

Article

Highly accurate protein structure prediction with AlphaFold

<https://doi.org/10.1038/s41586-021-03819-2>

Received: 11 May 2021

Accepted: 12 July 2021

Published online: 15 July 2021

Open access



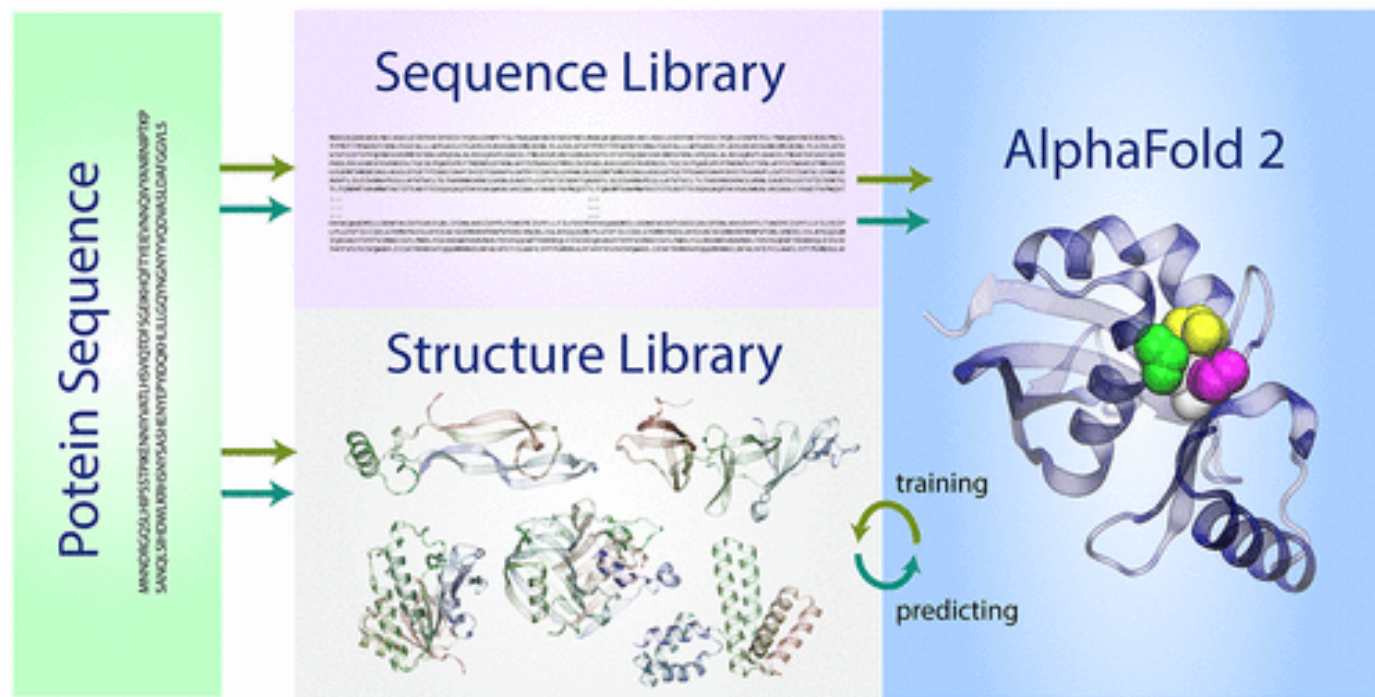
Check for updates

John Jumper^{1,4}✉, Richard Evans^{1,4}, Alexander Pritzel^{1,4}, Tim Green^{1,4}, Michael Figurnov^{1,4}, Olaf Ronneberger^{1,4}, Kathryn Tunyasuvunakool^{1,4}, Russ Bates^{1,4}, Augustin Židek^{1,4}, Anna Potapenko^{1,4}, Alex Bridgland^{1,4}, Clemens Meyer^{1,4}, Simon A. A. Kohl^{1,4}, Andrew J. Ballard^{1,4}, Andrew Cowie^{1,4}, Bernardino Romera-Paredes^{1,4}, Stanislav Nikolov^{1,4}, Rishub Jain^{1,4}, Jonas Adler¹, Trevor Back¹, Stig Petersen¹, David Reiman¹, Ellen Clancy¹, Michal Zielinski¹, Martin Steinegger^{2,3}, Michalina Pacholska¹, Tamas Berghammer¹, Sebastian Bodenstein¹, David Silver¹, Oriol Vinyals¹, Andrew W. Senior¹, Koray Kavukcuoglu¹, Pushmeet Kohli¹ & Demis Hassabis^{1,4}✉

Presenter: **Ly thi truc NGUYEN**

BIOPOP-Department of Agricultural Biotechnology,
Seoul National University.

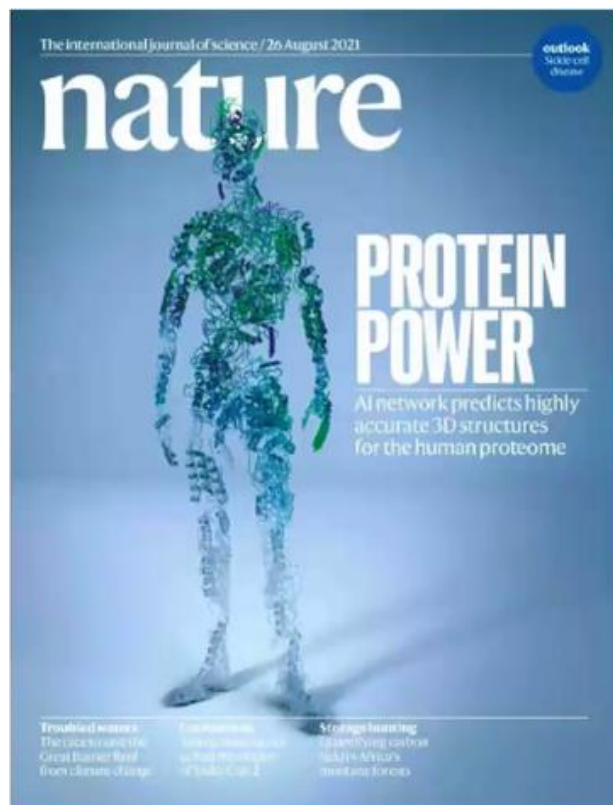
1. Introduction to protein structure prediction
 - Protein: an essential component for life
 - Why protein? What can do with protein structure?
2. Challenges: The 'protein folding problem'
3. How does AlphaFold2 works?
4. Remaining challenges
5. Implementation (using COLAB)



[nature](#) > [news](#) > article

NEWS | 22 July 2019

AI protein-folding algorithms solve structures faster than ever

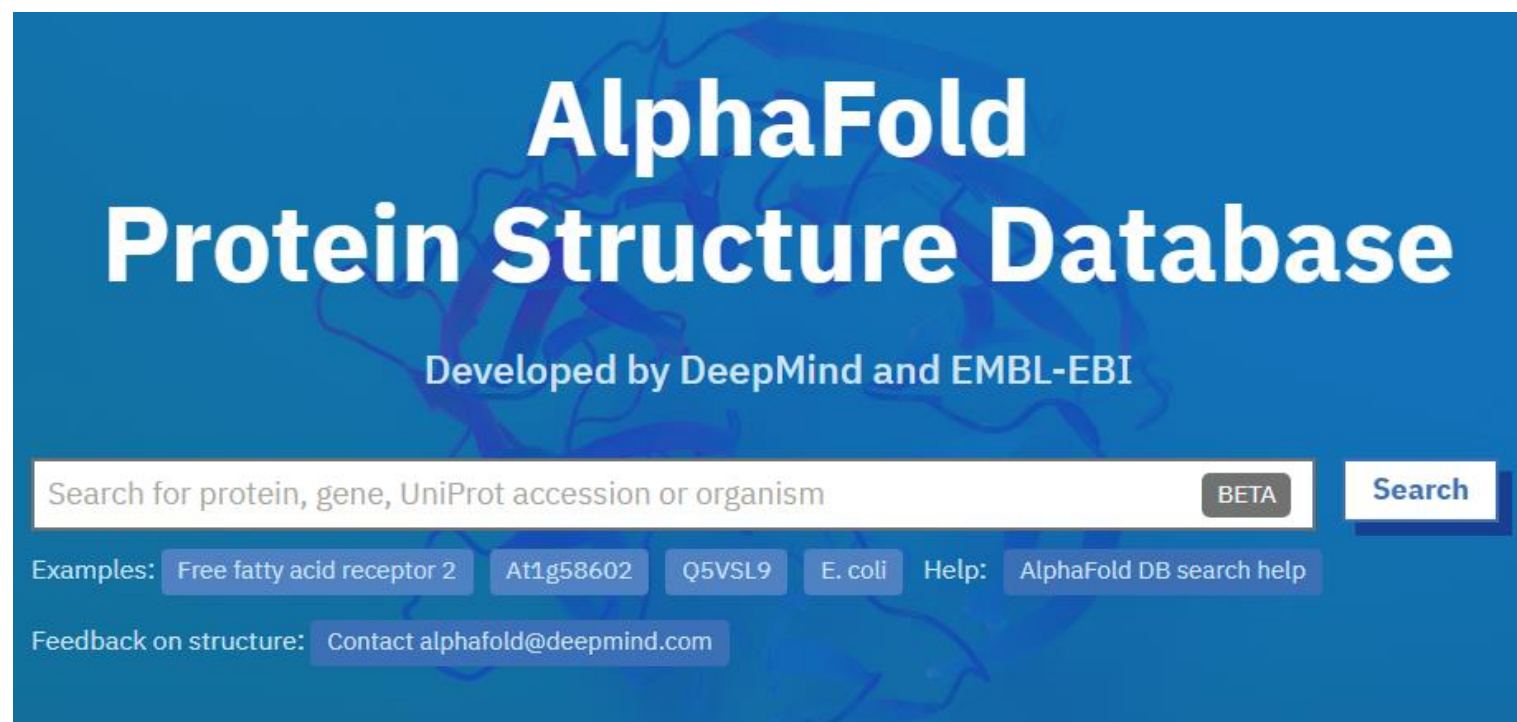


AlphaFold2 (J. Jumper et al, Nature, 2021)

[nature](#) > [news](#) > article

NEWS | 30 November 2020

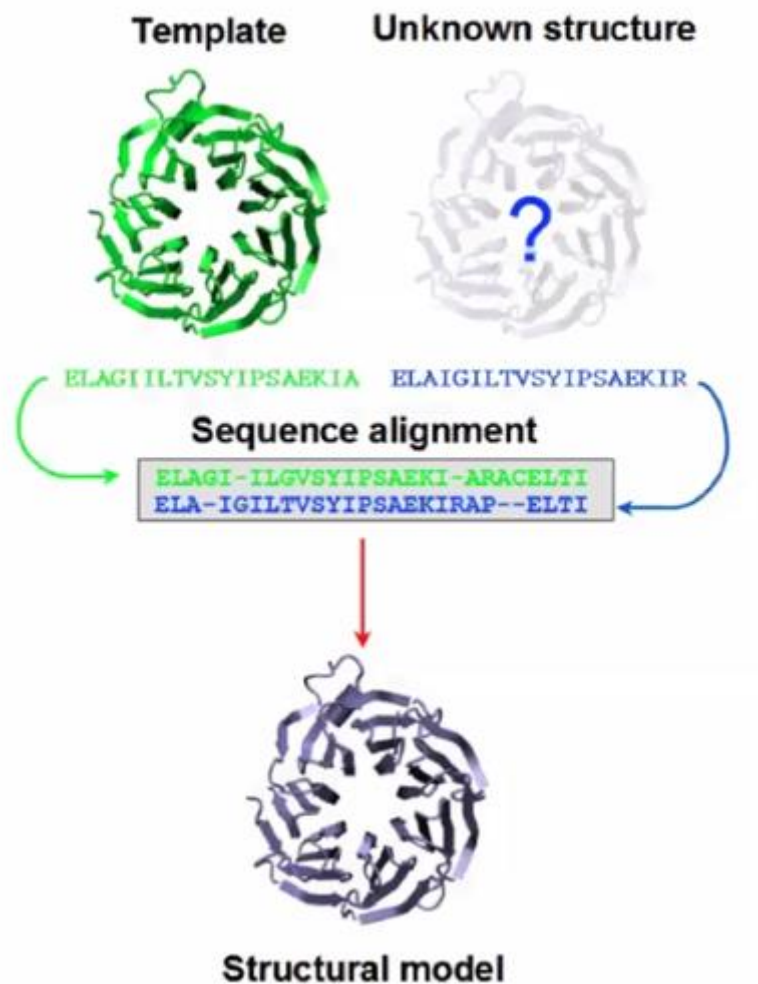
‘It will change everything’: DeepMind’s AI makes gigantic leap in solving protein structures



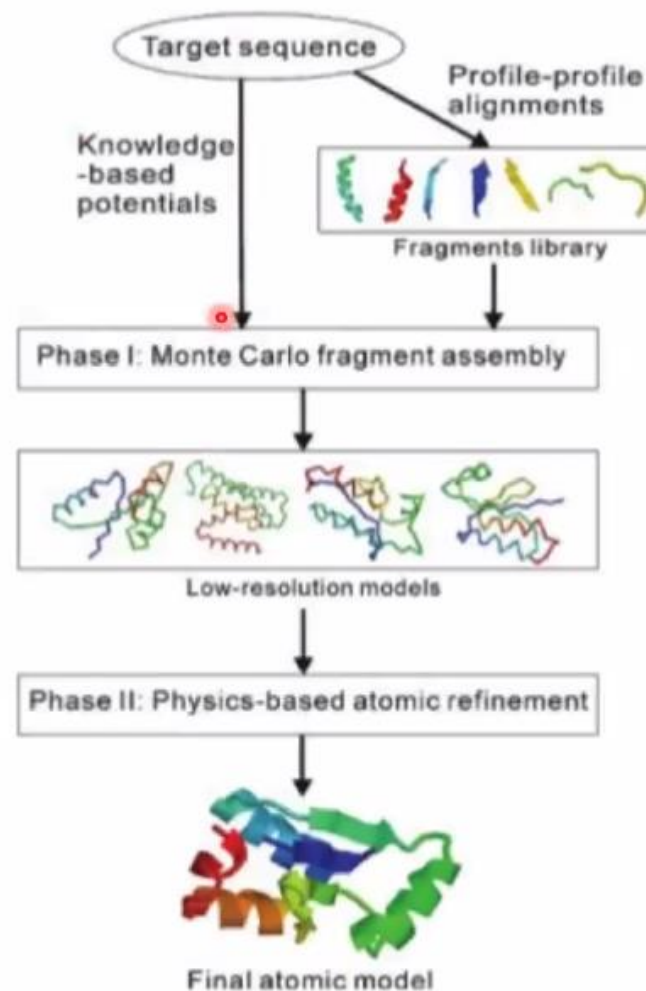
<https://alphafold.ebi.ac.uk/>

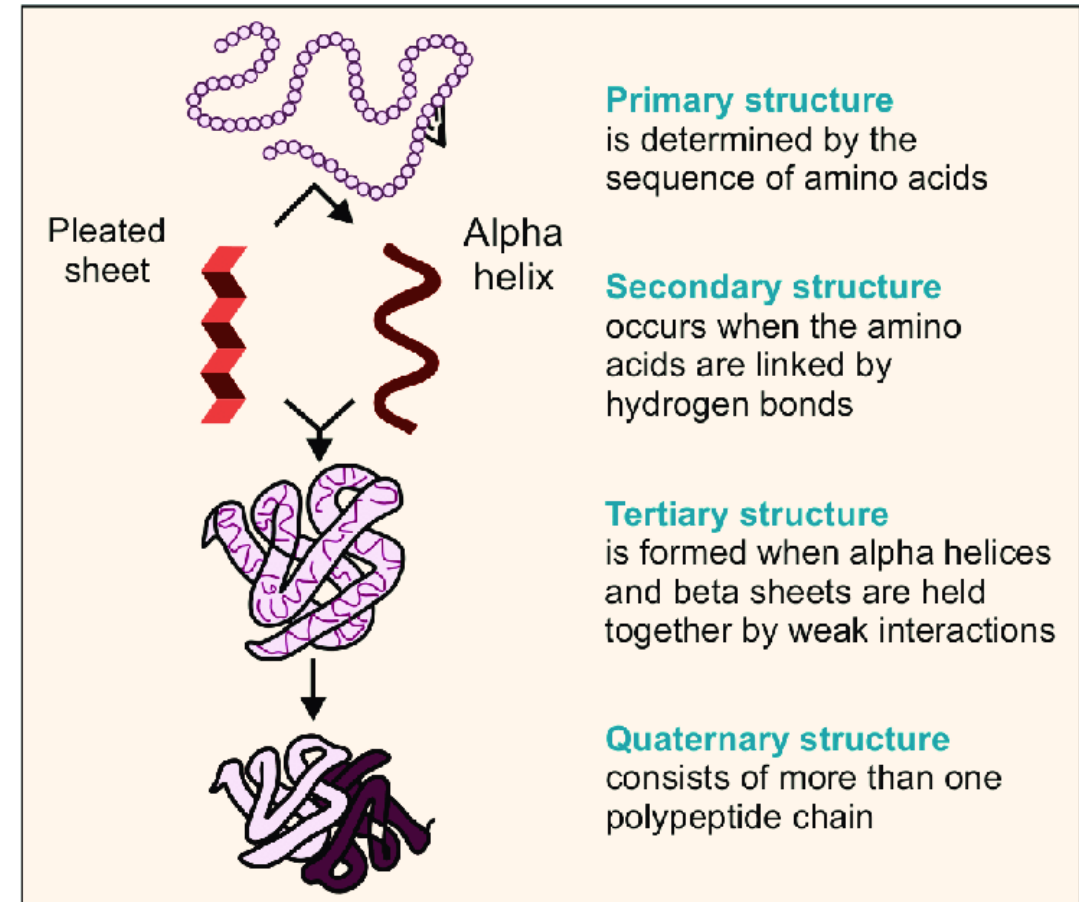
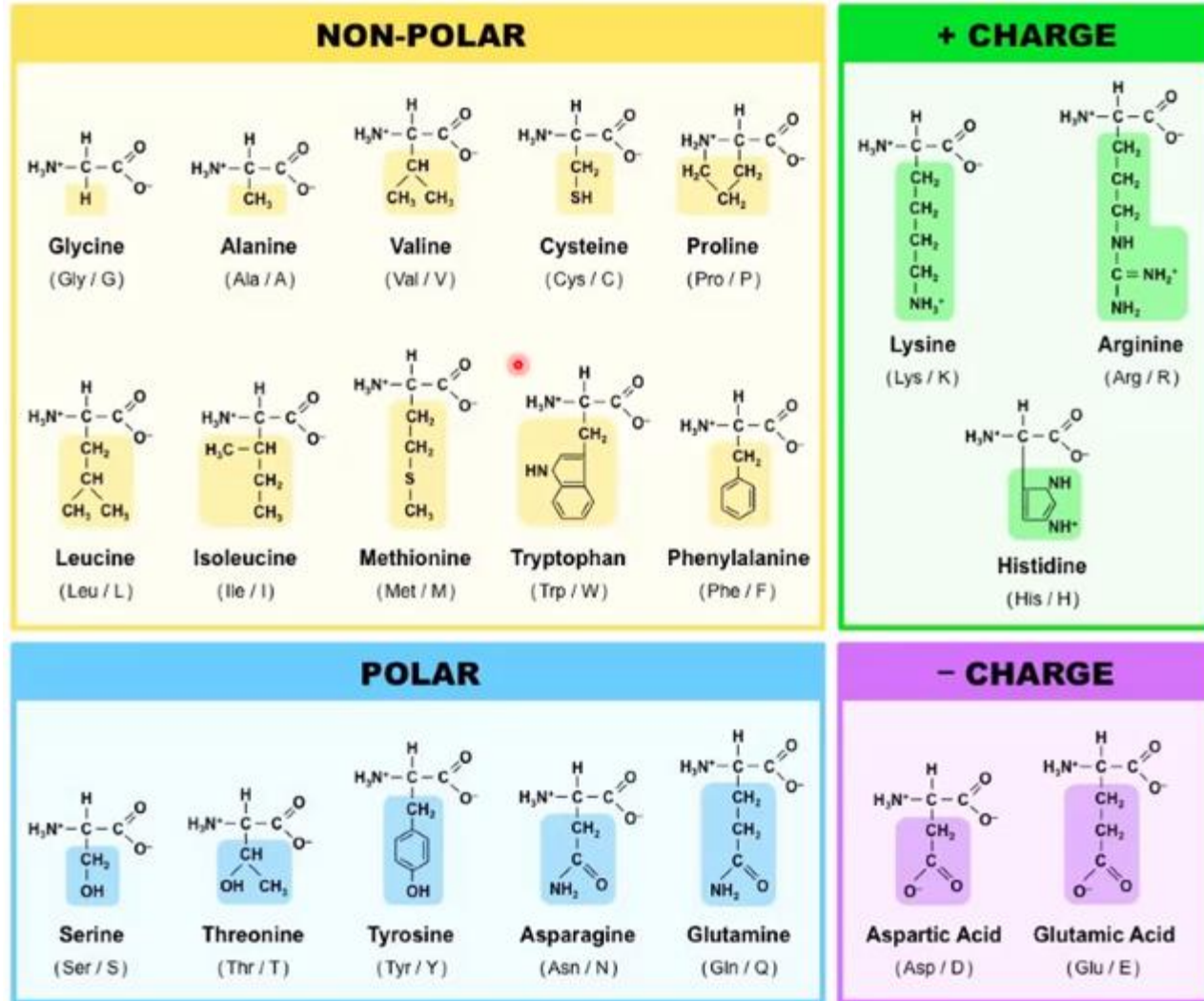
Basic assumption: Similar sequences \rightarrow similar structures

Template-based Modeling



(Template-)Free Modeling

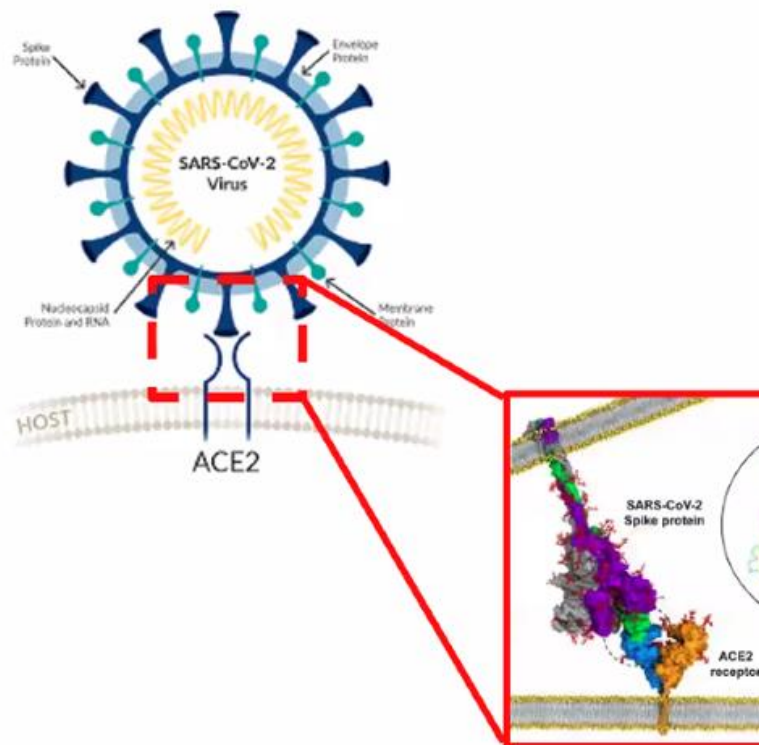




Levels of organizations of proteins

Protein: an essential component for life

Better understanding of biological



[Credit: University of California, Berkeley; Istanbul Techni

3:53

Thread

You Retweeted



Protein Data Bank
@PDBeurope

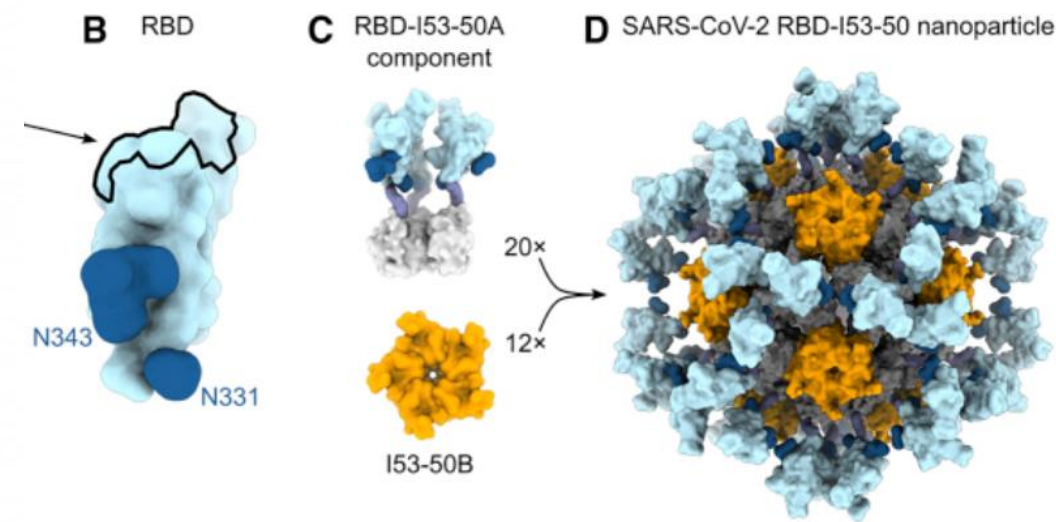
The first 2019-nCoV [#coronavirus](#) structure of the main protease has been released today (6LU7). This is a result of a huge effort from Prof. Rao's group [@ShanghaiTechUni](#). For more information on the structure and related data available in the PDB visit wwpdb.org/news/news?year...



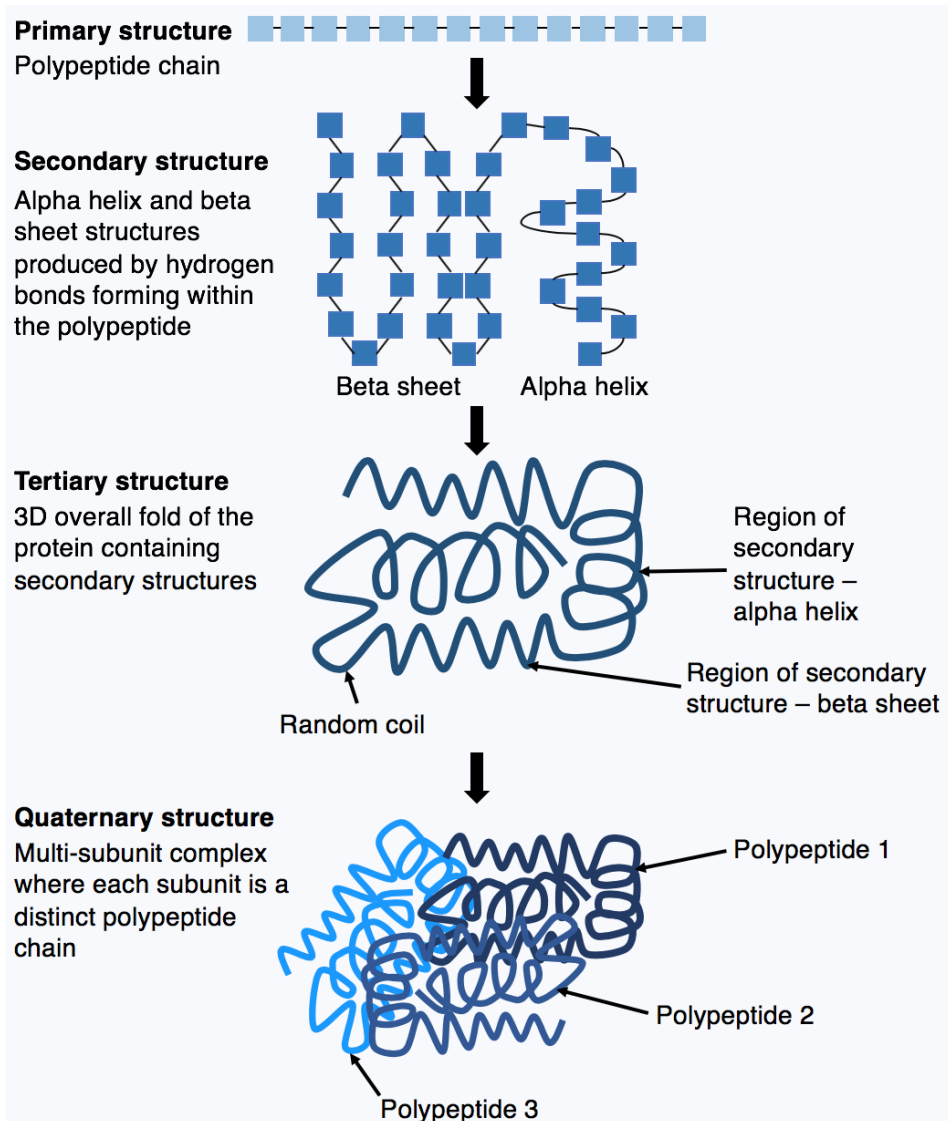
1:15 PM · 05 Feb 20 · Buffer

Tweet your reply

oping drugs / vaccines / biosensors



Walls, A.C., et al 2020 cell, 183(5) 1367-1382



The “protein folding problem” consists of three related puzzles:

- What is the folding code?
- What is the folding mechanism?
- Can we predict the native structure of a protein from its amino acid sequences?



AlphaFold has solved one of biology's grand challenges: “Protein folding”.

“We have been stuck on this one problem – how do proteins fold up – for nearly 50 years.”

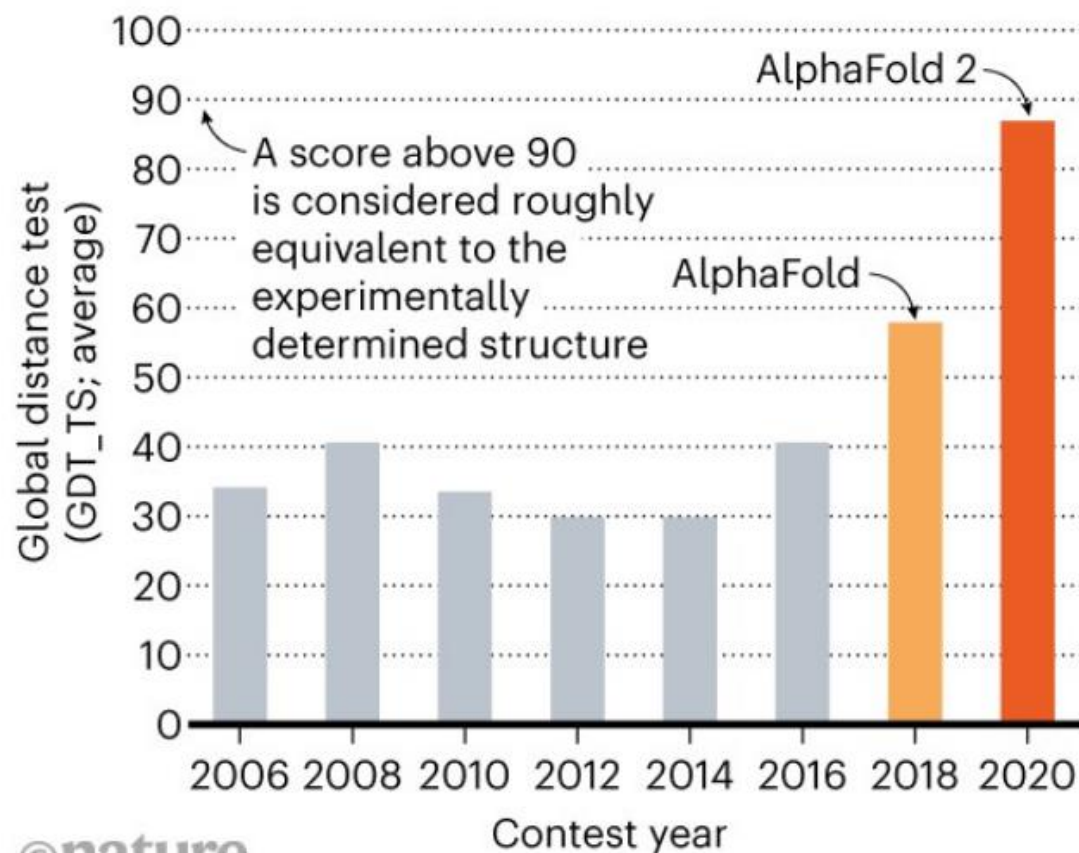
Professor John Moulton, Co-founder and Chair of CASP, University of Maryland



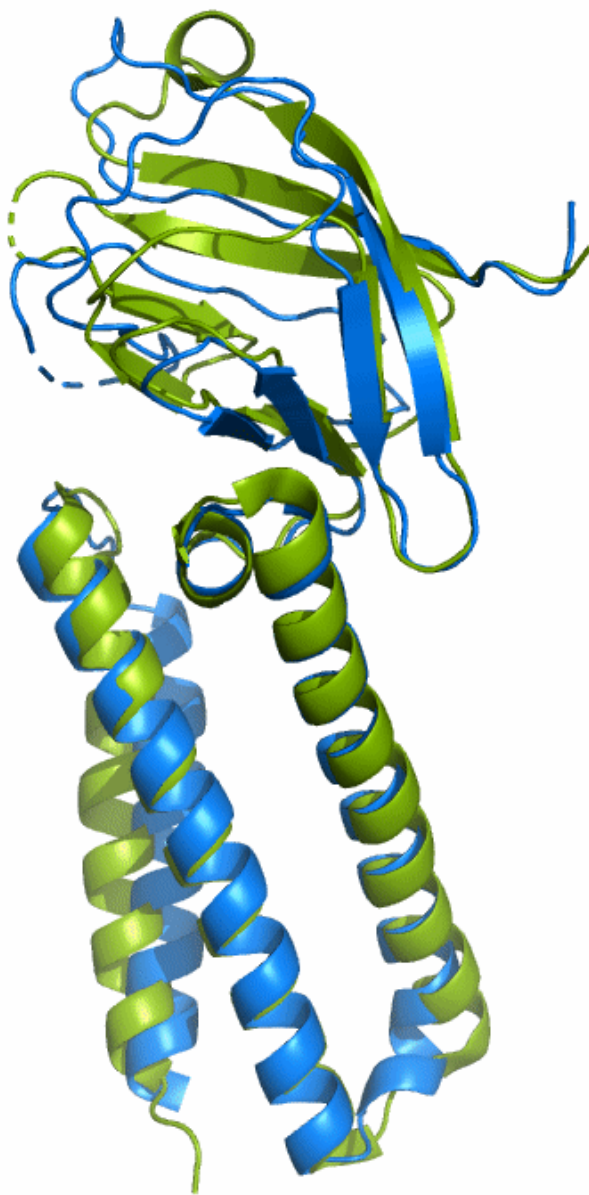
- an AI system developed by DeepMind
- Predicts a protein's 3D structure from its amino acid sequence.
- AlphaFold2 can predict the shape of a protein in minutes.
- Achieve accuracy competitive with experiment.
- A significant breakthrough and highlights the impact AI can have on science.

STRUCTURE SOLVER

DeepMind's AlphaFold 2 algorithm significantly outperformed other teams at the CASP14 protein-folding contest — and its previous version's performance at the last CASP.



©nature



The best **CASP**_commons prediction by AlphaFold is indicated in **blue**, and the experimental structure is in **green**.

Source: DeepMind

- Dec. 2018: Alphafold1 won CASP13 (CASP: Critical Assessment of protein Structure Prediction)

[nature](#) > [articles](#) > [article](#)

- Jan. 2020: Alphafold1 Published

Article | [Published: 15 January 2020](#)

Improved protein structure prediction using potentials from deep learning

[Andrew W. Senior](#) , [Richard Evans](#), [John Jumper](#), [James Kirkpatrick](#), [Laurent Sifre](#), [Tim Green](#), [Chongli Qin](#), [Augustin Žídek](#), [Alexander W. R. Nelson](#), [Alex Bridgland](#), [Hugo Penedones](#), [Stig Petersen](#), [Karen Simonyan](#), [Steve Crossan](#), [Pushmeet Kohli](#), [David T. Jones](#), [David Silver](#), [Koray Kavukcuoglu](#) & [Demis Hassabis](#)

[Nature](#) **577**, 706–710 (2020) | [Cite this article](#)

121k Accesses | **657** Citations | **629** Altmetric | [Metrics](#)

- Nov. 2020: Alphafold2 won CASP14

- Jul. 2021: Alphafold2 Published, 20.000 human proteins published

[nature](#) > [articles](#) > [article](#)

Article | [Open Access](#) | [Published: 15 July 2021](#)

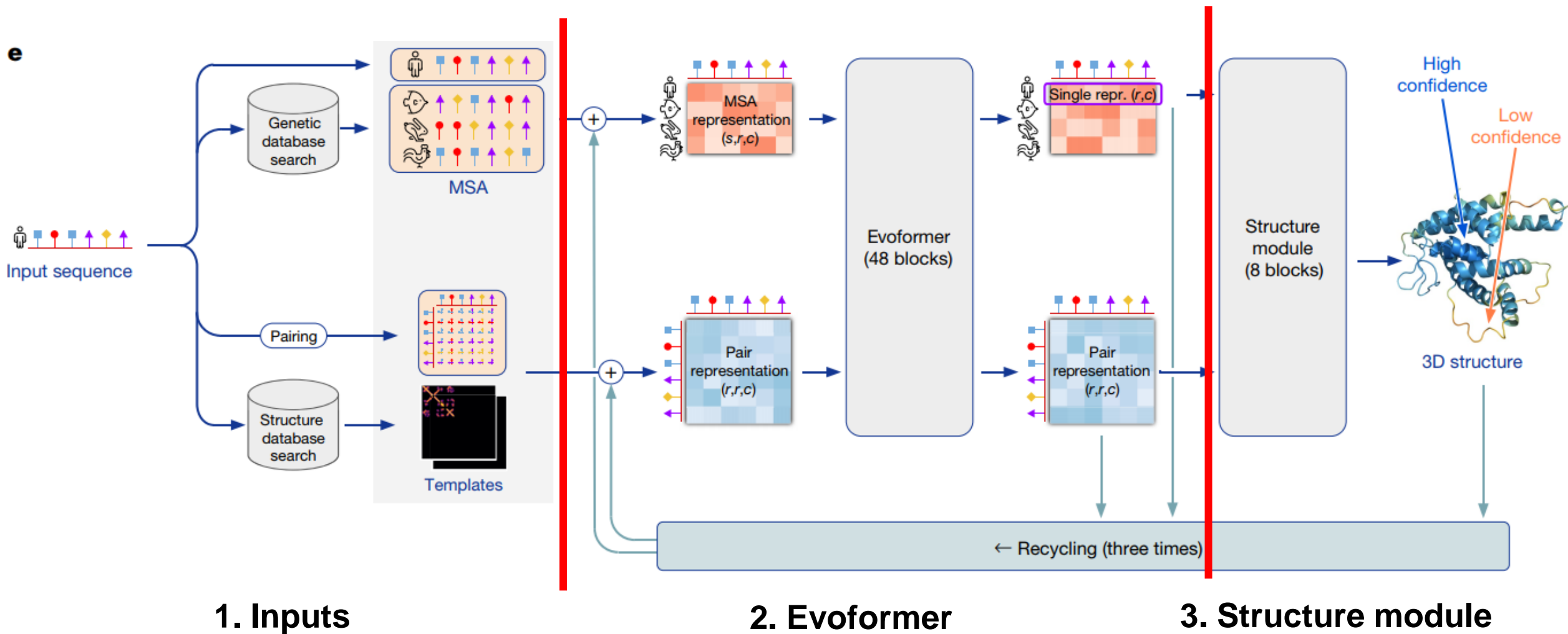
Highly accurate protein structure prediction with AlphaFold

[John Jumper](#) , [Richard Evans](#), [...] [Demis Hassabis](#) 

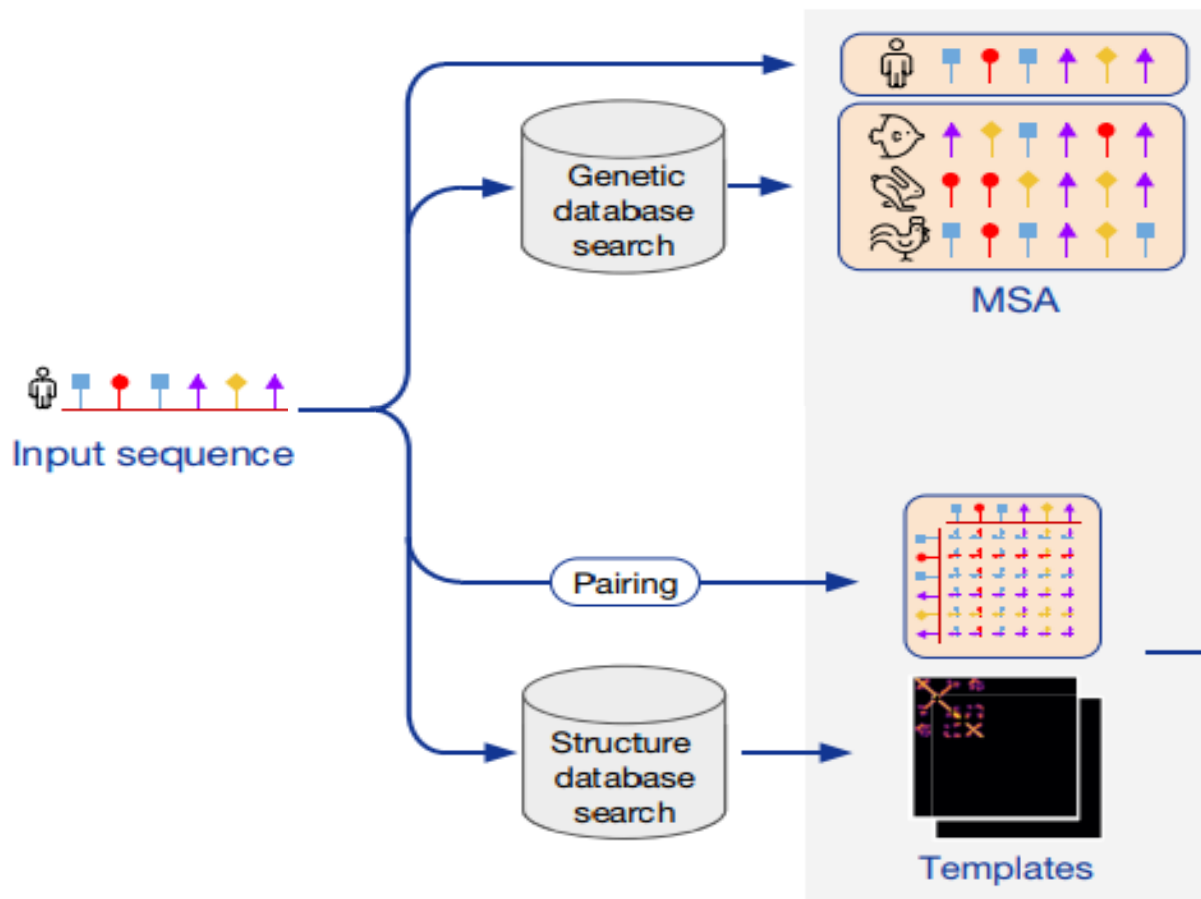
[Nature](#) **596**, 583–589 (2021) | [Cite this article](#)

444k Accesses | **283** Citations | **2798** Altmetric | [Metrics](#)

Schematic



1. Inputs



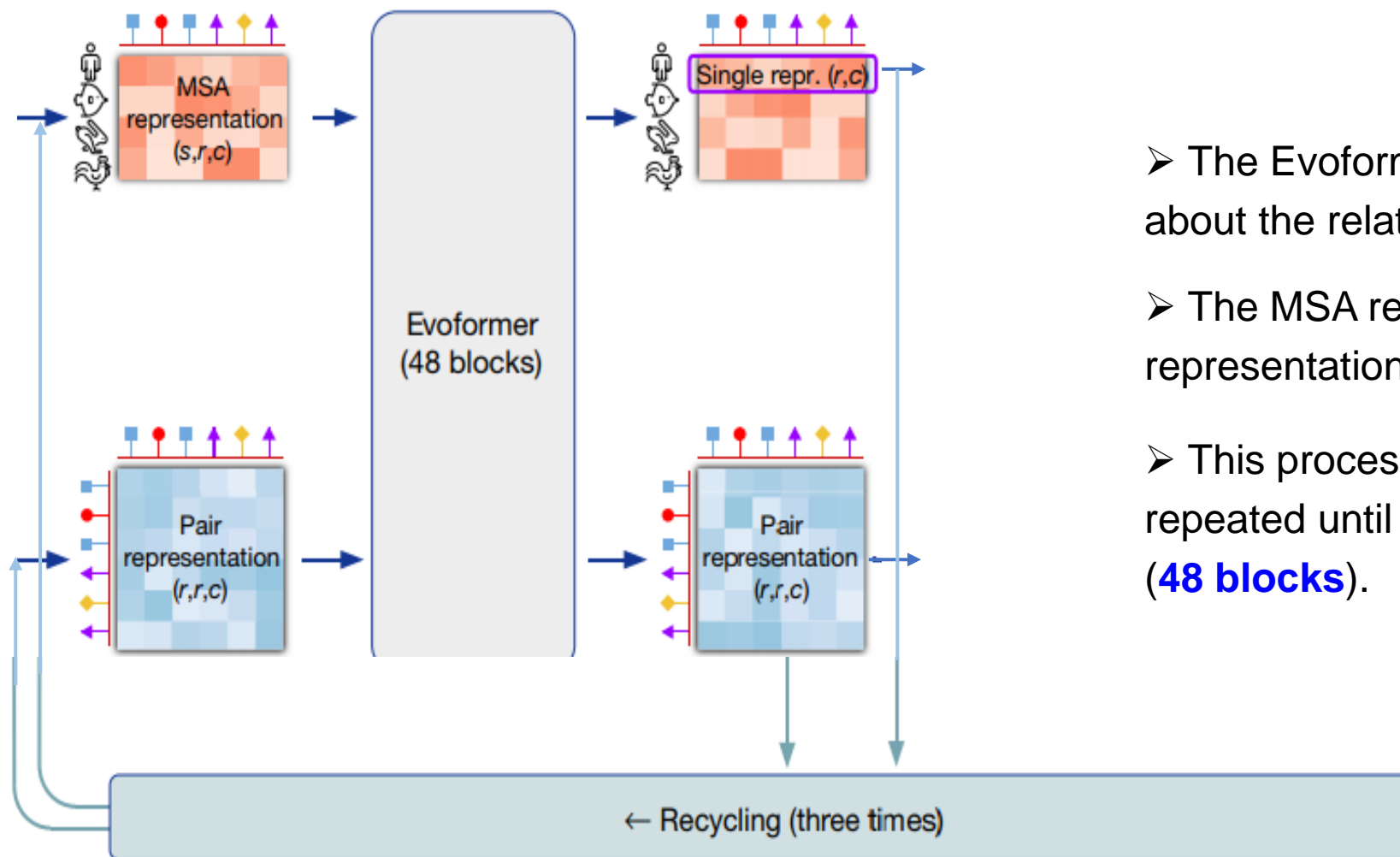
➤ A key AF2 input is the MSA, containing sequences evolutionarily related to the target.

➤ Related sequences are found using standard tools and public databases.

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

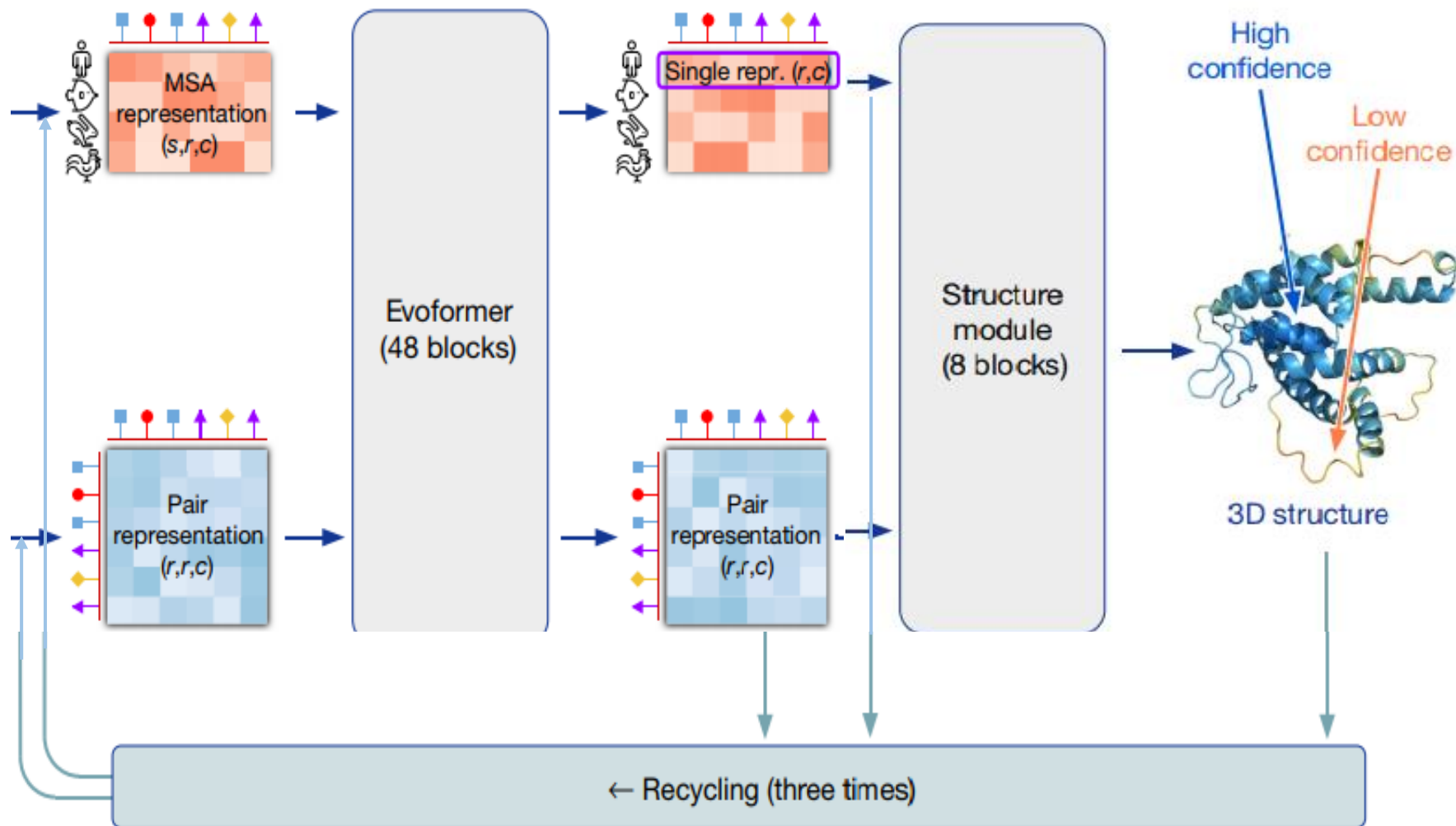
➤ AlphaFold can also use template structures from the PDB, found using standard tools. However, it often produces accurate predictions without a template.

2. Evoformer



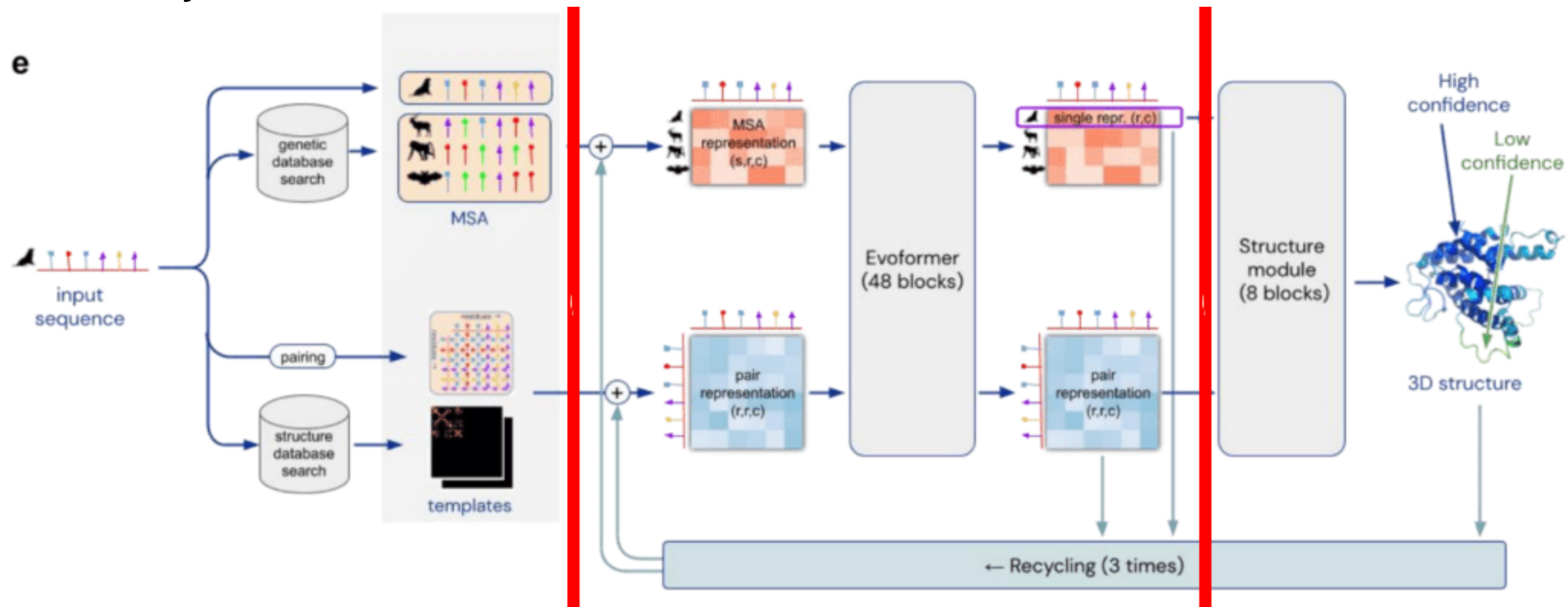
- The Evoformer blocks extract information about the relationship between residues.
- The MSA representation can update the pair representation and vice versa.
- This process is organised in blocks that are repeated until a specified number of cycles (**48 blocks**).

3. Structure module



- Feeding certain outputs back through the network again improves performance.

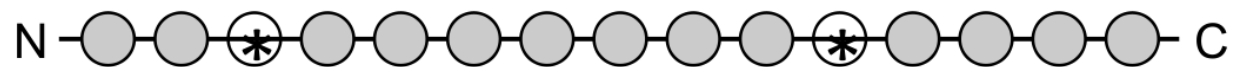
Summary



1. Inputs

2. Evoformer

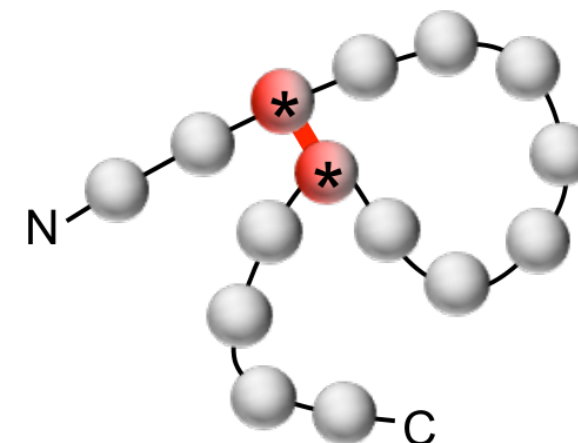
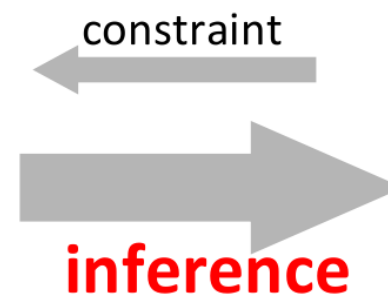
3. Structure module



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

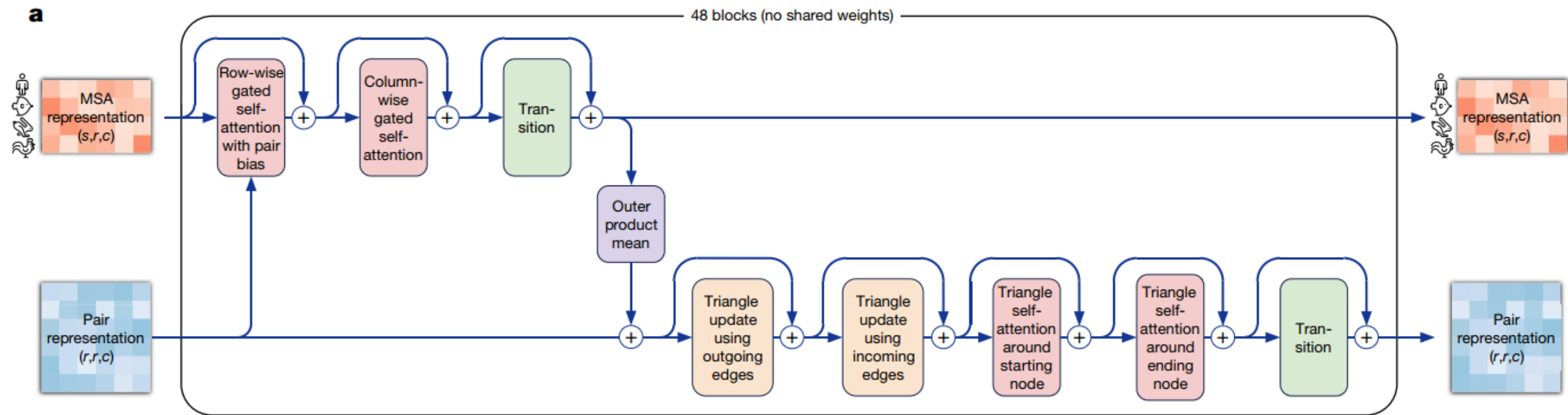


correlated

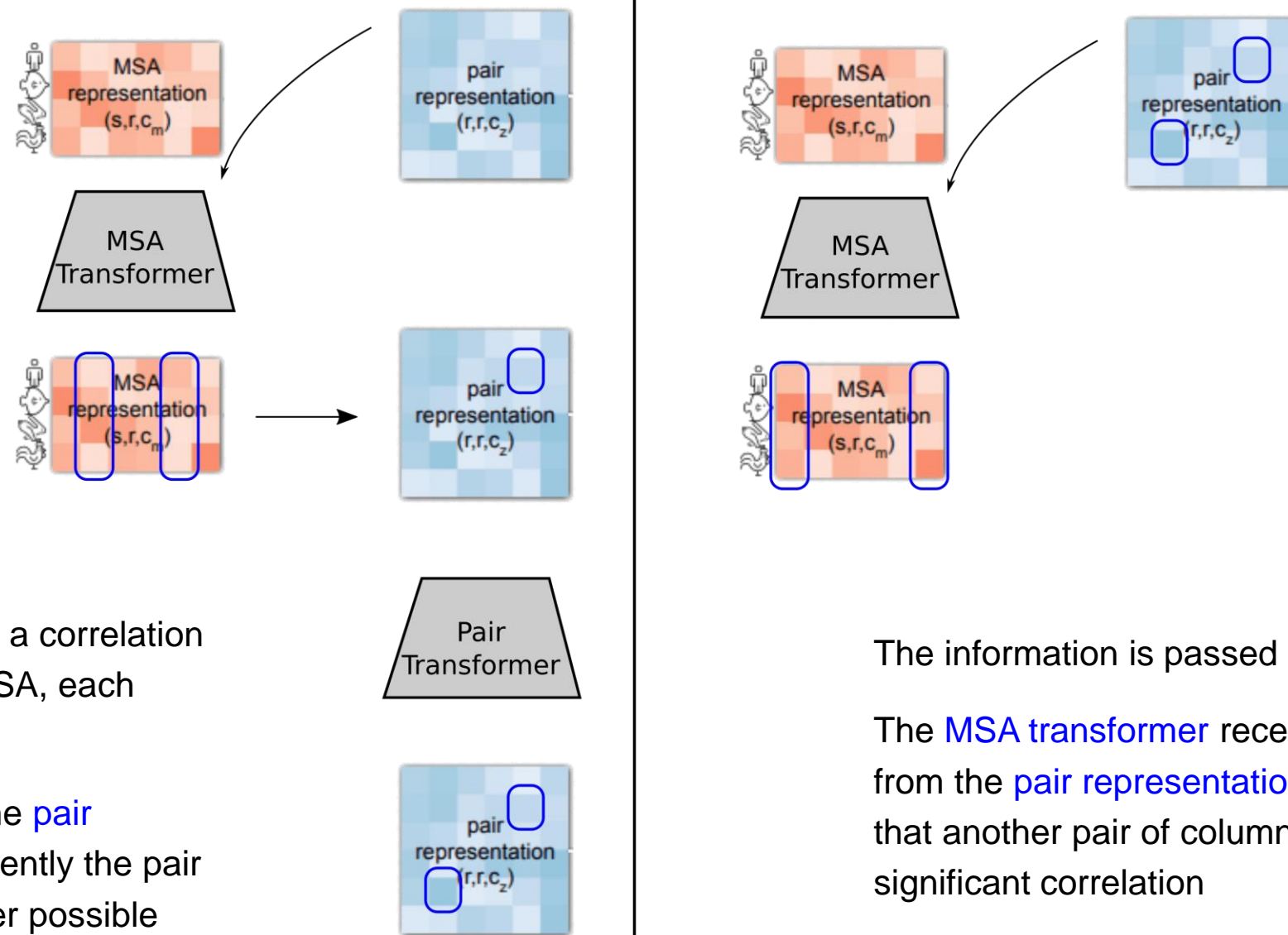


contact in 3D

Correlated mutations carry information about distance relationships in protein structure.



Architectural details - Evoformer block

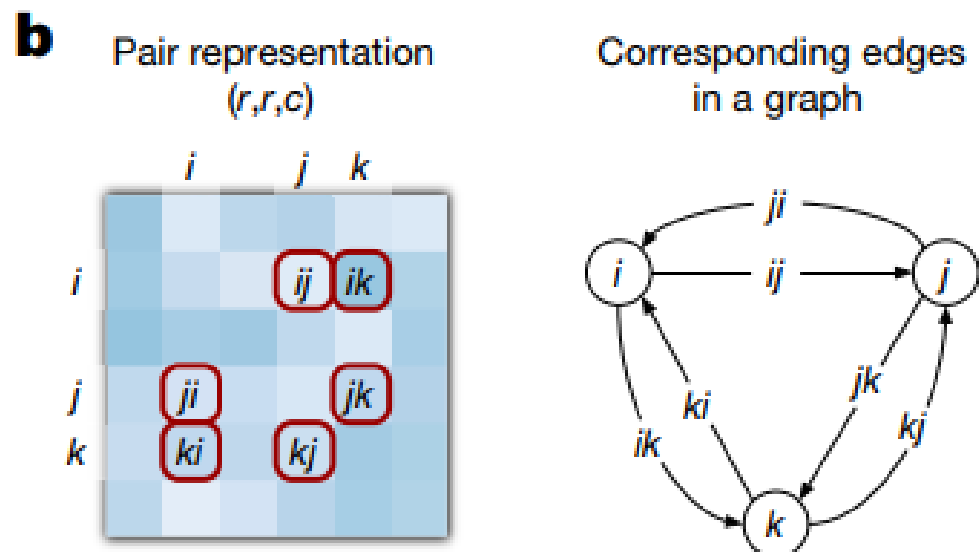


The **MSA transformer** identifies a correlation between two columns of the MSA, each corresponding to a residue.

This information is passed to the **pair representation**, where subsequently the pair representation identifies another possible interaction.

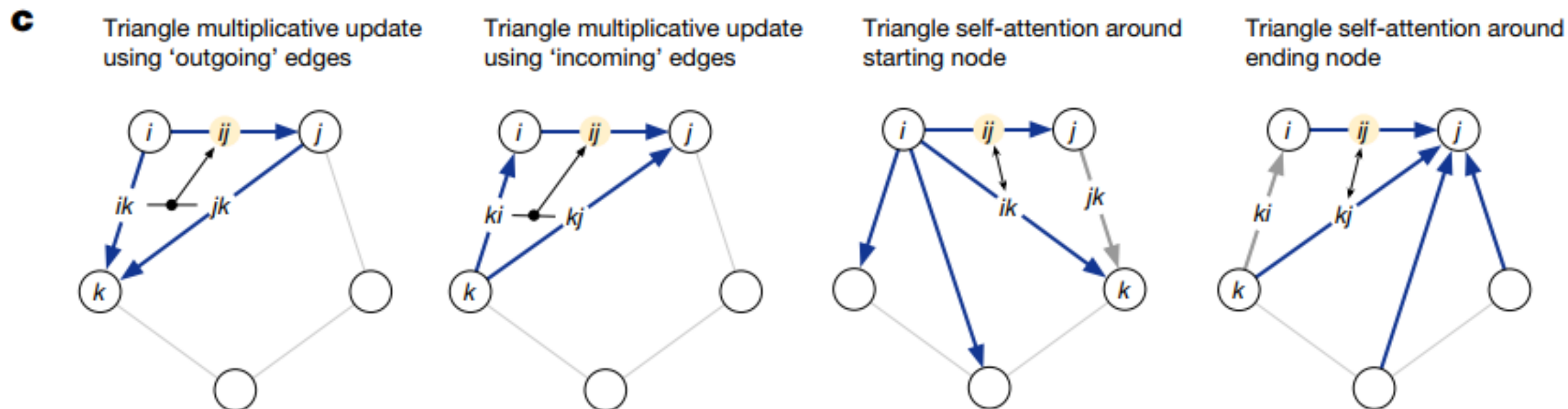
The information is passed back to the MSA.

The **MSA transformer** receives an input from the **pair representation**, and observes that another pair of columns exhibits a significant correlation

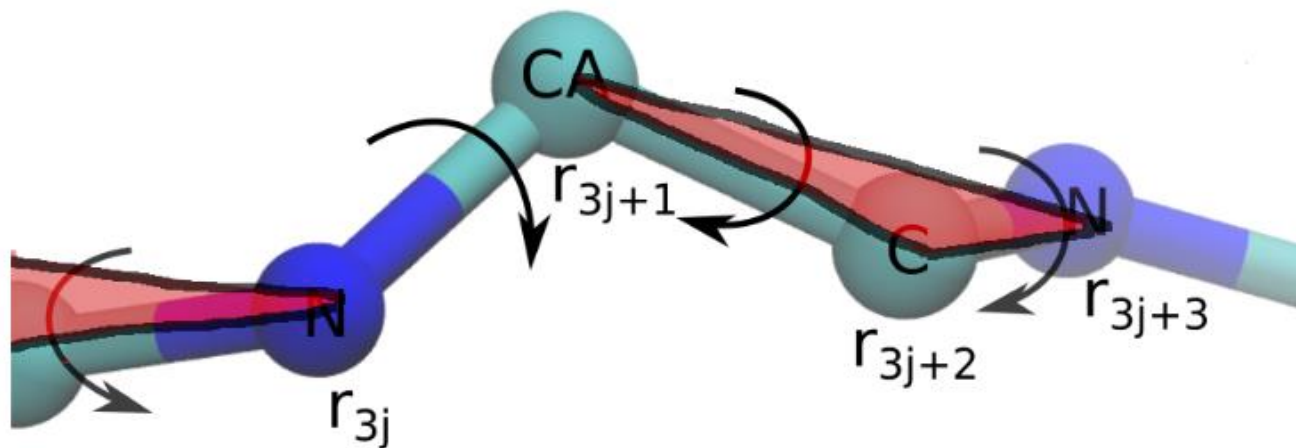


b. The pair representation interpreted as directed edges in a graph.

c. Triangle multiplicative update and triangle self-attention. The circles represent residues. Entries in the pair representation are illustrated as directed edges and in each diagram, the edge being updated is ij .



Rigid bodies representation of a protein¹



$$\mathbf{M} = \begin{pmatrix} \begin{matrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{matrix} & \begin{matrix} a_{14} \\ a_{24} \\ a_{34} \end{matrix} \\ \begin{matrix} 0 & 0 & 0 \end{matrix} & 1 \end{pmatrix}$$

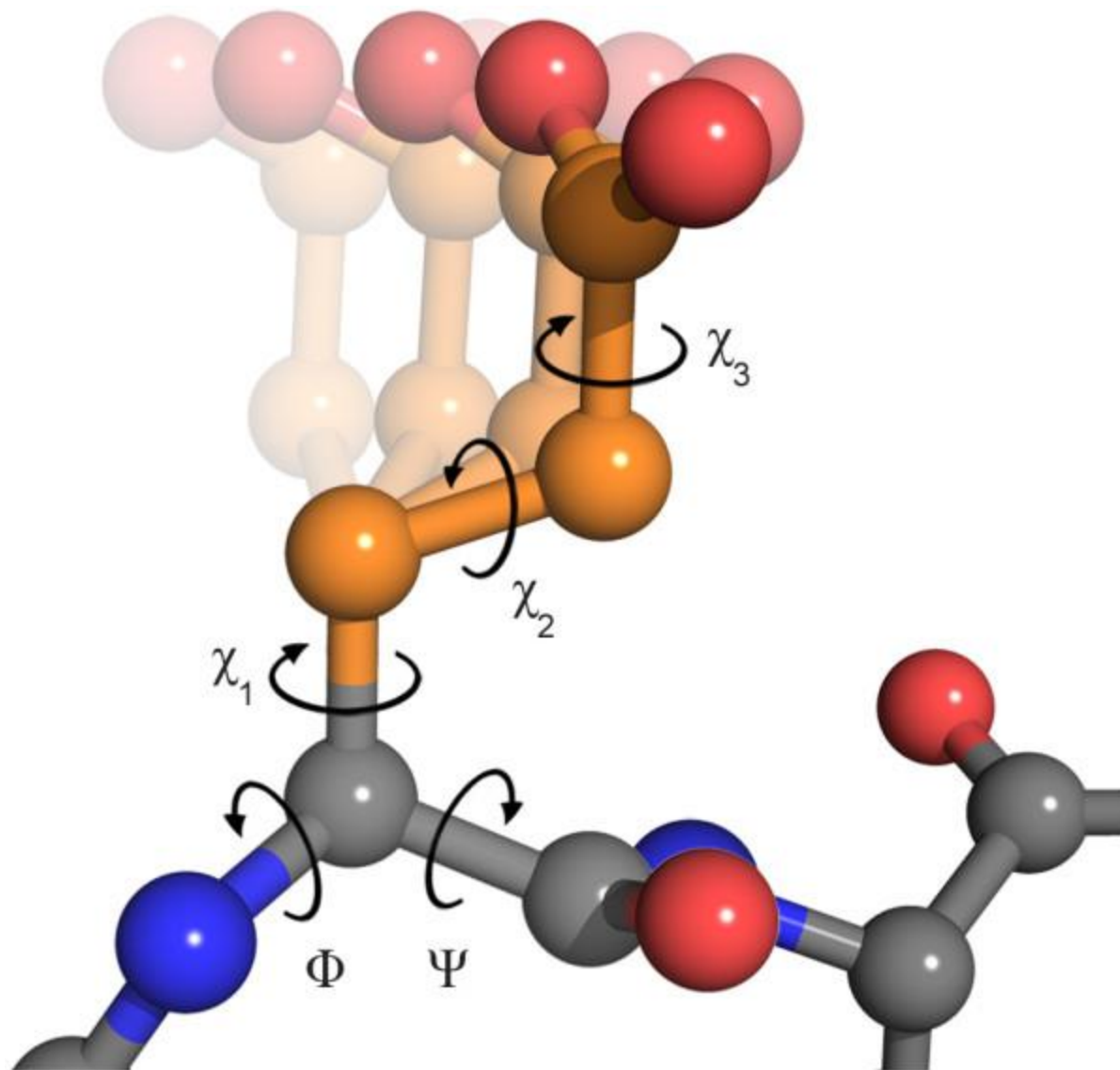
Mathematical representation of an affine transformation matrix. The 3×3 part corresponds to a rotation matrix, and the 1×3 column is the displacement vector. Multiplication of an affine matrix times a vector is equivalent to displacing and subsequently rotating.²

1) <https://lupoglaz.github.io/OpenFold2/protein.html>

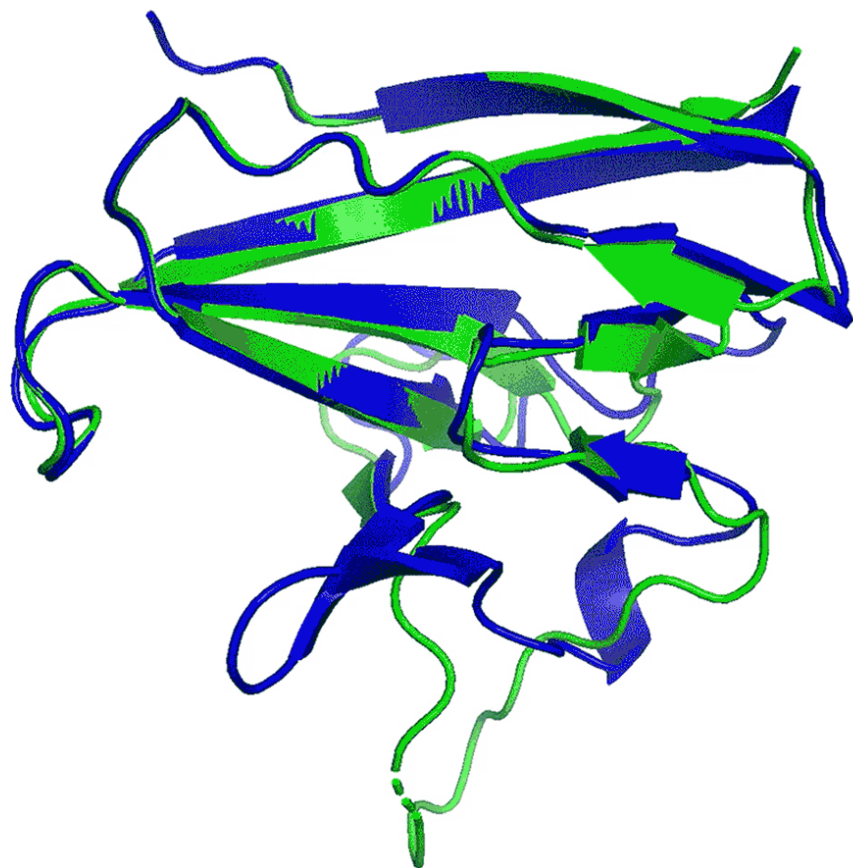
2) <https://www.brainvoyager.com/bv/doc/UsersGuide/CoordsAndTransforms/SpatialTransformationMatrices.html>



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



- Predicted several protein structures of the SARS-CoV-2 virus (ORF3a and ORF8 proteins)



ORF8, a protein associated with the coronavirus that causes COVID-19, was predicted by AF2 (blue), and the experimental structure is in green.

- AlphaFold Protein Structure Database <https://alphafold.ebi.ac.uk/>



<https://alphafold.ebi.ac.uk/>

“This computational work represents a stunning advance on the protein-folding problem, a 50-year-old grand challenge in biology. It has occurred decades before many people in the field would have predicted. It will be exciting to see the many ways in which it will fundamentally change biological research.”

Professor Venki Ramakrishnan, Nobel laureate and President of the Royal Society

“AlphaFold is a once in a generation advance, predicting protein structures with incredible speed and precision. This leap forward demonstrates how computational methods are poised to transform research in biology and hold much promise for accelerating the drug discovery process.”

Arthur d. Levinson, Phd, Founder & CEO Calico, Former Chairman & CEO, Genentech

bioRxiv preprint doi: <https://doi.org/10.1101/2021.10.04.463034>; this version posted October 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



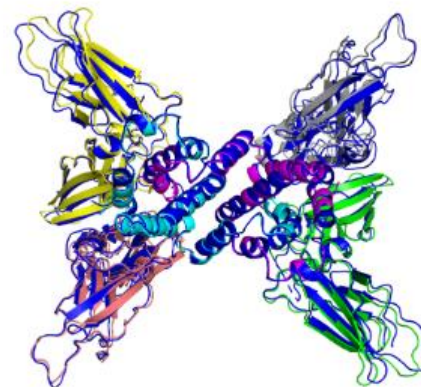
2021-10-04

Protein complex prediction with AlphaFold-Multimer

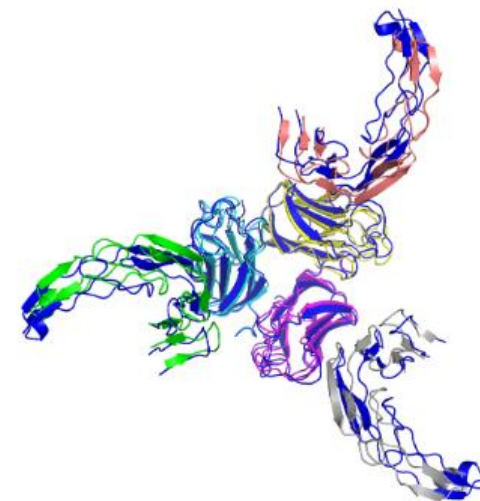
Richard Evans^{1*}, Michael O'Neill^{1*}, Alexander Pritzel^{1*}, Natasha Antropova^{1*}, Andrew Senior¹, Tim Green¹, Augustin Židek¹, Russ Bates¹, Sam Blackwell¹, Jason Yim¹, Olaf Ronneberger¹, Sebastian Bodenstein¹, Michal Zielinski¹, Alex Bridgland¹, Anna Potapenko¹, Andrew Cowie¹, Kathryn Tunyasuvunakool¹, Rishub Jain¹, Ellen Clancy¹, Pushmeet Kohli¹, John Jumper^{1*} and Demis Hassabis^{1*}

¹DeepMind, London, UK, *These authors contributed equally

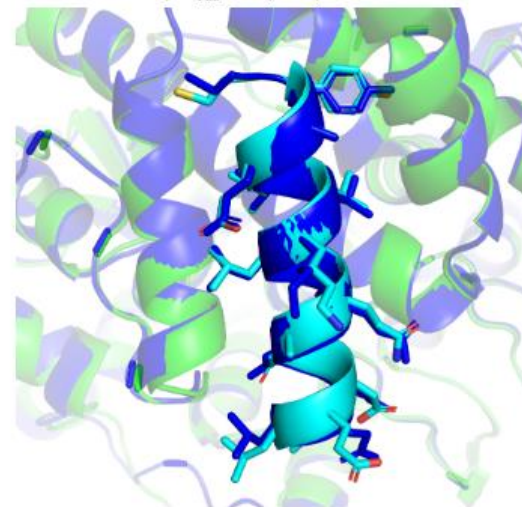
Structure examples predicted with the AlphaFold-Multimer. Visualised are the ground truth structures (blue) and predicted structures (coloured by chain).



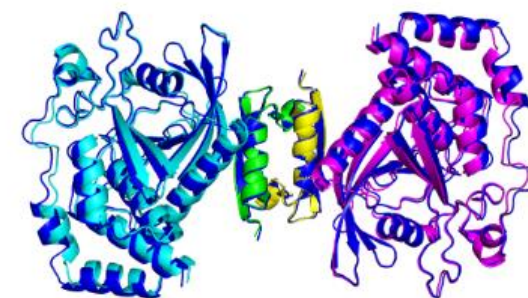
(a) A2B2C2 heteromer
TM-score = 98.0, $N_{\text{res}} = 1,246$, PDB ID = 6E3K



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



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



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

- “**black box**” problem – cannot figure out exactly what concepts the algorithm is “learning” about proteins → hasn’t directly increased our understanding of how protein sequence relates to structure.
- Accuracy drops substantially when the mean alignment depth is less than ~30 sequences.
- 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
- Expand AlphaFold Database to cover UniRef
- Prediction of mutational effects
 - Changes in structure and stability
 - Binding affinity

Cellular tumor antigen p53-AlphaFold structure prediction

Input Sequence:

```
>sp|P04637|P53_HUMAN Cellular tumor antigen p53 OS=Homo sapiens OX=9606 GN=TP53 PE=1 SV=4
MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLSPDDIEQWFTEDPGP
DEAPRMPEAAPPVAPAPAAPTPAAPAPAPSWPLSSSVPSQKTYQGSYGFRGLHSGTAK
SVTCTYSPALNKMFCQLAKTCPVQLWVDSTPPPGTRVRAMAIYKQSQHMTEVVRRCPHHE
RCSDSDGLAPPQHLIRVEGNLRVEYLDDRNTFRHSVVVPYEPPEVGSDCTTIHYNMCNS
SCMGGMNRRPILTIITLEDSSGNLLGRNSFEVRVCACPGRDRRTEENLRKKGEPHHELP
PGSTKRALPNNTSSSPQPKKKPLDGEYFTLQIRGRERFEMFRELNEALELKDAQAGKEPG
GSRAHSSHLKSKKGQSTSRHKKLMFKTEGPDSD
```

<https://colab.research.google.com/drive/1D70qlEO9EnfmZkUxDEgKeY5NYNPOKMcG#scrollTo=kOblAo-xetgx>

Thank you for your attention!