphaFold Database
- Offers an initial model for Cryo-EM structure determination
- Inspired others to use coevolution to predict structure

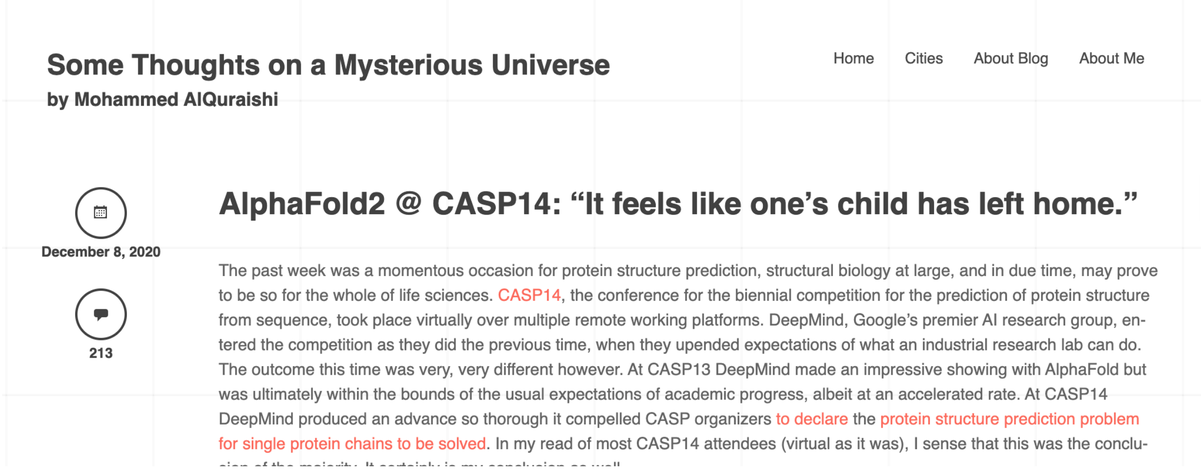
Limitations

- Limited to single chain predictions
    - Context in biology is critical
- Not able to predict PTMs, DNA, ligands, etc.
- Predecessor to AlphaFold 3’s diffusion model
- Limited accuracy predicting dissimilar protein sequences
    - Protein Data Bank is limited
- Static structure predictions (T1024)

Next steps

- Can we predict all protein conformations?
    - Protein Data Bank only contains static structures
    - pLDDT not perfect for defining domain flexibility
- Can we design useful proteins?
    - Antibodies
    - Inhibitors
    - Genome editors

More about AlphaFold2



Some Thoughts on a Mysterious Universe

by Mohammed AlQuraishi

Home Cities About Blog About Me

December 8, 2020

213

AlphaFold2 @ CASP14: “It feels like one’s child has left home.”

The past week was a momentous occasion for protein structure prediction, structural biology at large, and in due time, may prove to be so for the whole of life sciences. CASP14, the conference for the biennial competition for the prediction of protein structure from sequence, took place virtually over multiple remote working platforms. DeepMind, Google’s premier AI research group, entered the competition as they did the previous time, when they upended expectations of what an industrial research lab can do. The outcome this time was very, very different however. At CASP13 DeepMind made an impressive showing with AlphaFold but was ultimately within the bounds of the usual expectations of academic progress, albeit at an accelerated rate. At CASP14 DeepMind produced an advance so thorough it compelled CASP organizers to declare the protein structure prediction problem for single protein chains to be solved. In my read of most CASP14 attendees (virtual as it was), I sense that this was the conclu-

Reaction to CASP14

Some Thoughts on a Mysterious Universe

by Mohammed AlQuraishi

Home Cities About Blog About Me

July 25, 2021

12

The AlphaFold2 Method Paper: A Fount of Good Ideas

Just over a week ago the long-awaited AlphaFold2 (AF2) method paper and associated code finally came out, putting to rest questions that I and many others raised about public disclosure of AF2. Already, the code is being pushed in all sorts of interesting ways, and three days ago the companion paper and database were published, where AF2 was applied to the human proteome and 20 other model organisms. All in all I am very happy with how DeepMind handled this. I reviewed the papers and had some chance to mull over the AF2 model architecture during the past couple of months (it was humorous to see people suggest that the open sourcing of AF2 was in response to RoseTTAFold—it was in fact DeepMind’s plan well before RoseTTAFold was preprinted.) In this post I will summarize my main takeaways about what makes AF2 interesting or surprising. This post is *not* a high-level summary of AF2—for that I suggest reading the main text of the paper, which *is* a well-written high-level summary, or this blog post by Carlos Outeiral. In fact, I suggest that you read the paper, including the supplementary information (SI), before reading this post, as I am going to assume familiarity with the model. My focus here is really on technical aspects of the architecture, with an eye toward generalizable lessons that can be applied to other molecular problems.

Reaction to AlphaFold2 publication

References

[1] Jumper, J., Evans, R., Pritzel, A. *et al.* Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589 (2021).

[2] Dill, K. A., Ozkan, S. B., Weikl, T. R., Chodera, J. D. & Voelz, V. A. The protein folding problem. *Annu. Rev. Biophys.* 37, 289–316 (2008).

[3] Kohl, S. Highly Accurate Protein Structure Prediction with AlphaFold. Heidelberg AI Talk (YouTube livestream, 5 May 2022).

[4] Eddy, S. R. Accelerated profile HMM searches. PLoS Comput. Biol. *7*, e1002195 (2011).

[5] Remmert, M., Biegert, A., Hauser, A. & Söding, J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat. Methods *9*, 173–175 (2012).

[6] wwPDB Consortium. Protein Data Bank: the single global archive for 3D macromolecular structure data. Nucleic Acids Res. *47*, D520–D528 (2019).

[7] Xie, Q., Luong, M.-T., Hovy, E. & Le, Q. V. Self-training with Noisy Student improves ImageNet classification. In Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) 10687–10698 (2020).

[8] Jumper, J. *et al.* Applying and improving AlphaFold at CASP14. Proteins Struct. Funct. Bioinform. *89*, 1711–1721 (2021).

[9] Tsaban, T. *et al.* Harnessing protein folding neural networks for peptide–protein docking. Nat. Commun. *13*, 176 (2022).

[10] Evans, R. *et al.* Protein complex prediction with AlphaFold-Multimer. bioRxiv 2021.10.04.463034 (2021).

[11] Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. Nat. Methods *19*, 679–682 (2022).