AlphaFold 2

Kovács Gábor
research / engineering @ Turbine.AI

AlphaFold 'pushes science forward' by releasing structures of almost all human proteins

DeepMind's AI predicted over 365,000 protein structures, which are now freely available online

by ***Emily Harwitz***

July 29, 2021 | A version of this story appeared in **Volume 99, Issue 28**

DeepMind

BLOG POST
RESEARCH

30 NOV 2020

AlphaFold: a solution to a 50-year-old grand challenge in biology

AlphaFold 2 is here: what's behind the structure prediction miracle

Nature has now released that AlphaFold 2 paper, after eight long months of waiting. The main text reports more or less what we have known for nearly a year, with some added tidbits, although it is accompanied by a painstaking description of the architecture in the supplementary information. Perhaps more importantly, the authors have released the entirety of the code, including all details to run the pipeline, on Github. And there is no small print this time: you can run inference on *any* protein (I've checked!).

Oct 3, 2021, 07:34pm EDT | 54,391 views

AlphaFold Is The Most Important Achievement In AI —Ever

Rob Toews Contributor

AI

I write about the big picture of artificial intelligence.

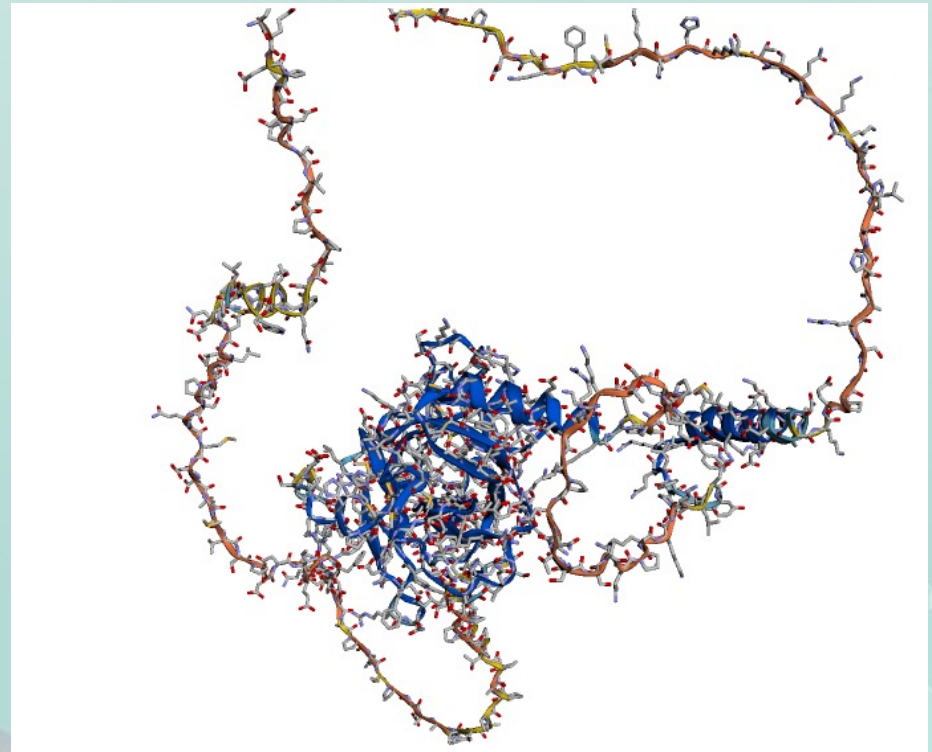
Follow

DeepMind open-sources AlphaFold 2 for protein structure predictions

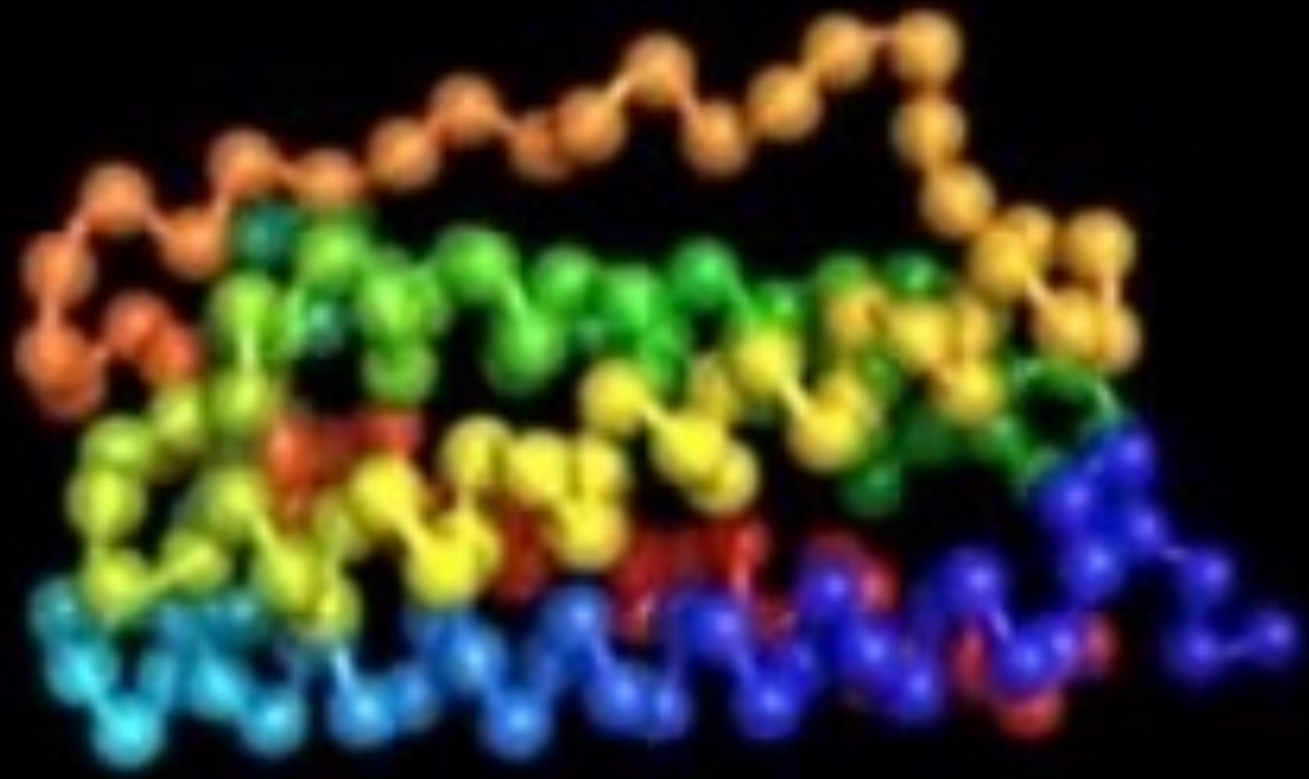
Why is AlphaFold a big deal?

- An **end-to-end** neural network for the protein folding problem
- Proteins are everywhere (digestion, muscles, firing neurons, immune system....)
- Predict 3D structure from amino acid (AA) sequence

```
sequence: "MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLSPDDIEQWFTEDPGPDEAPRMPEAAPPVAPAPAAPTPAAPAPAPS"
```

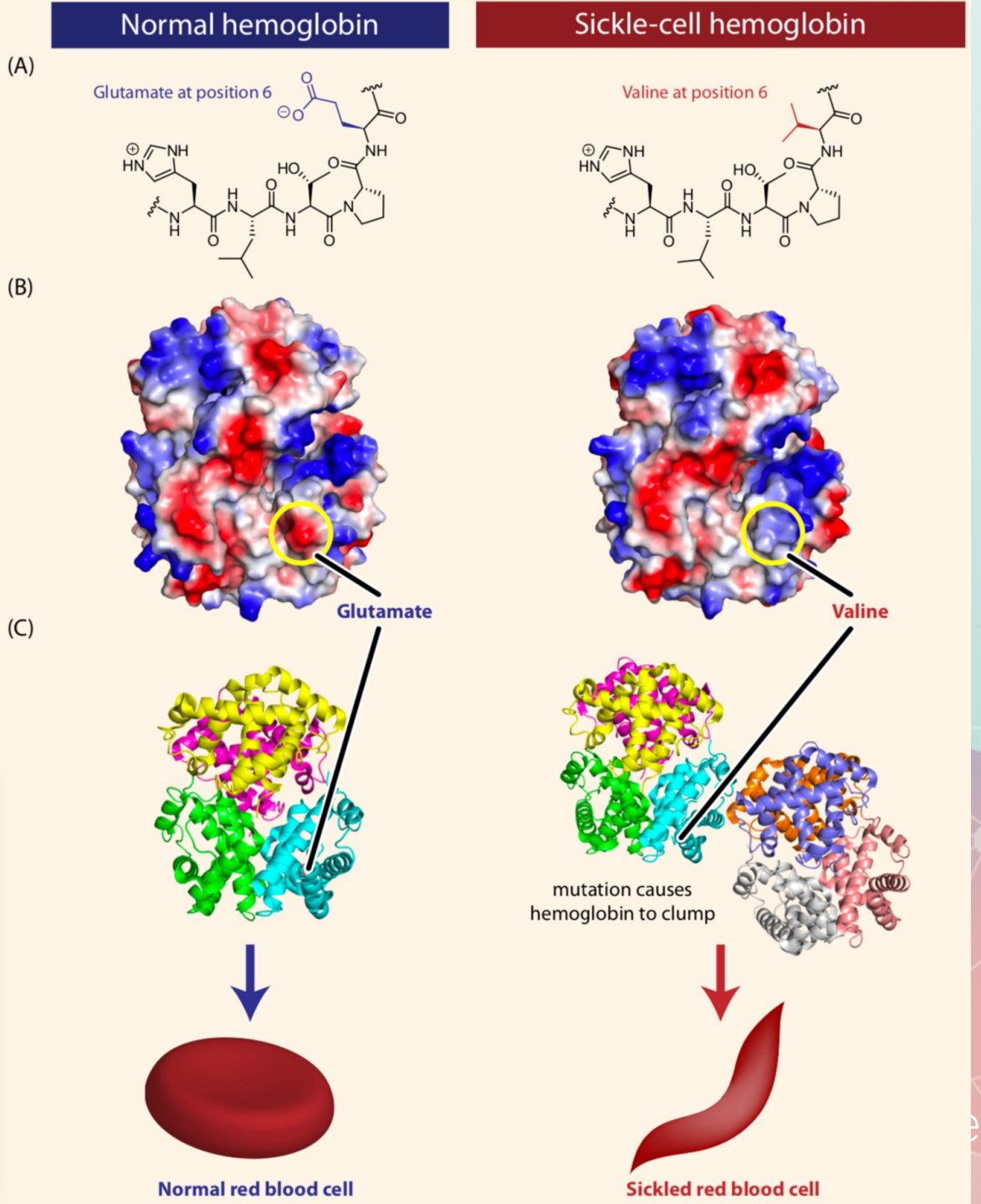


turbine



Why is AlphaFold a big deal?

- Why is it important?
  - Shape -> implies function
  - Design new drugs
  - Outside medicine: design proteins that digest plastic…



Why is AlphaFold a big deal?

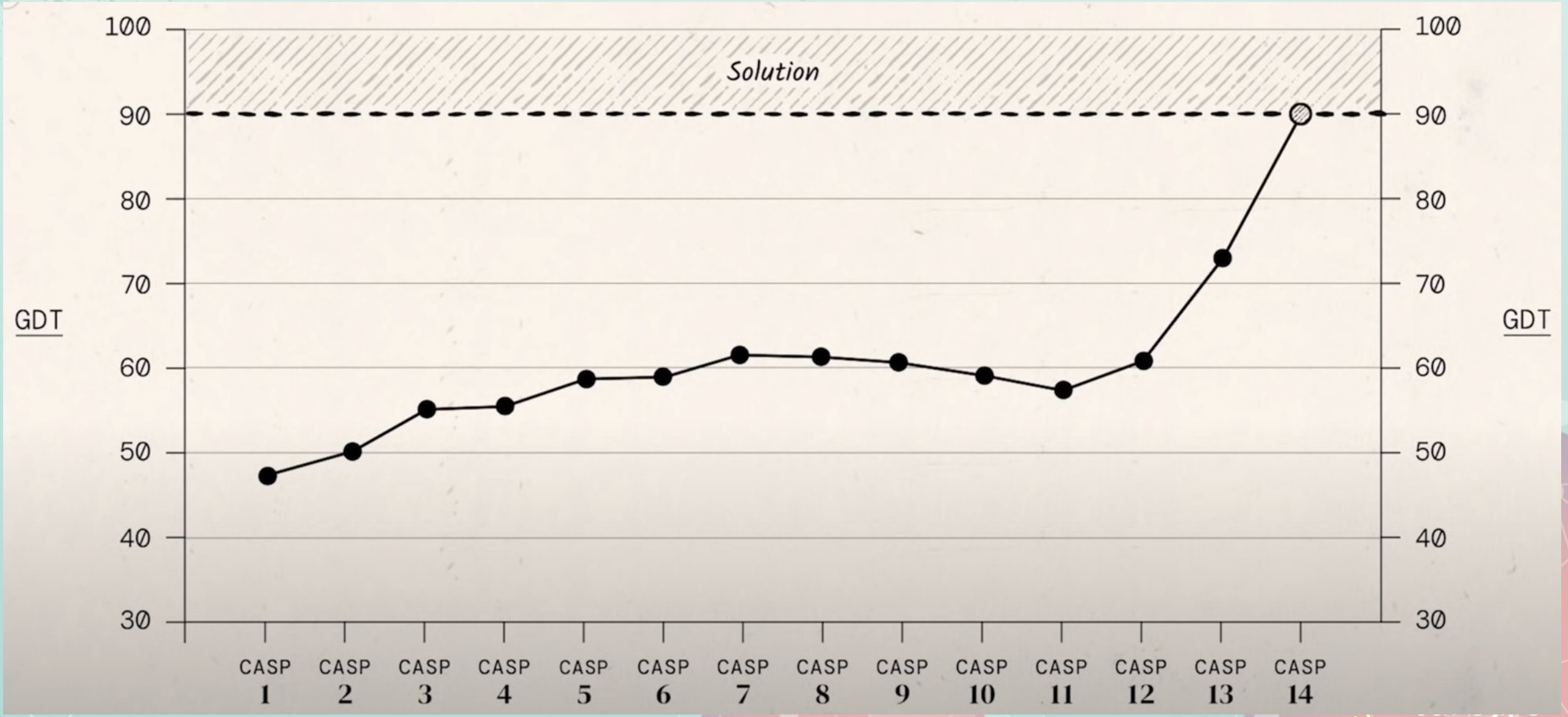
- Why is it hard?
  - Measuring structure experimentally: slow & expensive (X-ray crystallography)
  - ~100K measured --> 100s of billions potential configurations
  - Levinthal's paradox
    - 10^300 conformations
    - Happens in msec scale in life
    - Nature doesn't brute force folding…

Why is AlphaFold a big deal?

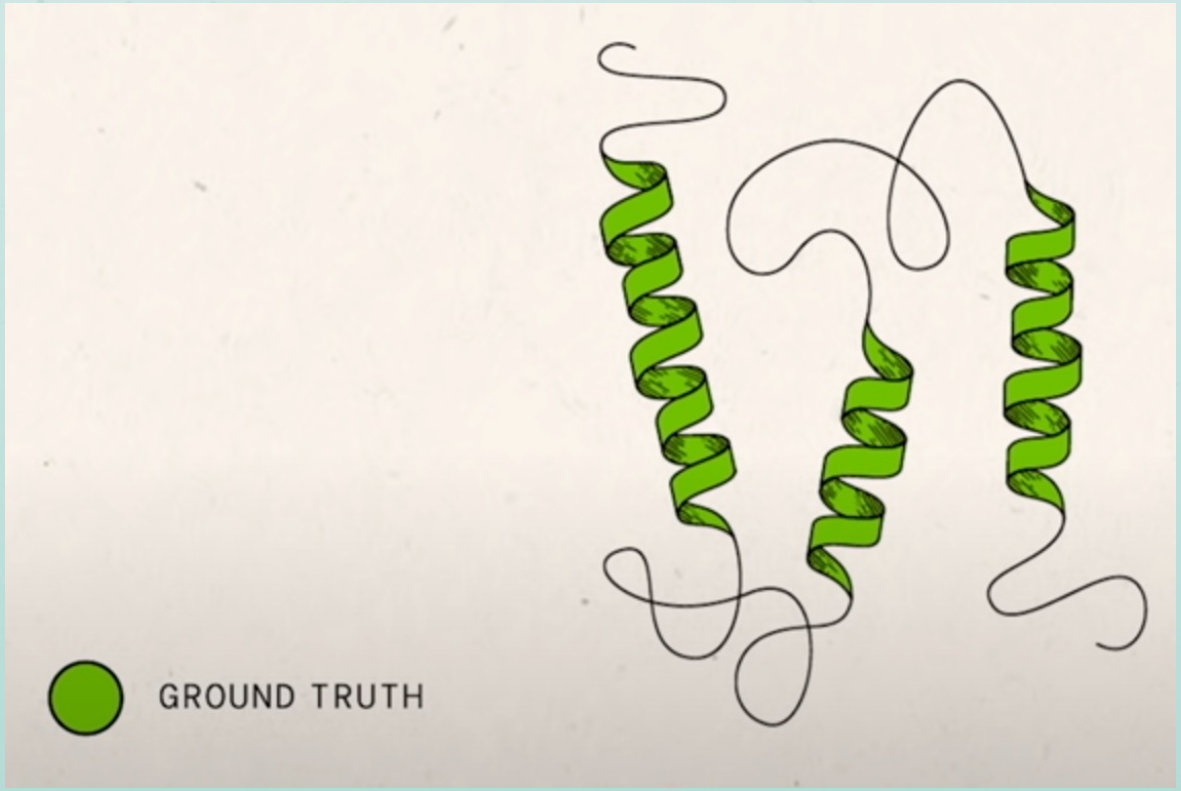
- Predictive model classes
  - **Physical** interactions-based methods
  - **Evolutionary history**-based methods
  - Since 2018: **neural networks** emerge on this field with dominating results

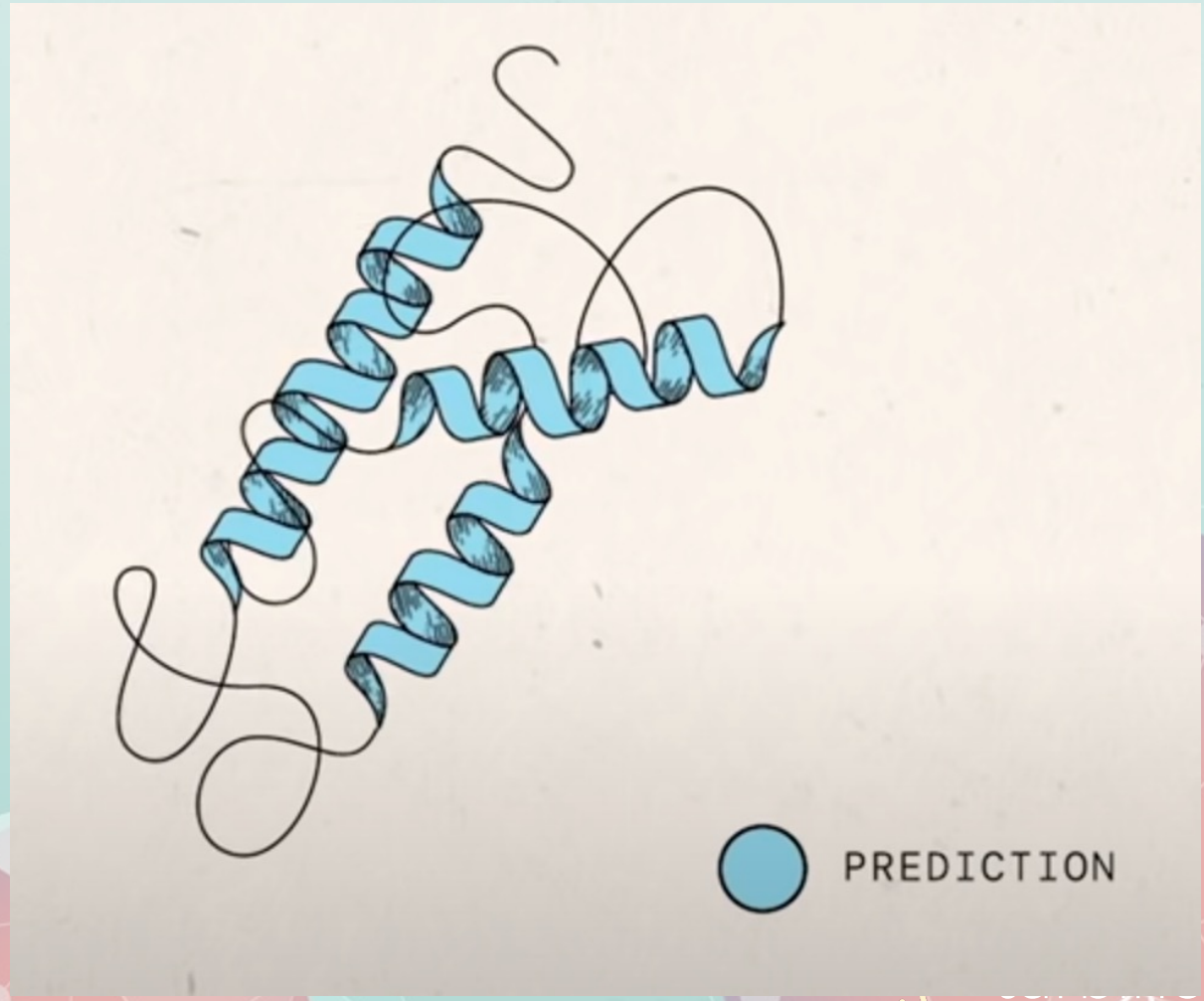
- CASP - Critical Assessment of Techniques for Protein Structure Prediction
  - Biennial challenge for models to predict yet unpublished prot. structures
  - DeepMind's AlphaFold 2 model wins CASP 14 event... by a lot

Why do they say: "problem solved" ?

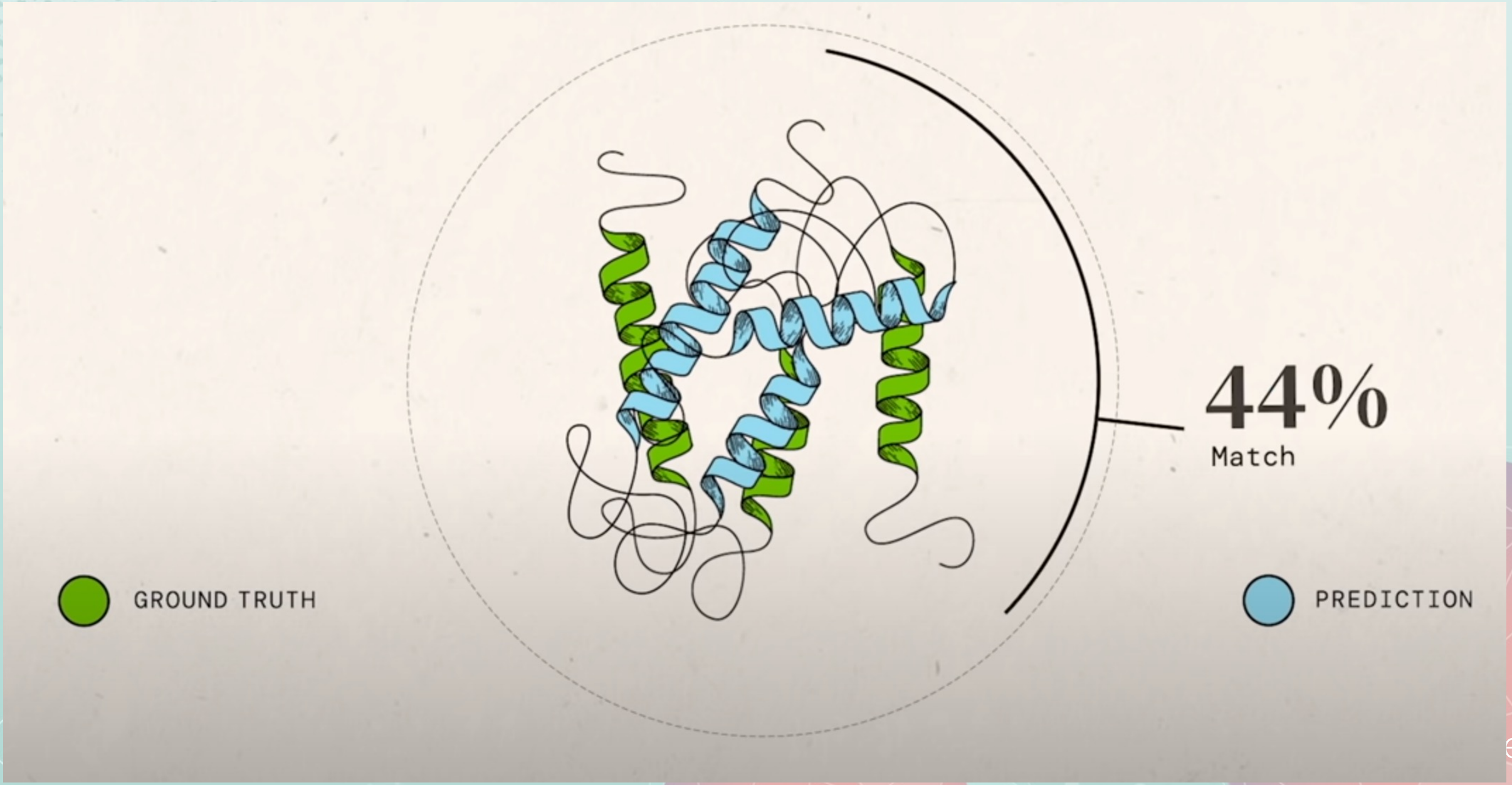


Why do they say: "problem solved" ?

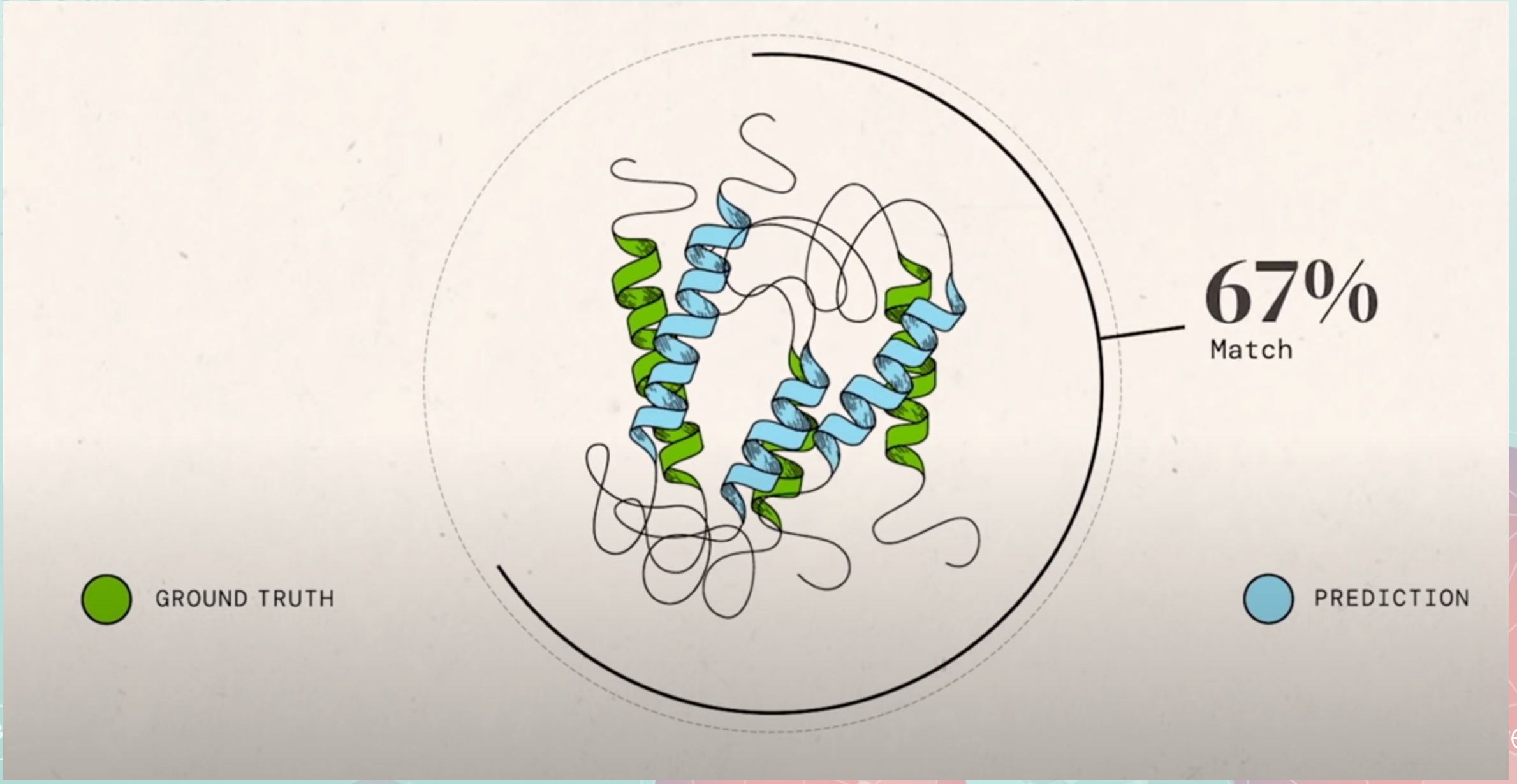




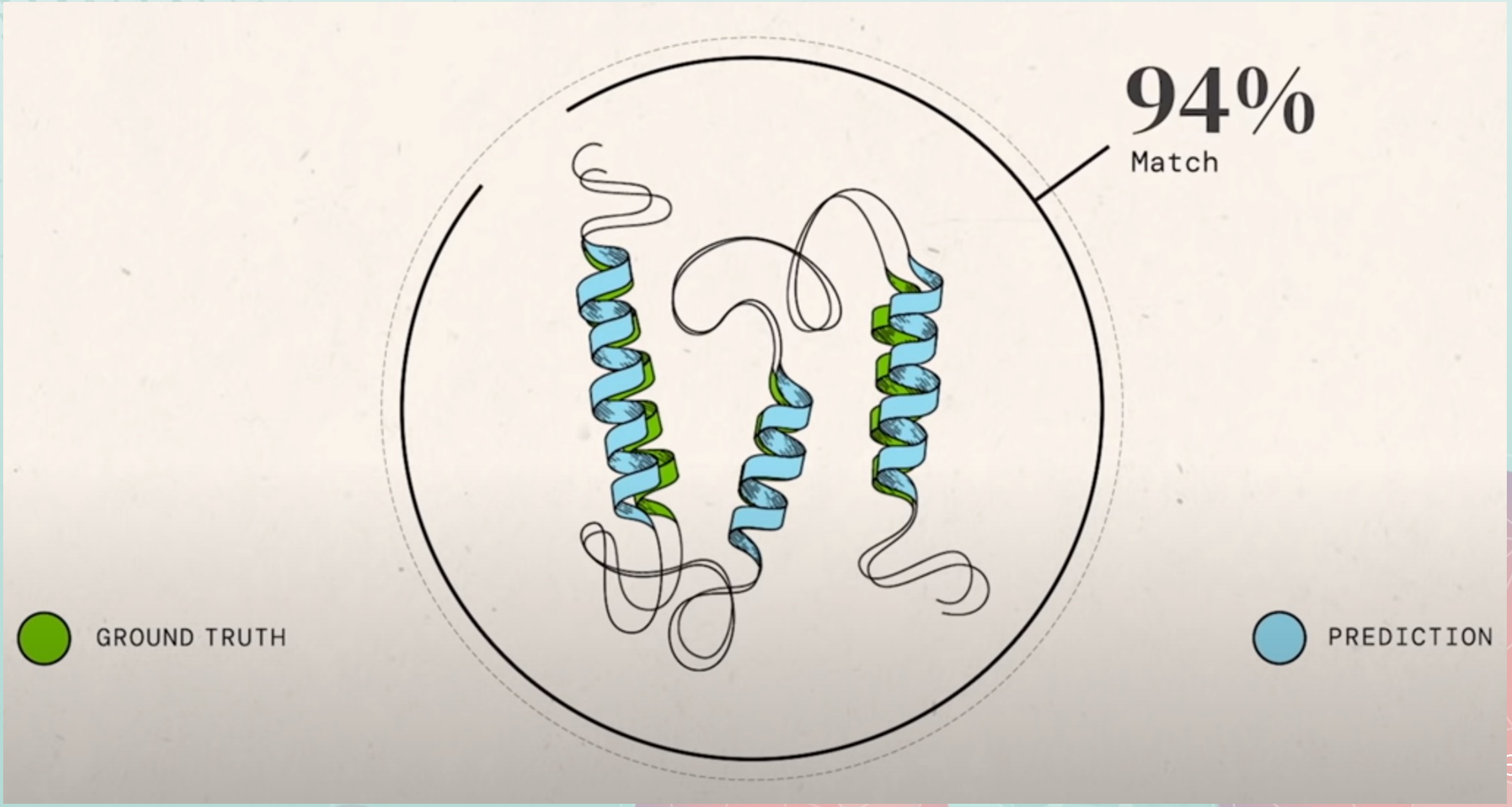
Why do they say: "problem solved" ? - Global distance test (GDT)



Why do they say: "problem solved" ? - Global distance test (GDT)

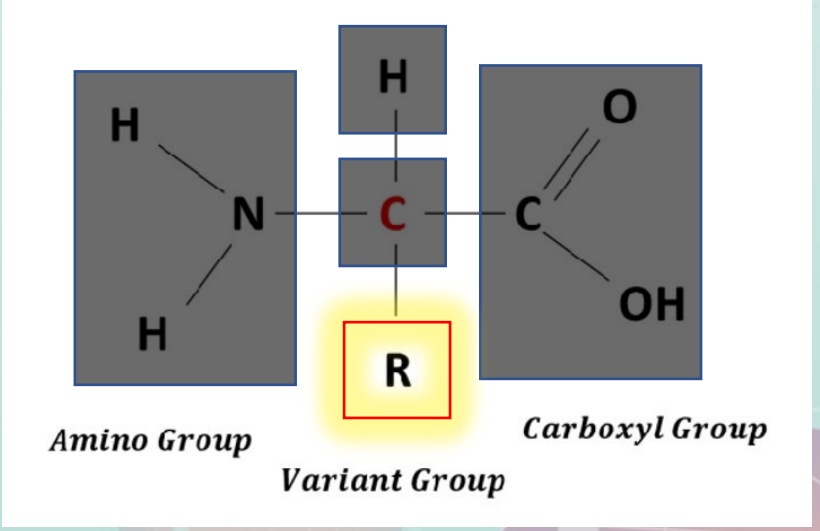


Why do they say: "problem solved" ? - Global distance test (GDT)



Biology background – a layman's view

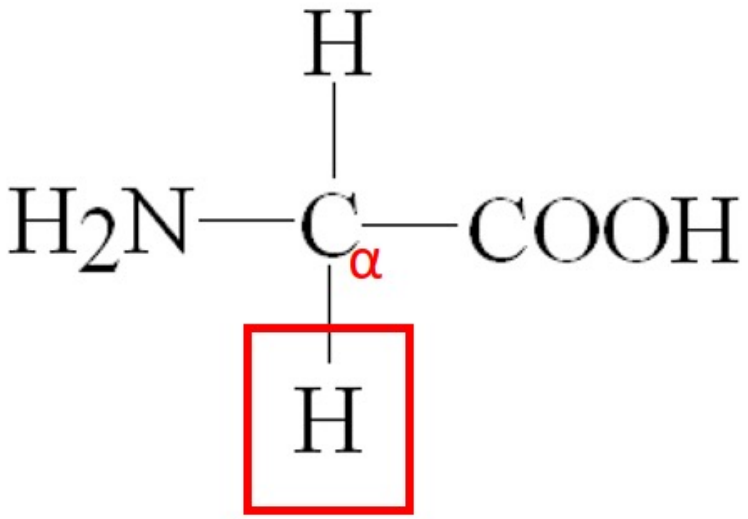
- Inputs are proteins, defined by their amino acid (AA) sequences.
- But what is an amino acid?



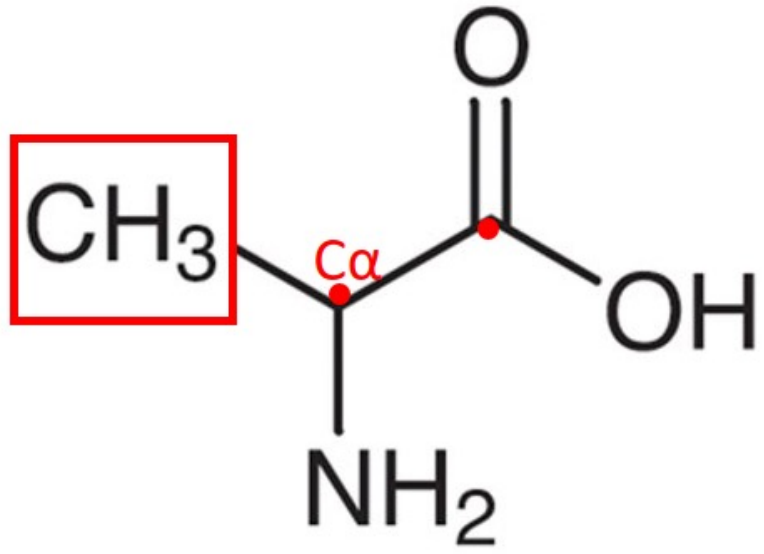
- Cα - alpha carbon
- NH2 – amino group
- COOH – carboxyl group
- R – residue / R group

- R groups are very important. They determine each AA's **identity**
- Variations of R groups determine characteristics of molecules that form when AAs are joined together

Biology background – a layman's view



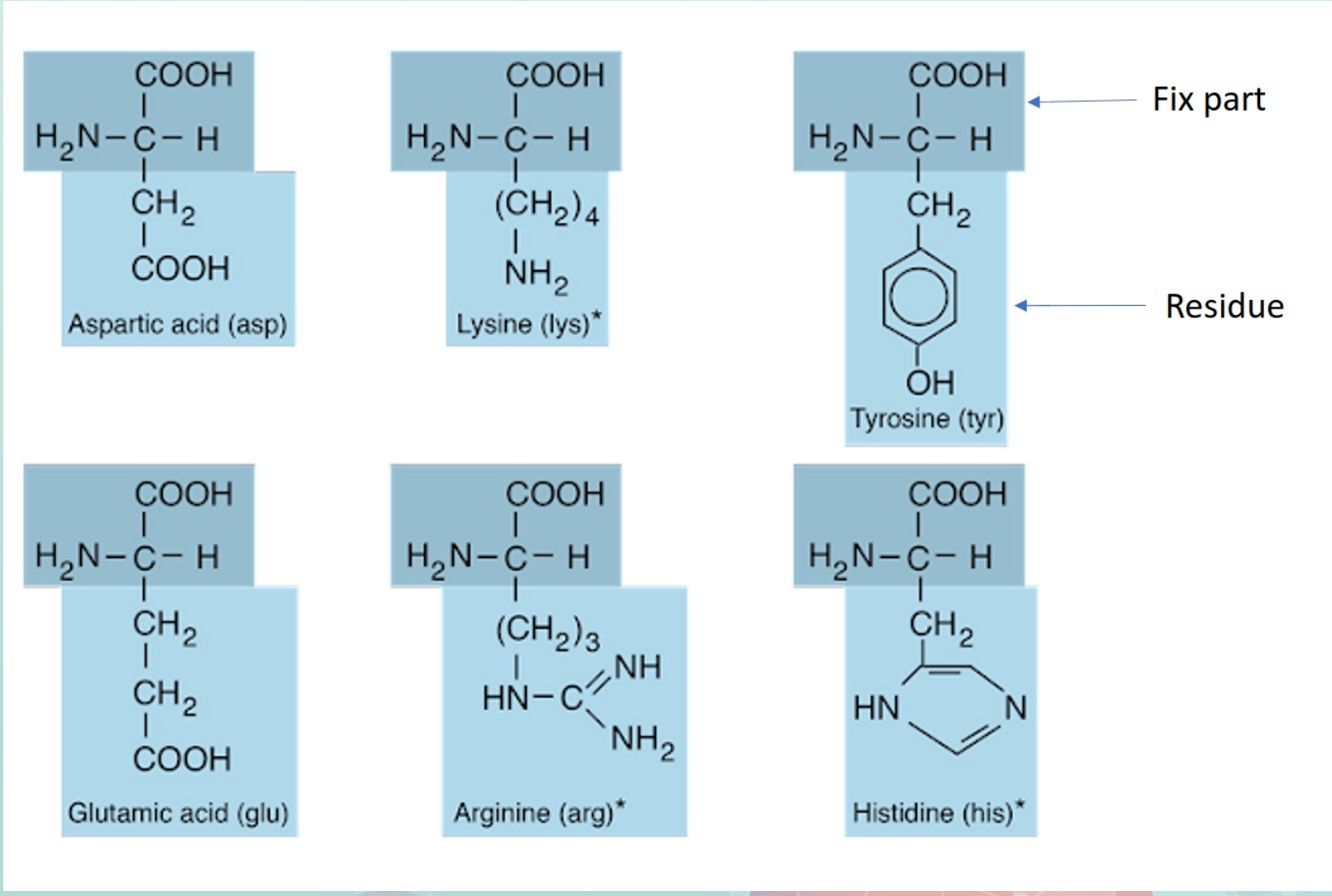
glycine



alanine

turbine

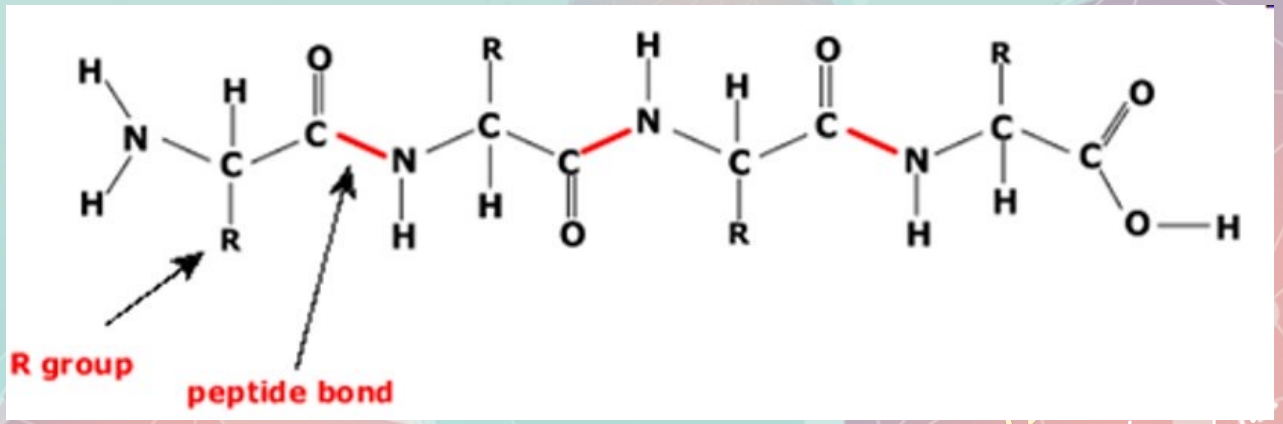
Amino acids: "alphabet" --- Proteins: "very long words"



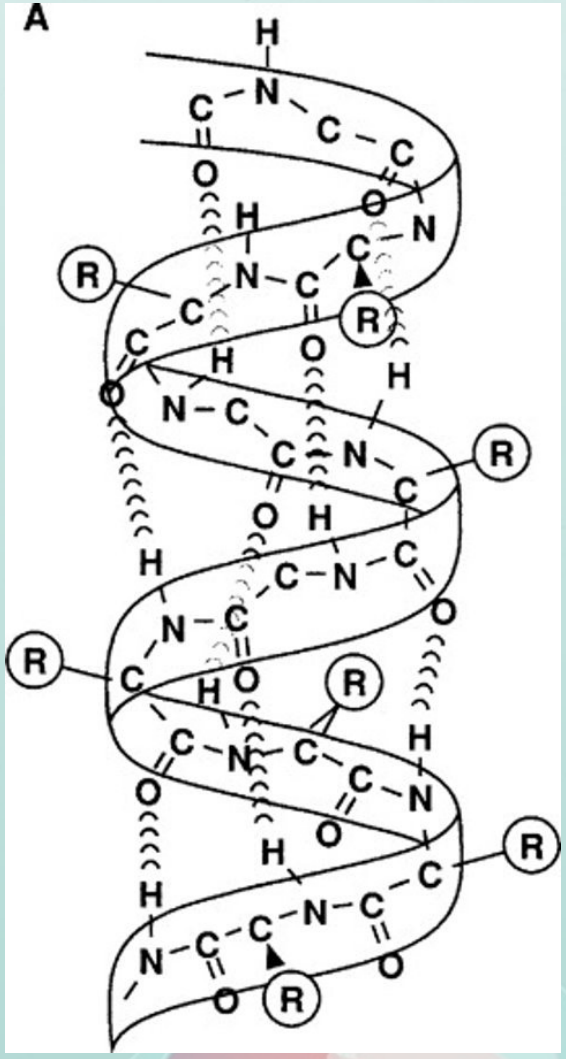
turbine

From AA sequence --> to 3D structure

- Backbone is essentially planar & rigid due to π bond (covalent subtype)
- σ bond between Cα and R group -> can freely rotate
- Through these σ bonds' rotations local structures are formed
- R groups of 2 AAs can interact by hydrogen bond / electrostatically



Secondary structure

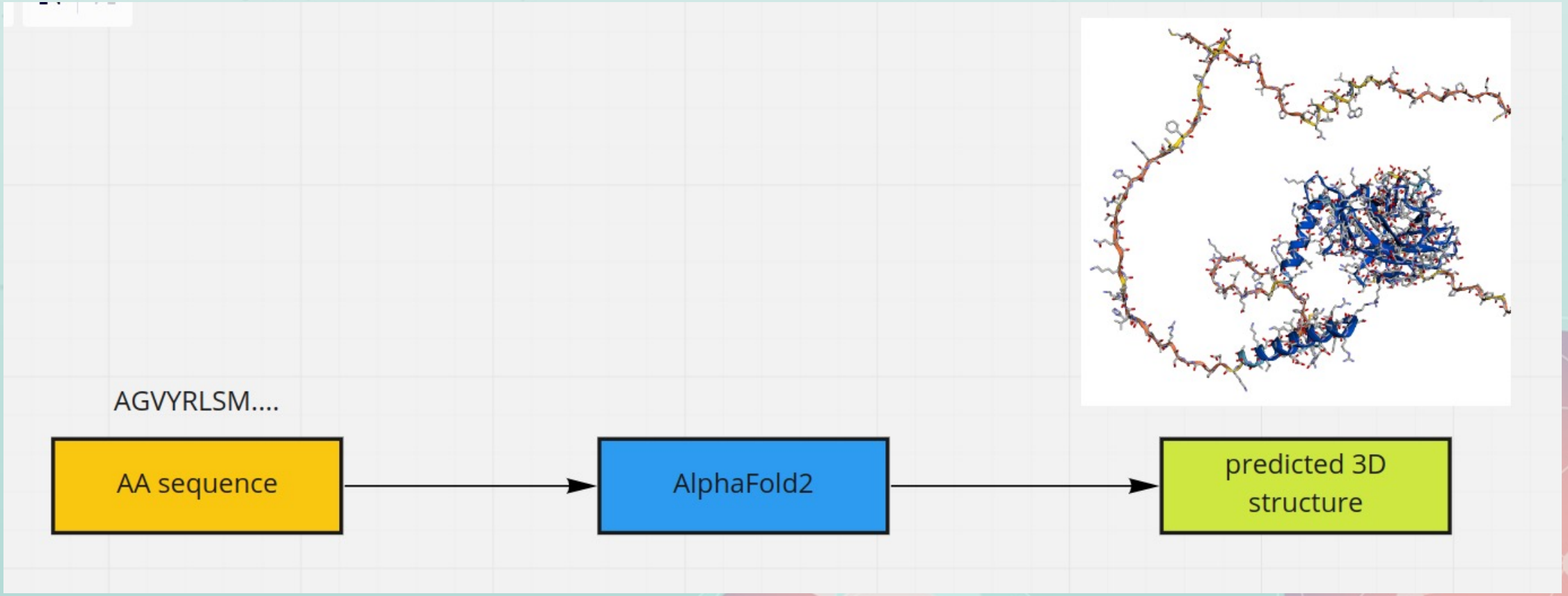


alpha helix

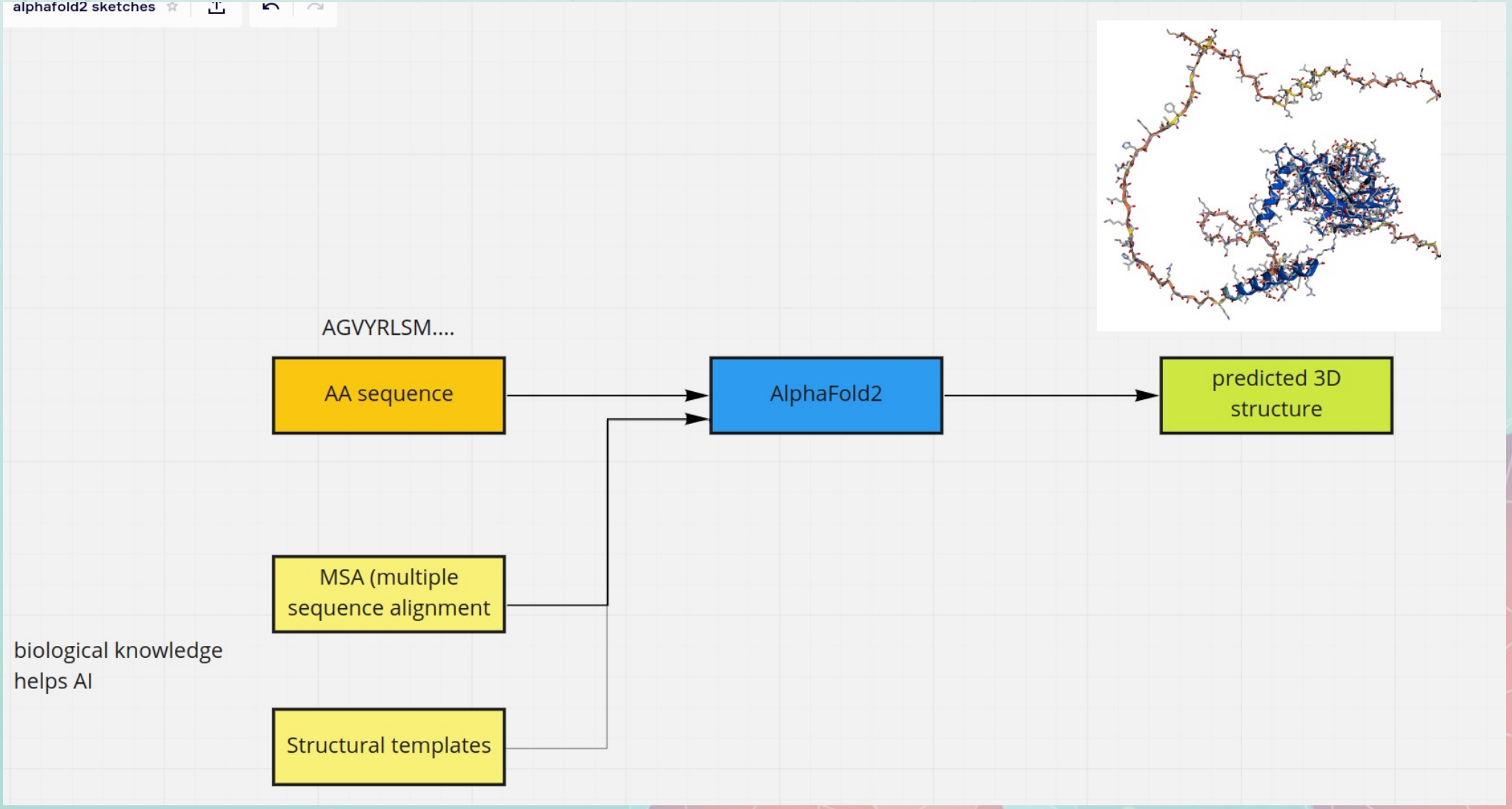
turbine

Tertiary structure

AlphaFold 2 architecture



AlphaFold 2 architecture



AlphaFold 2 inputs – MSA – Multiple Sequence Alignment

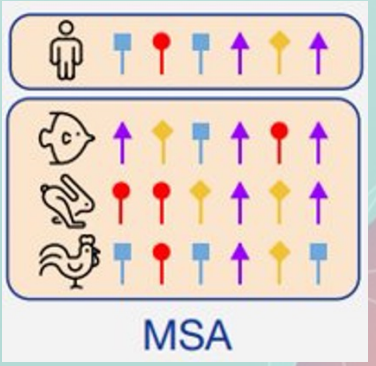
Query sequence to fold: PAWKFIQLLYP....

Bioinfo databases

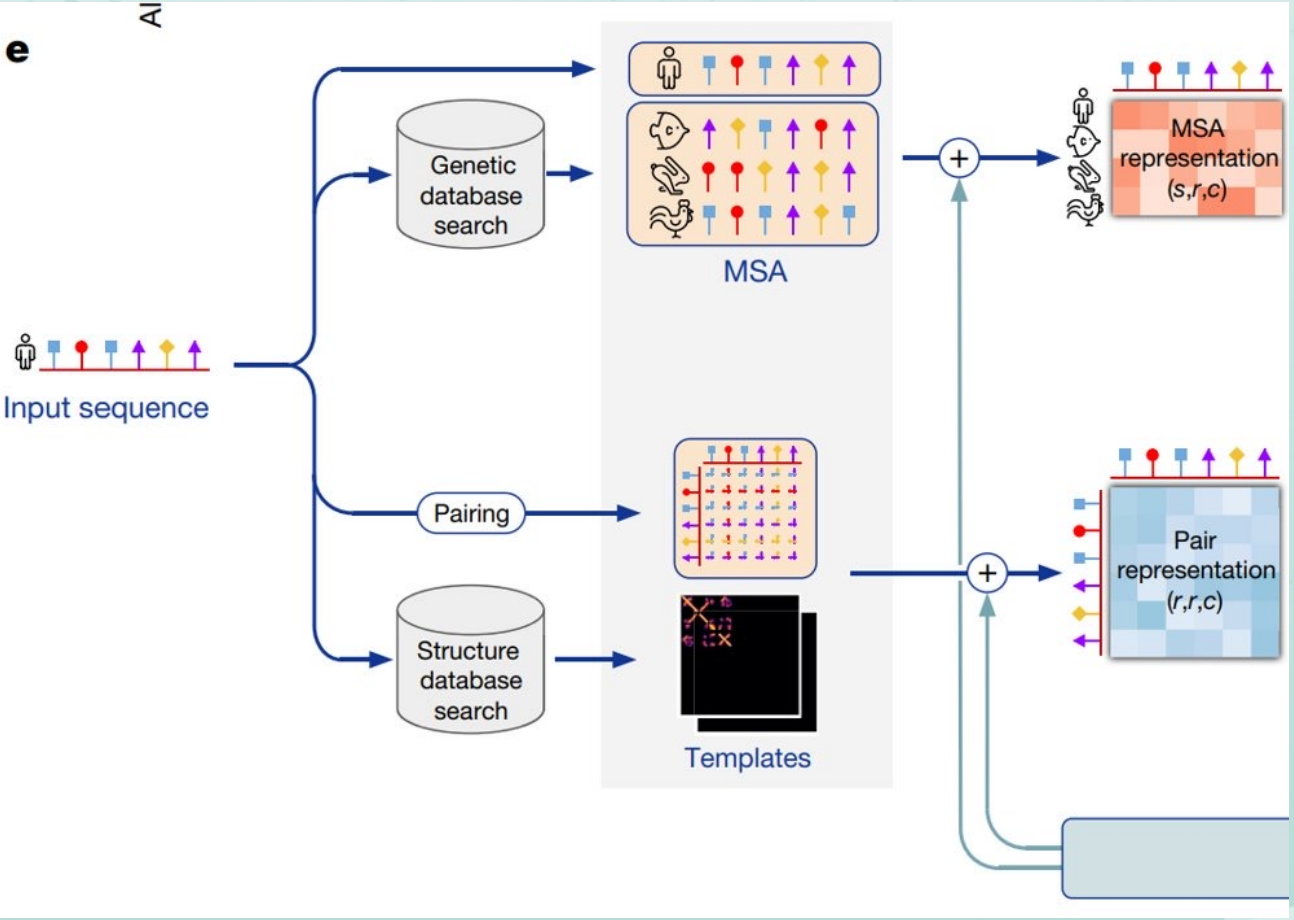
Result set. Each record is a similar matching pattern

```
Q5E940_BOVIN ----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
  RLA0_HUMAN ----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
  RLA0_MOUSE ----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
    RLA0_RAT ----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
  RLA0_CHICK ----------MPREDRATWKSNYFMKIIQLLDDYPKCFVVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
  RLA0_RANSY ----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--SALE
Q7ZUG3_BRARE ----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQTIRLSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
```

Convert into a matrix to feed into a neural network

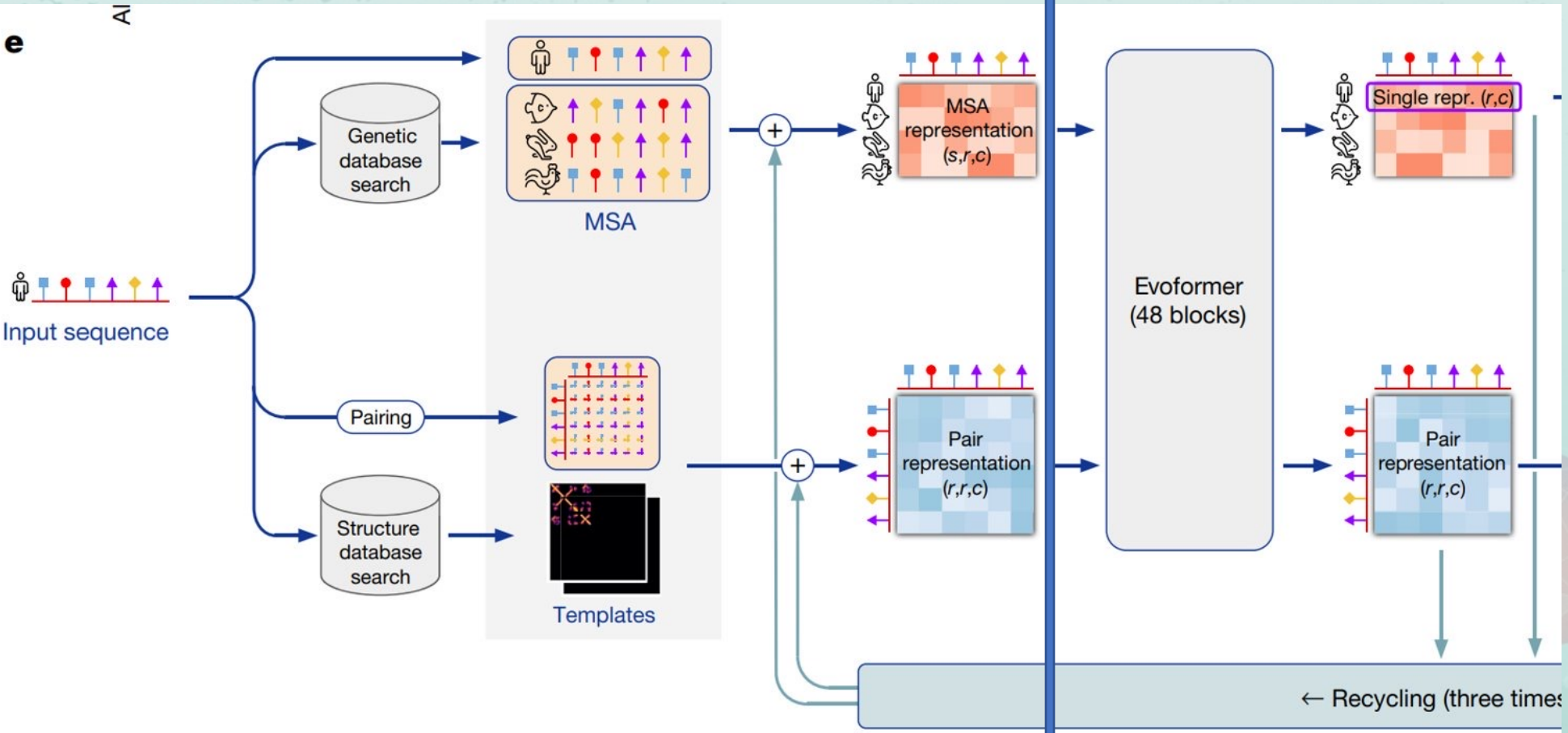


Feature representation



Feature representation

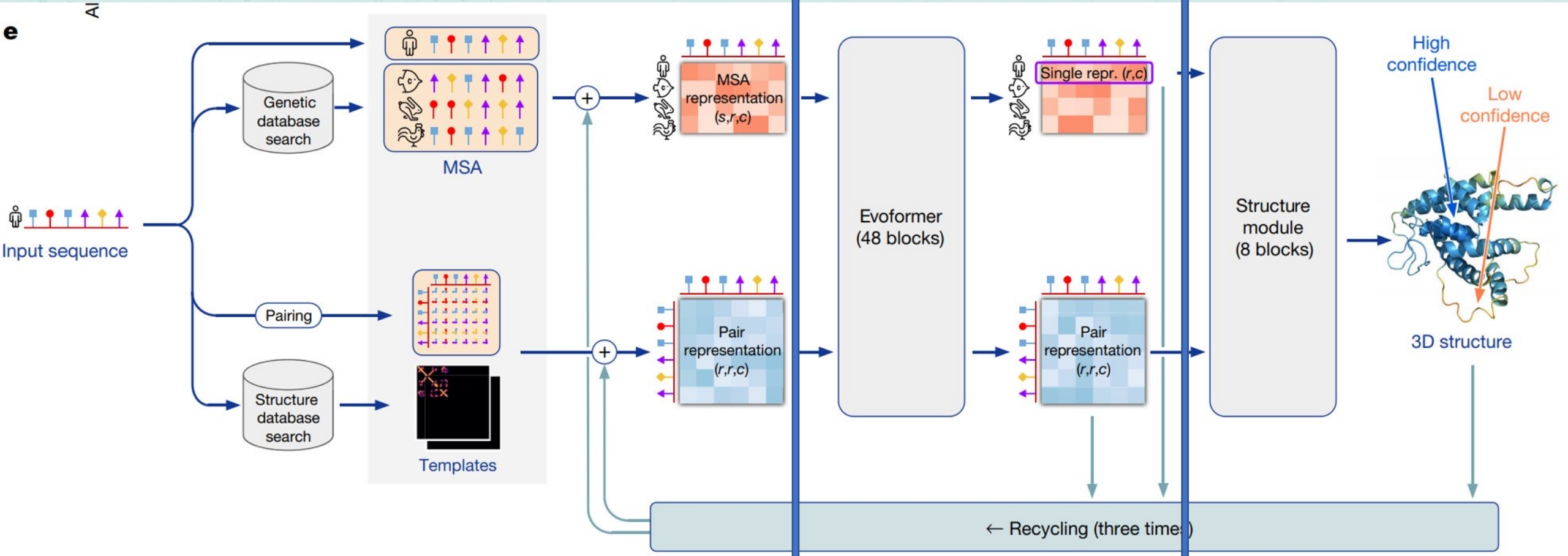
Evoformer



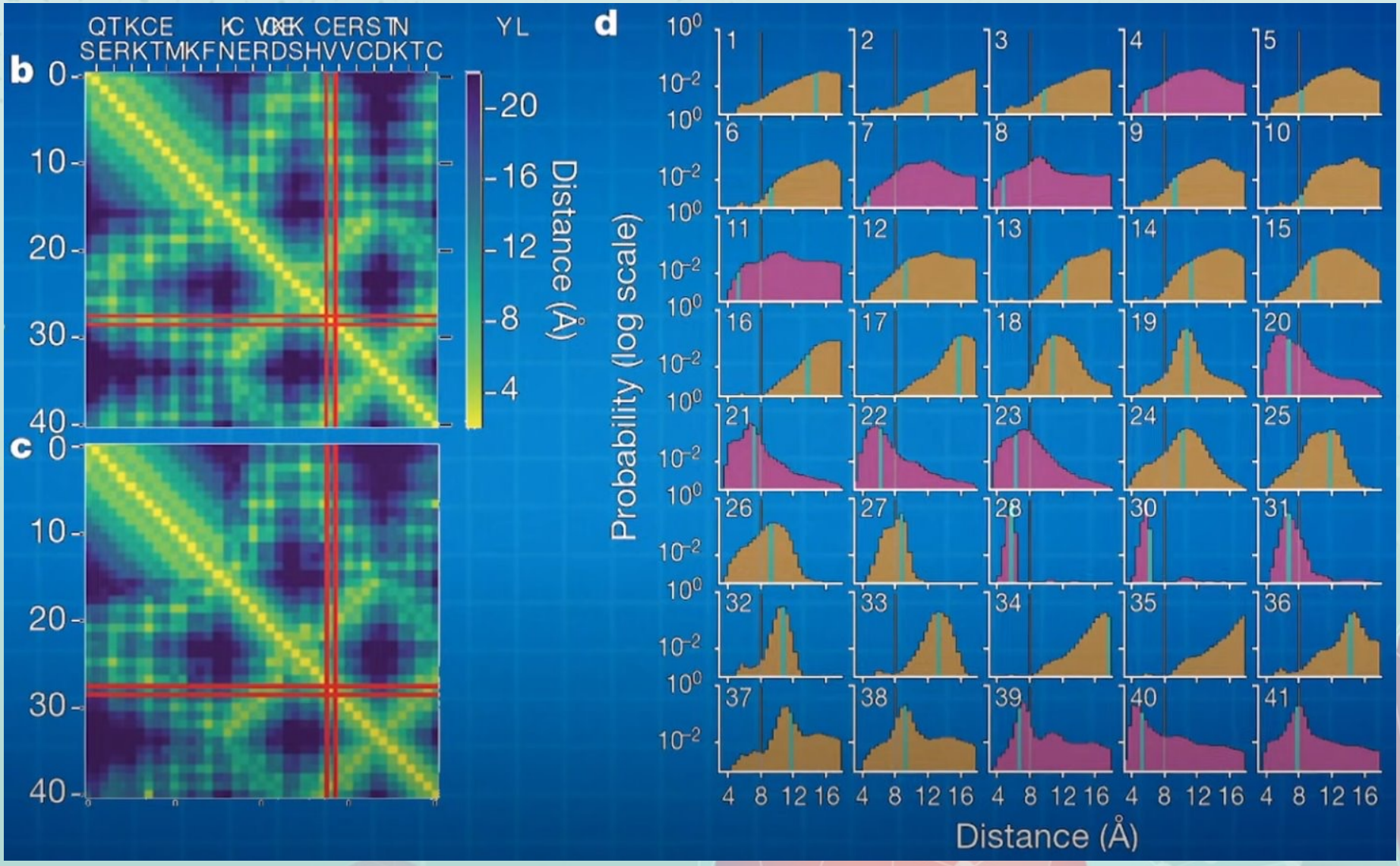
Feature representation

Evoformer

Structure module

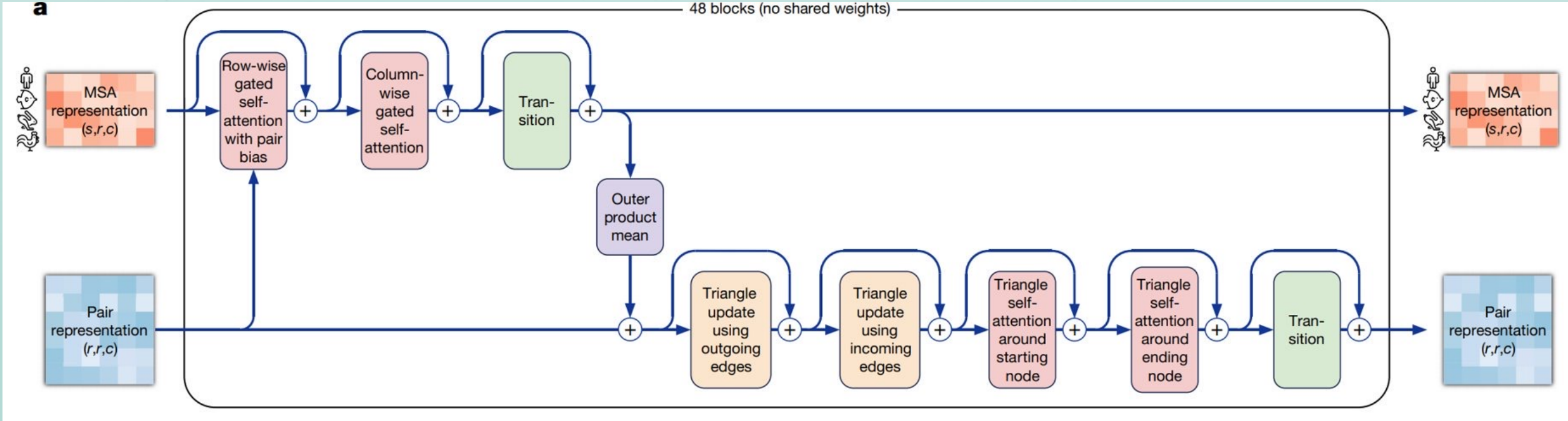


Step 2 – Pairwise representations (distograms)

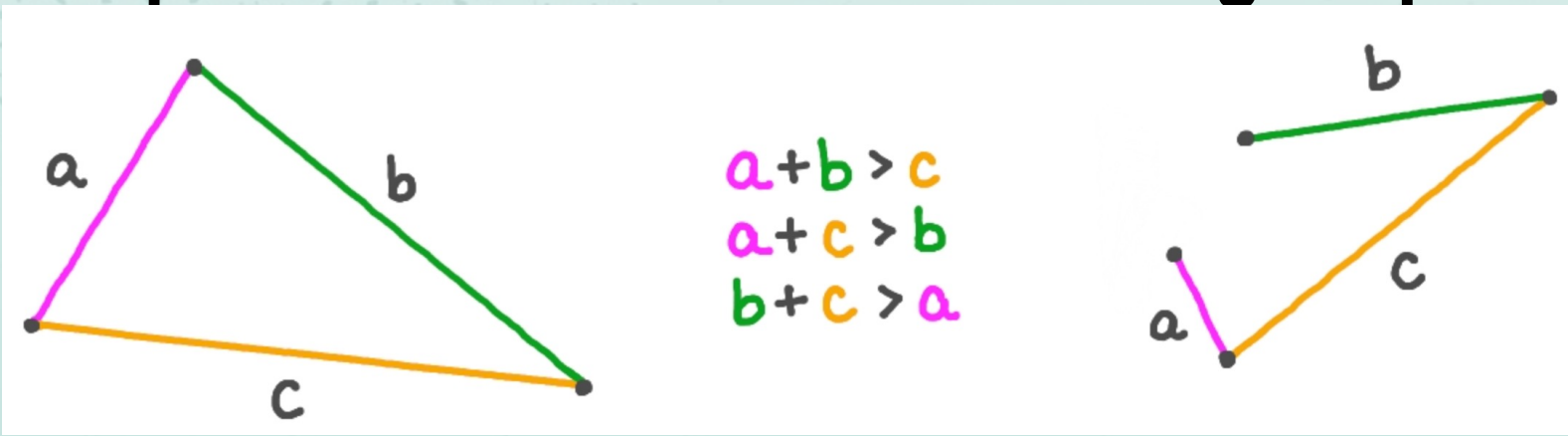


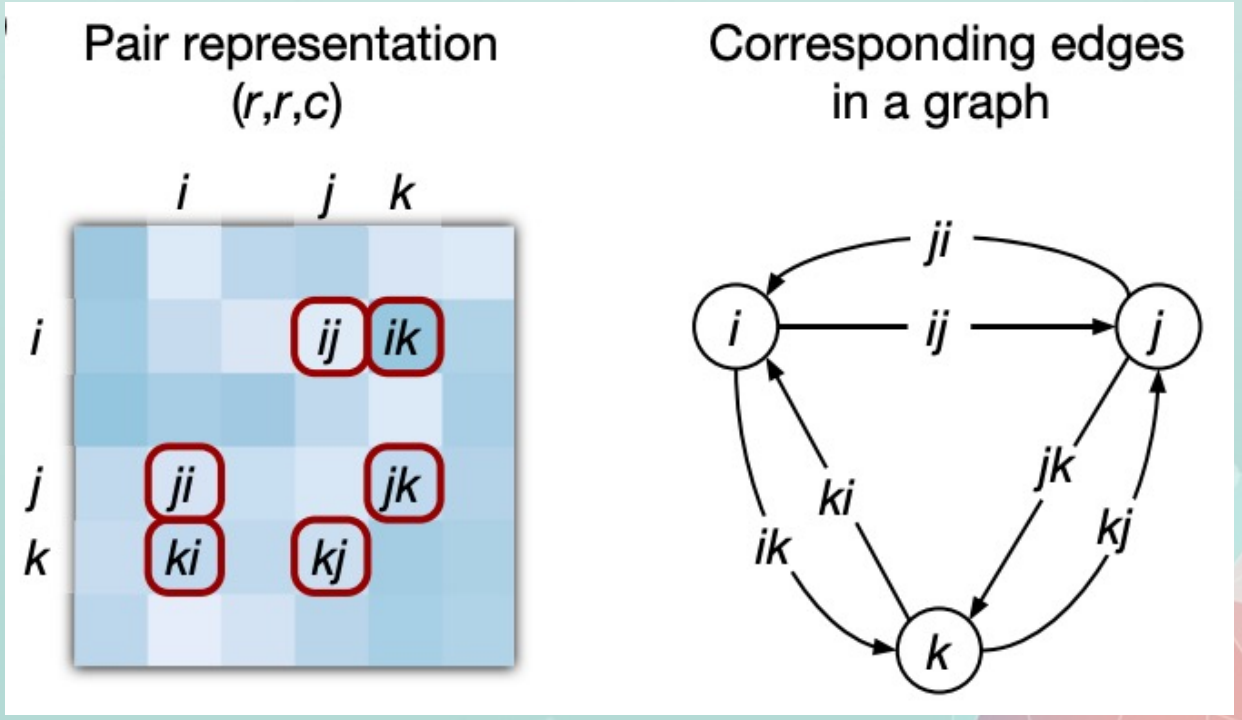
Step 2 – Evoformer stack – inside 1 block

- Attention vs convolution
- 2 way of information flow for the 2 modalities
- Row-wise & column-wise multihead self-attention



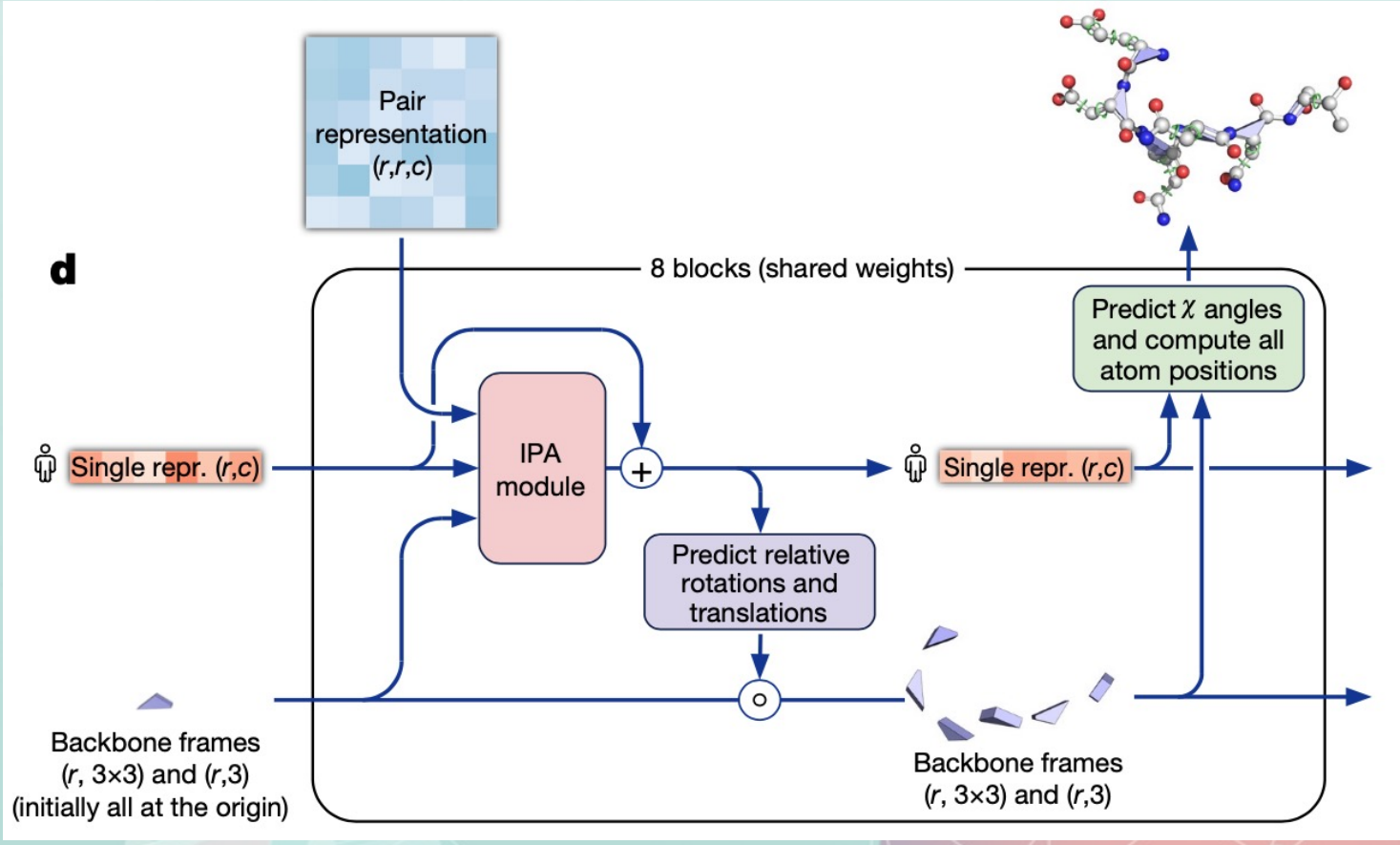
Step 2 – Evoformer stack – triangle update

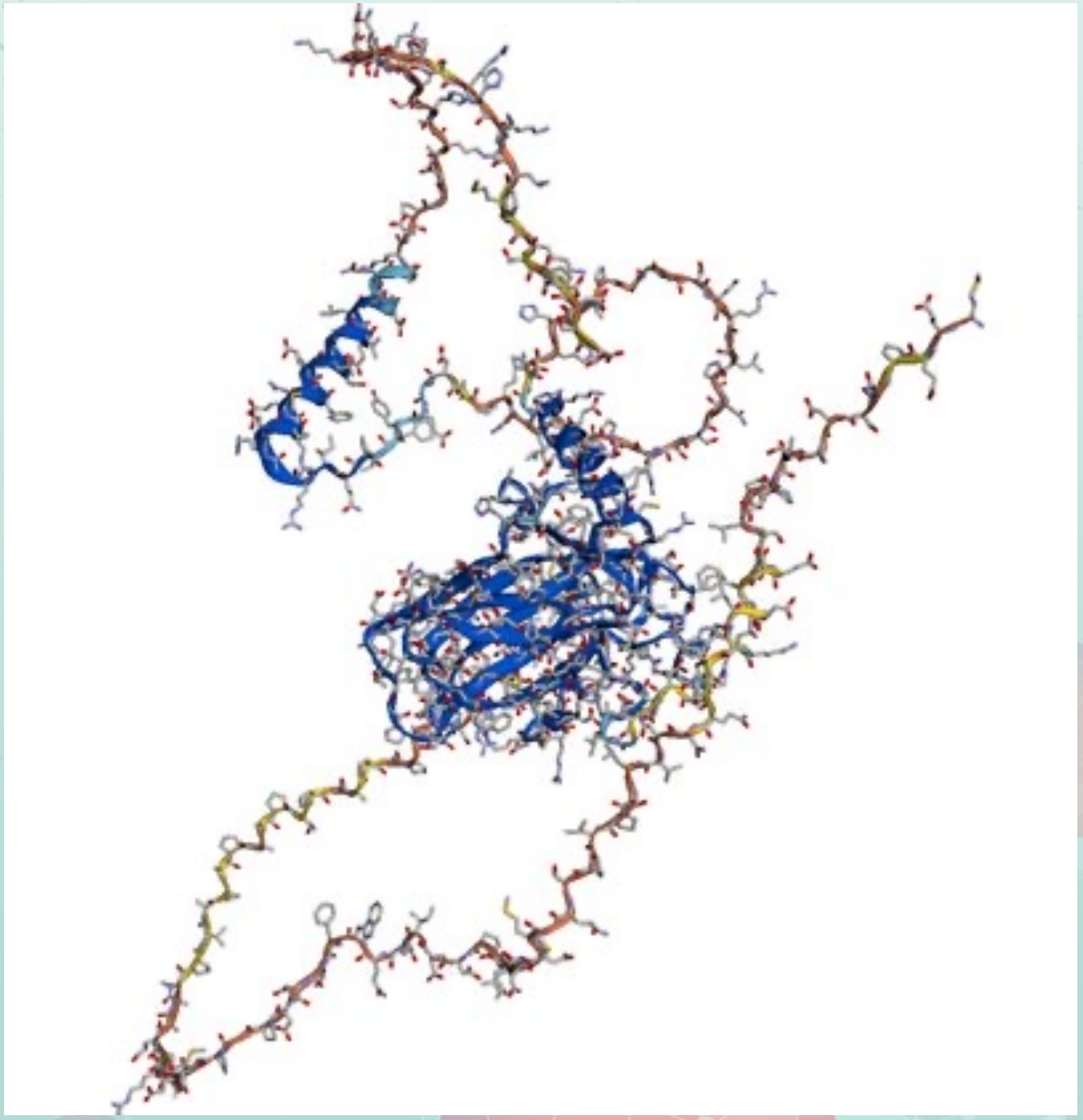




Step 3 – Structure module

- Try to predict 3D positions of residues
- black hole initialization
- Starting point: all residues in the same spatial position (0,0,0)
- Iterative refinement
- Done by the structure module's 8 layers of recurrent attention blocks
- Refinement of predicted 3D structure step by step
  - Step 1: seq representation update
  - Step 2: invariant update on residue positions





Techniques in training

- Recycling
- Self-distillation
- Intentionally wrong templates
- Multiple components of loss function
  - **Locally good** predicted atom coordinates (Frame Aligned Point Error)
  - cross entropy between true and predicted **pairwise distance** representations
  - **Random masking** applied to input MSA masks. 1 task during training: restore the masked items correctly. Something like in language models <-- local representations get stronger
  - tries to **predict how confident 3D** structure prediction will be

$$\mathcal{L} = \begin{cases} 0.5\mathcal{L}_{\text{FAPE}} + 0.5\mathcal{L}_{\text{aux}} + 0.3\mathcal{L}_{\text{dist}} + 2.0\mathcal{L}_{\text{msa}} + 0.01\mathcal{L}_{\text{conf}} & \text{training} \\ 0.5\mathcal{L}_{\text{FAPE}} + 0.5\mathcal{L}_{\text{aux}} + 0.3\mathcal{L}_{\text{dist}} + 2.0\mathcal{L}_{\text{msa}} + 0.01\mathcal{L}_{\text{conf}} + 0.01\mathcal{L}_{\text{exp resolved}} + 1.0\mathcal{L}_{\text{viol}} & \text{fine-tuning} \end{cases},$$

Q & A

Title | Sub-title

Cuccos

Bold cuccos