

# 8/31/2022

- Nick Miller presentation
- Exercise 3: Modeling Cytochrome P450 structure with ColabFold, part I
- Structure prediction principles
- Exercise 3: Modeling Cytochrome P450 structure with ColabFold, part II

# Structure Prediction

- This lecture is intended to help you achieve the following learning objective: Predict protein structure based on the sequence of amino acids. Express confidence in the quality of a structure prediction.
- It will introduce
  - motivations of structure prediction
  - how structure space < sequence space
  - making predictions that maximize template information
  - the new deep learning methods, AlphaFold2 and its faster cousin, ColabFold
- At the end of this lecture, we will have a discussion about:
  - What factors increase/decrease
    - your confidence in a predicted structure?
    - the influences of the prediction algorithm?
  - Can current structure prediction methods be used to predict the effect of a mutations or buffer conditions?

# **Exercise 3: Modeling Cytochrome P450 structure with ColabFold, part I**

[colab](#)

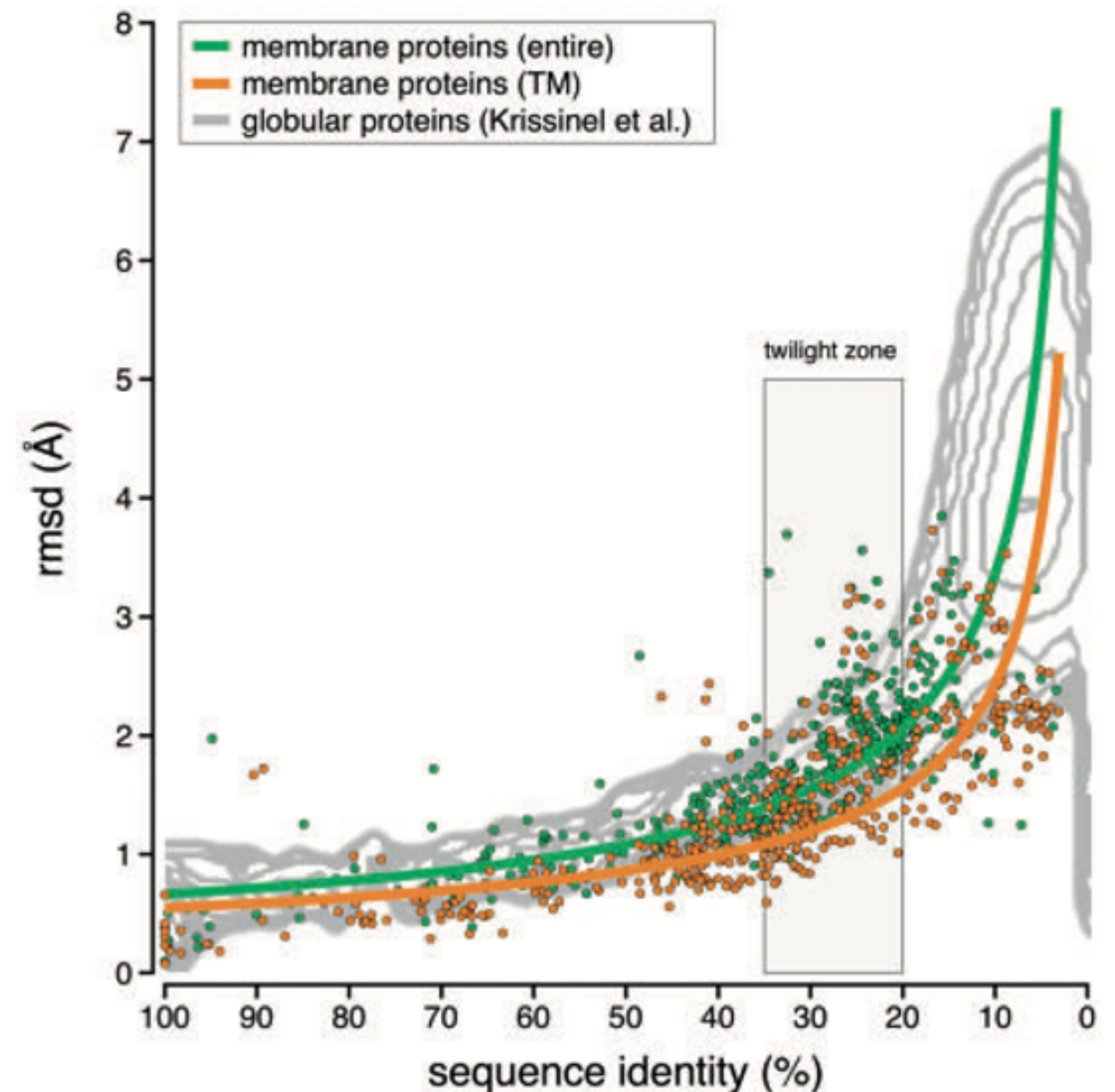
# Structure Prediction Principles

# Why perform structure prediction?

- Predict
  - functional differences between homologs (e.g. isoforms or different species)
  - effects of mutations
  - binding sites and drug interactions
- Design the above
- It's easier than structure determination
- Even with rapid expansion of the PDB, many-fold more sequences have been obtained
- Fortunately...

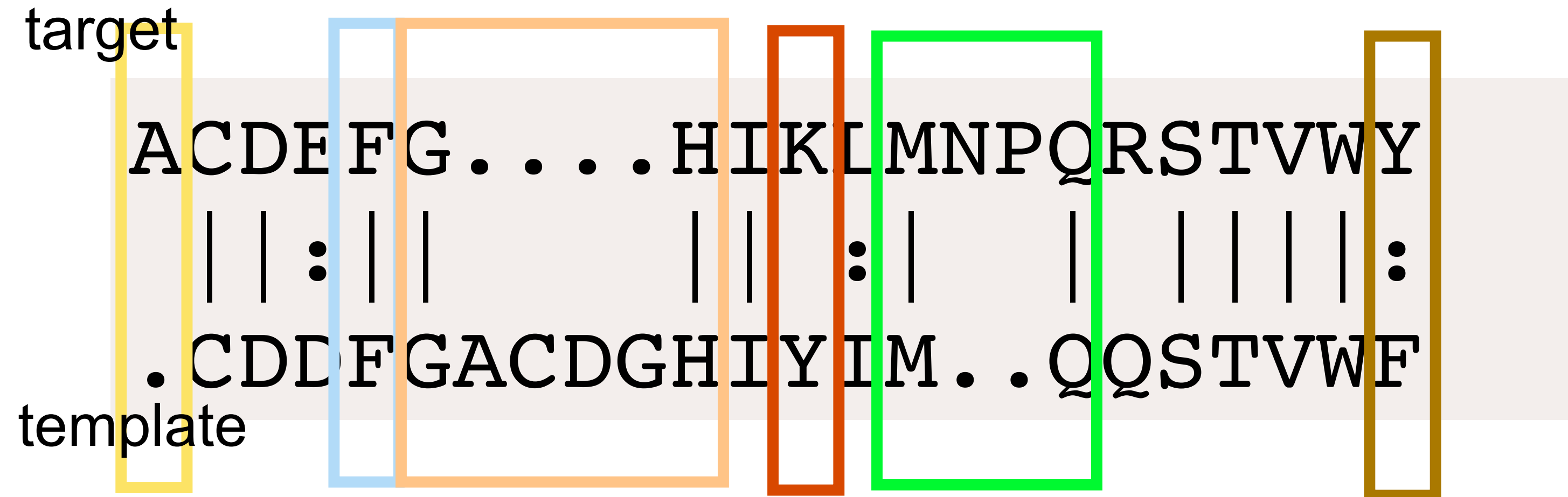
# Structure < sequence space

- The number of unique folds is far less than the number of sequences
- Similar sequences have similar structure - in general, sequence identity only needs to be 30-40%!
- Therefore we can model a sequence of unknown structure based on a homolog with known structure
- But how?



# Evolutionary significance of an alignment

Given this alignment...



Biologist infers...

- The gene was extended by one residue at the N-terminus.
- The Phe is conserved.
- Four residue deletion occurred between G to H.
- A non-similar mutation Y->K occurred..
- A two-residue insertion occurred between M and Q.
- A similar mutation F->Y occurred..

Aligned positions share a common ancestral position.



# An alignment as modeling instructions

Given this alignment...



- Modeler program should...
- Add Ala to the N-terminal Cys using energy minimization.
- Keep the conserved Phe sidechain and backbone.
- Cut out the four residue insertion and connect G to H.
- Switch non-similar sidechains Y->K. Possibly move backbone.
- Possibly pick another alignment.
- Cut at M-Q, insert two residues, Asn-Pro.
- Switch similar sidechains F->Y. Keep backbone fixed.

Aligned positions share a common spatial position.



# Choosing Structure Prediction Software

- There are many software tools for protein structure prediction (see [https://en.wikipedia.org/wiki/List\\_of\\_protein\\_structure\\_prediction\\_software](https://en.wikipedia.org/wiki/List_of_protein_structure_prediction_software))
- How should you decide which to use?
  - Ease of use
    - Web server - easier for sporadic use
    - Downloadable and scriptable - easier for large-scale applications
  - Accuracy

# CASP

- “Critical Assessment of protein Structure Prediction” (CASP) experiments are *blinded* tests of the ability to predict structure from sequence. (see <http://www.predictioncenter.org/index.cgi>)
- “I-TASSER (as 'Zhang-Server') was ranked as the No 1 server for protein structure prediction in recent community-wide CASP7, CASP8, CASP9, CASP10, CASP11, CASP12, and CASP13 experiments.”
- “AlphaFold is an AI system developed by DeepMind that predicts a protein’s 3D structure from its amino acid sequence. It regularly achieves accuracy competitive with experiment.” “In CASP14, AlphaFold was the top-ranked protein structure prediction method by a large margin.”

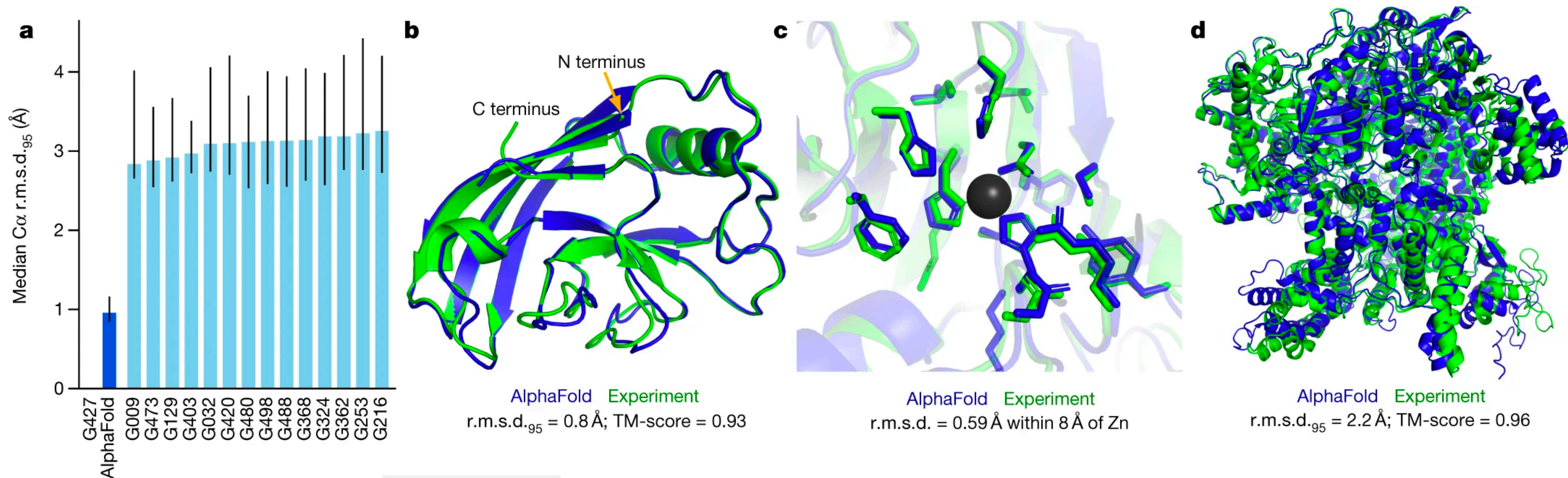
# Discuss

- What factors increase/decrease
  - your confidence in a predicted structure?
  - the influences of choice of prediction algorithm?
- Can homology modeling/threading be used to
  - predict the effect of a mutation
    - of a contact with a ligand in a binding site?
    - on a large-scale conformational change?
  - predict the effect of buffer conditions?

# AlphaFold



# AlphaFold 2 performance




<https://www.nature.com/articles/s41586-021-03819-2>








AI


# AlphaFold Is The Most Important Achievement In AI—Ever

**Rob Toews** Contributor   
*I write about the big picture of artificial intelligence.* [Follow](#)

Oct 3, 2021, 07:34pm EDT

 Listen to article 19 minutes 



  
  




DeepMind's AlphaFold represents the first time a significant scientific problem has been solved by ... [\[+\]](#) IMAGE SOURCE: PROTEINQUIRE

<https://www.forbes.com/sites/robtoews/2021/10/03/alphafold-is-the-most-important-achievement-in-ai-ever/>

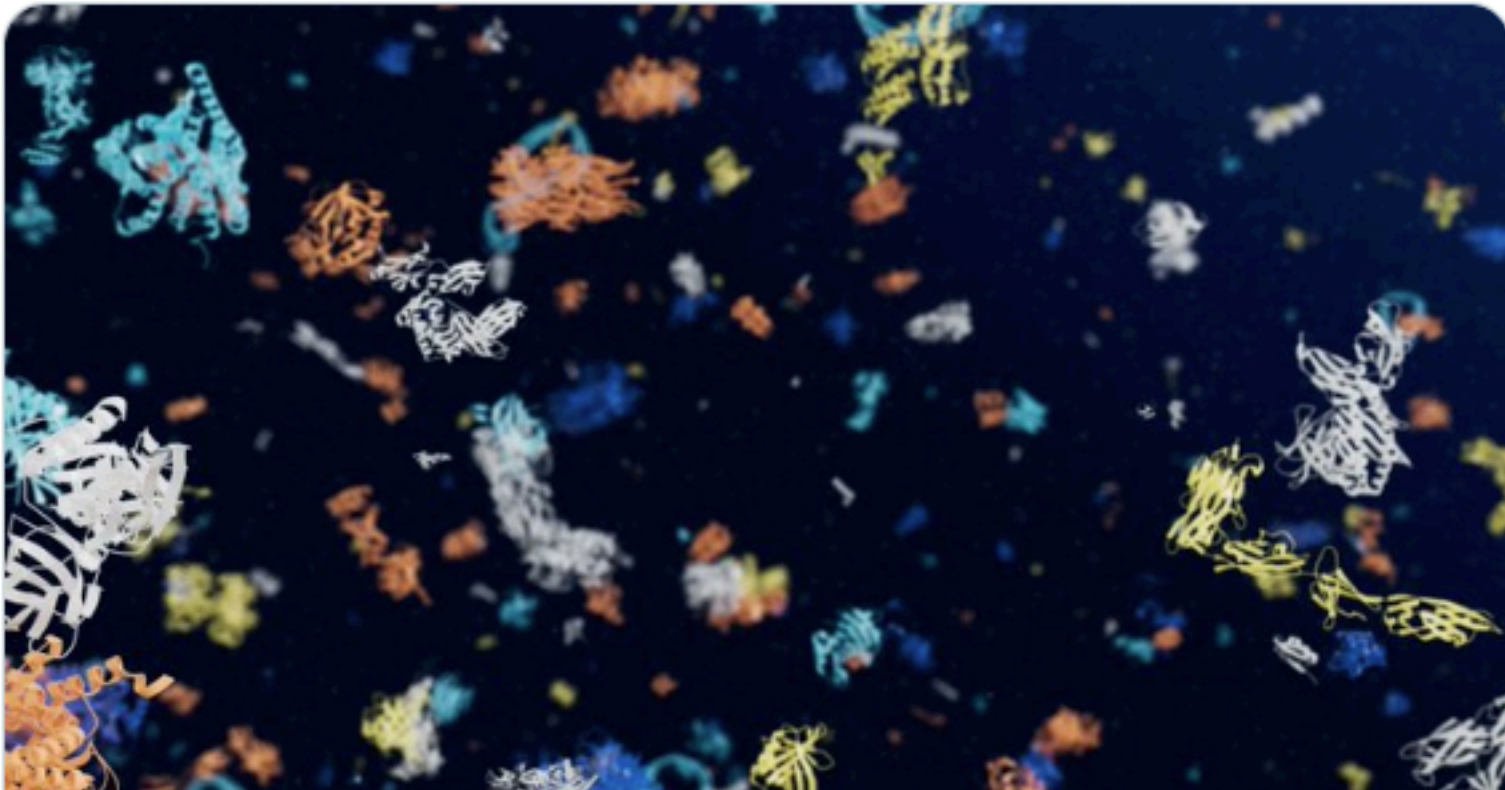
← **Tweet**

 **Demis Hassabis**   
@demishassabis

A year ago we open-sourced [#AlphaFold](#)

Today we're freely sharing the predicted structures of all 200M+ proteins known to science - almost the entire protein universe!

It's our gift to humanity, and a demonstration of the benefits AI can bring to society



deepmind.com  
AlphaFold reveals the structure of the protein universe  
Today, in partnership with EMBL's European Bioinformatics Institute (EMBL-EBI), we're now releasing predicted structures for nearly all catalogued proteins...

6:50 AM · Jul 28, 2022 · Twitter Web App

<https://twitter.com/demishassabis/status/1552622520406347777>



# AlphaFold and RoseTTA fold

nature

View all journals

Search  Login 

Explore content  About the journal  Publish with us 

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

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

## Highly accurate protein structure prediction with AlphaFold

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

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

**566k** Accesses | **983** Citations | **2993** Altmetric | [Metrics](#)

### Abstract

Proteins are essential to life, and understanding their structure can facilitate a mechanistic understanding of their function. Through an enormous experimental effort<sup>1,2,3,4</sup>, the structures of around 100,000 unique proteins have been determined<sup>5</sup>, but this represents a small fraction of the billions of known protein sequences<sup>6,7</sup>. Structural coverage is bottlenecked by the months to years of painstaking effort required to determine a single


RESEARCH ARTICLE | PROTEIN FOLDING

[f](#) [t](#) [in](#) [d](#) [w](#) [e](#)

## Accurate prediction of protein structures and interactions using a three-track neural network

[MINKYUNG BAEK](#) , [FRANK DIMAIO](#) , [IVAN ANISHCHENKO](#) , [JUSTAS DAUPARAS](#) , [SERGEY OVCHINNIKOV](#) , [GYU RIE LEE](#) , [JUE WANG](#) , [QIAN CONG](#) , [LISA N. KINCH](#) , [...] [DAVID BAKER](#)  [+23 authors](#) [Authors Info & Affiliations](#)

**SCIENCE** • 19 Aug 2021 • Vol 373, Issue 6557 • pp. 871–876 • DOI: 10.1126/science.abj8754

 19,728  17

   [GET ACCESS](#)

### Deep learning takes on protein folding

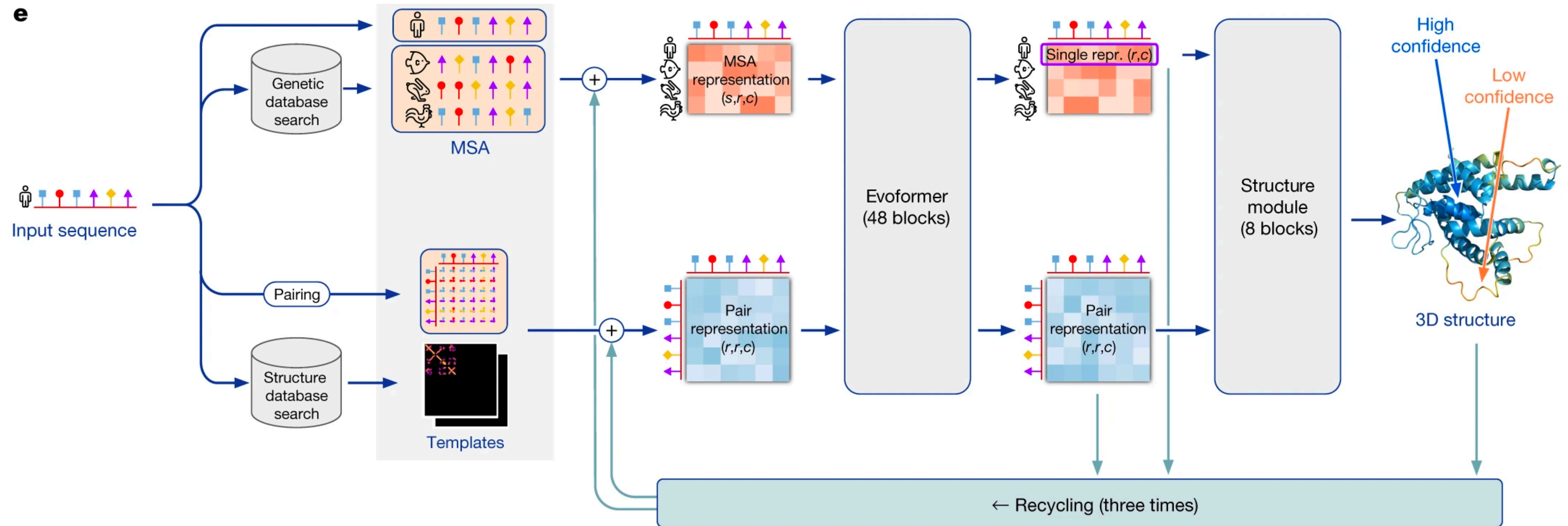
In 1972, Anfinsen won a Nobel prize for demonstrating a connection between a protein's amino acid sequence and its three-dimensional structure. Since 1994, scientists have competed in the biannual Critical Assessment of Structure Prediction (CASP) protein-folding challenge. Deep learning methods took center stage at CASP14, with DeepMind's AlphaFold2 achieving remarkable accuracy. Baek *et al.* explored network architectures based on the DeepMind framework. They used a three-track network to process sequence, distance, and coordinate information simultaneously and achieved accuracies approaching those of DeepMind. The method, RoseTTA fold, can solve challenging x-ray crystallography and cryo-electron microscopy modeling problems and generate accurate models of protein-protein complexes. —VV

Deep learning takes on protein folding  
Abstract  
Supplementary Material  
References and Notes

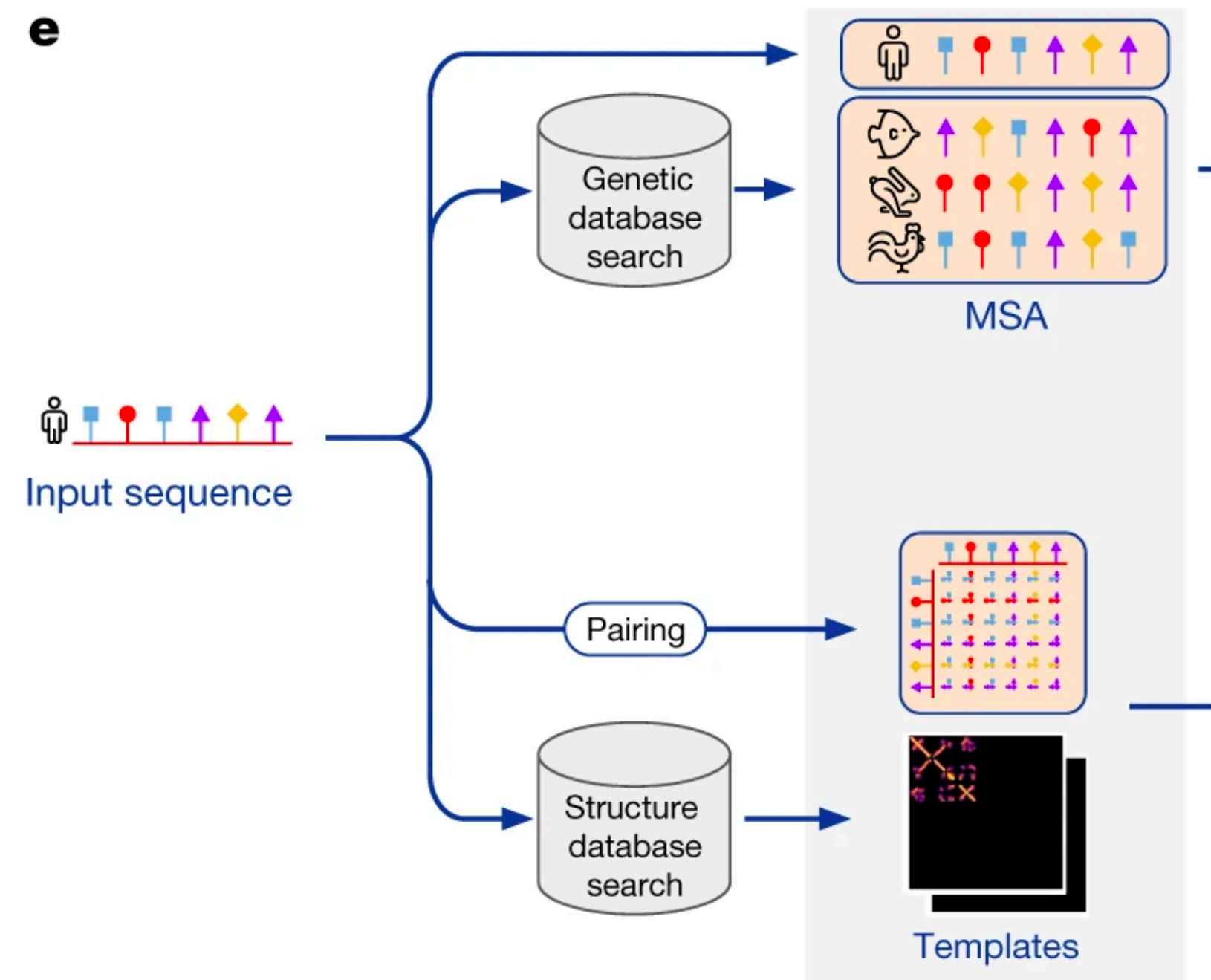
      



# How does it work? An overview



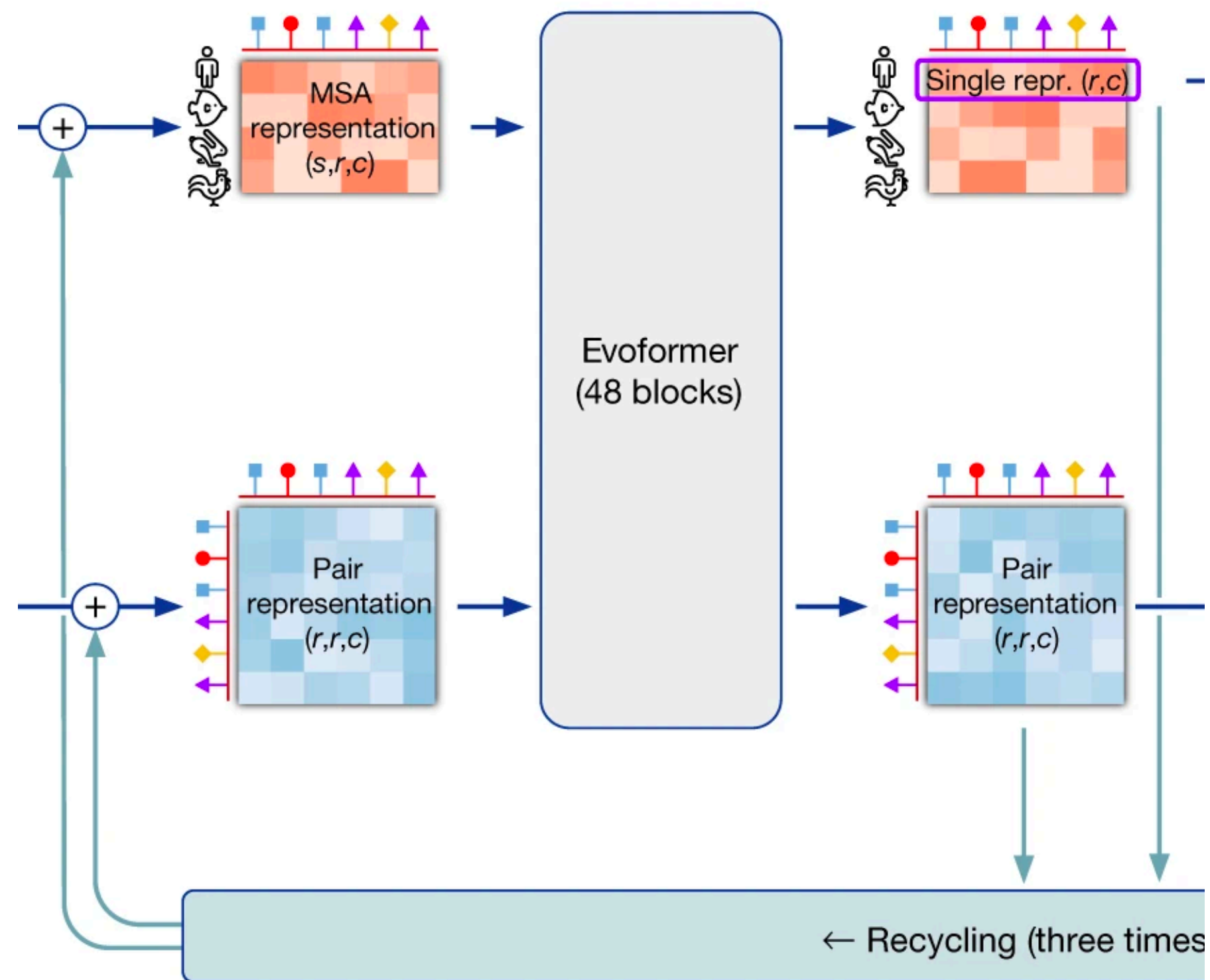
# Inputs



- At the most basic level, a machine learning model takes inputs and produces outputs
- For AlphaFold2, the inputs are
  - a multiple sequence alignment, a set of similar sequences from different organisms
    - MSA helps identify
      - which parts are most likely to mutate
      - correlations between mutations, or coevolution; amino acids that coevolve are likely to be close in physical space
  - pair representations of templates
    - pair representation includes distance between beta carbons, displacement of alpha carbons

<https://www.blopig.com/blog/2021/07/alphafold-2-is-here-whats-behind-the-structure-prediction-miracle/>

# Transformer Overview

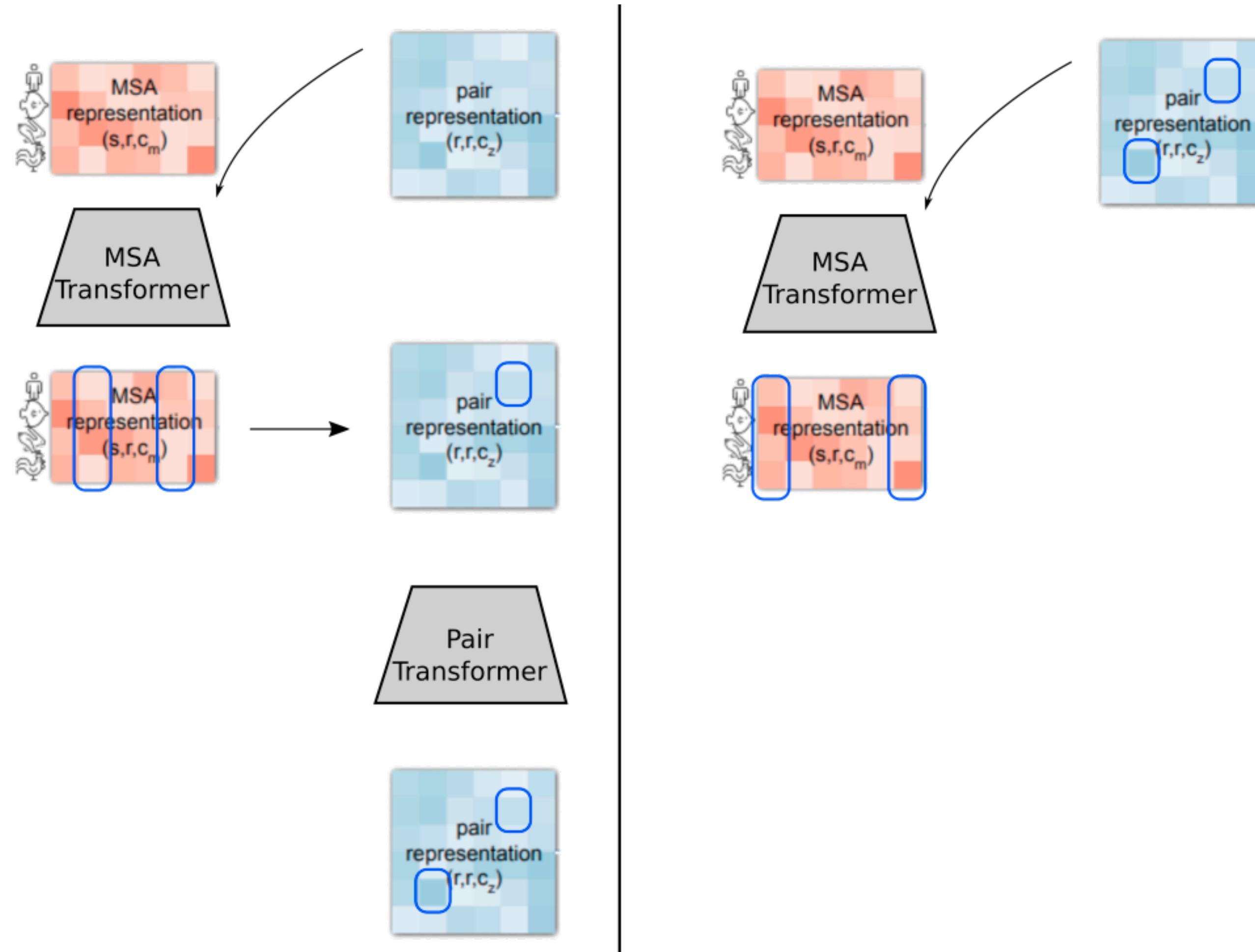


- Series of functions that
  - take input
  - generate outputs
  - exchange information between MSA and pair representations
- What is new? “Before AlphaFold 2, most deep learning models would take a multiple sequence alignment and output some inference about geometric proximity.”

<https://www.blopg.com/blog/2021/07/alphafold-2-is-here-whats-behind-the-structure-prediction-miracle/>



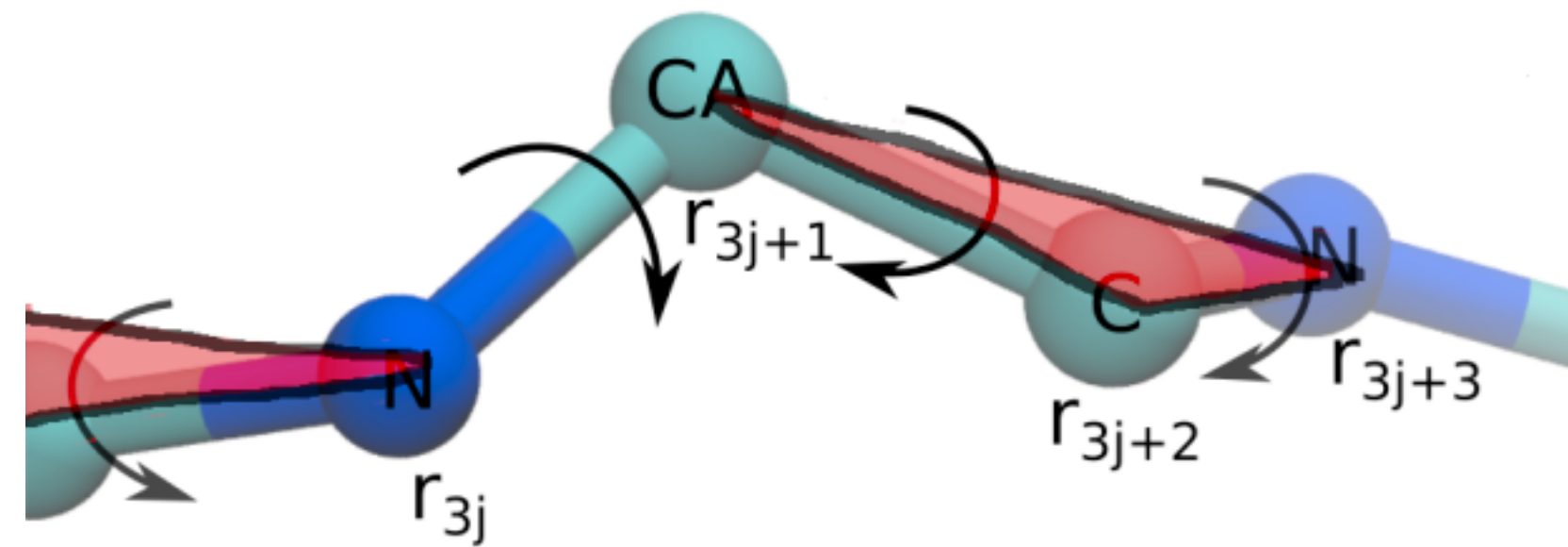
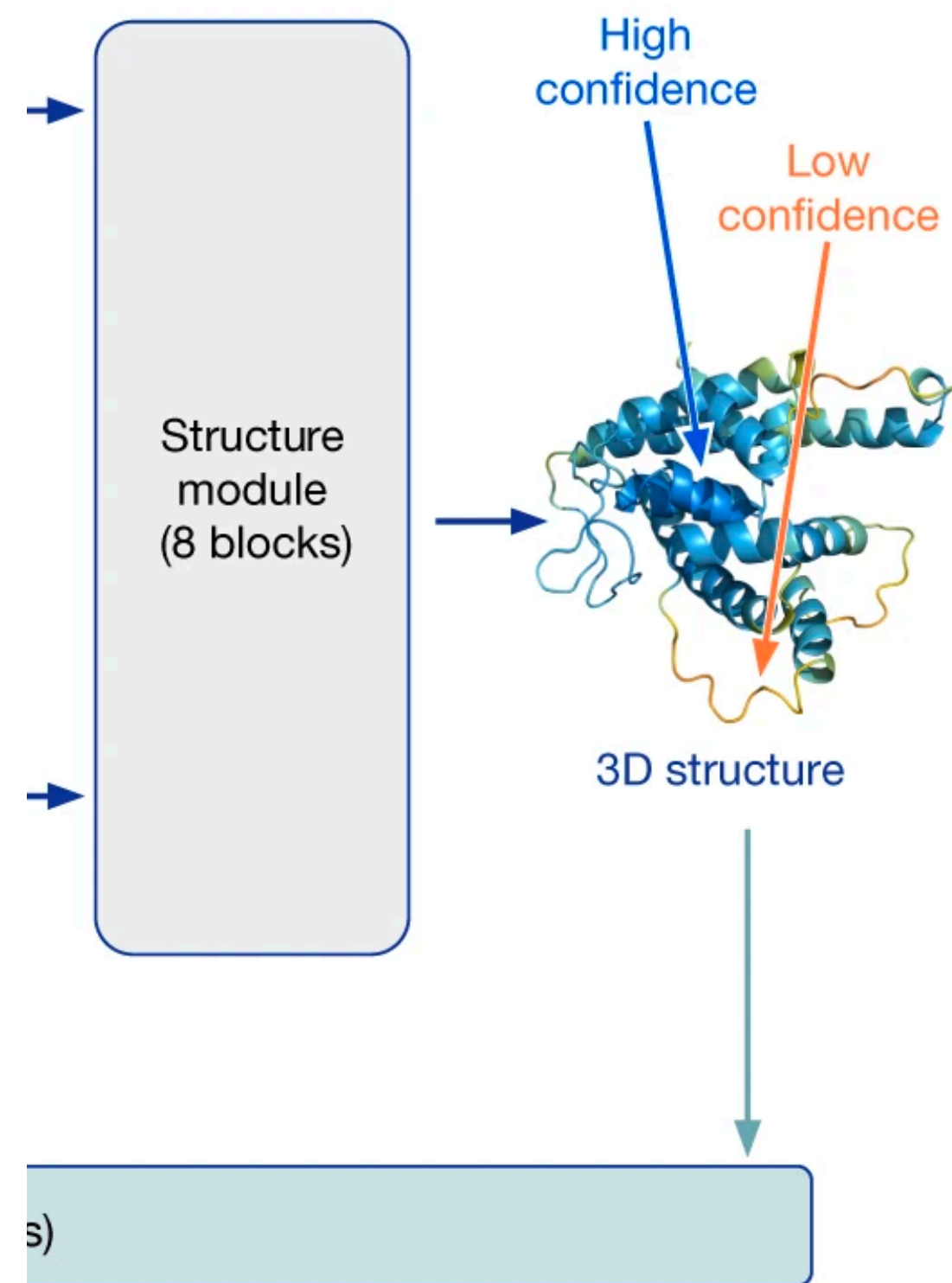
# The New Transformer Concept



Conceptualization of the Evoformer information. In the left diagram, 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. In the right diagram, 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.

<https://www.blopig.com/blog/2021/07/alphafold-2-is-here-whats-behind-the-structure-prediction-miracle/>

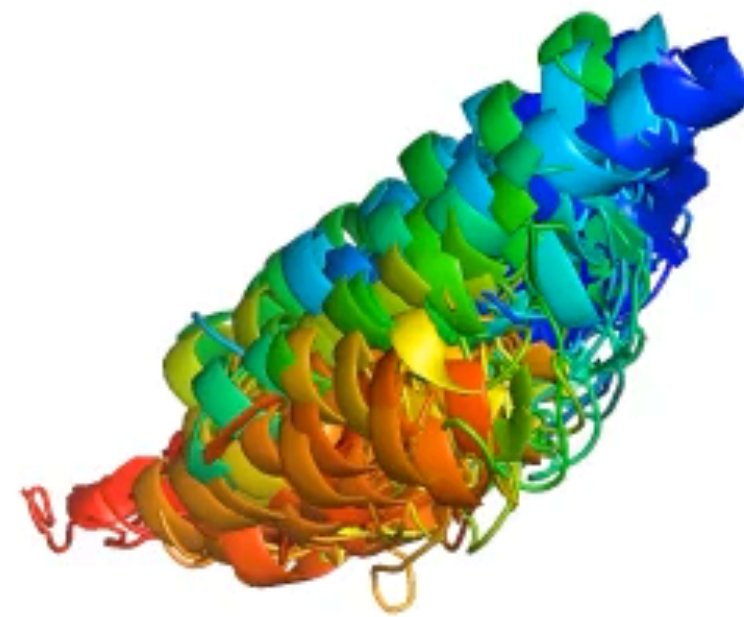
# Structure



- Takes transformed MSA and pair representations
- Generates Cartesian coordinates

<https://www.bloig.com/blog/2021/07/alphafold-2-is-here-whats-behind-the-structure-prediction-miracle/>

# Structure module for each Evoformer block



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

<https://www.nature.com/articles/s41586-021-03819-2#Sec20>

# Why does it work?

- “it is a very sophisticated fold recognition algorithm that exploits the completeness of the library of single domain PDB structures” - Jeffrey Skolnick et al
- “The incredible performance of this network seems down to DeepMind’s superb engineering” - Carlos Rubiera, Oxford



# Limitations

- The same as other prediction methods
- Good input information (sequences and structures) is necessary
- Does not
  - predict changes with conditions (buffer, ions, ligands) or dynamics
  - include molecules that are not amino acids

# **Exercise 3: Modeling Cytochrome P450 structure with ColabFold, part II**

[colab](#)

# References

- Lecture 19 of BIOL 4550 by Chris Bystroff of Rensselaer Polytechnic Institute
- Lab 04 of IIBM3202 Molecular Modeling and Simulation from the Institute for Biological and Engineering at Pontificia Universidad Catolica de Chile
- Chothia C & Lesk AM (1986) \_EMBO J\_ 5(4), 823–826
- AlphaFold paper: <https://www.nature.com/articles/s41586-021-03819-2>
- Oxford Protein Informatics Group Blog Post: <https://www.blopig.com/blog/2021/07/alphafold-2-is-here-whats-behind-the-structure-prediction-miracle/>
- Skolnick paper: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8592092/pdf/nihms-1751143.pdf>
- RoseTTA fold paper: <https://doi.org/10.1126/science.abj8754>