AlphaFold 2

A paradigm shift in structural biology

Francisco J. Palmero Moya

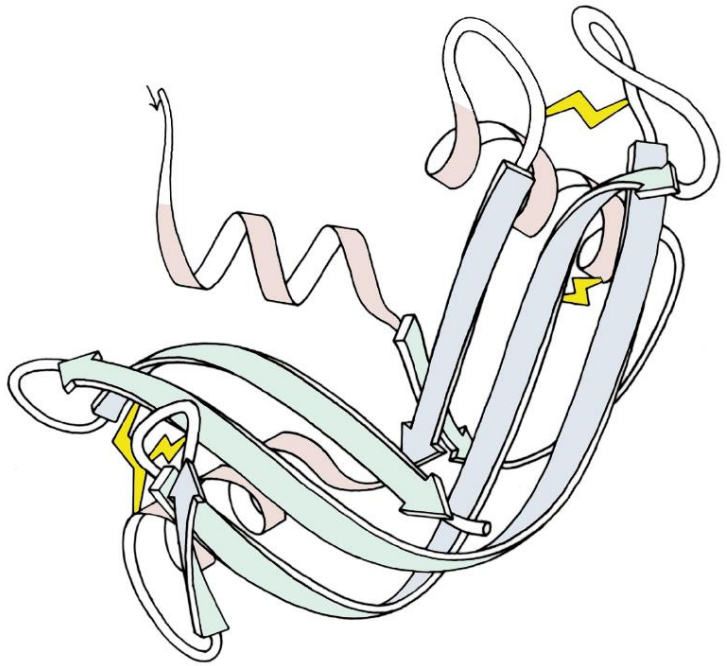
Nynke Dekker Lab

13/10/2023

TU Delft

Context:
Protein folding process

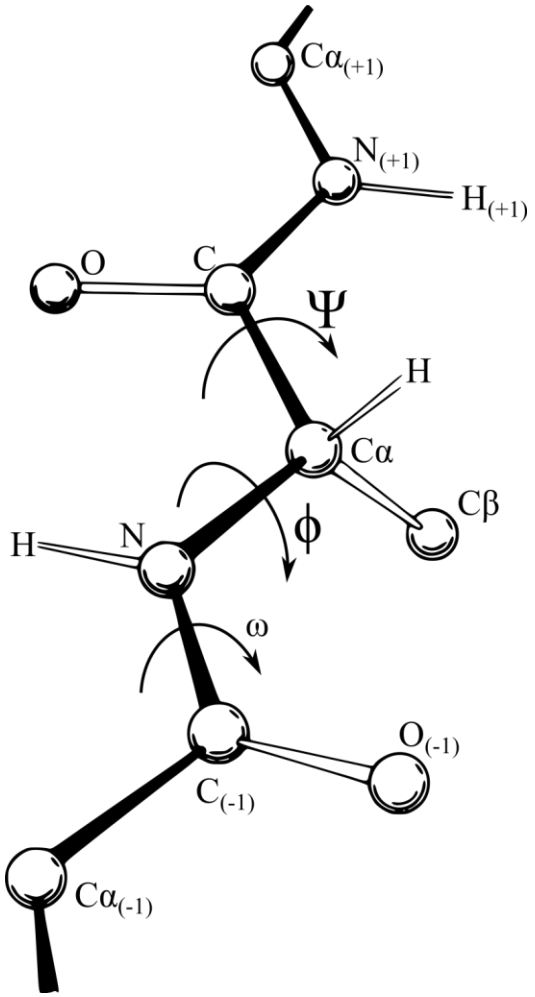
- Proteins are the workhorses of the cell, carrying out essential functions which are mainly determined by its structure.
- They are initially synthesized as linear chains of amino acids that spontaneously fold into complex 3D structures.
- The resulting 3D structure is determined by the amino-acid sequence (Anfinsen's dogma).
- Understanding and simulating the protein folding process has been an important challenge for computational biology.

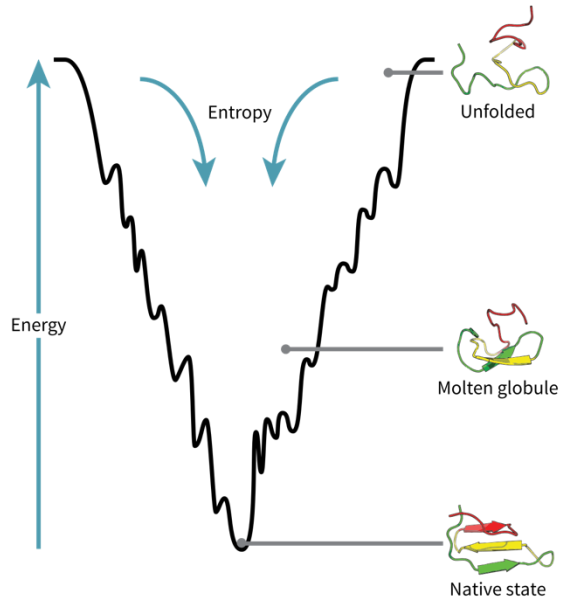


Ribonuclease A

Context:
Protein folding problem

- Proteins have an enormous number of possible conformations due to the very large degrees of freedom.
- A random sequential sampling of all these conformations would take an impractically long time (Levinthal's Paradox).
- It was proposed that proteins don't fold randomly but instead progress through a series of stable intermediate states.
- The configuration space of a protein during folding can be visualized as an energy landscape.
- Proteins are thought to fold via various pathways and intermediates rather than a single mechanism.





Aim:
Protein structure prediction

Structures of around 100,000 unique proteins have been determined, but this represents a small fraction of the billions of known protein sequences.

Some experimental techniques employed for studying protein folding are:

X-ray crystallography, fluoresce spectroscopy, nuclear magnetic resonance, etc.

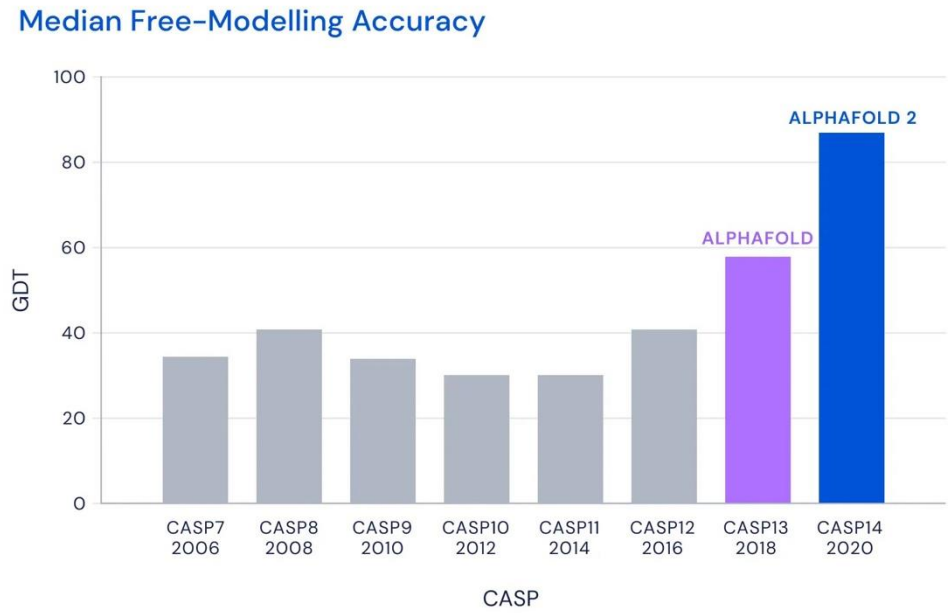
Accurate computational approaches are needed to address this gap. AlphaFold 2 provides a computational method that can regularly predict protein structure with atomic accuracy.

Contents

Critical Assessment of Structure Prediction

CASP provides research groups with an opportunity to objectively test their structure prediction methods.

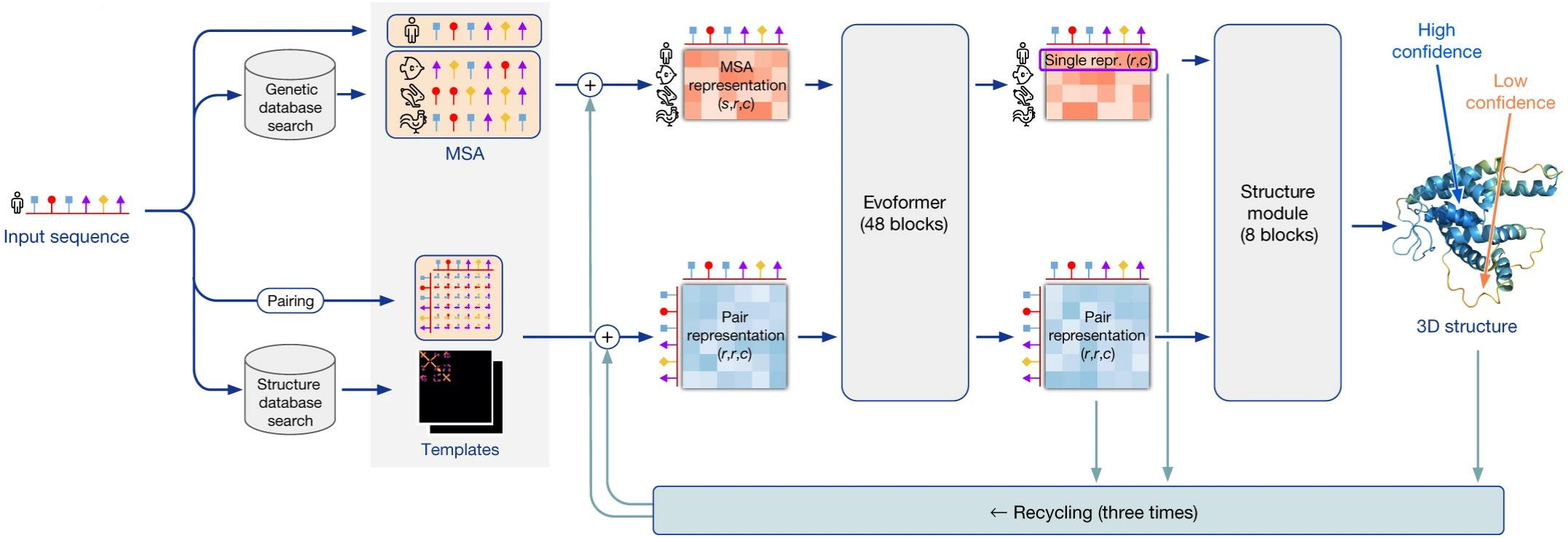
Neither the predictors nor the organizers and assessors know the structures of the target proteins at the time when predictions are made.



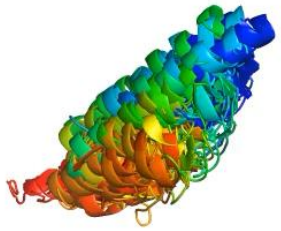
AlphaFold 2

How does it work?

AlphaFold 2 Network Topology



Intermediate structure trajectory : RNA polymerase of crAss-like phage



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

Inputs

Sequence databases

- UniRef90[6] (JackHMMER[3])
- BFD[5] (HHblits[4])
- MGnify clusters[2] (JackHMMER[3])

Structural databases

- PDB[1] (training)
- PDB70 clustering (hhsearch[4])

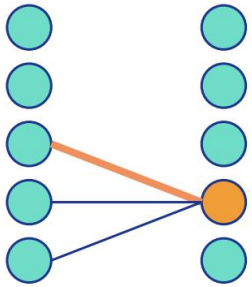
All publicly available data.

[1] *Berman et al.*, Nature Structural Biology (2003) doi:10.1038/nsb1203-980
[2] *Mitchell et al.*, Nucleic Acids Research (2019) doi:10.1093/nar/gkz1035
[3] *Potter et al.*, Nucleic Acids Research (2018) doi:10.1093/nar/gky448
[4] *Steinegger et al.*, BMC Bioinformatics (2019) doi:10.1186/s12859-019-3019-7
[5] *Steinegger et al.*, Nature Methods (2019) doi:10.1038/s41592-019-0437-4
[6] *Suzek et al.*, Bioinformatics (2015) doi:10.1093/bioinformatics/btu739

Visualisations:
The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC.
AS Rose, et al., Bioinformatics (2018) doi:10.1093/bioinformatics/bty419

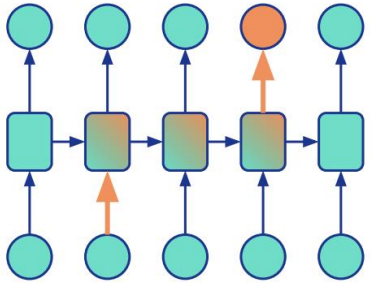
Law of the instrument

Law of the instrument



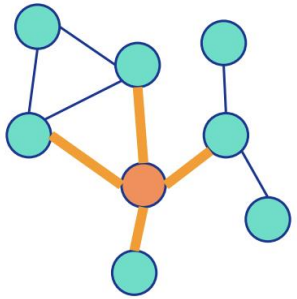
Convolutional Networks (e.g. computer vision)

- data in regular grid
- information flow to local neighbours



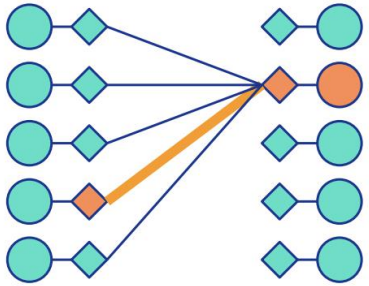
Recurrent Networks (e.g. language)

- data in ordered sequence
- information flow sequentially



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

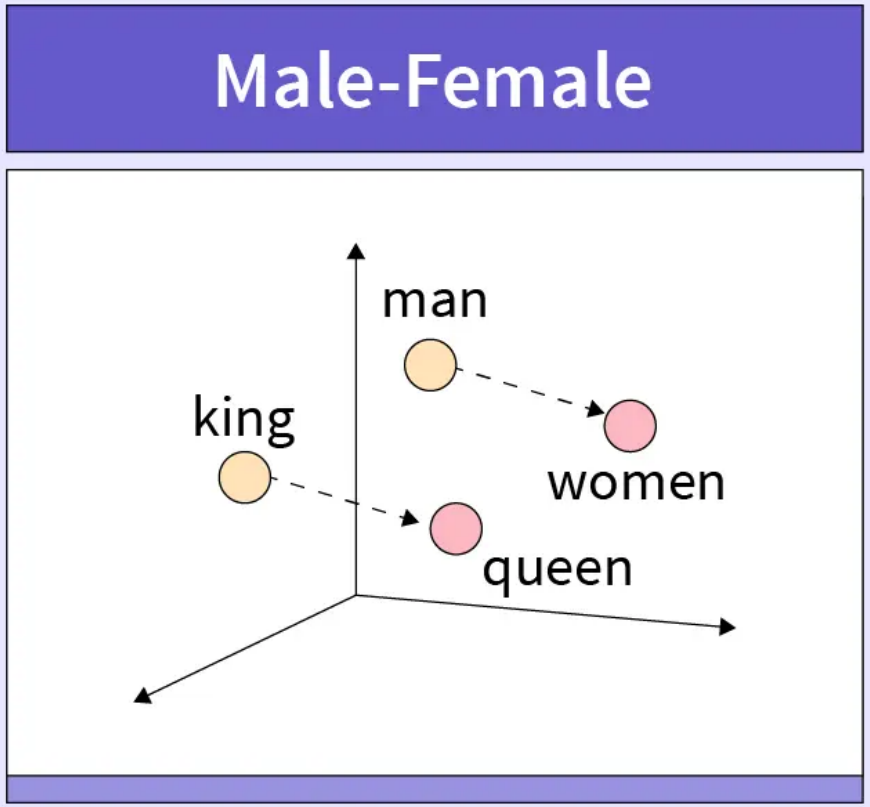
- data in fixed graph structure
- information flow along fixed edges



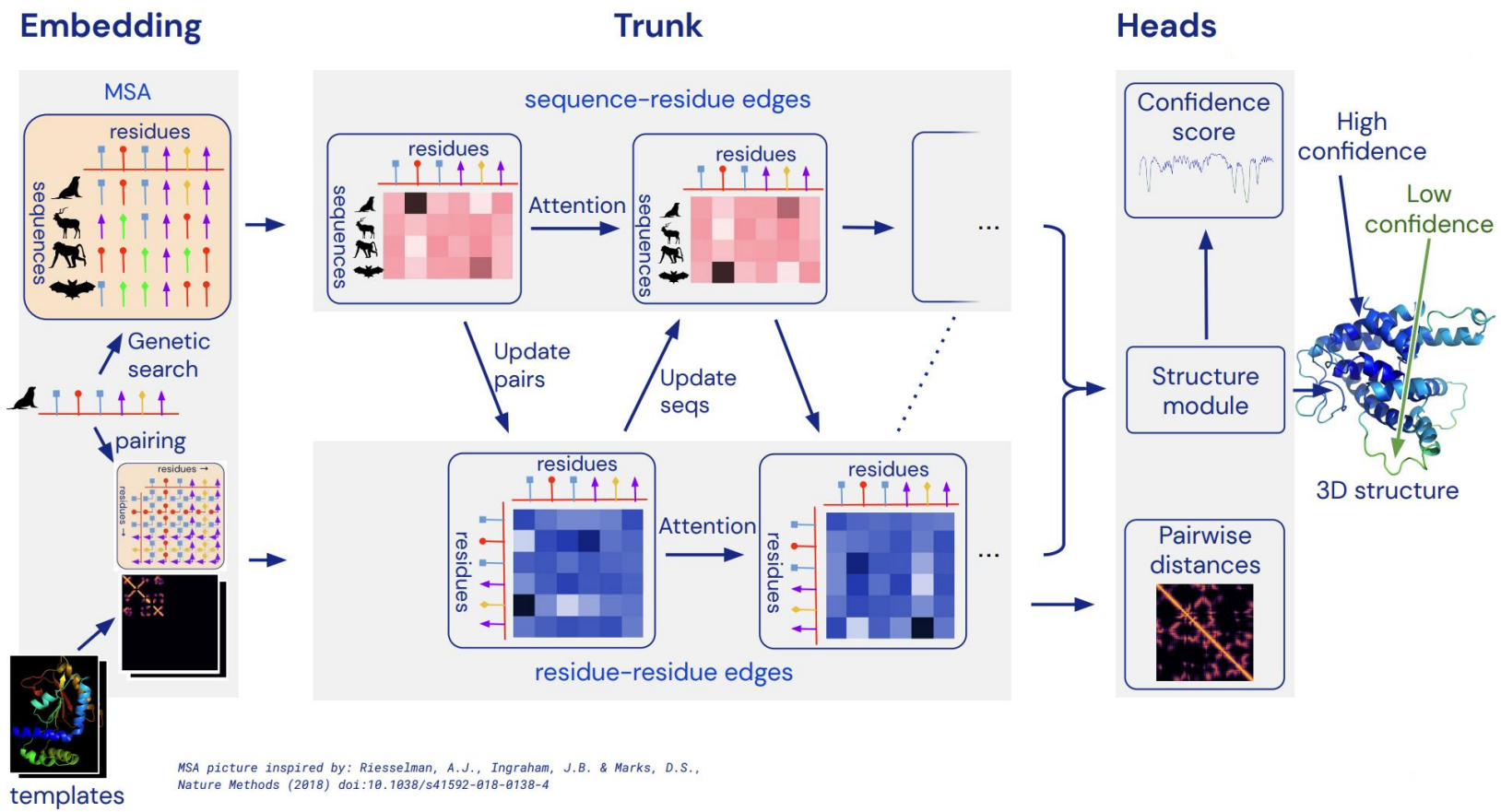
Attention Module (e.g. language)

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

Embedding

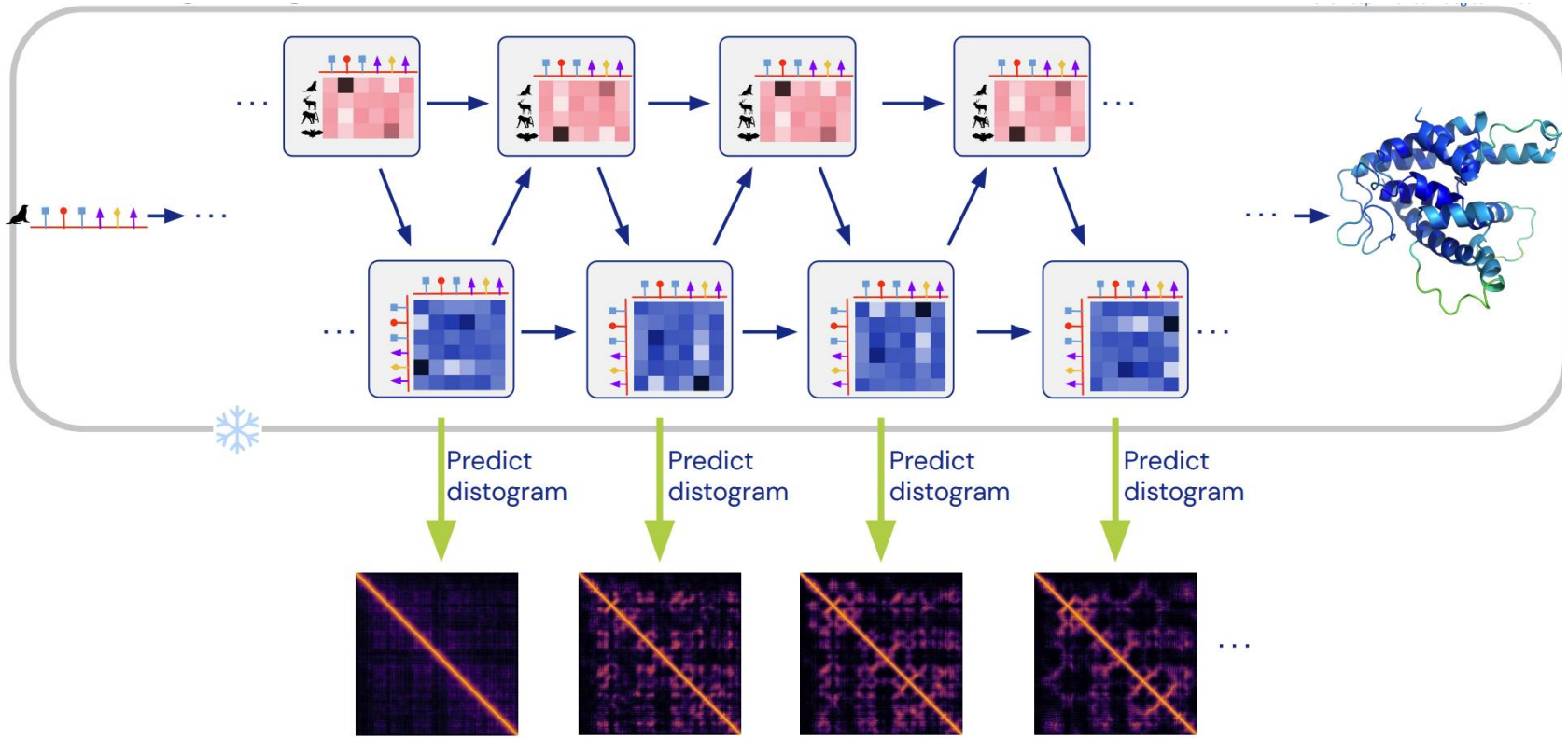


Evoformer

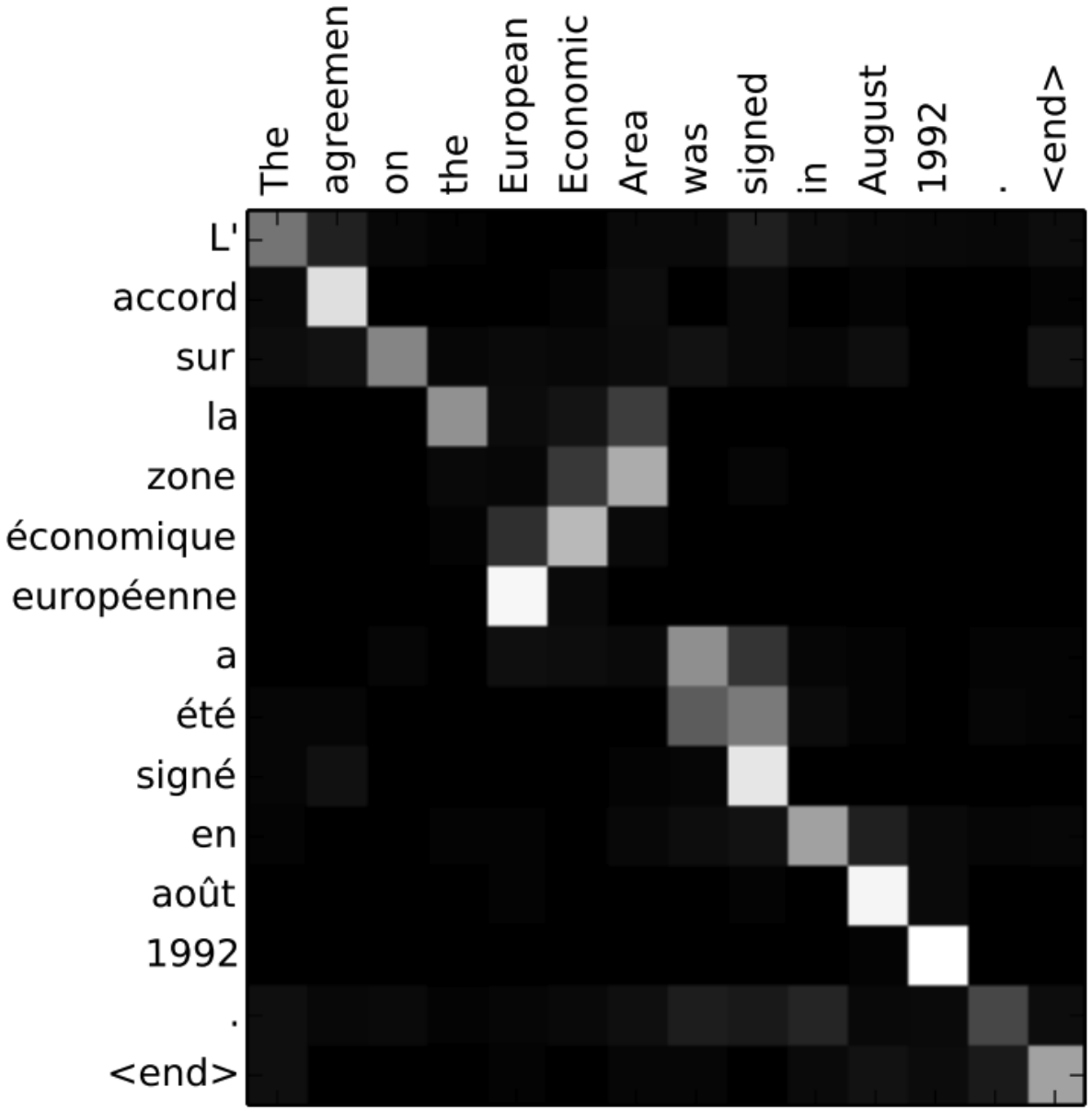


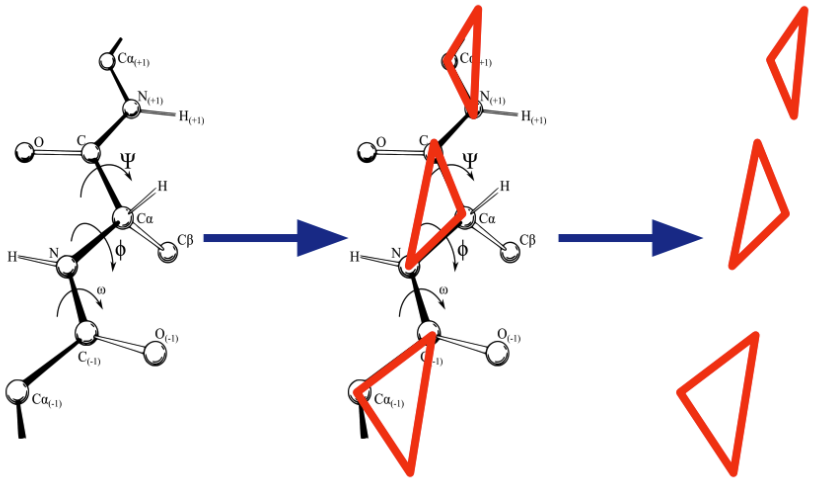
MSA picture inspired by: Riesselman, A.J., Ingraham, J.B. & Marks, D.S., Nature Methods (2018) doi:10.1038/s41592-018-0138-4

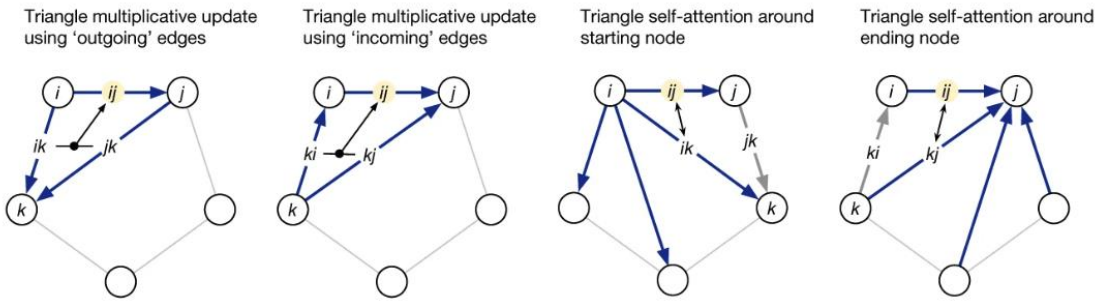
Interrogating the network



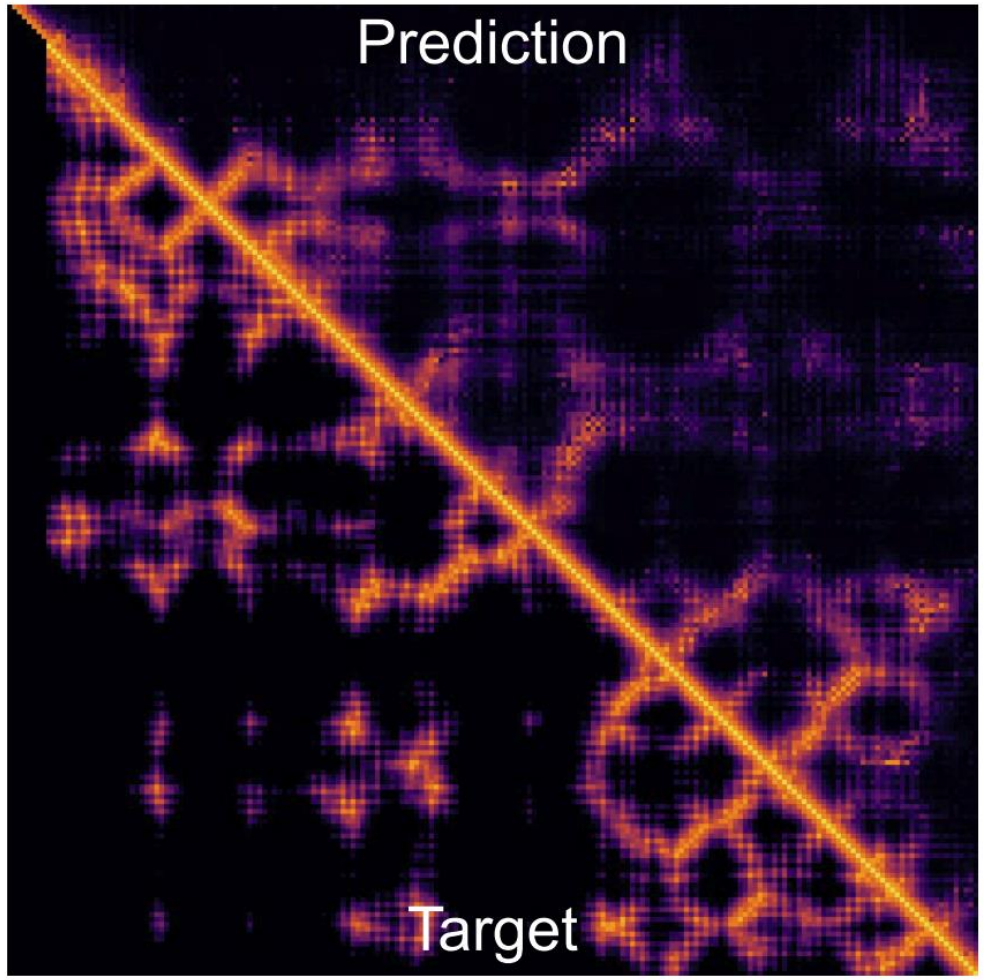
Transformers: Attention is all you need







Structure module


Prediction
Target

Metrics

Frame Aligned Point Error

- Similar to root-mean-squared deviation
- It is not invariant to reflection and thus prevents creating proteins of the wrong chirality.

Distogram loss

- Compare distogram prediction with ground truth

MSA masking

- Some symbols in MSA are masked out and the model is asked to predict them.

Impact

Access & Citations

1.28m	8669	9279
Article Accesses	Web of Science	CrossRef

Citation counts are provided from Web of Science and CrossRef. The counts may vary by service, and are reliant on the availability of their data. Counts will update daily once available.

Online attention

- 5321 tweeters
- 300 news outlets
- 11 Video uploaders
- 12886 Mendeley
- 74 blogs
- 7 Redditors
- 26 Wikipedia page
- 5 Facebook pages
- 15 F1000

This article is in the 99th percentile (ranked 97th) of the 426,750 tracked articles of a similar age in all journals and the 98th percentile (ranked 16th) of the 927 tracked articles of a similar age in *Nature*

View more on Altmetric

Altmetric calculates a score based on the online attention an article receives. Each coloured thread in the circle represents a different type of online attention. The number in the centre is the Altmetric score. Social media and mainstream news media are the main sources that calculate the score. Reference managers such as Mendeley are also tracked but do not contribute to the score. Older articles often score higher because they have had more time to get noticed. To account for this, Altmetric has included the context data for other articles of a similar age.

Impact

The DeepMind team published their method and made it freely available.

There is an online (free) tool called ColabFold where you can make predictions by making use of Google Computers.

CASP 15 (2022) was plenty of AlphaFold similar structures but DeepMind did not participate.

Conclusions

- High-accuracy.
- It became a popular tool in biology.
- It may be over-engineered, meaning a simplified version achieved similar results.
- Open the doors to other problems such as the reverse problem: given a known structure which are the possible sequences that could generate it.

Thank you!

Francisco J. Palmero Moya