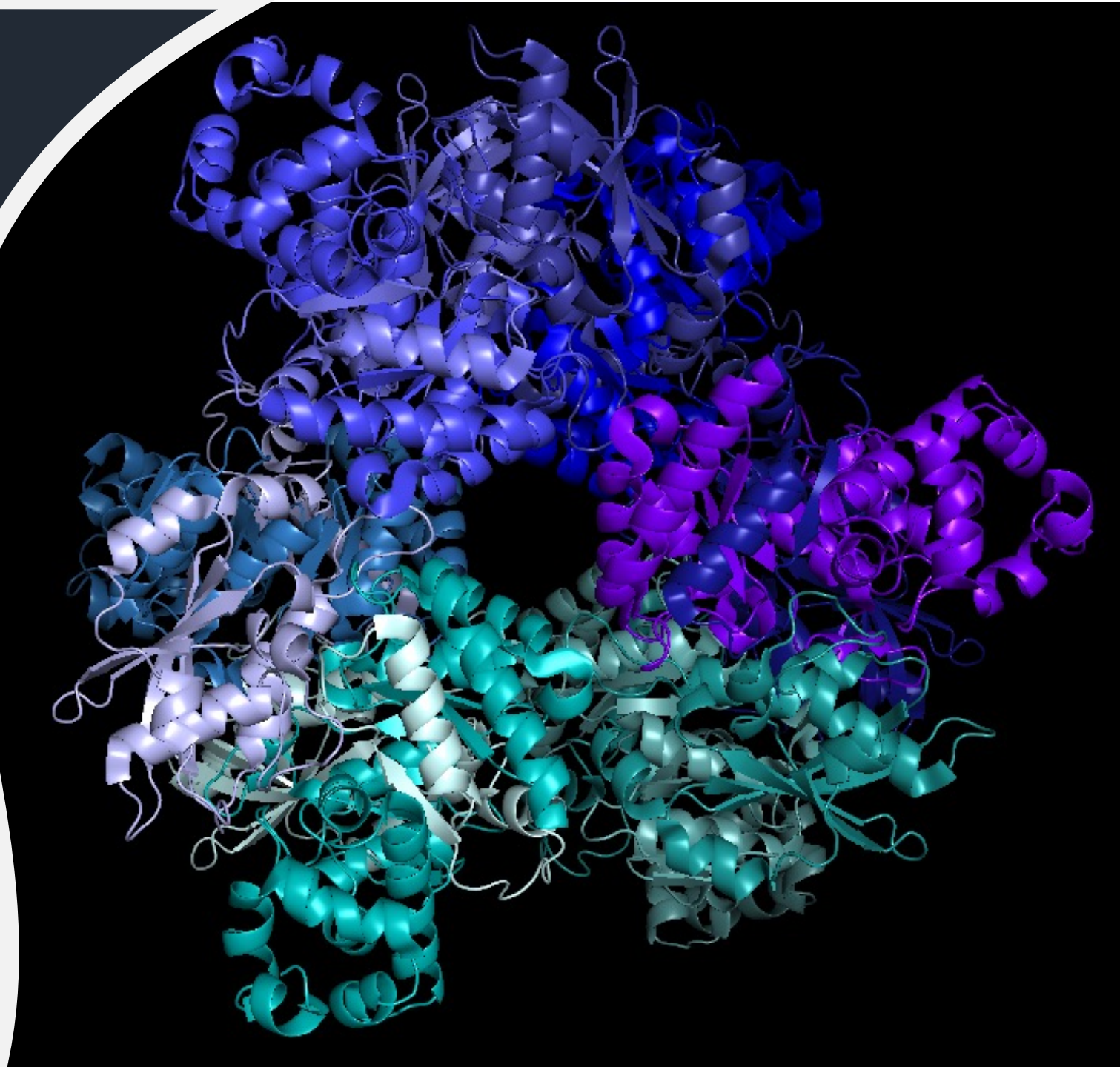
Introduction to Protein Structure Prediction With AlphaFold 2

Jason Laird

Bioinformatics Scientist



The Research Technology Team

Delilah Maloney
High Performance Computing Specialist

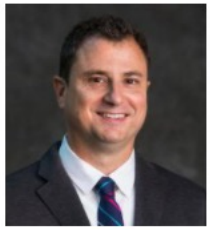
Kyle Monahan
Senior Data Science Specialist

Shawn Doughty
Manager, Research Computing

Jason Laird
Bioinformatics Scientist

Chris Barnett
Senior Geospatial Analyst

Tom Phimmasen
Senior Data Consultant



Patrick Florance
Director, Academic Data Services

Jake Perl
Digital Humanities NLP Specialist

Carolyn Talmadge
Senior GIS Specialist

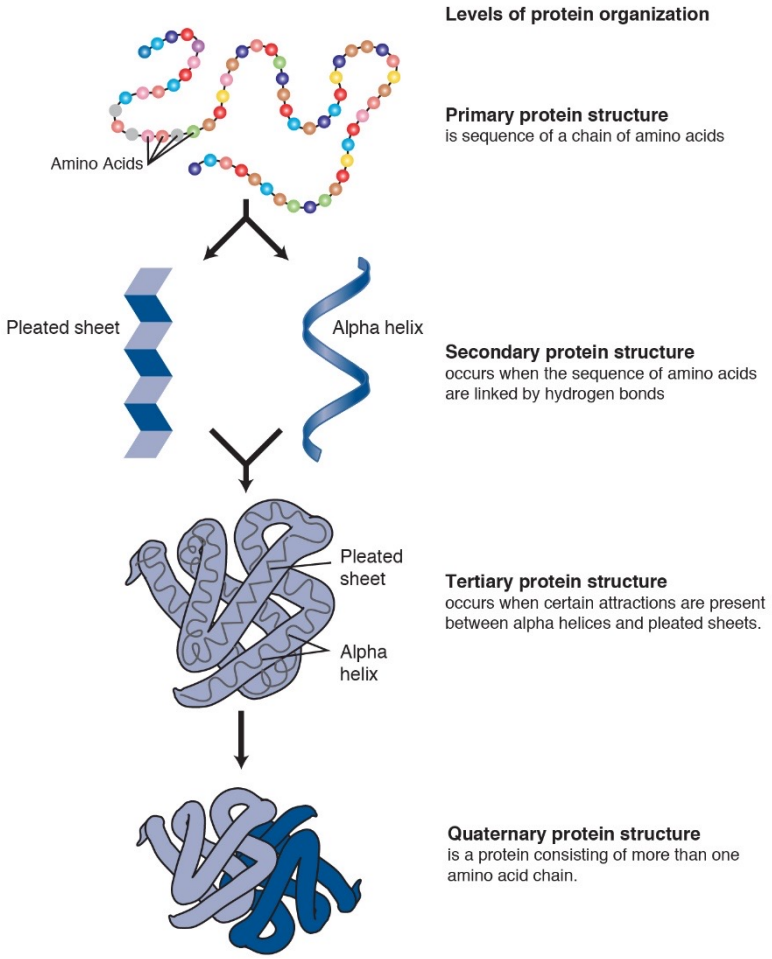
Uku-Kaspar Uustalu
Data Science Specialist

- ✓ Consultation on Projects and Grants
- ✓ High Performance Compute Cluster
- ✓ Workshops

https://it.tufts.edu/research-technology

Protein Organization

- Primary Structure: amino acid sequence
- Secondary Structure: amino acid sequences linked by hydrogen bonds
- Tertiary Structure: organization of secondary structures
- Quaternary Structure: organization of multiple amino acid chains

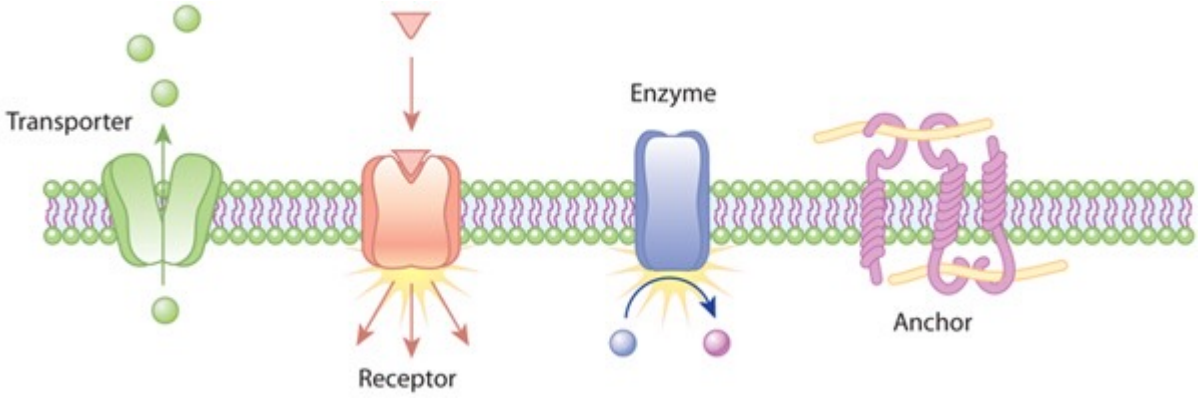


https://www.genome.gov/genetics-glossary/Protein

The Importance of Protein Structure

- Can help determine what a protein does
- Often more conserved than the amino acid sequences that form them

Examples of Different Proteins



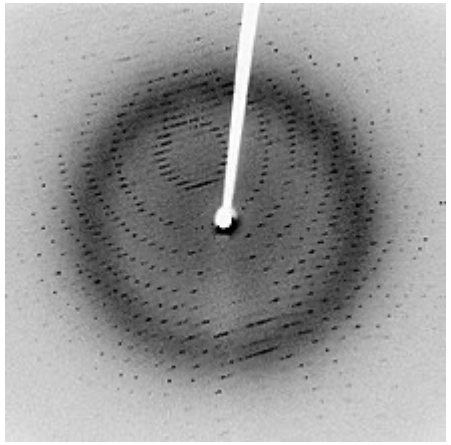
https://www.nature.com/scitable/topicpage/protein-function-14123348/

https://www.ncbi.nlm.nih.gov/books/NBK26820/

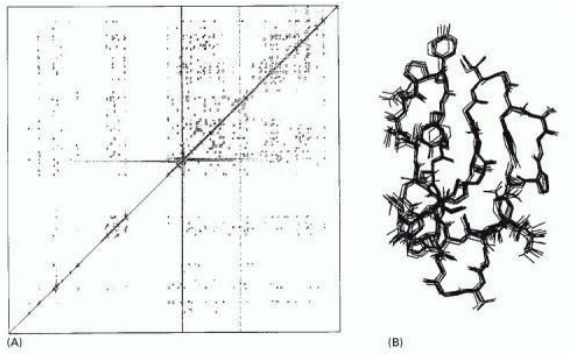
Laboratory Means To Determine Protein Structure

- X-ray Crystallography
- Nuclear Magnetic Resonance (NMR) Spectroscopy
- 3D Electron Microscopy

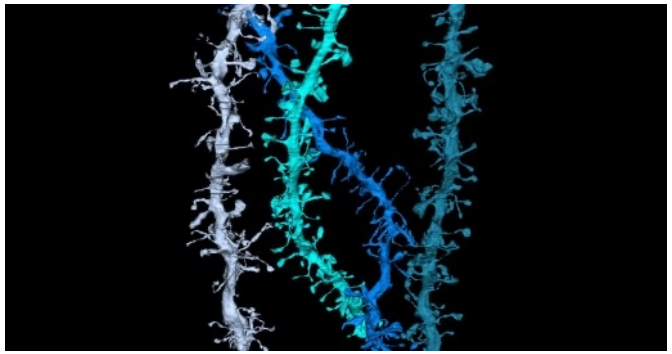
X-ray Crystallography



NMR Spectroscopy



3D Electron Microscopy



https://directorsblog.nih.gov/tag/serial-scanning-3d-electron-microscopy/

https://www.ncbi.nlm.nih.gov/books/NBK26820/

https://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/methods-for-determining-structure

https://simple.wikipedia.org/wiki/X-ray_crystallography

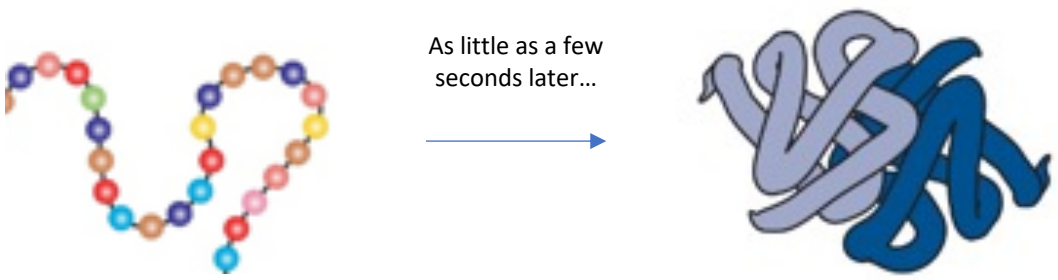
The Protein Structure Problem

- 100,000,000 known distinct proteins
- Each has a unique structure that determines function
- Determining protein structure is time consuming
- Only a small fraction of exact 3D structures are known

https://deepmind.com/research/case-studies/alphafold

Levinthal's Paradox

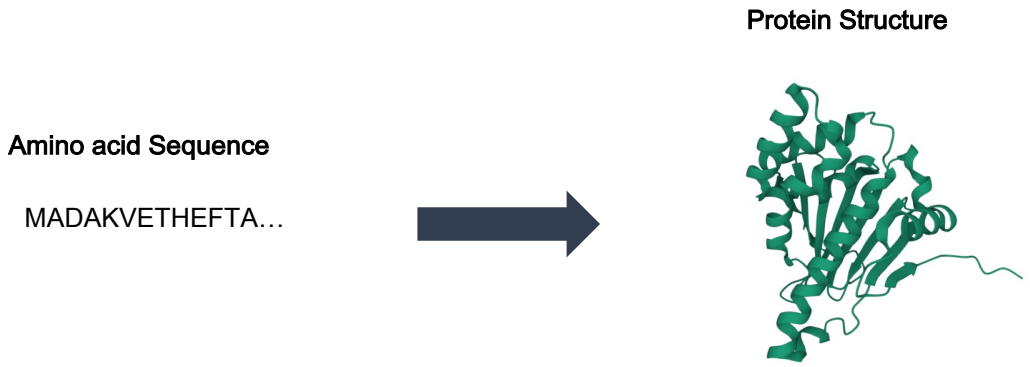
- Finding the native folded state of a protein by random searching of all possible configurations would take an enormous amount of time
- However, proteins can often fold within seconds
- Meaning some process must be guiding this folding



https://www.ncbi.nlm.nih.gov/pmc/articles/PMC48166/

Using Sequence To Predict Structure

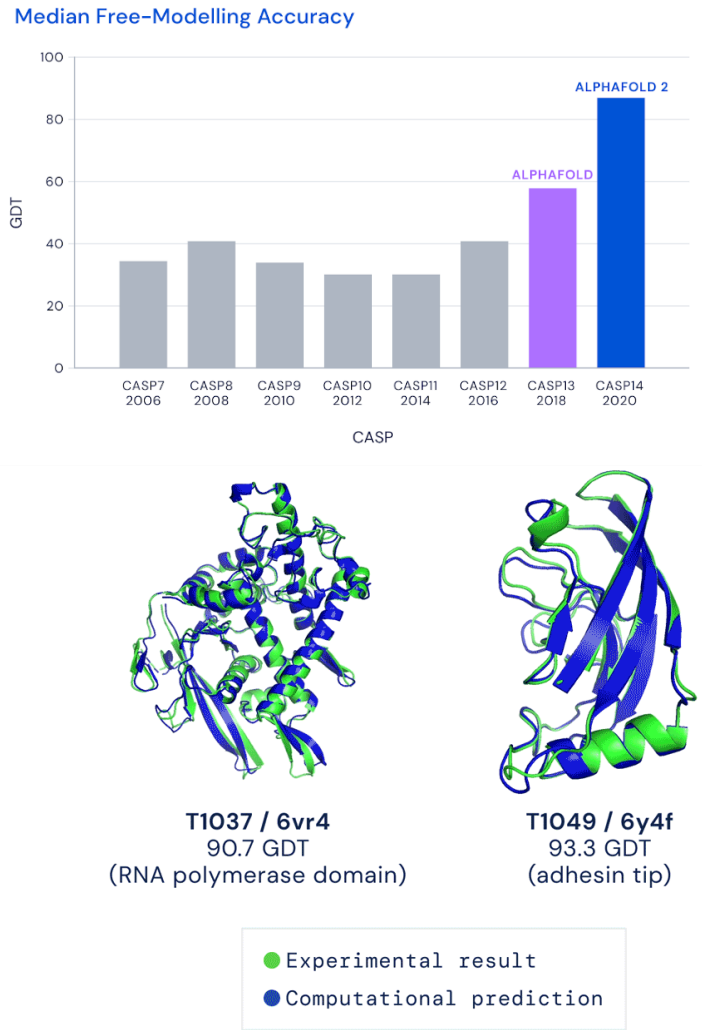
- Instead of laboratory experimentation, there have been massive efforts to use a protein's sequence to determine structure
- In 1994, the Critical Assessment of Structure Protein (CASP) was established as a biennial assessment of methods to predict structure from sequence



https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology
https://predictioncenter.org/

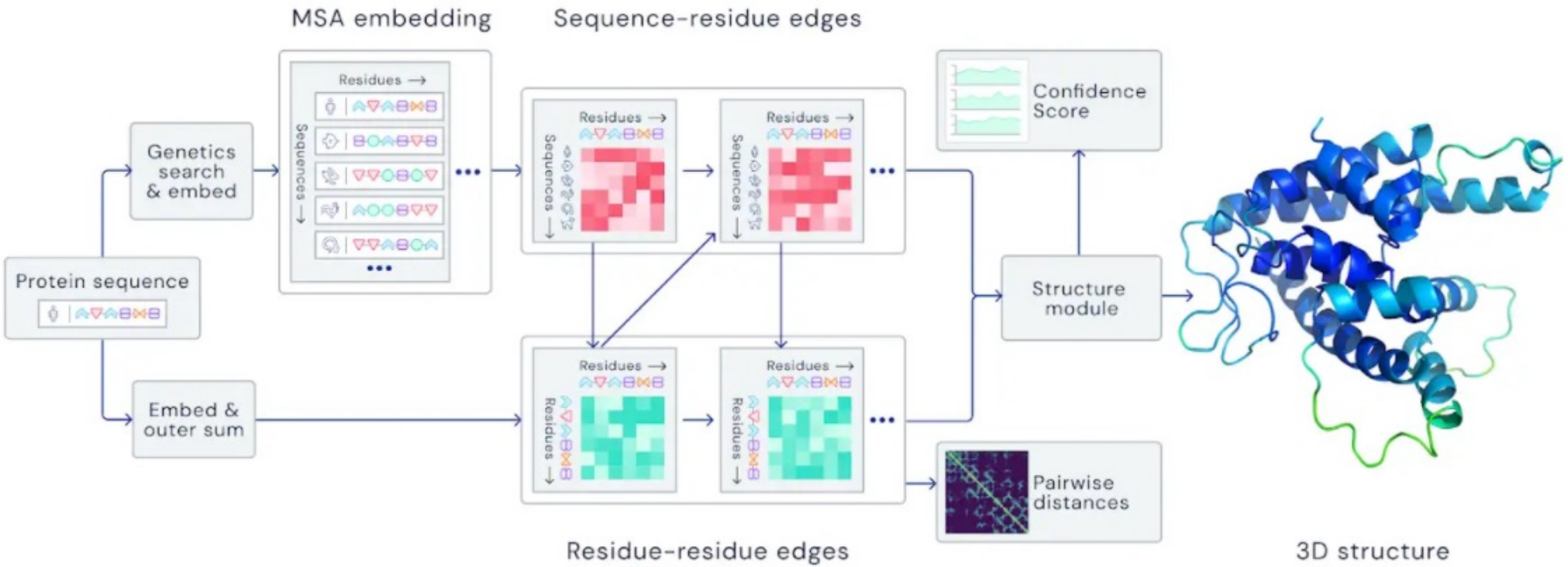
Enter AlphaFold 2

- Google's DeepMind team Entered AlphaFold 2 in CASP14
- Achieved a median Global Distance Test Score of 92.4
- AlphaFold 2 works by finding similar sequences to the query, extracts the information using a neural network, then passes that information to another neural network that construct a theoretical structure



https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology

The AlphaFold 2 Workflow



https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology

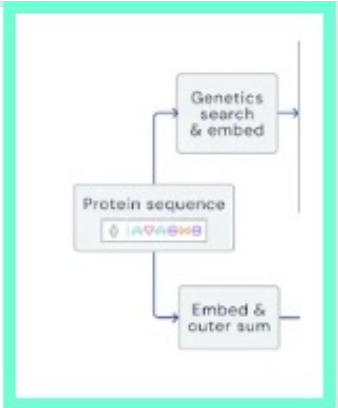
Protein Sequence Information

- Protein sequence information stored as a FASTA file. Consists of:

Header

Sequence

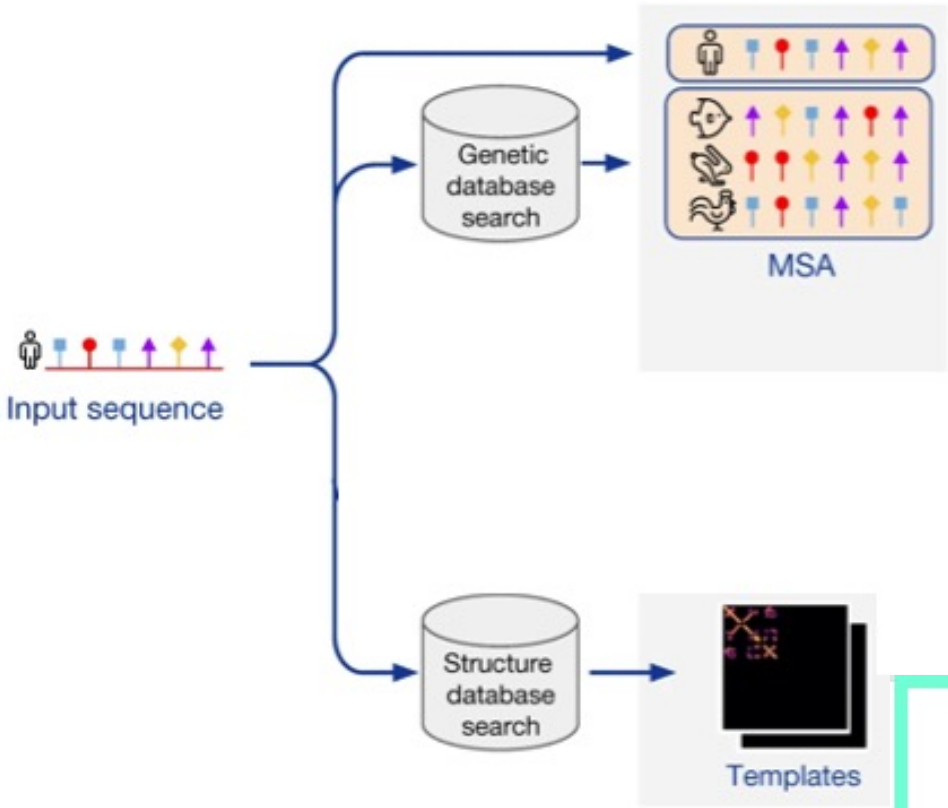
```
>sp|P46598|HSP90_CANAL Heat shock protein 90 homolog OS=Candida albicans
(strain SC5314 / ATCC MYA-2876) OX=237561 GN=HSP90 PE=1 SV=1

MADAKVETHEFTAEISQLMSLIINTVYSNKEIFLRELISNASDALDKIRYQALSDPSQLE
SEPELFIRIIPQKDQKVLEIRDSGIGMTKADLVNNLGTIAKSGTKSFMEALSAGADVSMI
GQFGVGFYSLFLVADHVQVISKHNDDEQYVWESNAGGKFTVTLDETNERLGRGTMLRLFL
KEDQLEYLEEKRIKEVVKKHSEFVAYPIQLVVTKEVEKEVPETEE
```

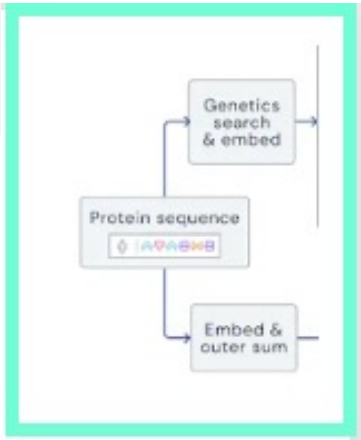


https://en.wikipedia.org/wiki/FASTA_format

Searching for Similar Sequences

- This query sequence is compared to:
  - **UniRef90 database:** to find similar sequences
  - **PDB70** to find similar structures
- Sequences that are too similar to our query are filtered out so that we don't just build a replicate based on that sequence
- These sequences are arranged as an MSA

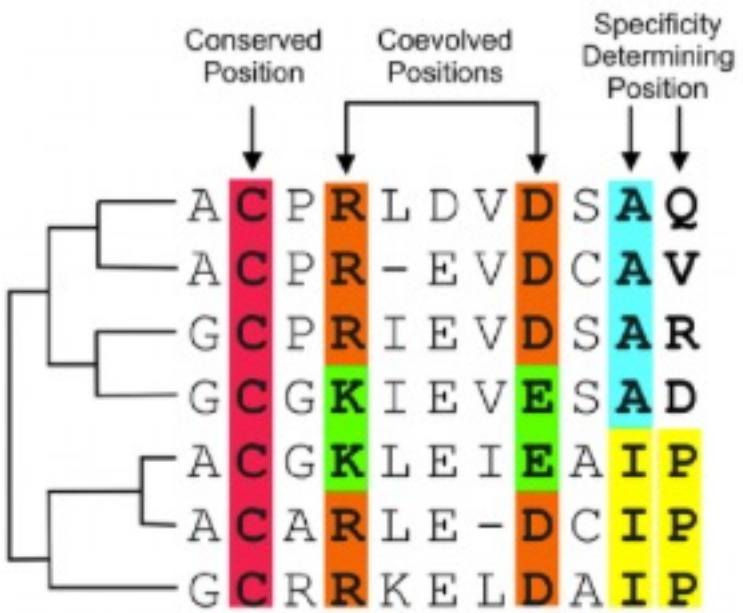


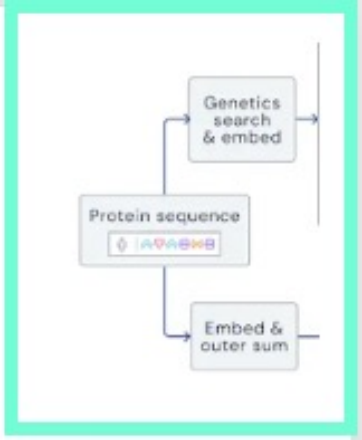


https://www.blopig.com/blog/2021/07/alphafold-2-is-here-whats-behind-the-structure-prediction-miracle/
/

Multiple Sequence Alignment (MSA)

- An MSA is an array of sequences
- These sequences are ***aligned*** with one another as to best match similar regions
- These sequences don't always line up perfectly and as such we see:
  - **Conserved positions:** where the letter does not change
  - **Coevolved positions:** where the letter will change with another letter
  - **Specificity Determining positions:** where the letter is consistently different



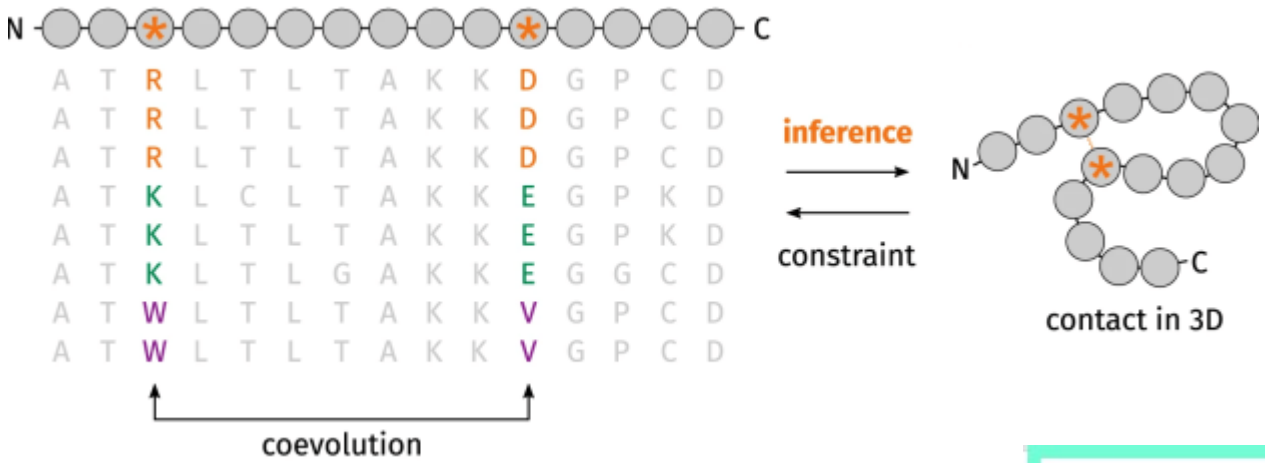


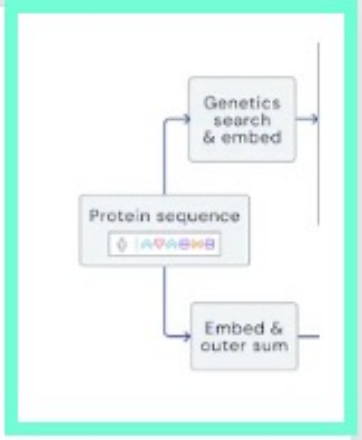
https://www.blopig.com/blog/2021/07/alphafold-2-is-here-whats-behind-the-structure-prediction-miracle/

https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-13-235

Why is an MSA Useful In Structure Prediction?

- The theory is that residues that coevolve are generally close to each other in the protein's folded state
- So, by assessing what residues change together we get an idea of where they might be spatially!

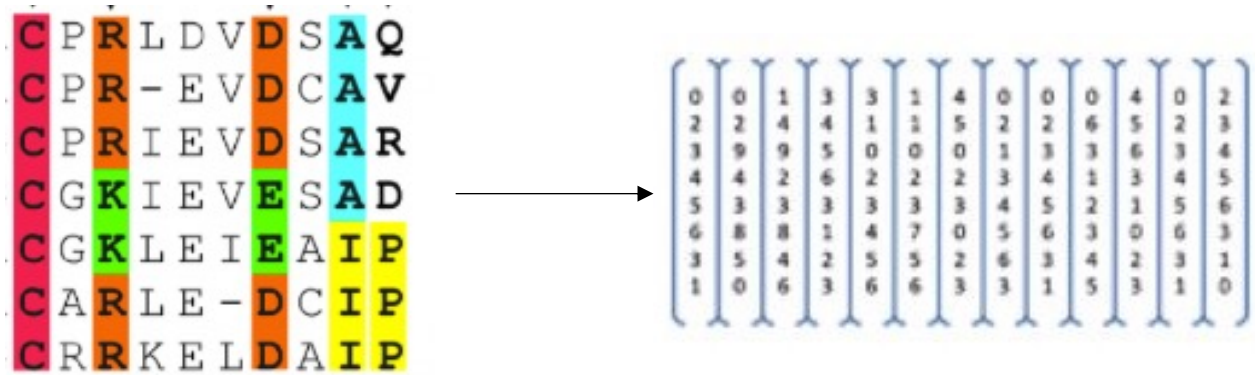


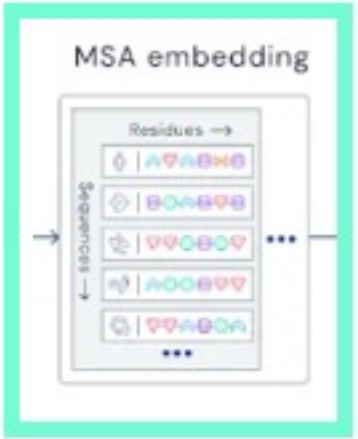


https://www.pnas.org/content/114/34/9122

MSA Embedding

- An MSA is still essentially an array of letters
- To be more computer friendly these letters are ***embedded*** as numbers using their positional information
- AlphaFold embeds these letter values as numeric ones and terms this the MSA representation





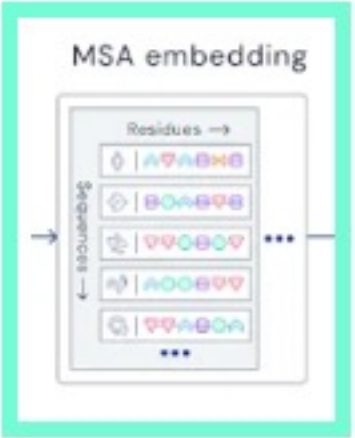
https://www.blopig.com/blog/2021/07/alphafold-2-is-here-whats-behind-the-structure-prediction-miracle/

/

Embedding Example

- Take for example the sentence "**I ate an apple and played the piano**"
- This string is embedded by positional information.
- **e.g. ate** was the second word so there is a 1 in the second column at row **"ate"**

	1	2	3	4	5	6	7	8
I	1	0	0	0	0	0	0	0
ate	0	1	0	0	0	0	0	0
an	0	0	1	0	0	0	0	0
apple	0	0	0	1	0	0	0	0
and	0	0	0	0	1	0	0	0
played	0	0	0	0	0	1	0	0
the	0	0	0	0	0	0	1	0
piano	0	0	0	0	0	0	0	1



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

Pair Representation

- Similar Structures were also queried for using our protein sequence.
- These structure files (A.K.A Crystallographic Information Files (CIF)) contain 3D coordinates for a protein's atoms in space
- These coordinates are used to initialize a pairwise distance matrix between residues that AlphaFold calls the pair representation

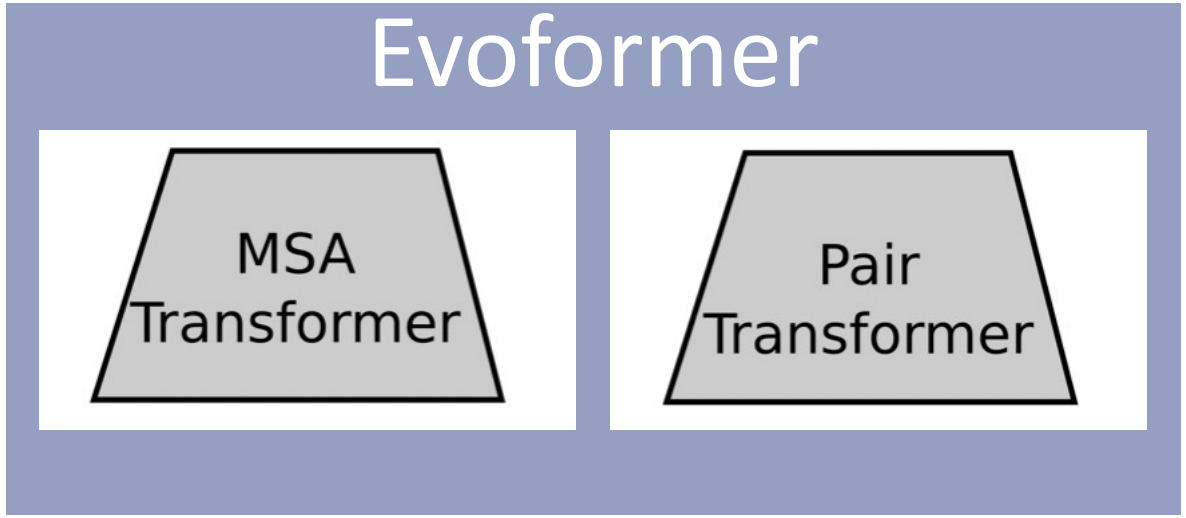
Residue	x	y	z
C	-3.8	-1.12	0.57
N	0.96	-1.07	-0.89
H	0.05	0.95	0.39

	C	N	H
C	0	-1.12	.94
N	0.54	0	3.1
H	0.05	1.32	0

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

The Evoformer

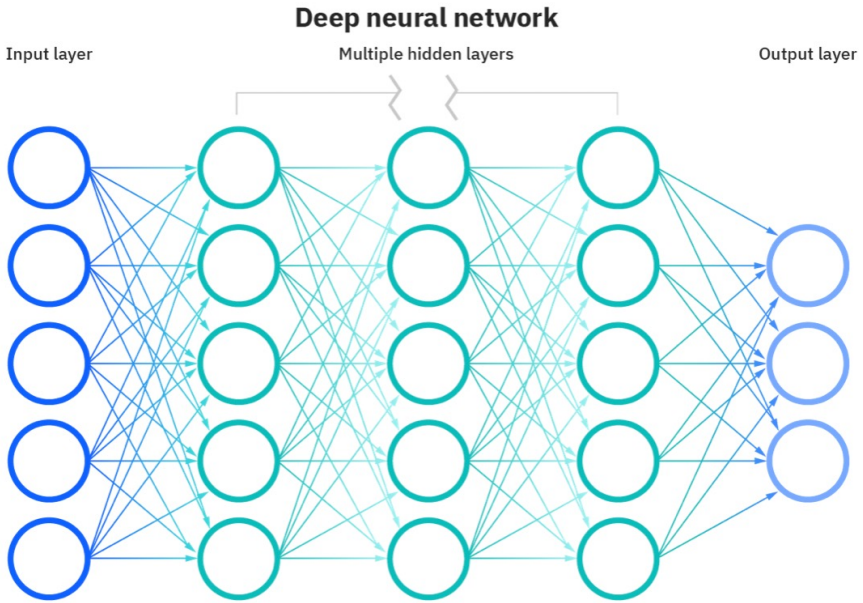
- The MSA representation and the pair representation are fed into in special type of neural network that AlphaFold terms the Evoformer
- The Evoformer is a combination of two special types of neural networks called Transformers



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

What Are Neural Networks?

- Neural Networks are machine learning algorithms that mimic the way neurons communicate
- They usually consist an input, hidden and output layer
- Each node has a threshold and if the output of the node isn't above that threshold it doesn't communicate with the next node



https://www.ibm.com/cloud/learn/neural-networks

What Is In A Node?

- Each node can be thought of as a linear regression model with input data, weights, a bias term and an output
- The weights are assigned as to weight importance – the larger the weight the more important the variable

A Single Node

$$= \sum_{i=1}^{m} w_i x_i + bias = w_1 x_1 + w_2 x_2 + w_3 x_3 + bias$$

https://www.ibm.com/cloud/learn/neural-networks

To Communicate Or Not Communicate?

- Each node will have an output based on this regression function
- That output is then fed into something called an activation function
- The output of this activation function is compared to some threshold
- If the threshold is met it ”fires” and communicates with the next layer

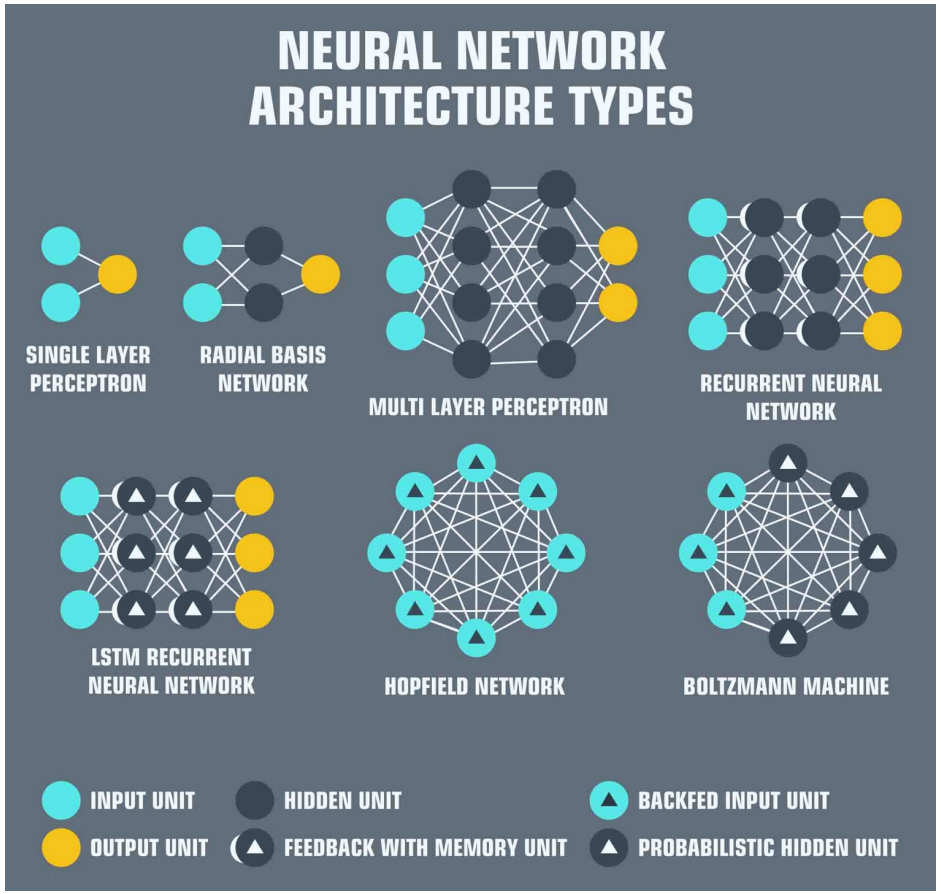
$$\sum_{i=1}^{m} w_i x_i + bias = w_1 x_1 + w_2 x_2 + w_3 x_3 + bias$$

$$\text{output} = \underbrace{f(x)}_{\text{Activation Function}} = \begin{cases} 1 \text{ if } \sum w_1 x_1 + b \geq 0 \\ 0 \text{ if } \sum w_1 x_1 + b < 0 \end{cases}$$

https://www.ibm.com/cloud/learn/neural-networks

Neural Network Customization

- There are different types of neural networks depending on what functions you use and how you organize node communication
- AlphaFold uses a Recurrent Neural Network

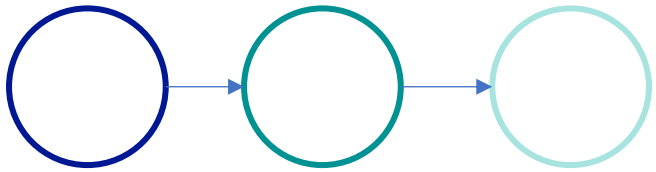


https://www.allerin.com/blog/3-types-of-neural-networks-that-ai-uses

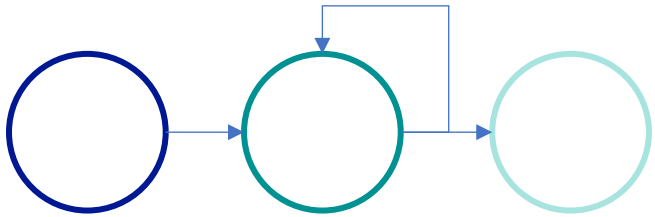
Recurrent Neural Network

- In a feed forward neural network you have input that is processed through a node and if that node is activated it communicates with the next node
- In a recurrent neural network, the output of a node can be used to inform and change the output of the node
- Naturally this comes at a memory cost when it tries to pull from old connections

Feed-Forward Neural Network



Recurrent Neural Network

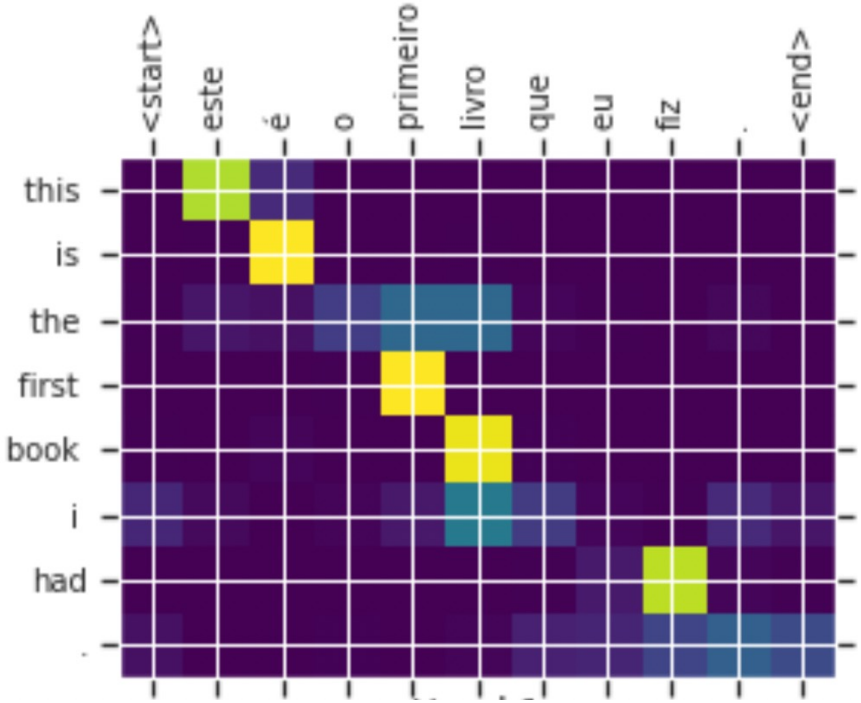


https://www.allerin.com/blog/3-types-of-neural-networks-that-ai-uses

Transformer And Attention

- To save on computational cost, Recurrent Neural Networks can have their attention limited
- Basically, values are scaled down to reveal which data points are worth paying ***attention*** to
- This focused recurrent neural network is called a Transformer

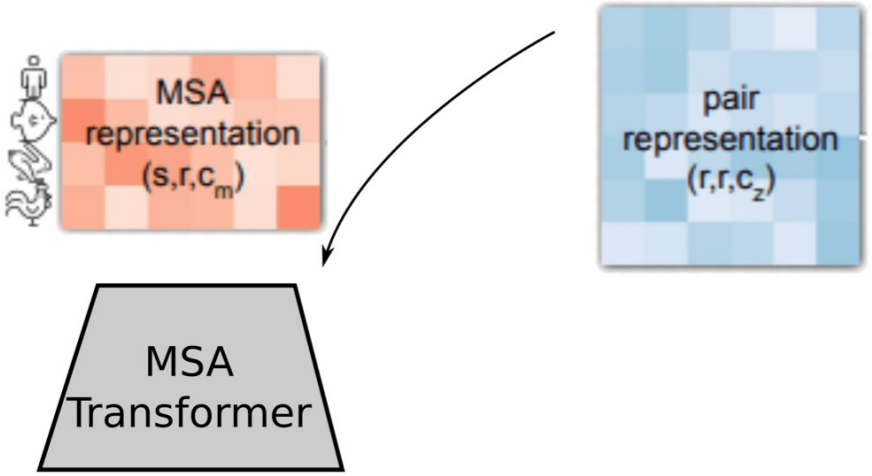
An Example Using Language Translation



https://www.allerin.com/blog/3-types-of-neural-networks-that-ai-uses

MSA Transformer

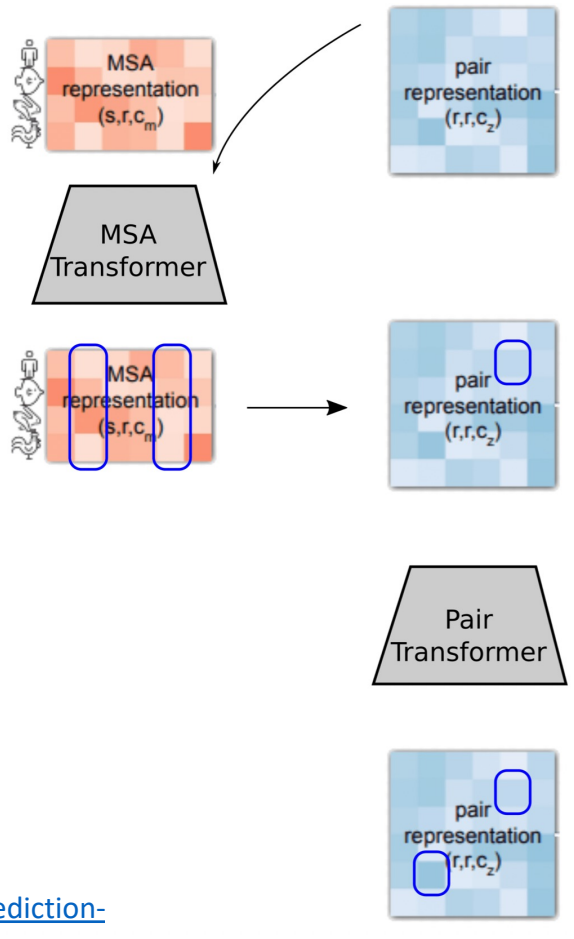
- The MSA Transformer limits its attention two ways:
  - Row-wise: to determine which residues are most related
  - Column-wise: to determine which sequences are most important
- The limited MSA along with the Pair Representation are then fed into the first head of the Evoformer

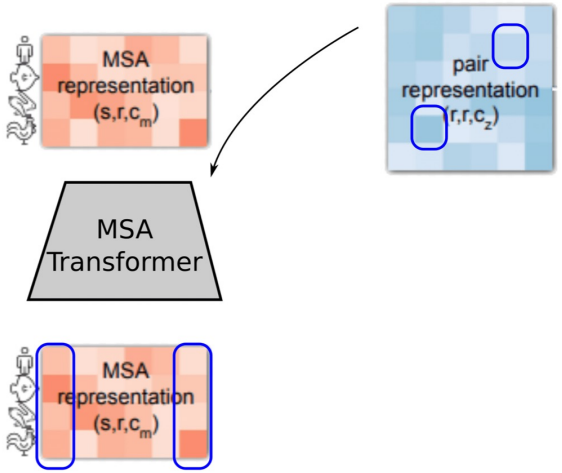


https://www.blopig.com/blog/2021/07/alphafold-2-is-here-whats-behind-the-structure-prediction-miracle/

Evoformer Part 1

- The first block of the Evoformer works to determine how close residues are
- start with correlations between two sets of residues, say A and B
- Highly correlation indicates these residues are close
- Now process is iterated - residue C is correlated with B
- So, B and C are close
- This process is repeated for all residues

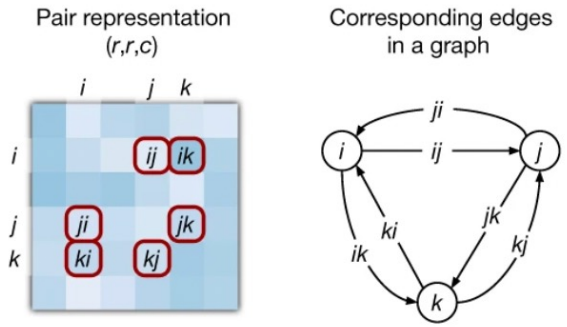


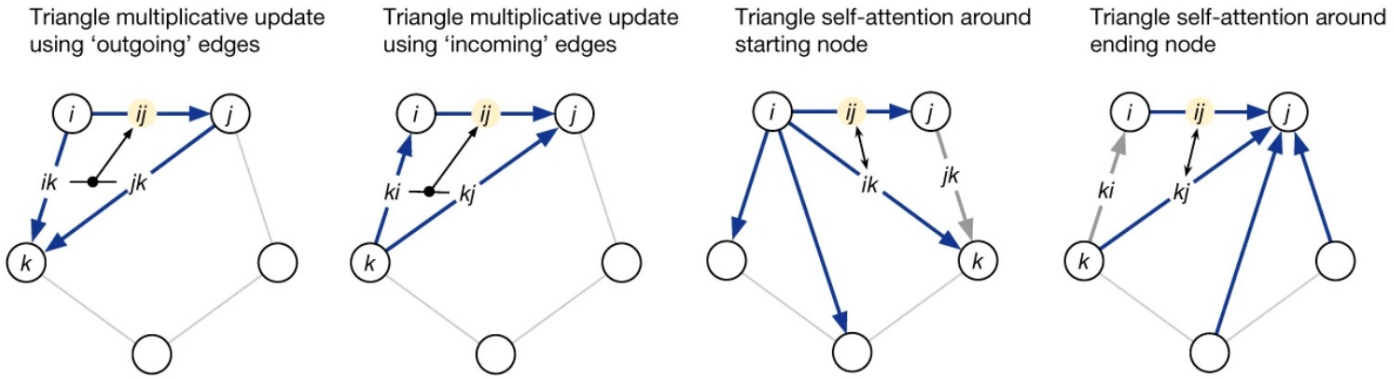


https://www.blopig.com/blog/2021/07/alphafold-2-is-here-whats-behind-the-structure-prediction-miracle/

Evoformer Part 2

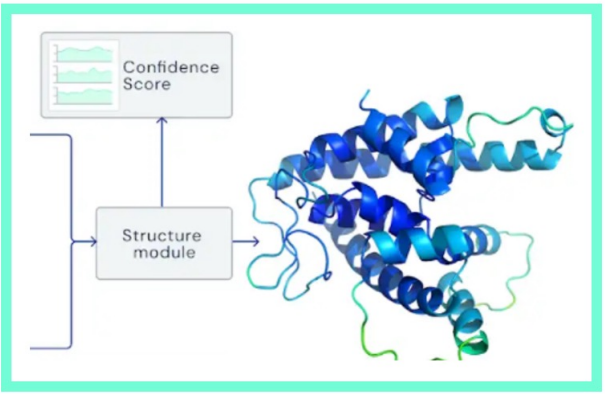
- The second block of the Evoformer works through pair wise distances between residues
- Here 3 residues are compared, and triangle inequality is enforced
- So, one side of the triangle must be less than or equal to the other two sides



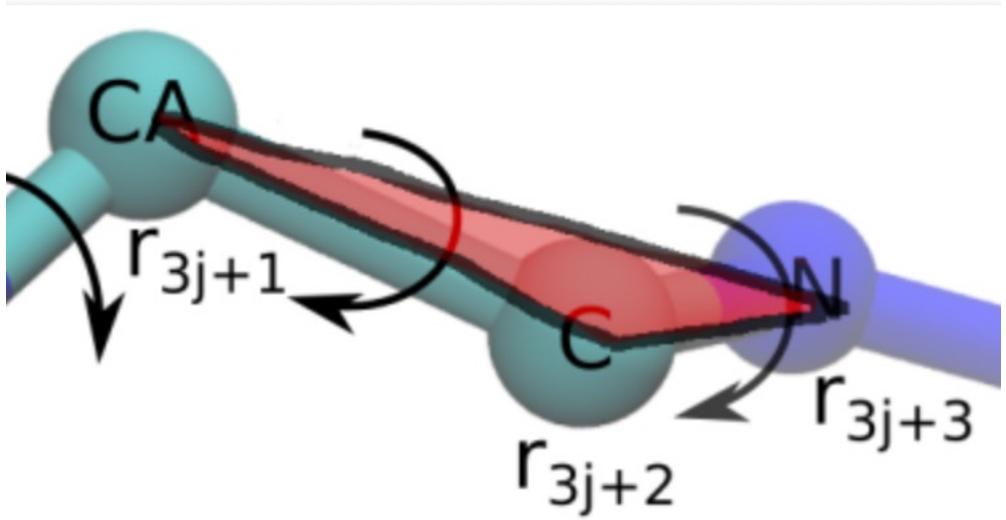


https://www.blopig.com/blog/2021/07/alphafold-2-is-here-whats-behind-the-structure-prediction-miracle/

Structure Module

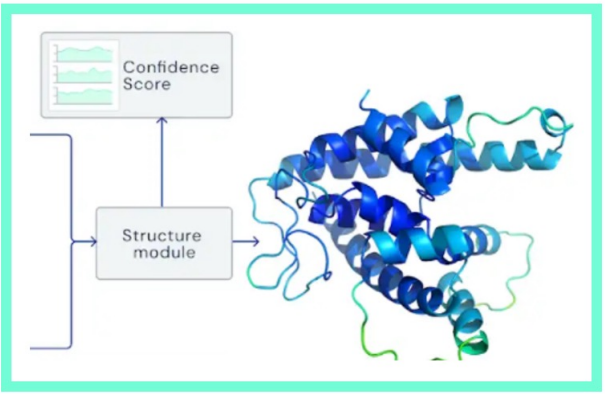


- The Evoformer outputs distances between residues, but residues are themselves three dimensional objects
- How are they oriented?
- Each residue starts as a “residue gas” or triangle between the Alpha Carbon, R-group Carbon, and the Nitrogen



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

Structure Module



- All residue gases start at the origin of the coordinate system
- Each position is defined as an affine matrix, or **xyz** coordinates for the three points of the triangle, which is multiplied by a displacement vector to "move" the residue gas to its final location

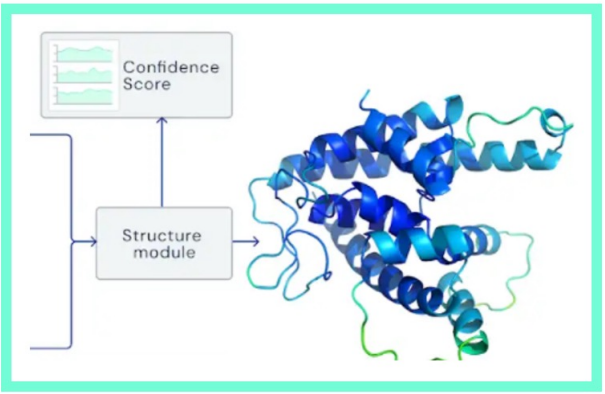
Displacement Vector

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

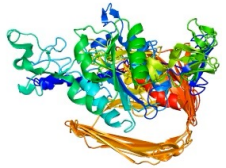
Affine Matrix

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

Structure Module – Invariant Point Attention



- The Structure Module also uses an attention mechanism called Invariant Point Attention
- This limits the data the model needs because points in 3D space are ***invariant*** to translation/rotation
- Basically, this means that no matter how you rotate/translate the final structure you still produce the same answer



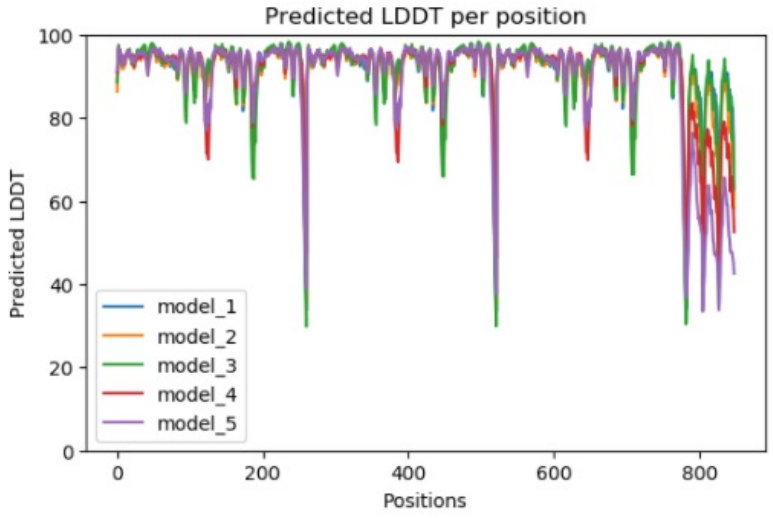
Recycling iteration 0, block 01
Secondary structure assigned from the final prediction

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

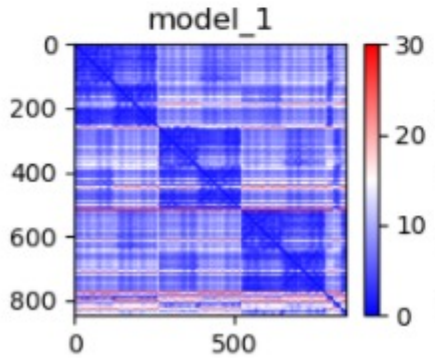
Assessing AlphaFold Accuracy

- We can assess the accuracy of the AlphaFold prediction using:
  - Predicted Local Distance Difference Test (pLDDT)
  - Predicted Alignment Error

Predicted Local Distance Difference Test (pLDDT)



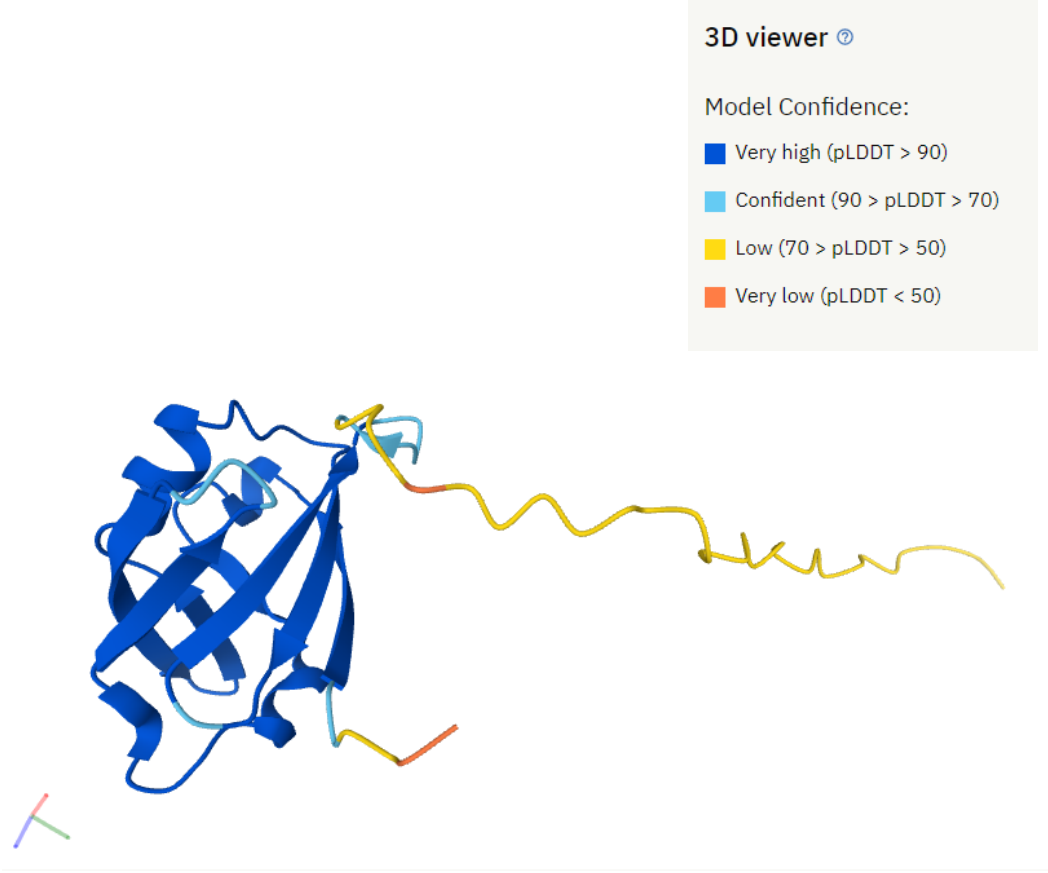
Predicted Alignment Error (PAE)



https://github.com/deepmind/alphafold

Predicted Local Distance Difference Test (pLDDT)

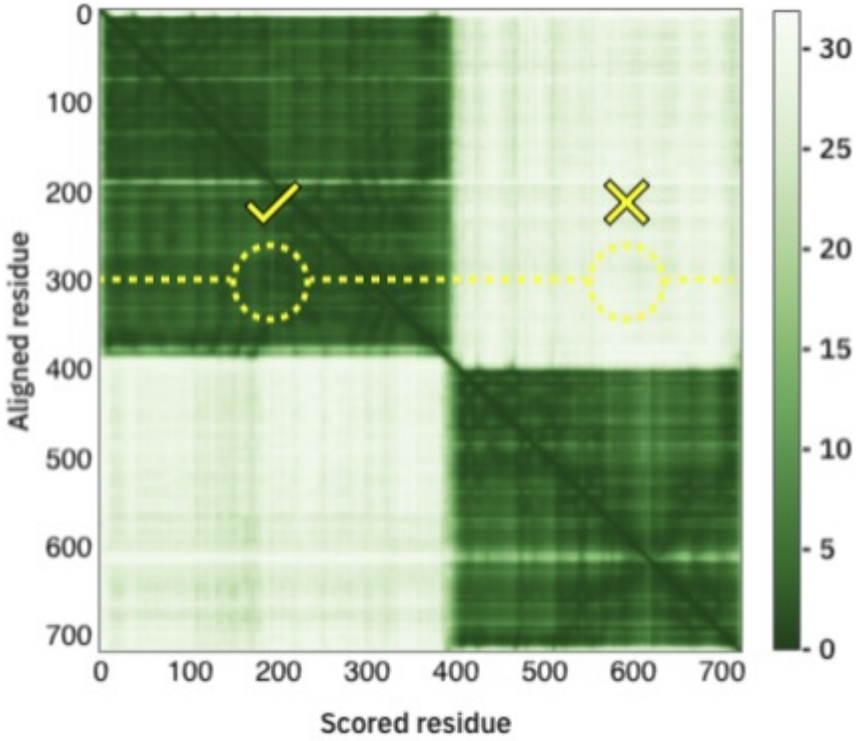
- per-residue confidence metric ranging from 0-100 (100 being the highest confidence)
- Regions below 50 could indicate disordered regions



https://alphafold.com/entry/Q9FX77

Predicted Alignment Error (PAE)

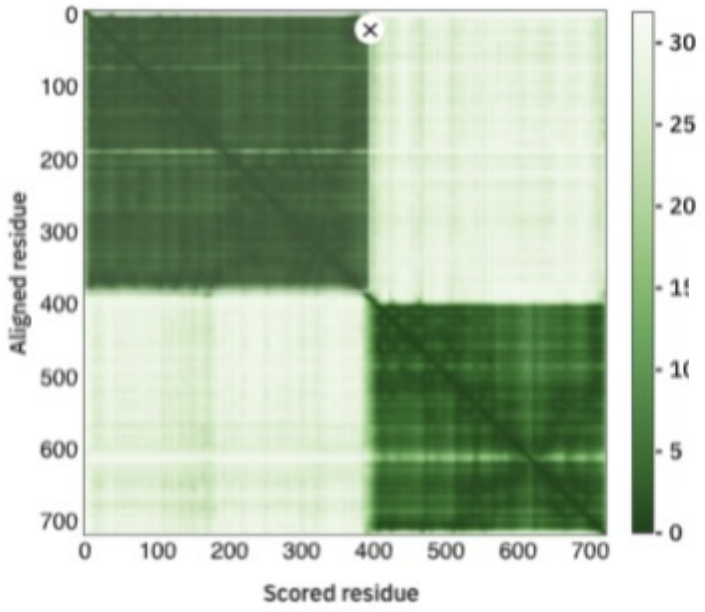
- The color at (x, y) corresponds to the expected distance error in residue x's position, when the prediction and true structure are aligned on residue y.
- So, in the example to the right:
  - The darker color indicates a lower error
  - When we are aligning on residue 300, we are more confident in the position of residue 200 and less confident in the position of residue 600

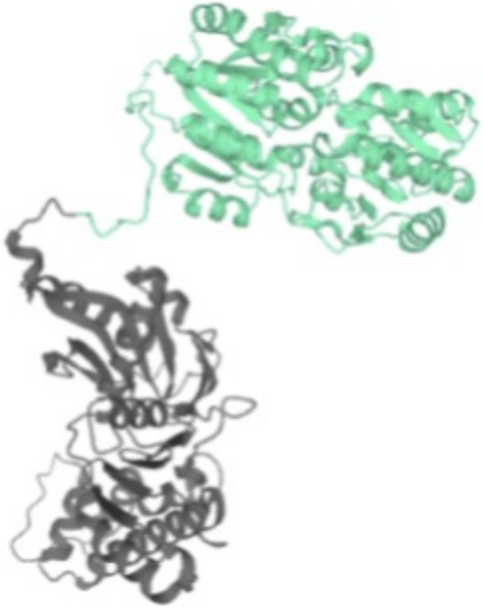


https://alphafold.com/entry/Q9FX77

Predicted Alignment Error (PAE) cont.

- The example in the previous slide came from a multimer prediction
- Here we see that the error is higher when assessing the position between the two chains





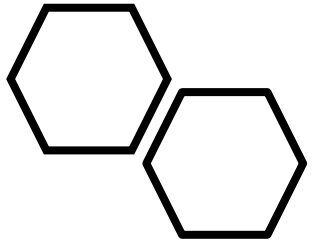
https://alphafold.com/entry/Q9FX77

Acknowledgement

Much of this tutorial has been adapted from the Oxford Protein Informatics Group's explanation on AlphaFold 2

References

1. https://www.genome.gov/genetics-glossary/Protein
2. https://www.nature.com/scitable/topicpage/protein-function-14123348/
3. https://www.ncbi.nlm.nih.gov/books/NBK26820/
4. https://directorsblog.nih.gov/tag/serial-scanning-3d-electron-microscopy/
5. https://www.ncbi.nlm.nih.gov/books/NBK26820/
6. https://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/methods-for-determining-structure
7. https://simple.wikipedia.org/wiki/X-ray_crystallography
8. https://deepmind.com/research/case-studies/alphafold
9. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC48166/
10. https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology
11. https://predictioncenter.org/
12. https://en.wikipedia.org/wiki/Neural_network
13. https://colah.github.io/posts/2015-08-Understanding-LSTMs/
14. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
15. https://towardsdatascience.com/transformer-neural-network-step-by-step-breakdown-of-the-beast-b3e096dc857f
16. https://en.wikipedia.org/wiki/FASTA_format
17. https://en.wikipedia.org/wiki/Multiple_sequence_alignment
18. https://www.pnas.org/content/114/34/9122
19. https://www.blopig.com/blog/2021/07/alphafold-2-is-here-whats-behind-the-structure-prediction-miracle/
20. https://github.com/deepmind/alphafold
21. https://alphafold.com/entry/Q9FX77



Next: Setup