DeepMind

Highly accurate protein structure prediction with AlphaFold

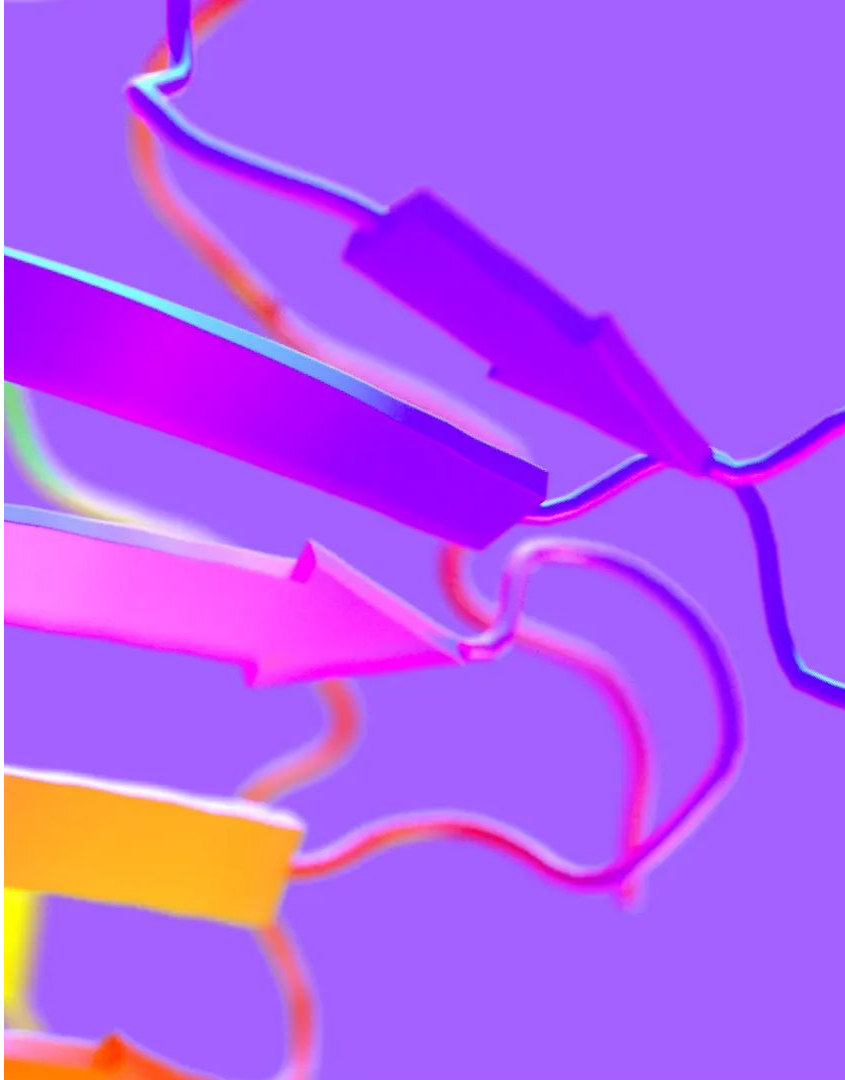
Michael Figurnov
Staff Research Scientist, DeepMind

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)

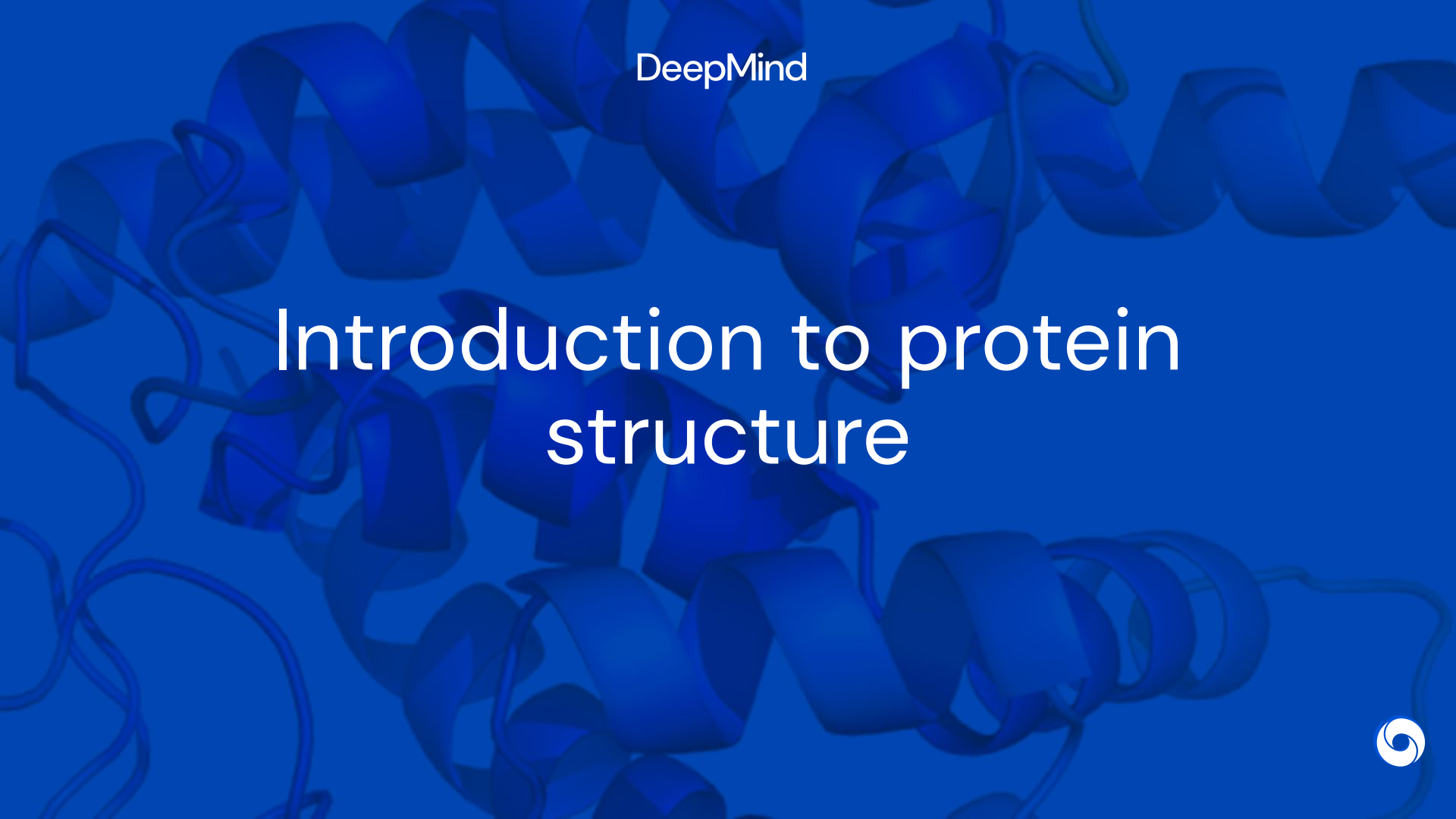
Agenda

- Introduction to protein structure
- How AlphaFold works
- AlphaFold and the biological community
- What's next?
- Lattice QCD work at DeepMind
- Q&A



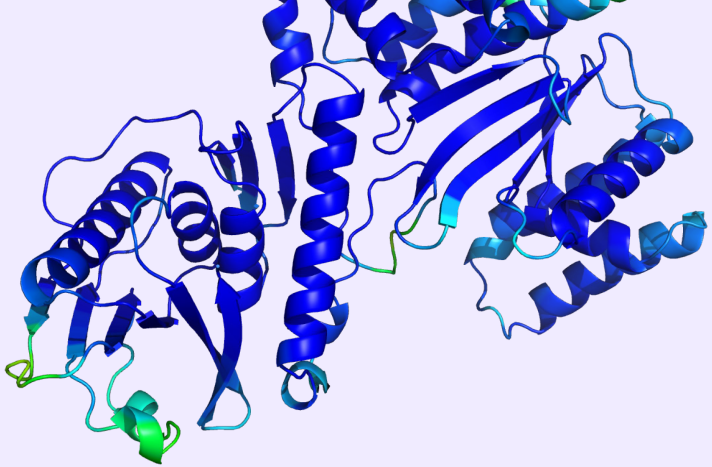
DeepMind

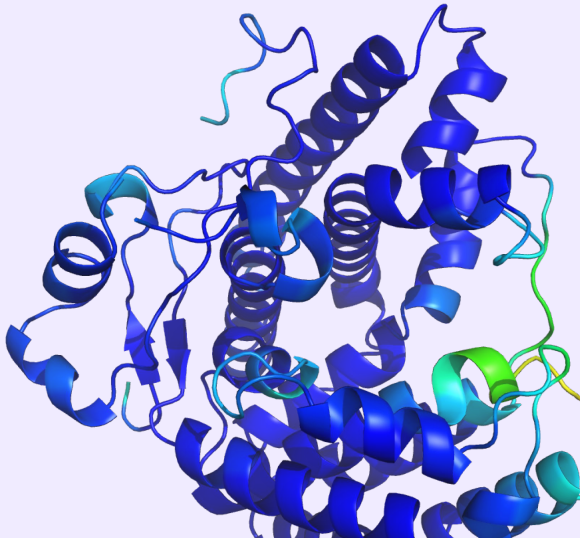
Introduction to protein structure



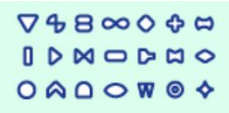
What is a protein?

- Proteins are molecular machines that are **essential to life**
- They have **many functions**: from our hair to our immune system
- Consist of **chains of amino acids** that fold into a 3D structure
- The exact **3D shape** is important for a protein's function
- Understanding protein structures is a **fundamental problem in biology**



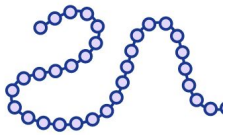


Protein structure: terminology



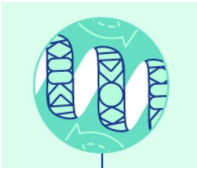
Vocabulary: 20 common amino acids / residues

- Small organic molecules with common groups
  + a **side chain** specific to each amino acid



Sequence: chain of 100s–1000s amino acids

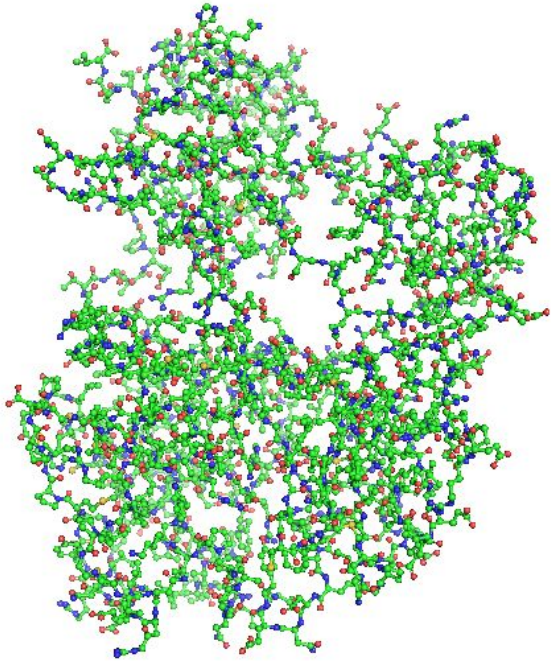
- Amino acids form peptide bonds and build up a protein chain
- **DNA sequences directly encode the amino acid sequence**

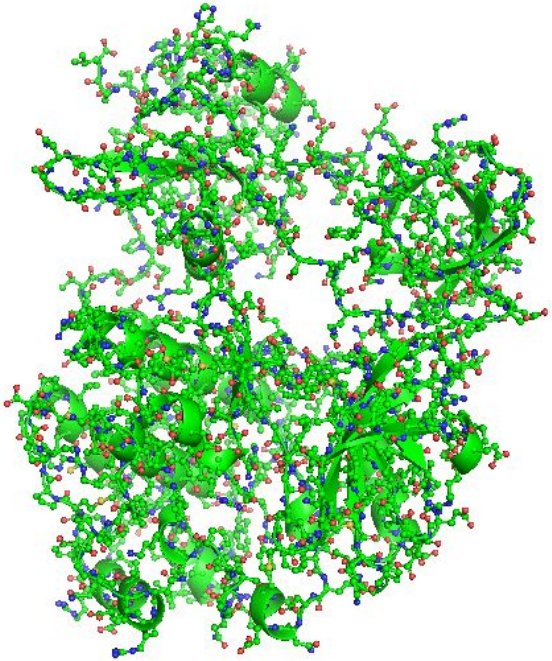


Structure: atom coordinates in 3D space (300 - 50,000 atoms)

- Unique 3D structure comes from physical interactions of amino acids.

Protein structure

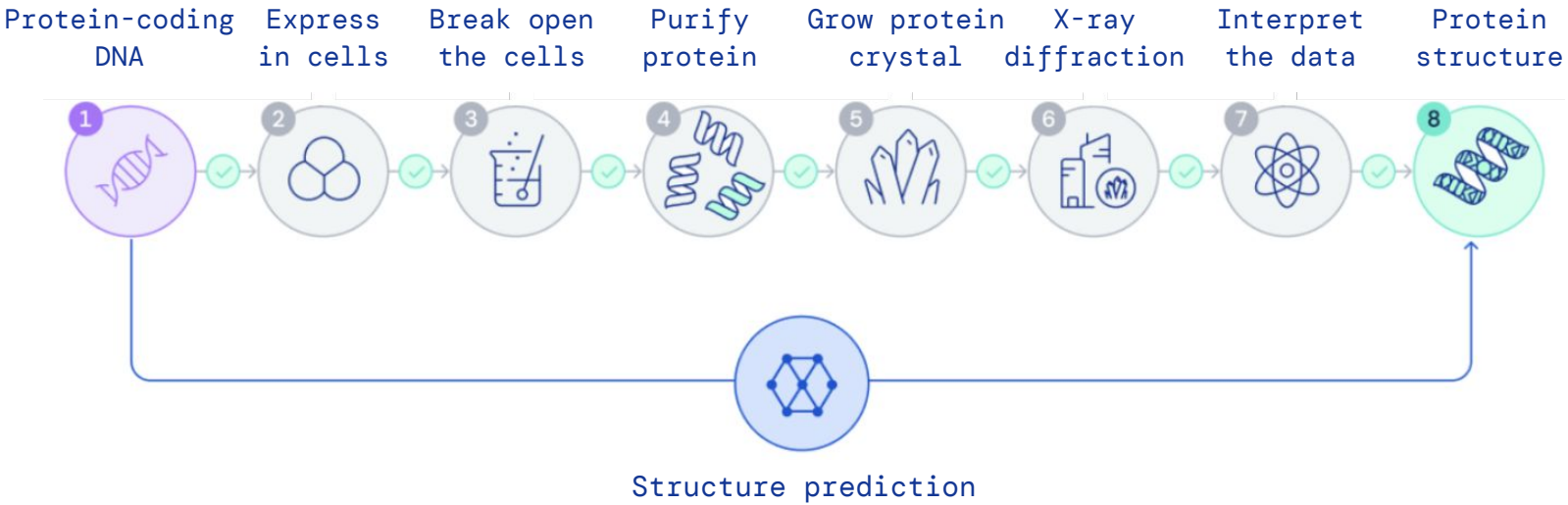




Why predict structures?

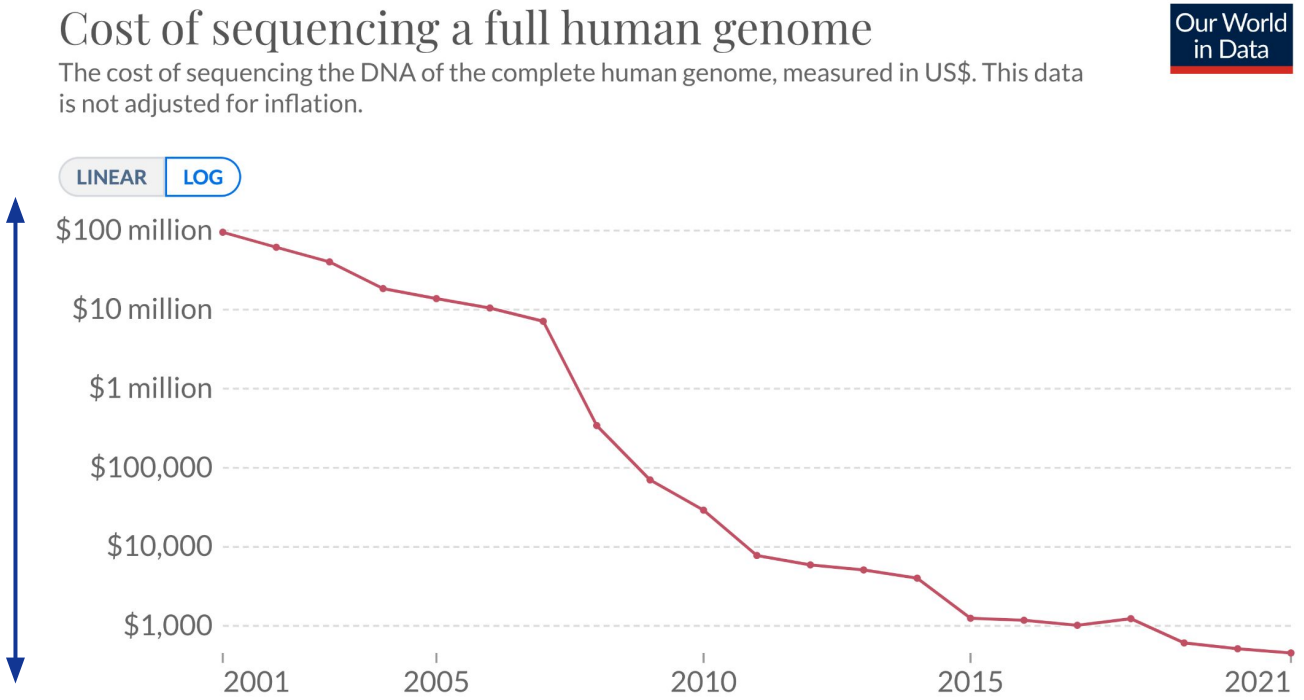
Predicting a protein's structure from its amino acid sequence has been a **grand challenge in biology** for the past 50 years.

- ➔ Experimental structure determination **takes months to years.**
- ➔ ~200,000 protein structures experimentally determined so far.
- ➔ Structure prediction can provide actionable information faster.



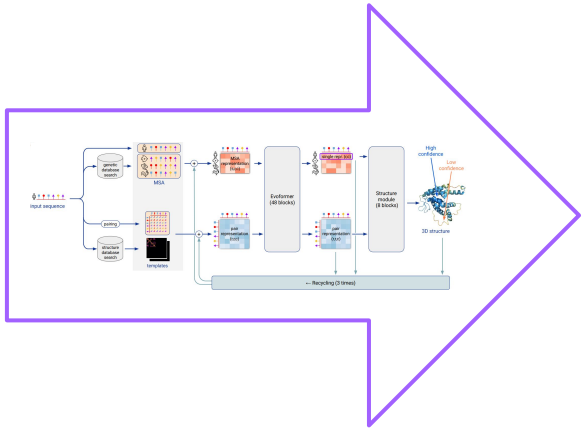
Reading DNA has become cheap

200,000x decrease in 20 years

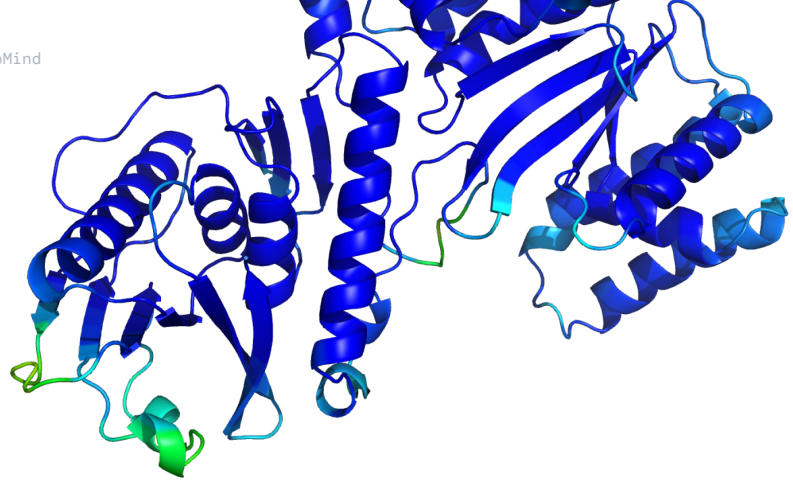


AlphaFold predicts highly accurate protein structures from amino acid sequences

SIFSYITESTGTPSNATYT
YVIERWDPETSGILNPCYG
WPVCYVTVNHKHTVNGTGG
NPAFQIARIEKLRTLAEVR
DVVLKNRSFPIEGQTTHRG
PSLNSNQECVGLFYQPNSS
GISPRGKLLPGSLCGIAPP
PVHHHHHH



T1049 / 6y4f
93.5 GDT
(adhesin tip)



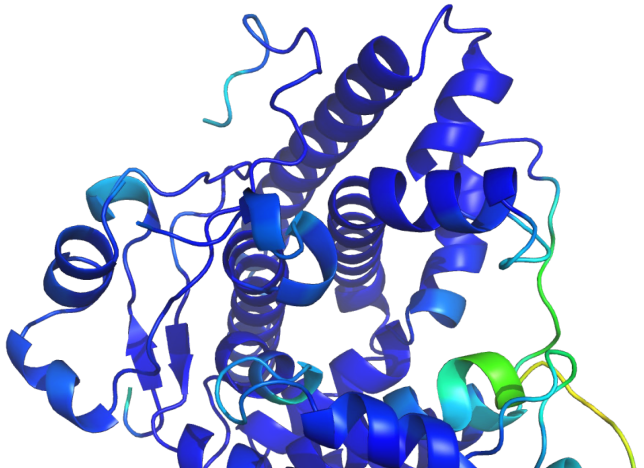
External independent benchmarks are critical

Run every 2 years since 1994 - gold standard

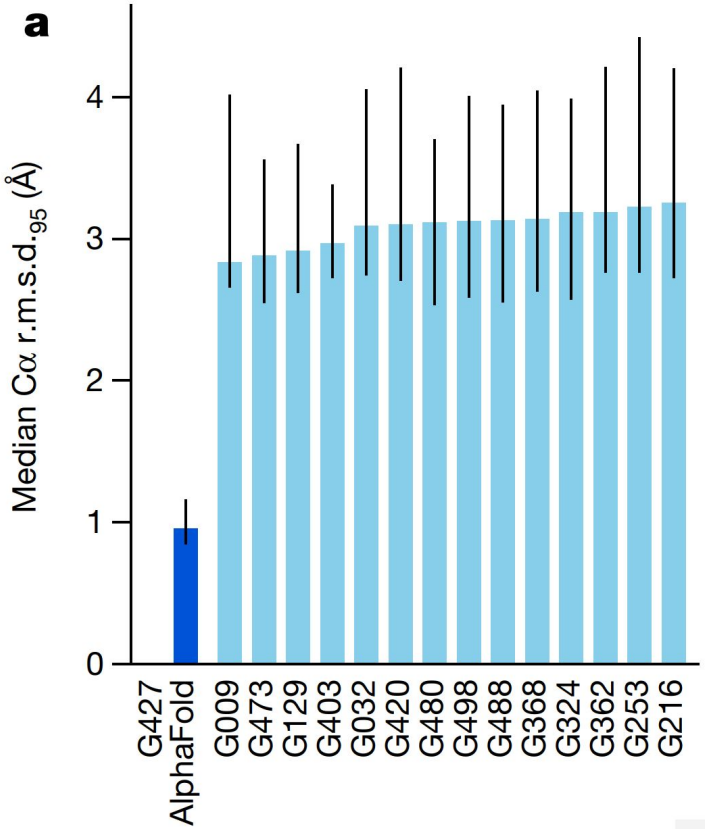
Blind prediction assessment

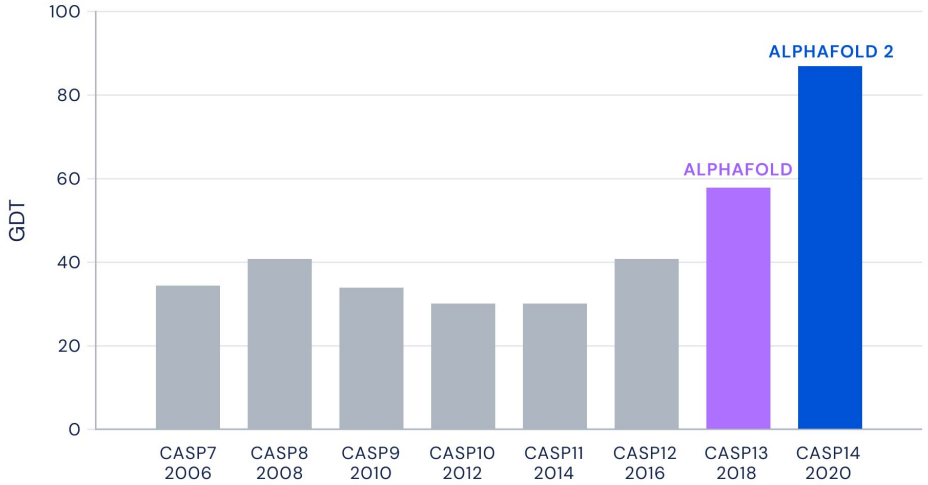
CASP

Critical Assessment of
protein Structure Prediction



AlphaFold at CASP14





T1037 / 6vr4

90.7 GDT, RNA polymerase domain

Making AlphaFold available

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

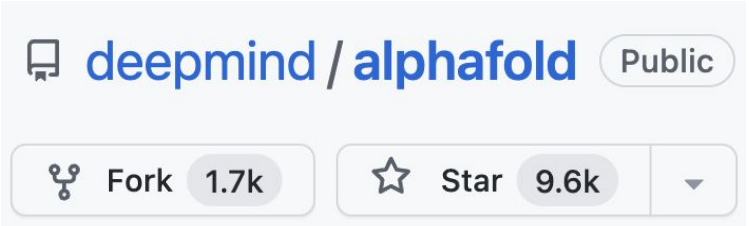
John Jumper[1,4,5], Richard Evans[1,4], Alexander Pritzel[1,4], Tim Green[1,4], Michael Figurnov[1,4], Olaf Ronneberger[1,4], Kathryn Tunyasuvunakool[1,4], Russ Bates[1,4], Augustin Žídek[1,4], Anna Potapenko[1,4], Alex Bridgland[1,4], Clemens Meyer[1,4], Simon A. A. Kohl[1,4], Andrew J. Ballard[1,4], Andrew Cowie[1,4], Bernardino Romera-Paredes[1,4], Stanislav Nikolov[1,4], Rishub Jain[1,4], Jonas Adler[1], Trevor Back[1], Stig Petersen[1], David Reiman[1], Ellen Clancy[1], Michal Zielinski[1], Martin Steinegger[2,3], Michalina Pacholska[1], Tamas Berghammer[1], Sebastian Bodenstein[1], David Silver[1], Oriol Vinyals[1], Andrew W. Senior[1], Koray Kavukcuoglu[1], Pushmeet Kohli[1] & Demis Hassabis[1,4]✉

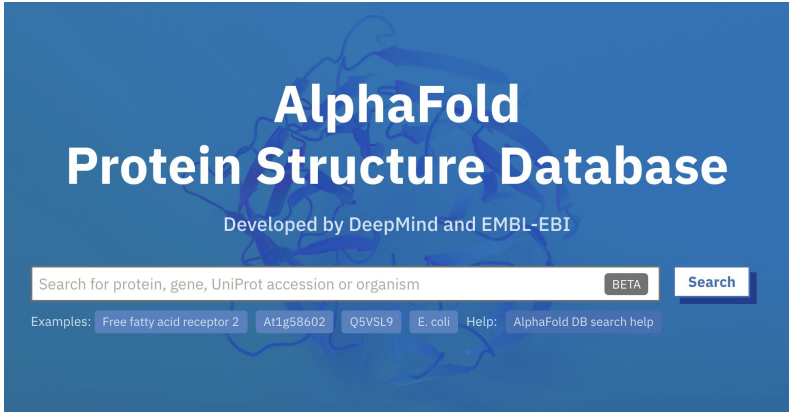
Proteins are essential to life, and understanding their structure can facilitate a mechanistic understanding of their function. Through an enormous experimental effort[1–4], the structures of around 100,000 unique proteins have been determined[5], but this represents a small fraction of the billions of known protein sequences[6,7]. Structural coverage is bottlenecked by the months to years of painstaking effort required to determine a single protein structure. Accurate computational approaches are needed to address this gap and to enable large-scale structural bioinformatics. Predicting the three-dimensional structure that a protein will adopt based solely on its amino acid sequence—the structure prediction component of the 'protein folding problem'[8]—has been an important open research problem for more than 50 years[9]. Despite recent progress[10–14], existing methods fall far short of atomic accuracy, especially when no homologous structure is available. Here we provide the first computational method that can regularly predict protein structures with atomic accuracy even in cases in which no similar structure is known. We validated an entirely redesigned version of our neural network-based model, AlphaFold, in the challenging 14th Critical Assessment of protein Structure Prediction (CASP14)[15], demonstrating accuracy competitive with experimental structures in a majority of cases and greatly outperforming other methods. Underpinning the latest version of AlphaFold is a novel machine learning approach that incorporates physical and biological knowledge about protein structure, leveraging multi-sequence alignments, into the design of the deep learning algorithm.

The development of computational methods to predict three-dimensional (3D) protein structures from the protein sequence has proceeded along two complementary paths that focus on either the physical interactions or the evolutionary history. The physical interaction programme heavily integrates our understanding of molecular driving forces into either thermodynamic or kinetic simulation of protein physics[16] or statistical approximations thereof[17]. Although theoretically very appealing, this approach has proved highly challenging for even moderate-sized proteins due to the computational intractability of molecular simulation, the context dependence of protein stability and the difficulty of producing sufficiently accurate models of protein physics. The evolutionary programme has provided an alternative in recent years, in which the constraints on protein structure are derived from bioinformatics analysis of the evolutionary history of proteins, homology to solved structures[18,19] and pairwise evolutionary correlations[20–23]. This bioinformatics approach has benefited greatly from the steady growth of experimental protein structures deposited in the Protein Data Bank (PDB)[5], the explosion of genomic sequencing and the rapid development of deep learning techniques to interpret these correlations. Despite these advances, contemporary physical and evolutionary-history-based approaches produce predictions that are far short of experimental accuracy in the majority of cases in which a close homologue has not been solved experimentally and this has limited their utility for many biological applications.

In this study, we develop the first, to our knowledge, computational approach capable of predicting protein structures to near experimental accuracy in a majority of cases. The neural network AlphaFold that we developed was entered into the CASP14 assessment (May–July 2020; entered under the team name 'AlphaFold2' and a completely different model from our CASP13 AlphaFold system[10]). The CASP assessment is carried out biennially using recently solved structures that have not been deposited in the PDB or publicly disclosed so that it is a blind test

[1]DeepMind, London, UK. [2]School of Biological Sciences, Seoul National University, Seoul, South Korea. [3]Artificial Intelligence Institute, Seoul National University, Seoul, South Korea. [4]These authors contributed equally: 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, Demis Hassabis.
✉e-mail: jumper@deepmind.com; dhcontact@deepmind.com

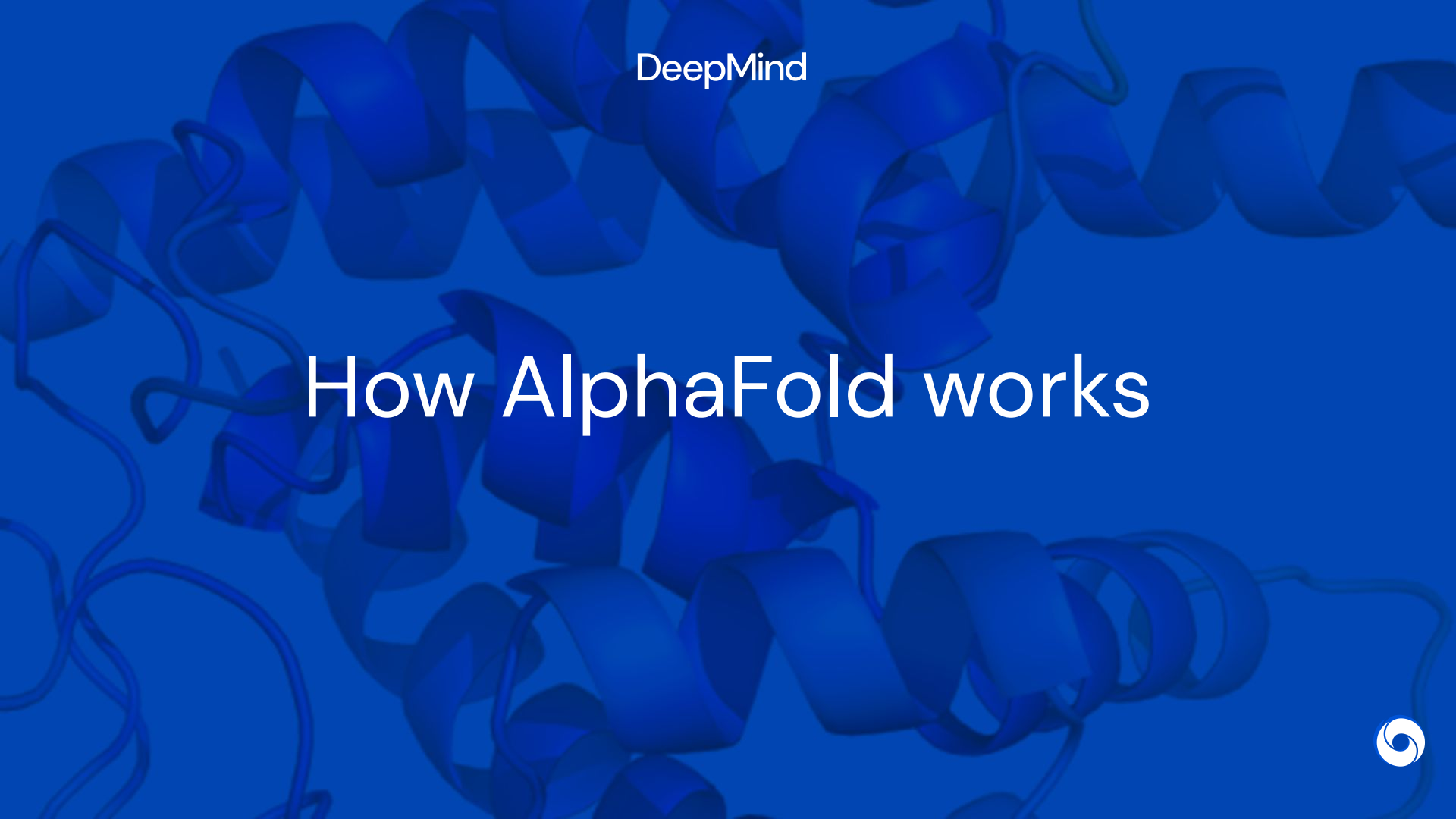




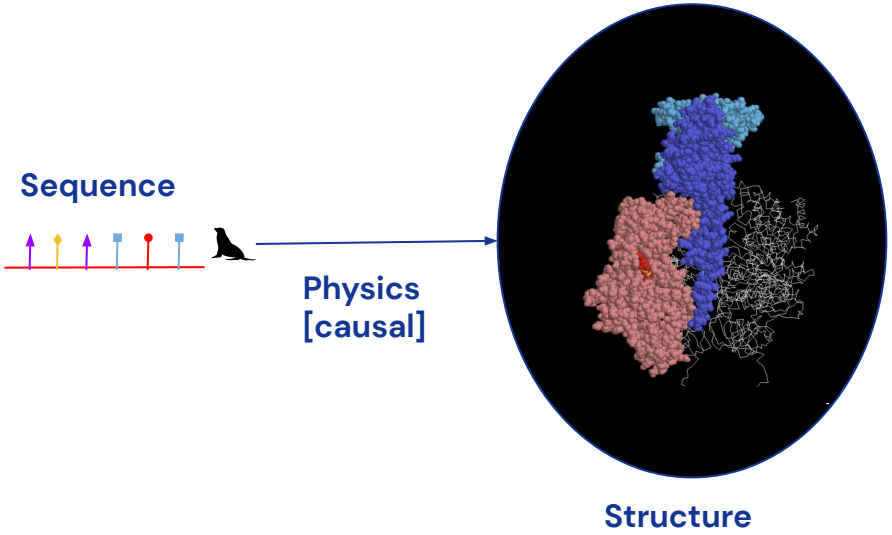
>200 million protein structures

DeepMind

How AlphaFold works

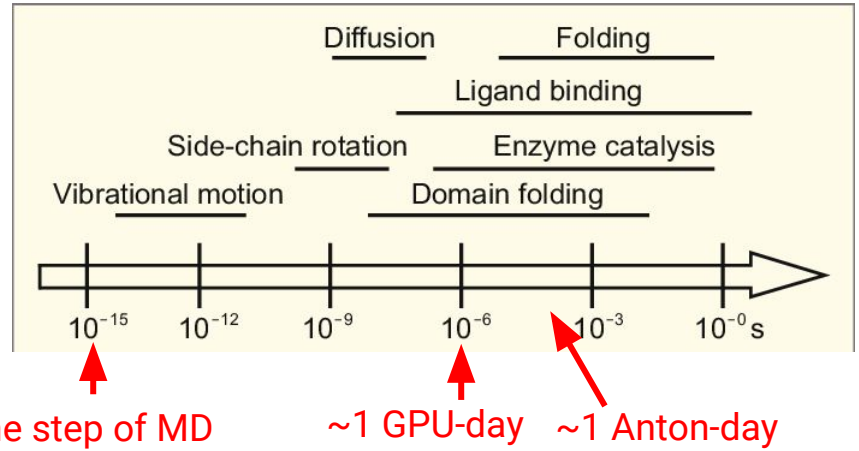


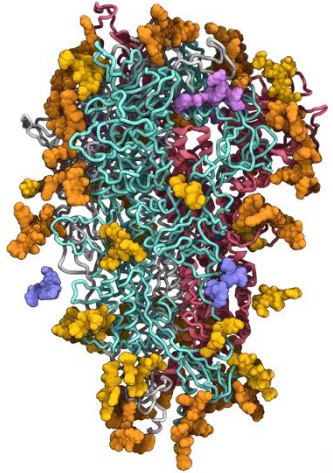
Why not physical simulations?



Why not physical simulations?
Reason 1: timescales

- **Molecular dynamics** (MD):
  - Langevin dynamics on a potential
  - Folding takes 10^{12}–10^{15} sequential steps – too slow

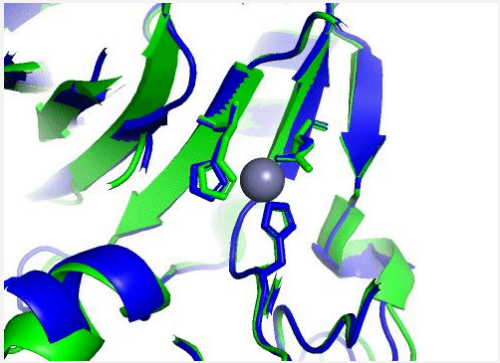




SARS-COV-2 spike, Anton simulation

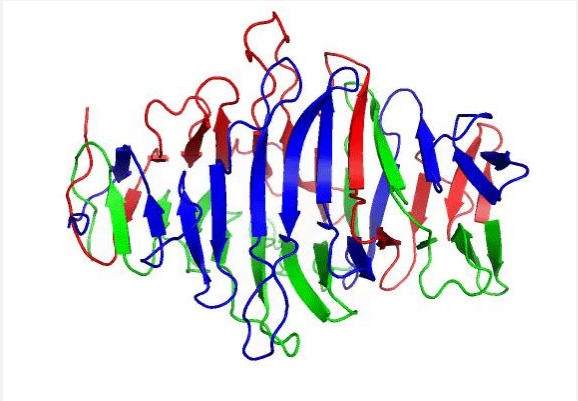
Why not physical simulations?
Reason 2: underspecified context

T1056 (zinc binding)

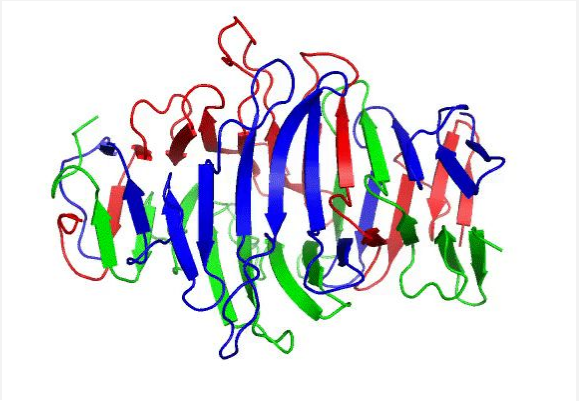


AlphaFold / **Experiment**

T1080 (trimer)



AlphaFold (monomer prediction x3)



Experimental structure

Determining Structure from Evolution - Intuition

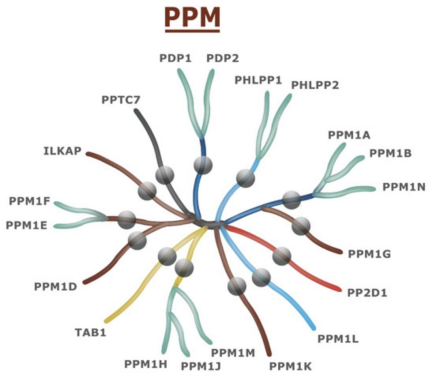
Evolutionary
Structure Prediction
[Bayesian]

Sequence

Physics
[causal]

Structure

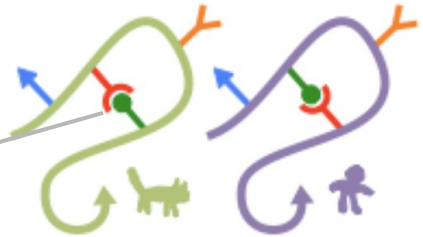
Function
[causal]



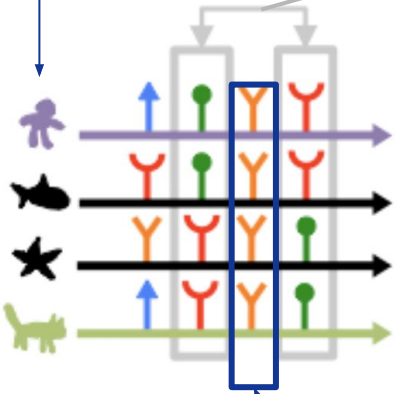
Evolutionary History

Determining Structure from Evolution - Intuition

Search genetics databases



Co-evolution: residues in contact must mutate together.



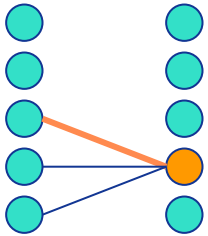
Multiple Sequence Alignment (MSA)

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

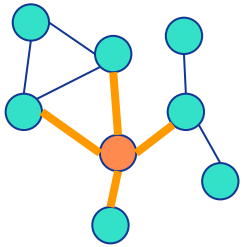
Coevolution cartoon by Sergey Ovchinnikov (https://jgi.doe.gov/seeking-structure-metagenome-sequences/cartoon-coevolution-sergey-o/)

Deep learning provides building blocks for approximating arbitrary functions

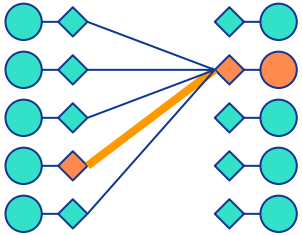
E.g.



Convolutional layers



Graph layers

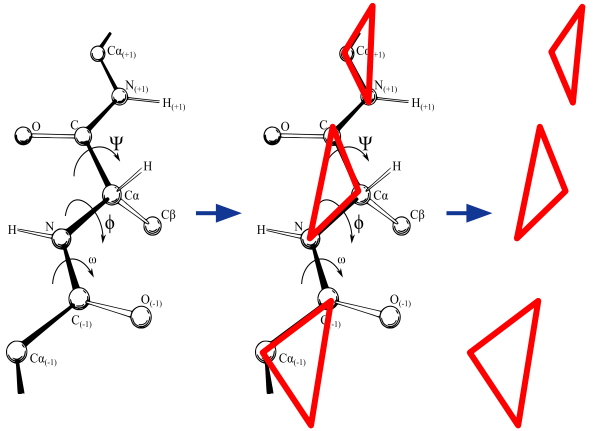


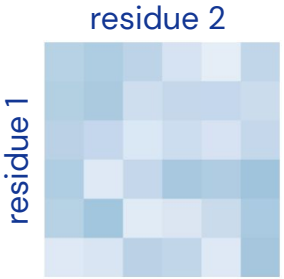
Attention layers

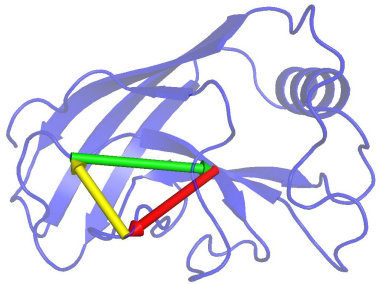
These blocks are very generic; they don’t take advantage of our scientific understanding of proteins and protein evolution

Design principles

- **End-to-end**
  - Network goes all the way from inputs to structure
  - Can learn about and optimise the whole process
- **Inductive biases**
  - Design reflects our knowledge of physics / geometry
  - Emphasis on pairs, not a sequence of residues
  - Output should be self-consistent







Protein backbone image by Dcrjsr, vectorised Adam Rędzikowski (Wikipedia)

AlphaFold: a deep learning model

Model

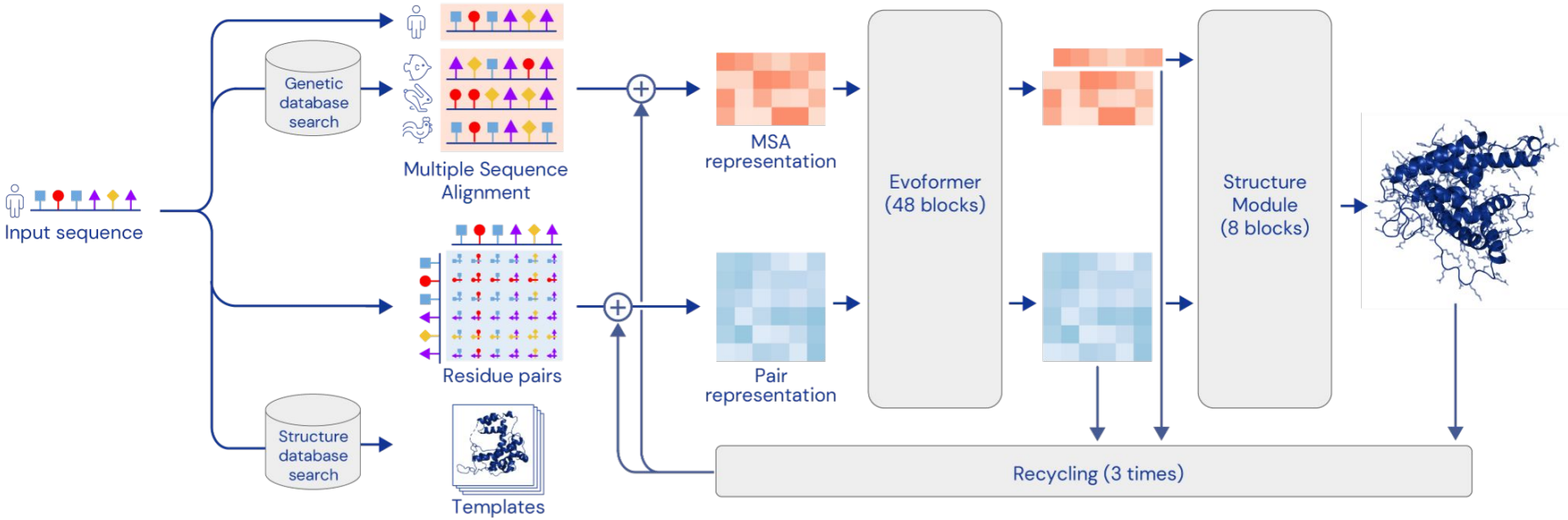
- Deep: 192 blocks (48 blocks x 4 cycles)
- 93 million parameters
    - Reasonably small; GPT-3 is 175 *billion* parameters
- Seconds to hours of runtime, depending on protein length

Training Data

- 170,000 3D structures

- Genetic databases (sequences without structure)

Model overview

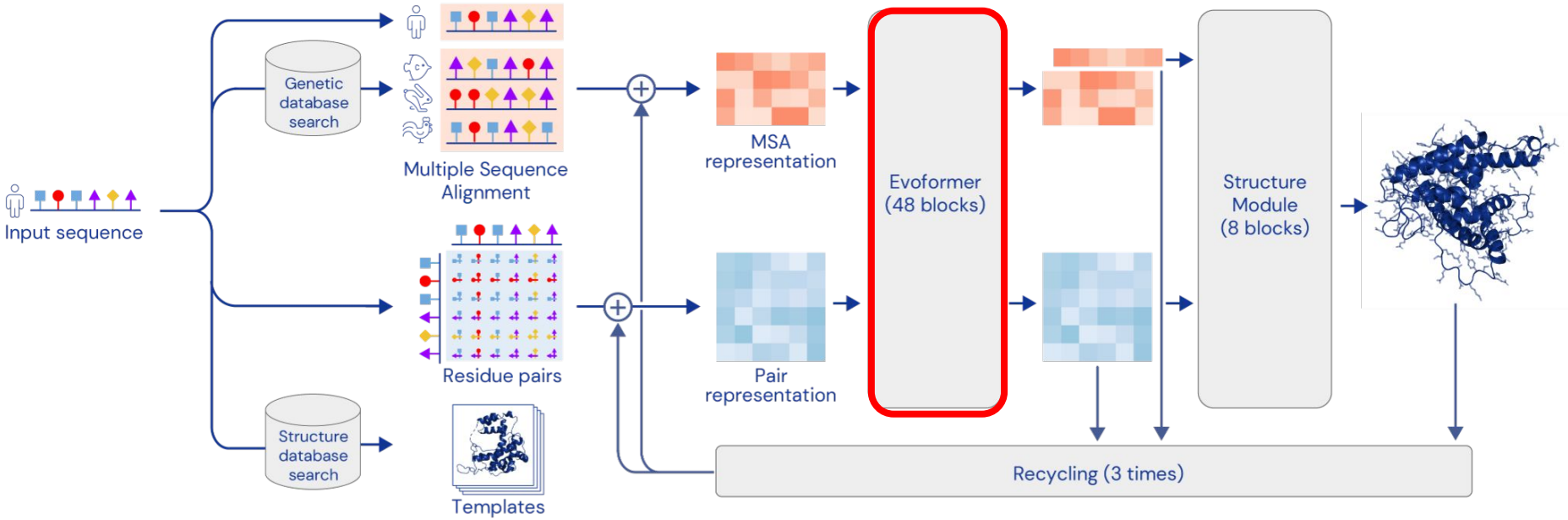


Incorporates evolutionary information

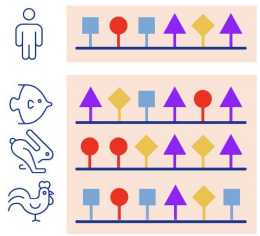
Determines relationship between residue pairs

End-to-end: outputs structure directly

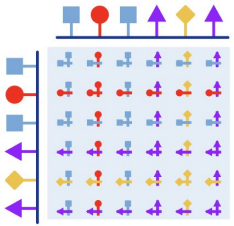
Evoformer


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



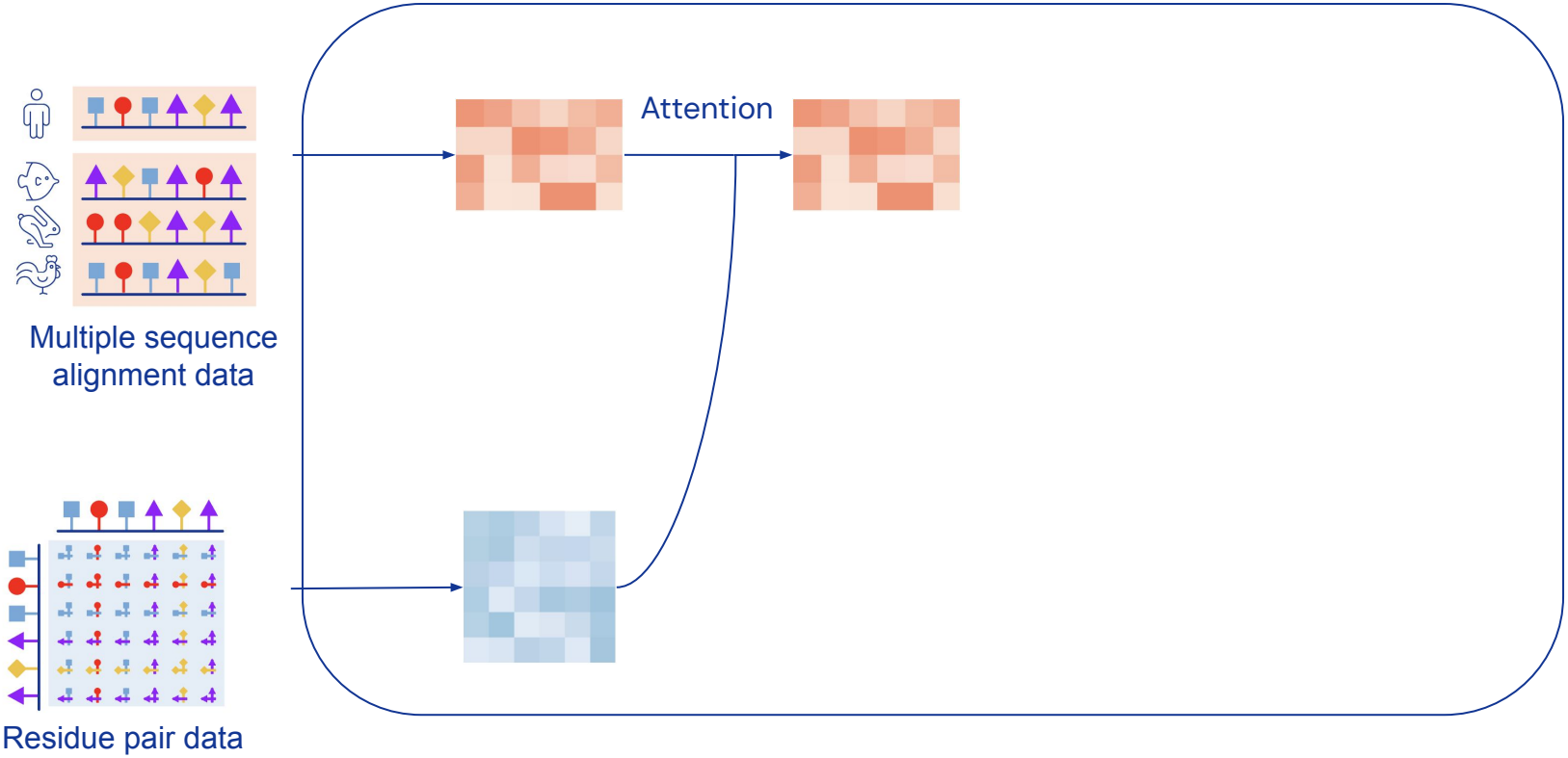
Multiple sequence alignment data



Residue pair data

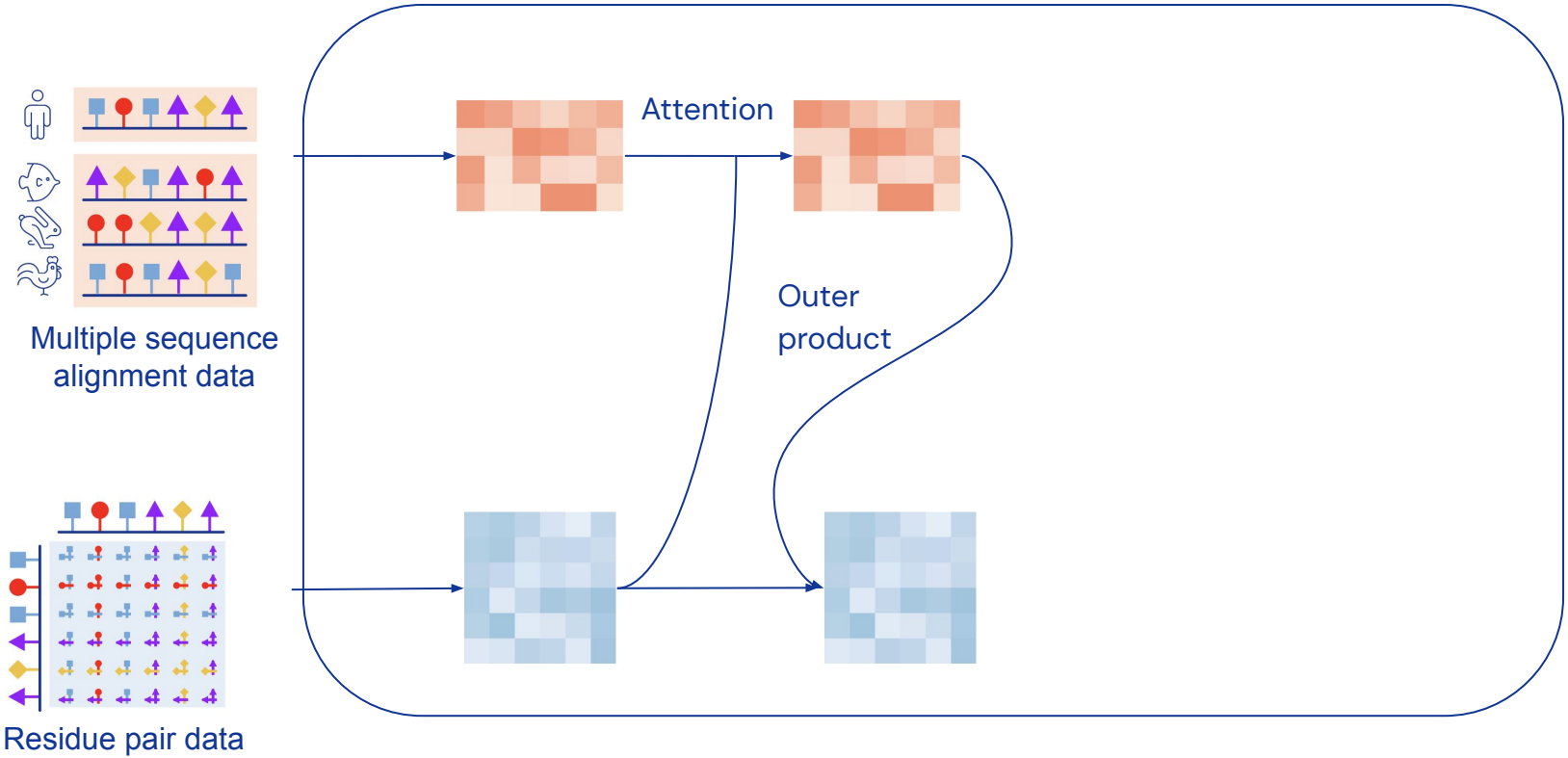
Multiple sequence alignment and residue pairs are the main concepts in Evoformer blocks

Evoformer



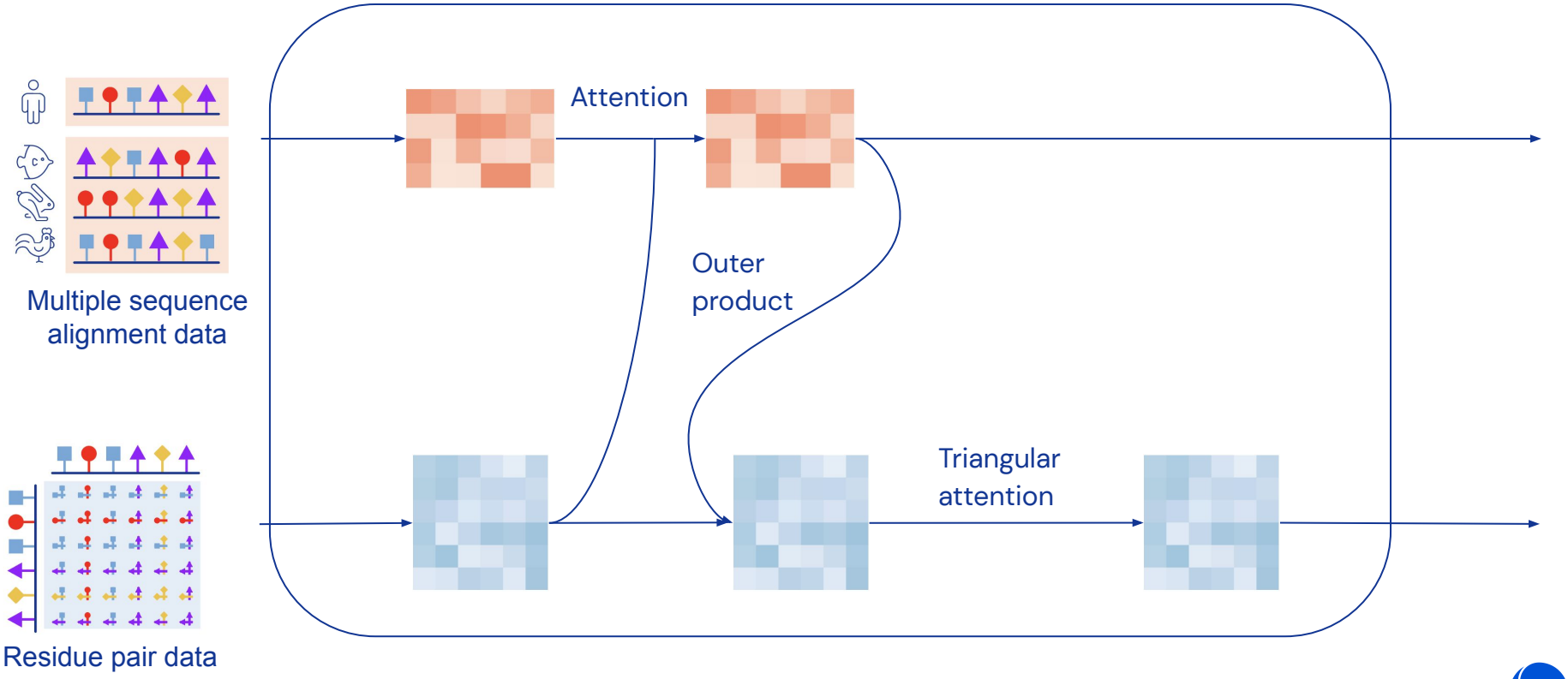
Attention is augmented by the network's belief about residue pairs

Evoformer



Outer product allows generalized correlation similar to co-evolution

Evoformer

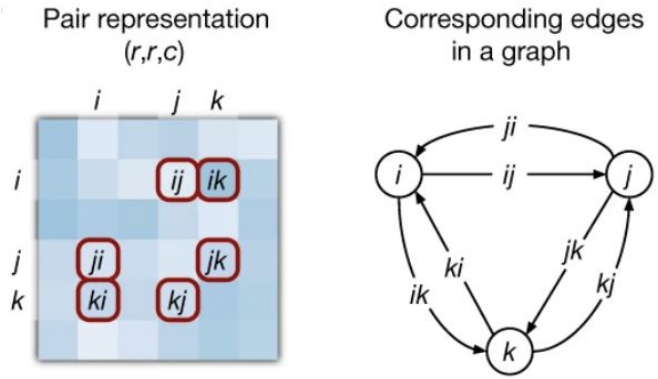


Residue pair interactions inspired by geometric interactions

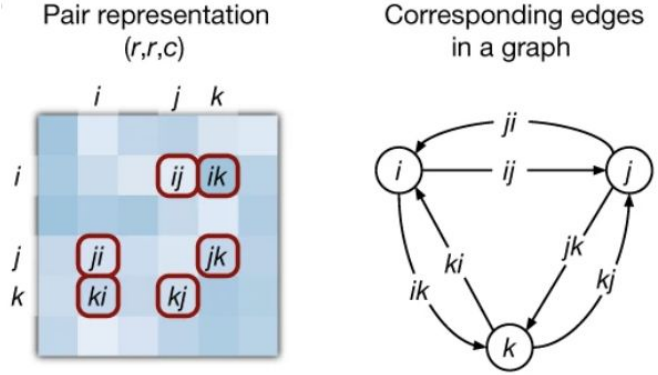
Triangular attention for residue pairs

➔ **Matrix view:**

- Each residue attends to all other residues **within a row**



Triangular attention for residue pairs



➔ **Matrix view:**

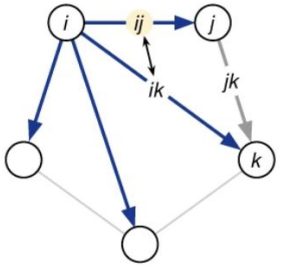
- Each residue attends to all other residues **within a row**

➔ **Graph view:**

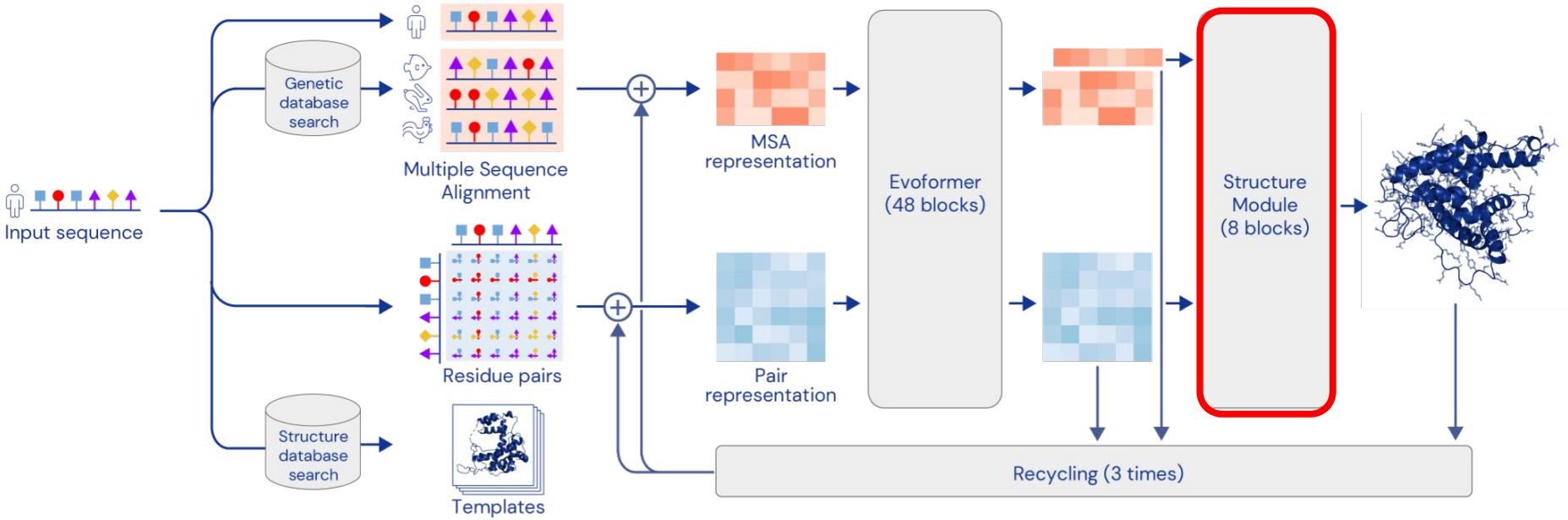
- Nodes = residues, edges = pairs of residues; represents residue spatial relations, e.g. distances in 3D.
- **Transitivity / triangle inequality:** update for query **ij** should depend on **ik** and **jk** (for residues **i, j, k**).

$$a^h_{ijk} = \text{softmax}_k \left(\frac{1}{\sqrt{c}} \, \mathbf{q}^{h\top}_{ij} \mathbf{k}^h_{ik} + b^h_{jk} \right)$$

This is missing in traditional self-attention: we add a bias term to bring the third edge in.



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

- Protein backbone = gas of 3-D rigid bodies (chain is learned!)

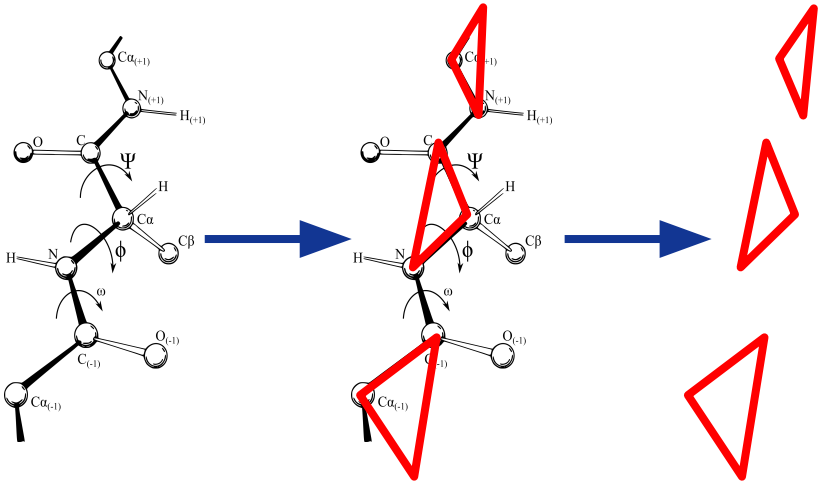
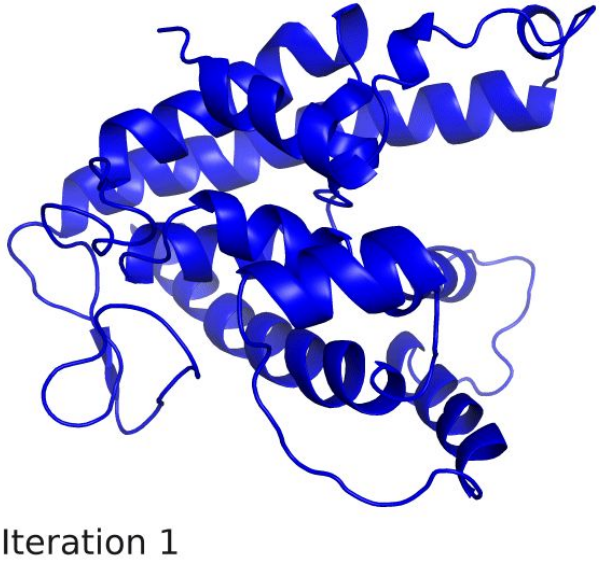


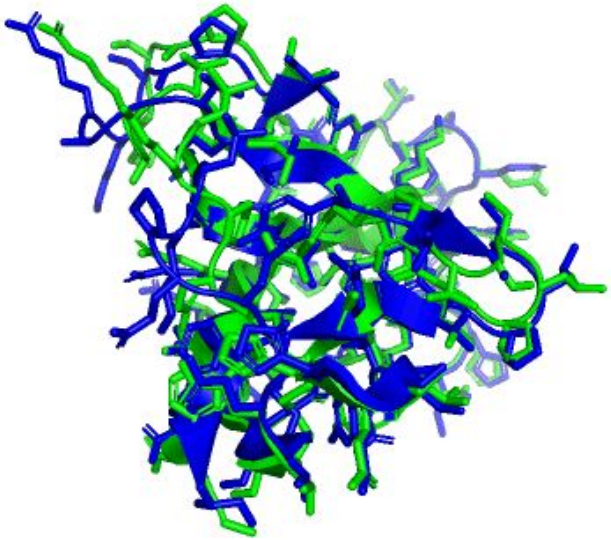
Image: Dcrjsr, vectorised Adam Rędzikowski (CC BY 3.0, Wikipedia)

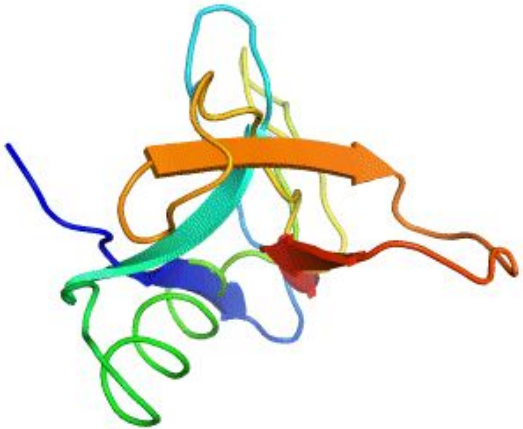
- **Geometric transformer architecture** updates the rigid bodies / backbone
  - Also builds the side chains from torsion angles



Target: T1041

Model interpretability - ORF8 - Sars-Cov2

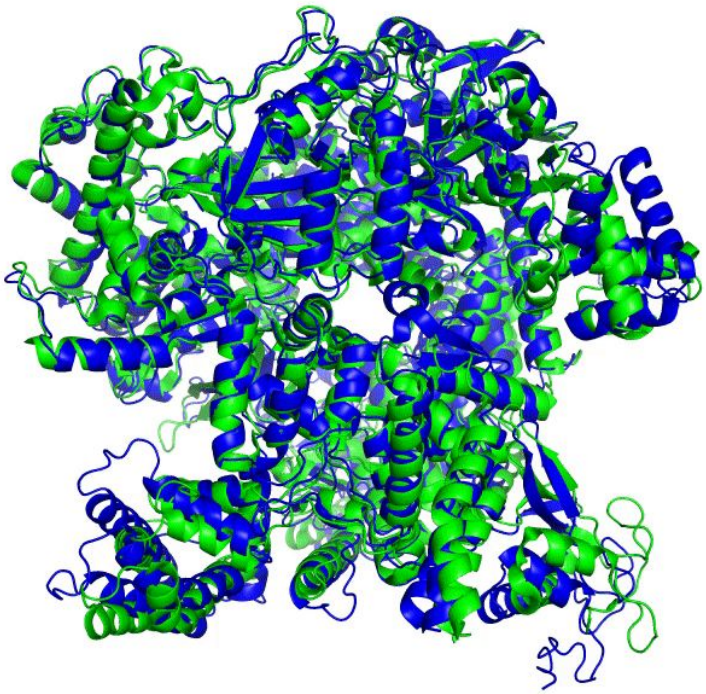


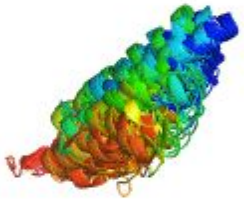


1

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

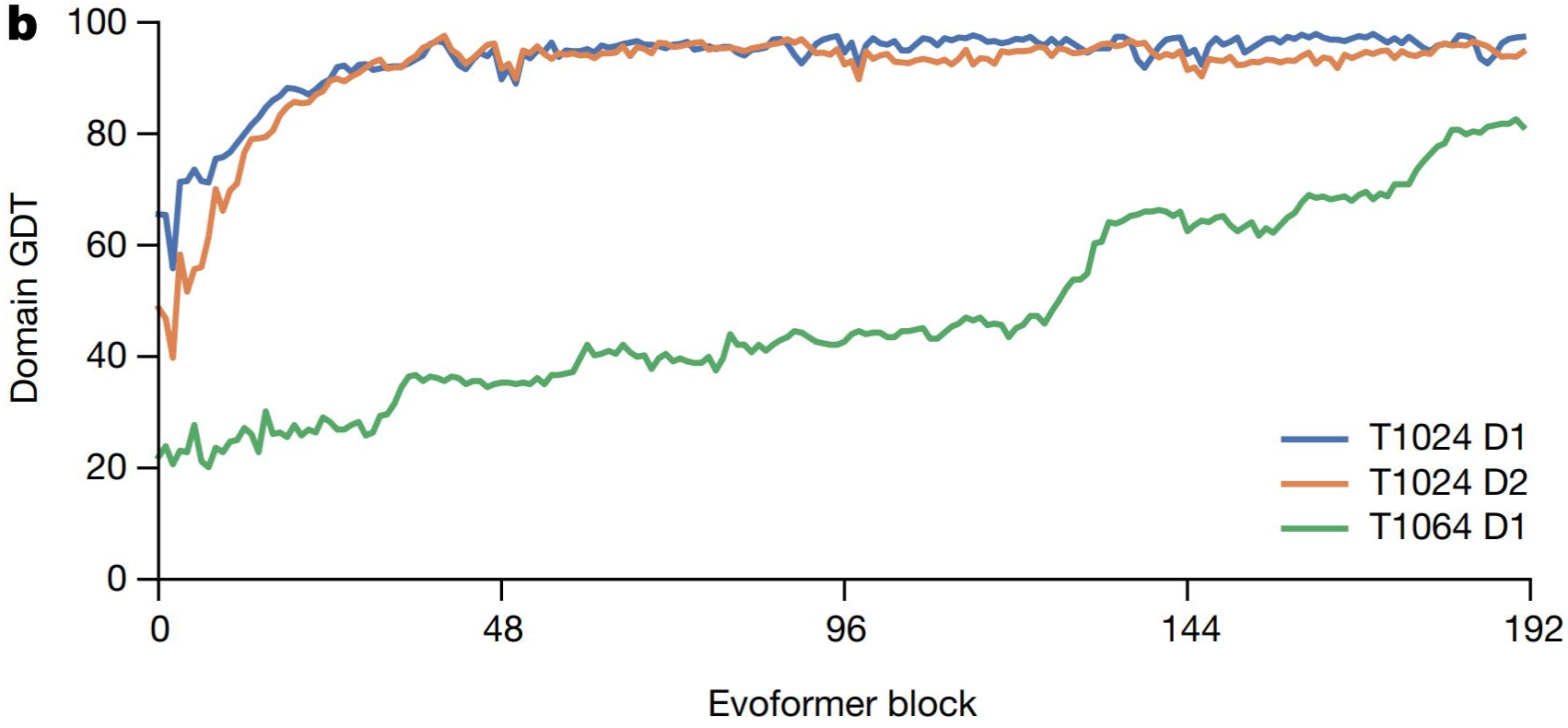
Model interpretability - T1044





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

Model interpretability - Role of depth



DeepMind

AlphaFold and the biology community

Kamil Górecki
@kamil_gorecki85

You really appreciate AlphaFold when you run it on a protein that for a year refused to get expressed and purified...

10:33 pm · 20 Jul 2021 · Twitter for iPhone

Pedro Beltrao
@pedrobeltrao

We joined a large community effort to assess diverse applications of AlphaFold 2 in the context of novel structural elements; missense variants; function and ligand binding sites; modelling of interactions and experimental structural data. Some highlights below:

> **bioRxiv** @biorxivpreprint · 12h
> A structural biology community assessment of AlphaFold 2 applications
> biorxiv.org/cgi/content/sh... #bioRxiv

7:35 AM · Sep 27, 2021 · Twitter Web App

Tristan Croll @CrollTristan · Jul 23

Upshot: while AlphaFold clearly isn't a *replacement* for experimental structures by any stretch, it's already very clear that it's going to make the task of *building* experimental structures both much easier and much less error prone. Welcome to the future! (fin)

3 35 287

Bálint Mészáros
@_BalintMeszaros

ProViz by @DaveyLab now has various AlphaFold-based disordered predictions and secondary structures

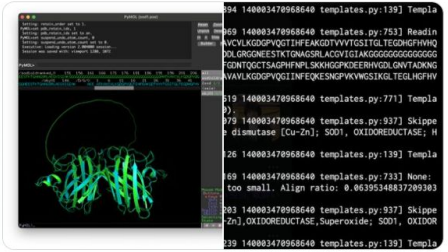
> **Short Linear Motif team** @DaveyLab · Aug 3
> @_BalintMeszaros and I have been staring at the @DeepMind @emblebi structure database non-stop. To see them in context we've updated ProViz to visualise AlphaFold2 data mapped to multiple sequence alignments and data from @uniprot @rcsbPDB @PfamDB tinyurl.com/ProVizAF.
> Show this thread

12:49 PM · Aug 3, 2021 · Twitter Web App

Yoshitaka Moriwaki
@Ag_smith

Prediction from AlphaFold2, SOD1 (Superoxide dismutase), max_template_date=1979-07-19 (No template).

Nevertheless AF2 could successfully predict the homodimer form if we input two SOD1 sequences with a polyG linker.

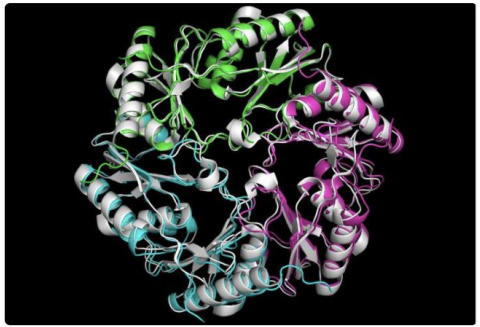


11:39 pm · 19 Jul 2021 · Twitter Web App

Sergey Ovchinnikov @sokrypton

Homooligomeric prediction in #alphafold works a little too good. So far worked on nearly every case we (me & @minkbaek) tried. Going beyond dimers! Seems @DeepMind accidentally "solved" the homooligomeric prediction problem (w/ MSA input) 😂 Give it a try:
https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb
https://pbs.twimg.com/media/E62EYFSXEAIIvCa.jpg

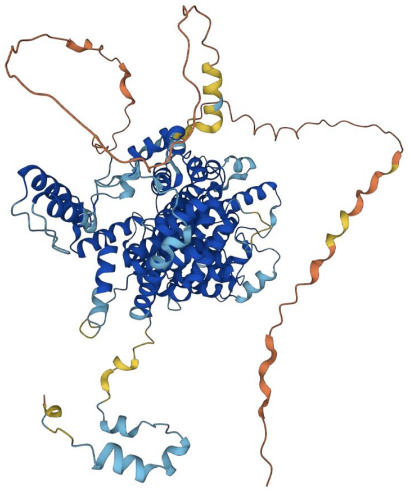
Twitter | Jul 21st (159 kB)



Communicating confidence is key

- AlphaFold outputs two confidence metrics: **pLDDT** and **PAE**
- Important for sharing predictions responsibly
- Supports new use cases:
  - Software integrations
  - Domain segmentation
  - Detecting interactions
  - Disorder prediction
  - Ranking structural models . . .

pLDDT



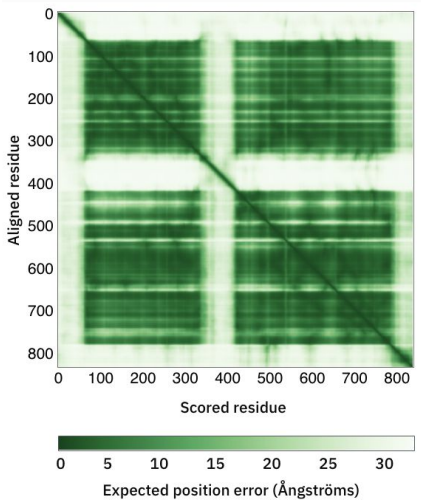
Very high (>90)

Confident (70-90)

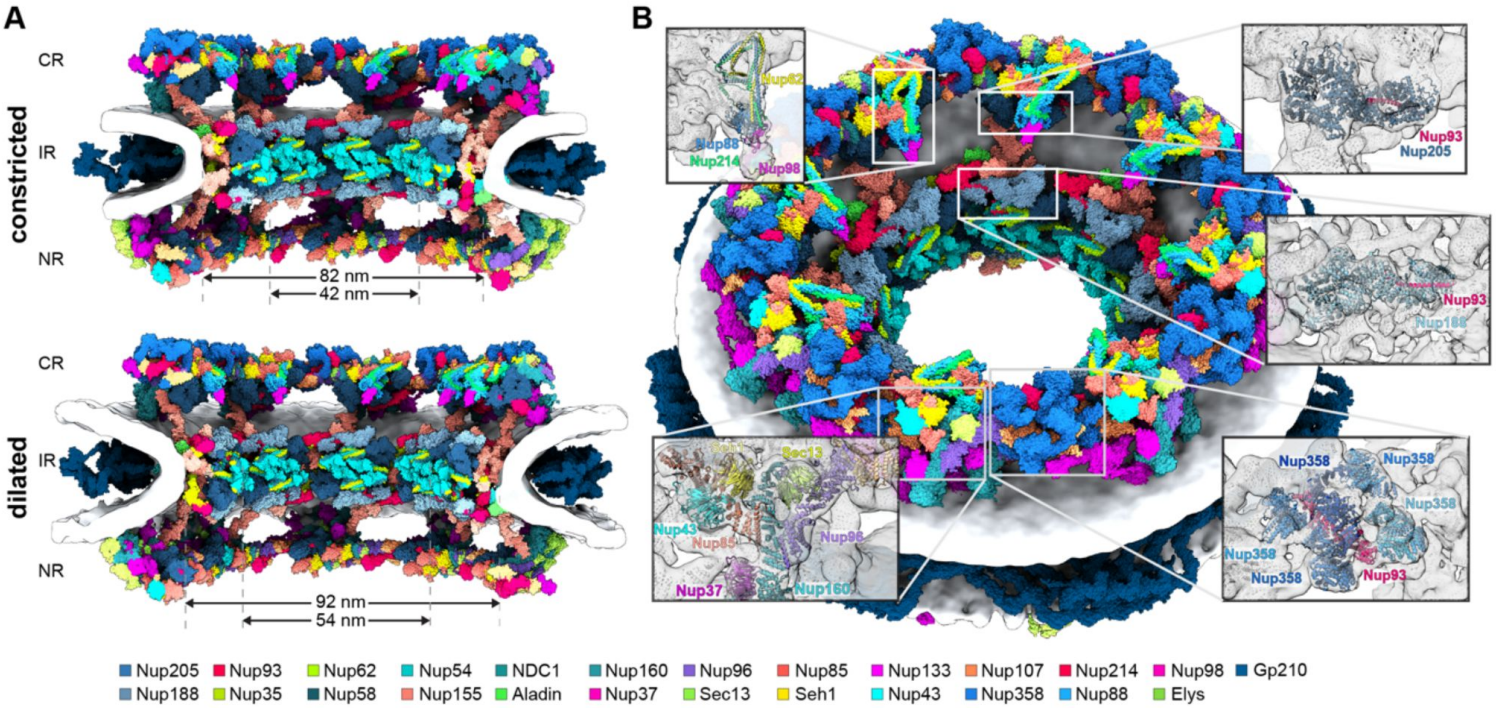
Low (50-70)

Very low (<50)

PAE



Studying molecular machines



Mosalaganti, S. et al. AI-based structure prediction empowers integrative structural analysis of human nuclear pores. Science (2022)

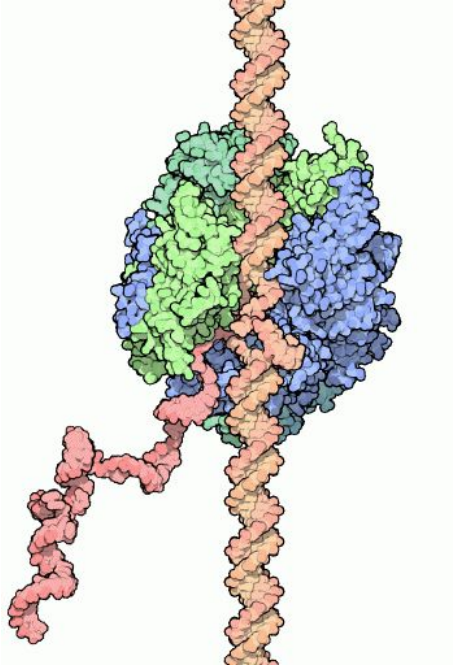
DeepMind

What's next?

What's next? Three major challenges

Proteins in the cellular context

- AlphaFold-Multimer: multiple protein chains
- Need to also consider non-protein components:
  DNA, RNA, ligands, water, ions…

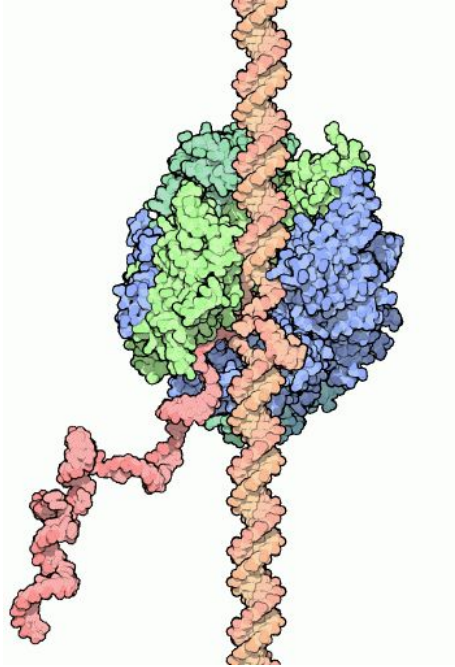


DESRES-ANTON-11021571 https://www.deshawresearch.com/downloads/download_trajectory_sarscov2.cgi/
RNA polymerase, PDB-101 https://pdb101.rcsb.org/motm/40

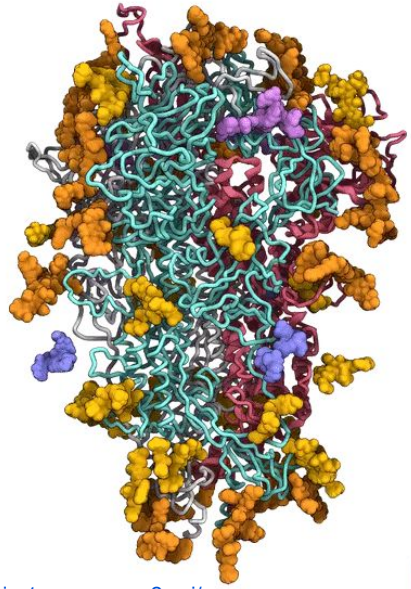
What's next? Three major challenges

Proteins in the cellular context

- AlphaFold-Multimer: multiple protein chains
- Need to also consider non-protein components:
  DNA, RNA, ligands, water, ions…



Protein dynamics



Effect of mutations on proteins

DESRES-ANTON-11021571 https://www.deshawresearch.com/downloads/download_trajectory_sarscov2.cgi/
RNA polymerase, PDB-101 https://pdb101.rcsb.org/motm/40

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

DeepMind

Lattice QCD work at DeepMind

Enabling Lattice QCD simulations to a wide community and accelerate progress in our understanding of matter

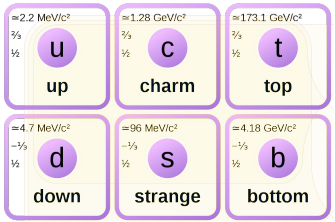
DeepMind

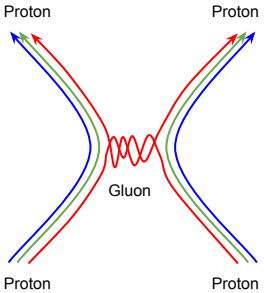
The problem; scientific viewpoint:

- Simulate quantum fluctuations of gauge fields (gluons, photons) and fermionic fields (quarks, electrons) in QFT using generative models
- Use the model's samples to compute physical observables (such as particle masses)

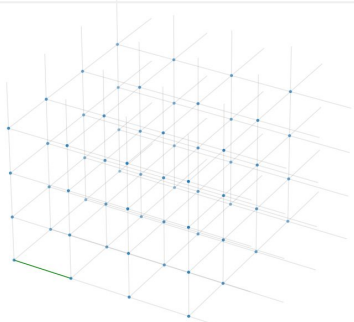
The problem; machine learning viewpoint:

- **Not a data fitting problem!**
- Inference problem:
  - Learn a target density in a 4D lattice of complex-valued matrices and vectors
  - Minimize a divergence using model sampler (e.g. reverse Kullback-Leibler)
- Requires enormous amount of memory due to lattice size





$$p(U) \propto e^{-\beta S[U]}$$



1. *Sampling using SU(N) gauge equivariant flows*, **Phys. Rev. D**, Boyda, Kanwar, Racaniere, Rezende, Albergo, Cranmer, Hackett, Shanahan
2. *Gauge-equivariant flow models for sampling in lattice field theories with pseudofermions*, **Phys. Rev. D**, Abbott, Albergo, Boyda, Cranmer, Hackett, Kanwar, Racanière, Rezende, Romero-López, Shanahan, Tian, Urban

DeepMind

Thank you!

Q&A