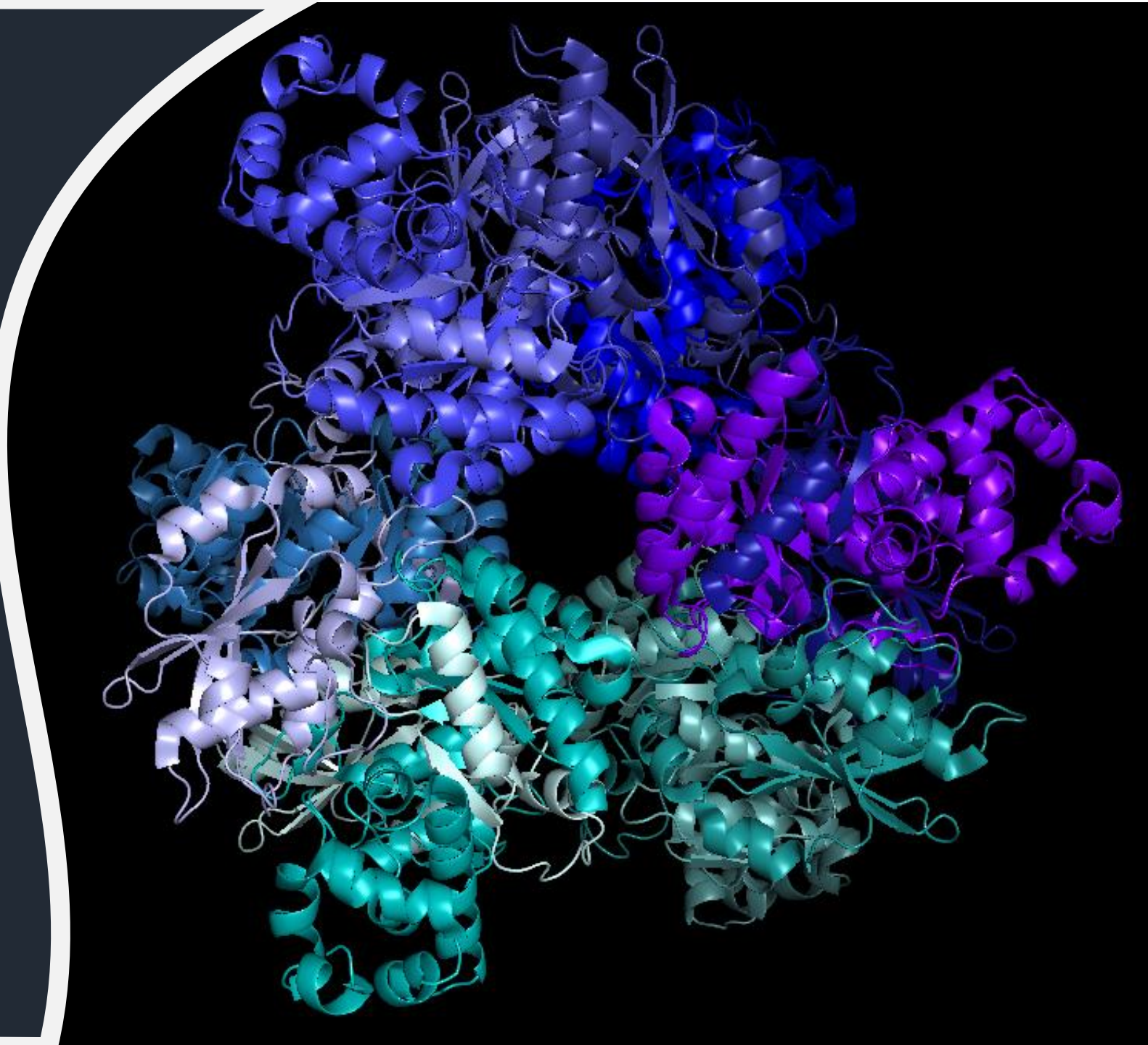
Introduction to Protein Structure Prediction With AlphaFold 2

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

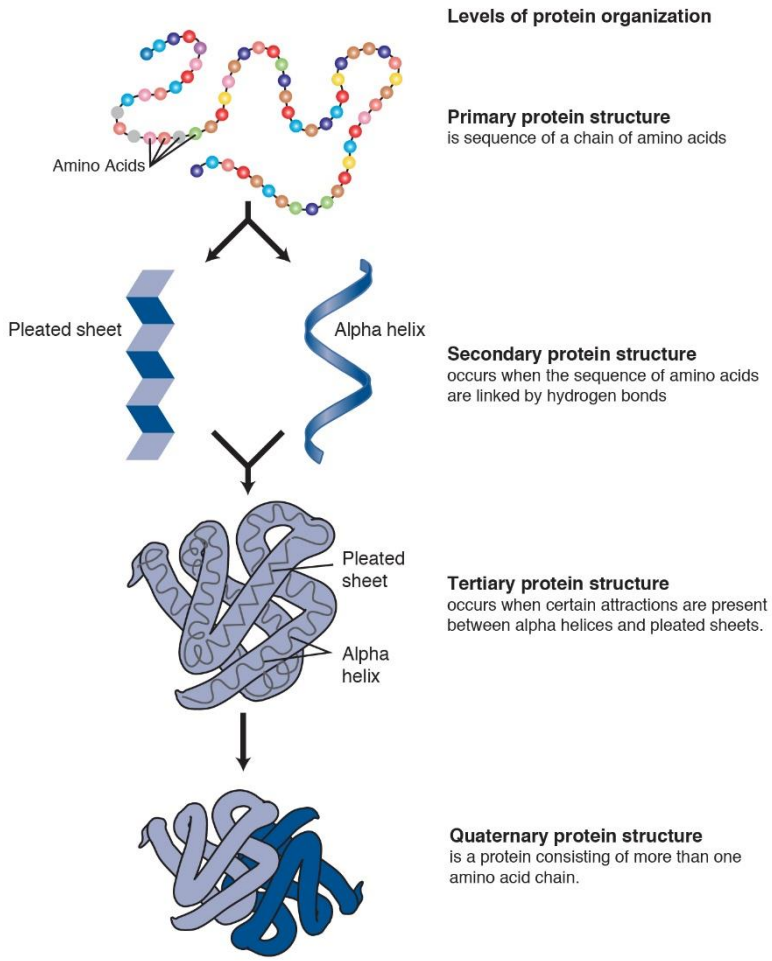
Uku-Kaspar Uustalu
Data Science Specialist

- ✓ 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

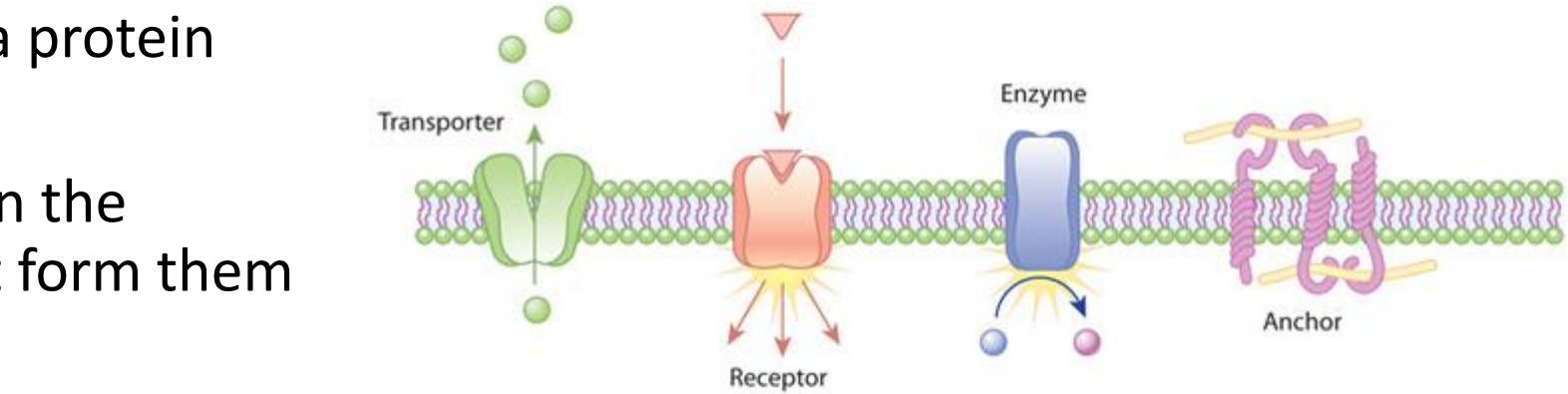


https://www.genome.gov/genetics-glossary/Protein

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



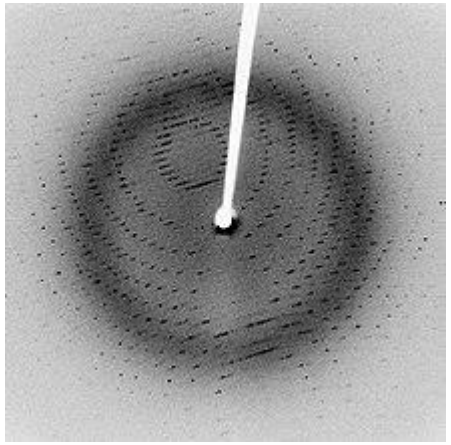
https://www.nature.com/scitable/topicpage/protein-function-14123348/

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

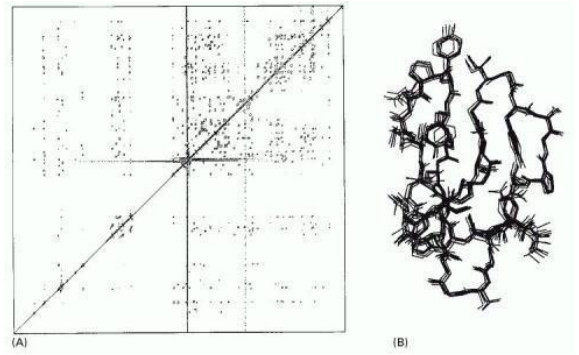
Laboratory Means To Determine Protein Structure

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

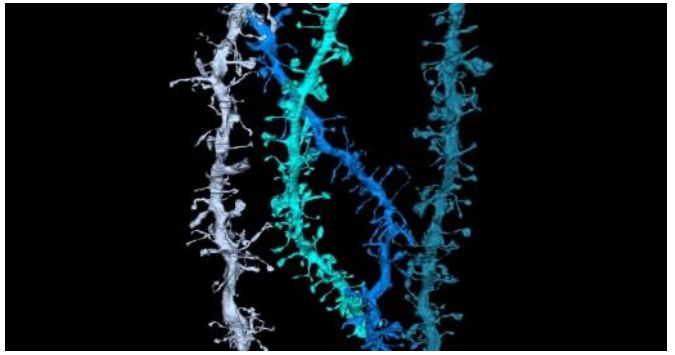
X-ray Crystallography



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

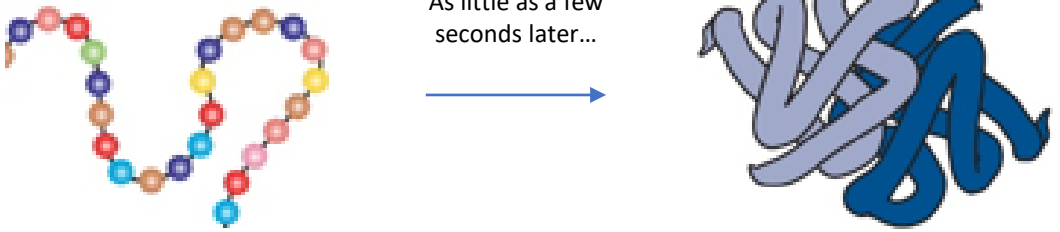
The Protein Structure Problem

- 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

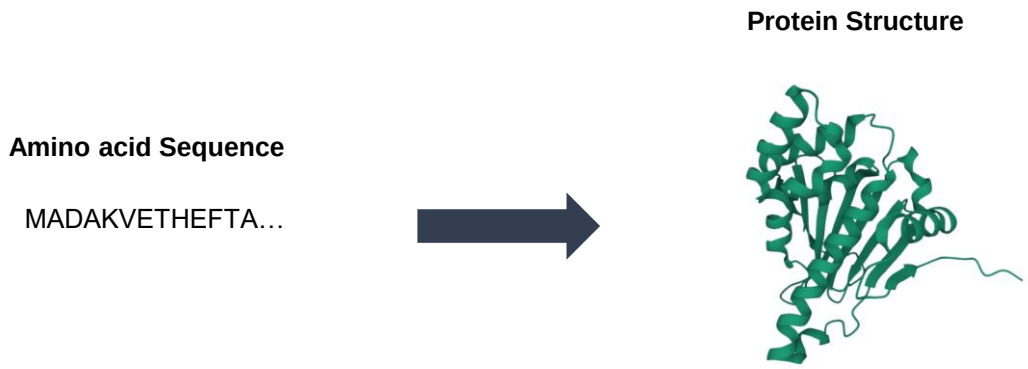
- Finding the native folded state of a protein by random searching of all possible configurations would take an enormous amount of time
- However, proteins can often fold within seconds
- Meaning some process must be guiding this folding



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
- 1994, the Critical Assessment of Structure Protein (CASP) was established as a biennial assessment of methods to predict structure from sequence

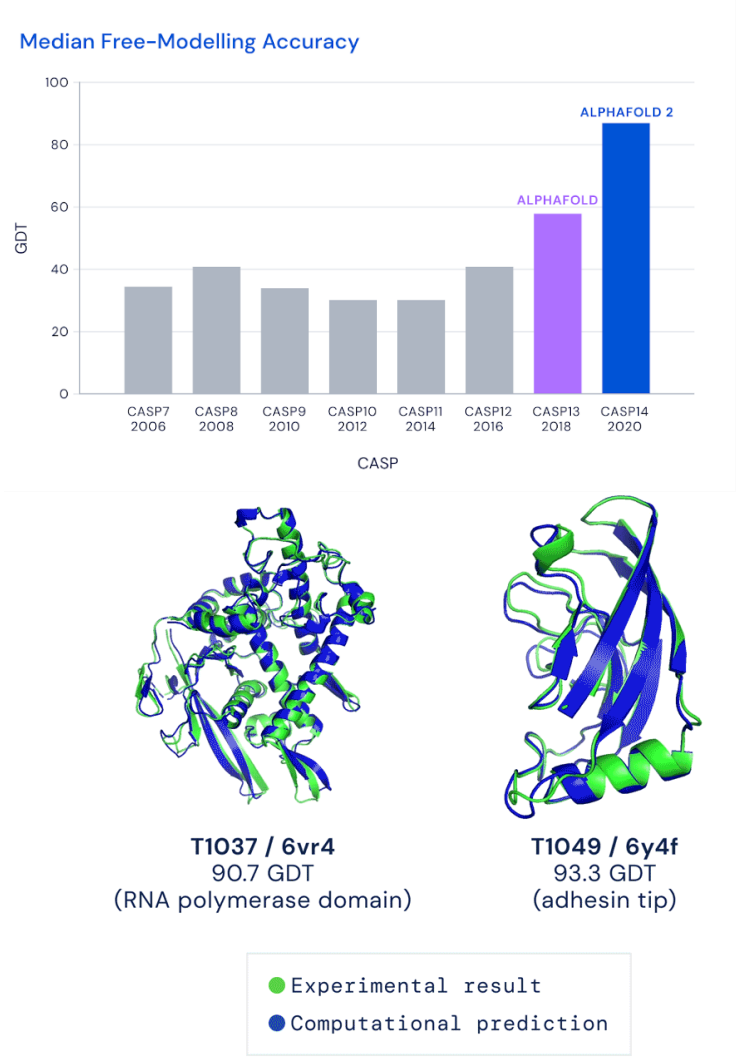


https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology

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

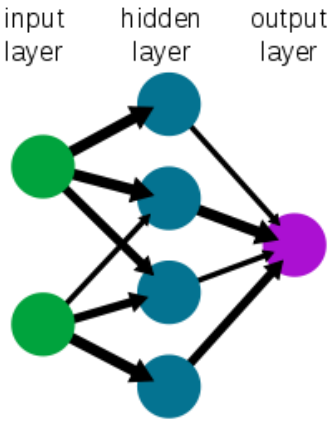


https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology

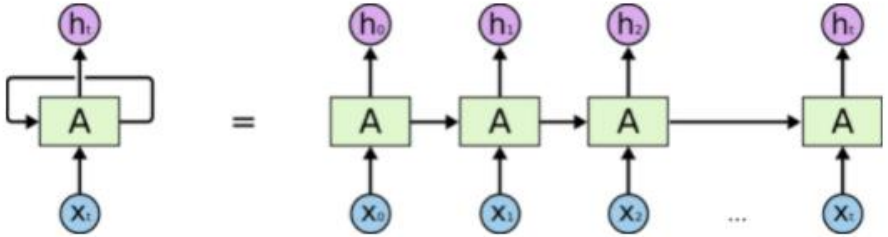
Simple v. Recurrent Neural Network

- A neural network is a machine learning algorithm commonly used in predictive modelling
  - Composed of an input layer, hidden layer, and an output layer
  - Traditionally learn from training
- A Recurrent Neural Network learns from training and from previous inputs
- However, the memory is poor when pulling from old connections

Simple Neural Network



Recurrent Neural Network



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

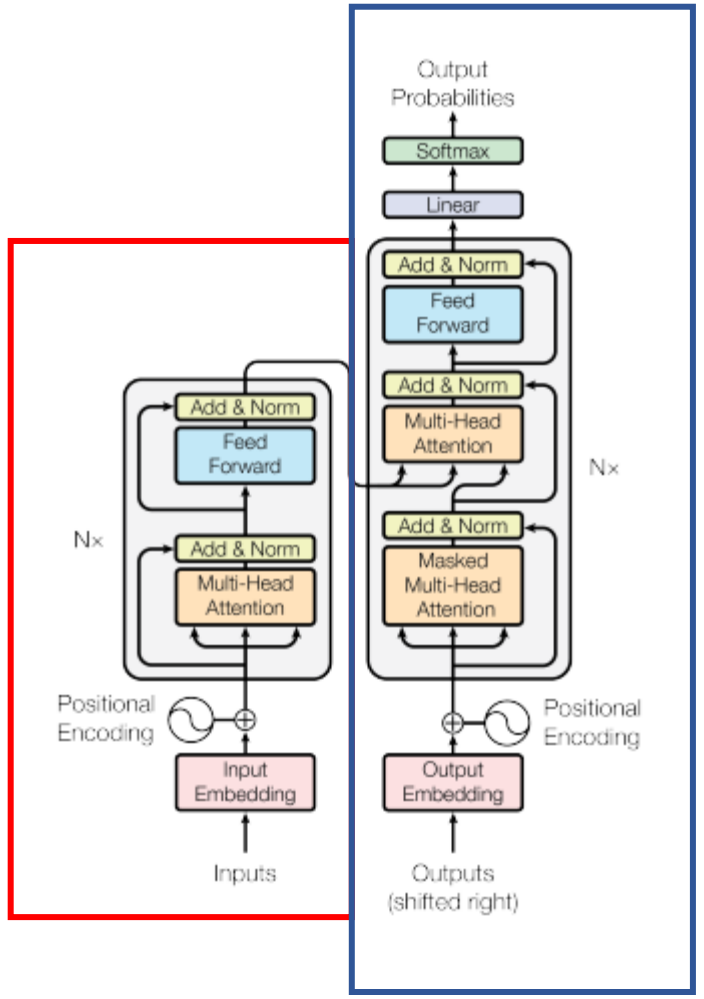
https://colah.github.io/posts/2015-08-Understanding-LSTMs/

The Transformer

- AlphaFold 2 uses an evolution of the Recurrent Neural Network called a Transformer
- The Transformers can be broken up into two blocks: the Encoder Block and the Decoder Block
- **Encoder Block:** turn sequences into vectors w/ positional information, the attention is limited by each character's interaction w/ the rest of the sequence
- **Decoder Block:** information from the previous block is converted to probability distributions

Encoder Block

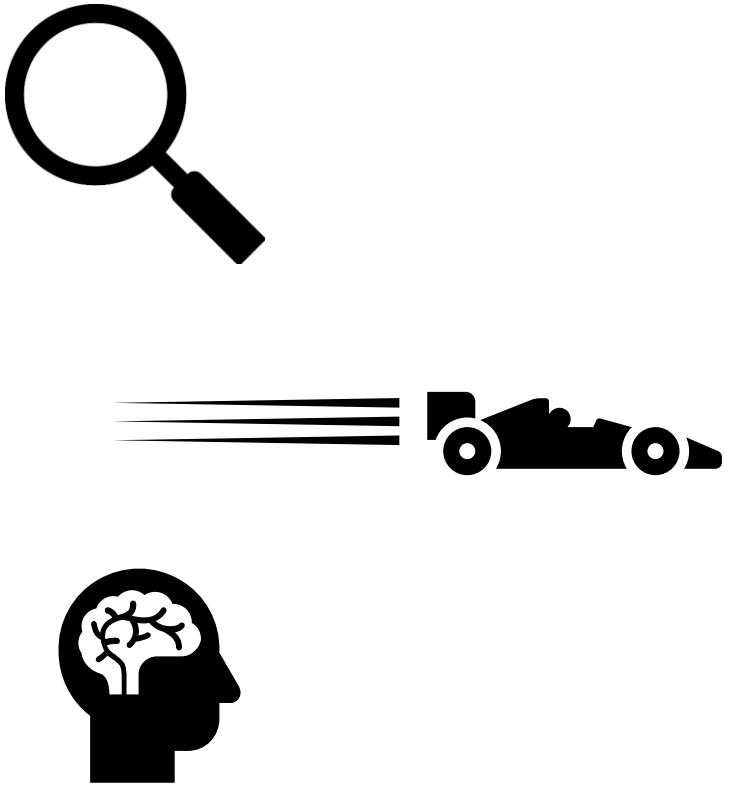
Decoder Block



https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

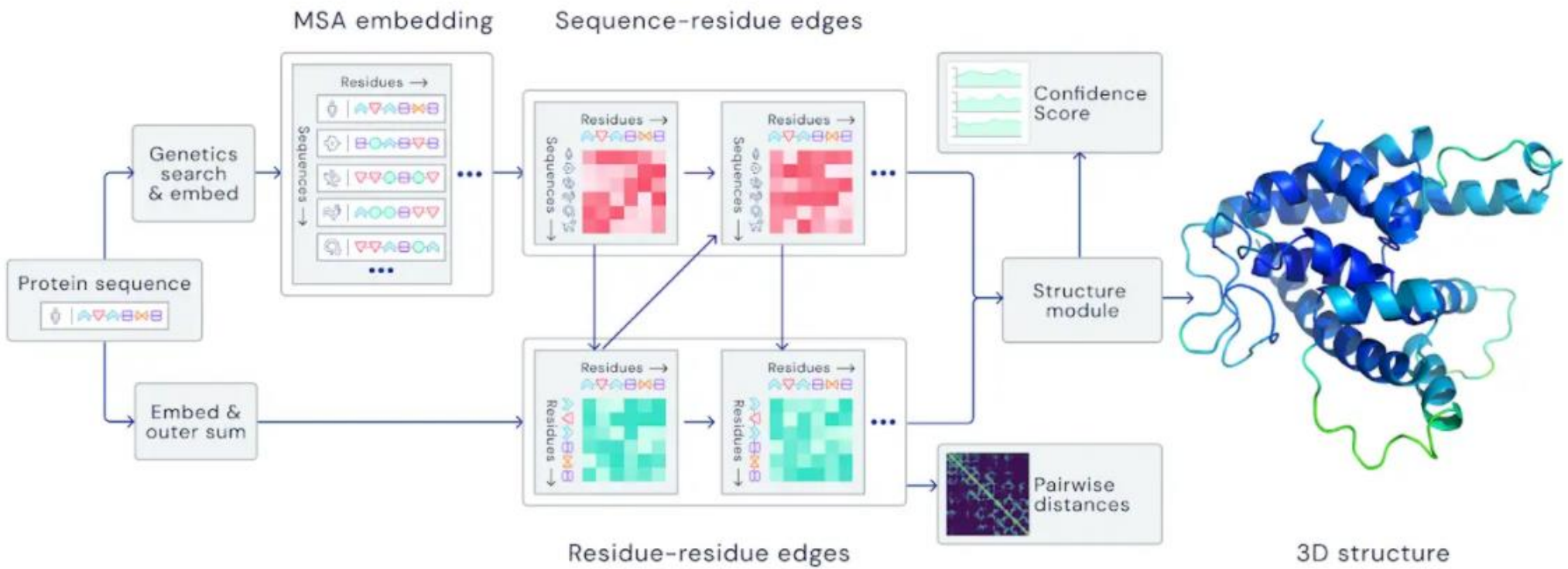
Transformer Benefits

- Limited attention means better memory utilization
- Faster model training times
- Overcome the issue of the memory being poor when pulling from old connections



https://towardsdatascience.com/transformer-neural-network-step-by-step-breakdown-of-the-beast-b3e096dc857f

The AlphaFold 2 Workflow



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

Protein Sequence Information

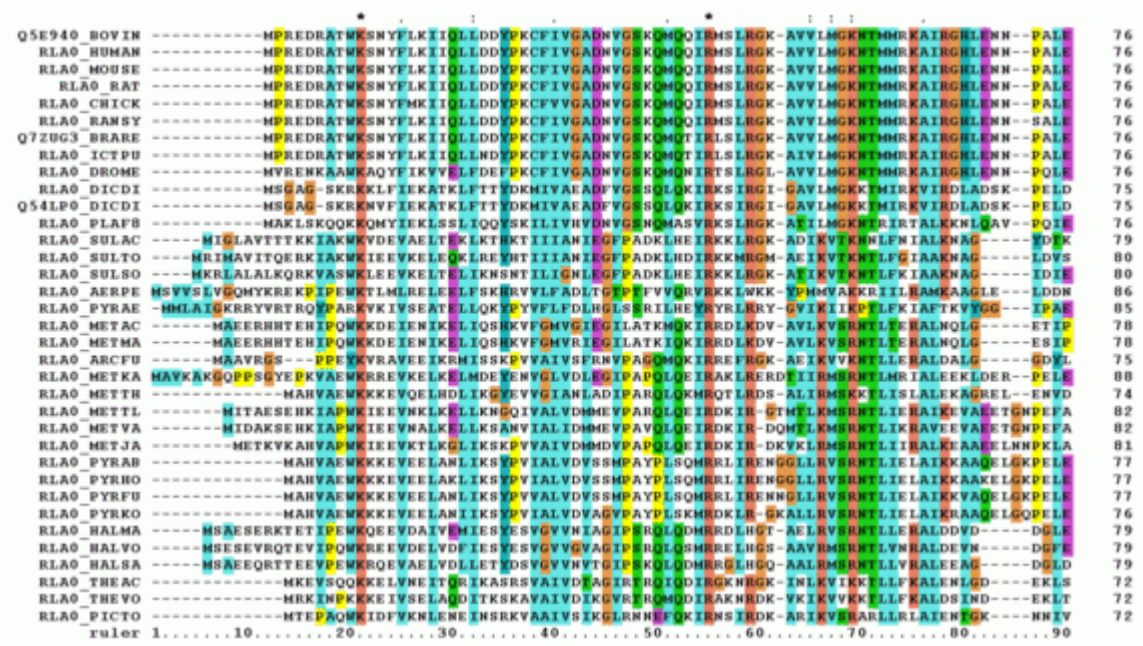
- Protein sequence information stored as a fasta file. Consists of a:
  - Header
  - sequence

```
>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
KEDQLEYLEEKRIKEVVKKHSEFVAYPIQLVVTKEVEKEVPETEEEDKAAEEDDKKPKLE
EVKDEEDEKKEKKTKTVKEEVTETEELNKTKPLWTRNPSDITQDEYNAFYKSISNDWEDP
LAVKHFSVEGQLEFRAILFVPKRAPFDAFESKKKKNNIKLYVRRVFITDDAEELIPEWLS
FIKGVVDSEDLPLNLSREMLQQNKILKVIRKNIVKKMIETFNEISEDQEQFNQFYTAFSK
NIKLGIHEDAQNRQSLAKLLRFYSTKSSEEMTSLSDYVTRMPEHQKNIYYITGESIKAVE
KSPFLDALKAKNFEVLFMVDPIDEYAMTQLKEFEDKKLVDITKDFELEESDEEKAAREKE
IKEYEPLTKALKDILGDQVEKVVVSYKLVDAPAAIRTGQFGWSANMERIMKAQALRDTTM
SSYMSSKKTFEISPSSPIIKELKKKVETDGAEDKTVKDLTTLLFDTALLTSGFTLDEPSN
FAHRINRLIALGLNIDDDSEETAVEPEATTTASTDEPAGESAMEEVD
```

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

Building a Multiple Sequence Alignment (MSA)

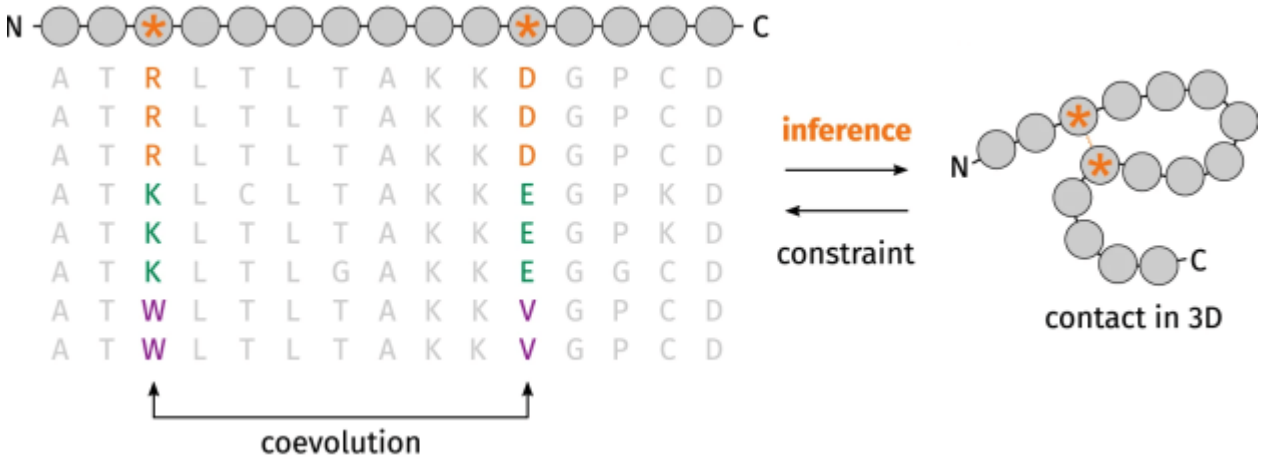
- The sequence is checked against a reference database of sequences - UniRef90 database
- Sequences with sections that align well to our query are then used as input



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

Coevolution of Residues

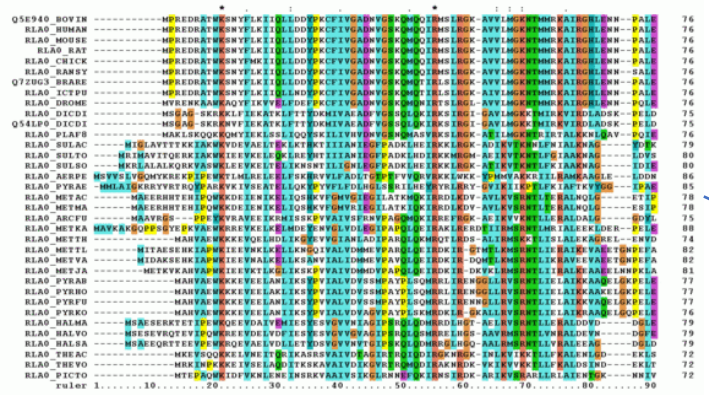
- So how does one go from an alignment to a structure?
- The theory is that residues that coevolve are generally close to each other in the protein's folded state

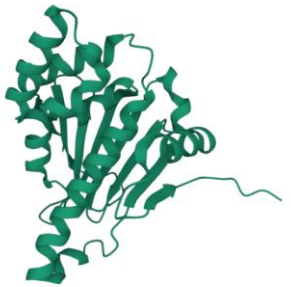


https://www.pnas.org/content/114/34/9122

Template Information

- The similar sequences pulled from also have structural information
- These templates can be converted into distance matrices to determine the distance between residues

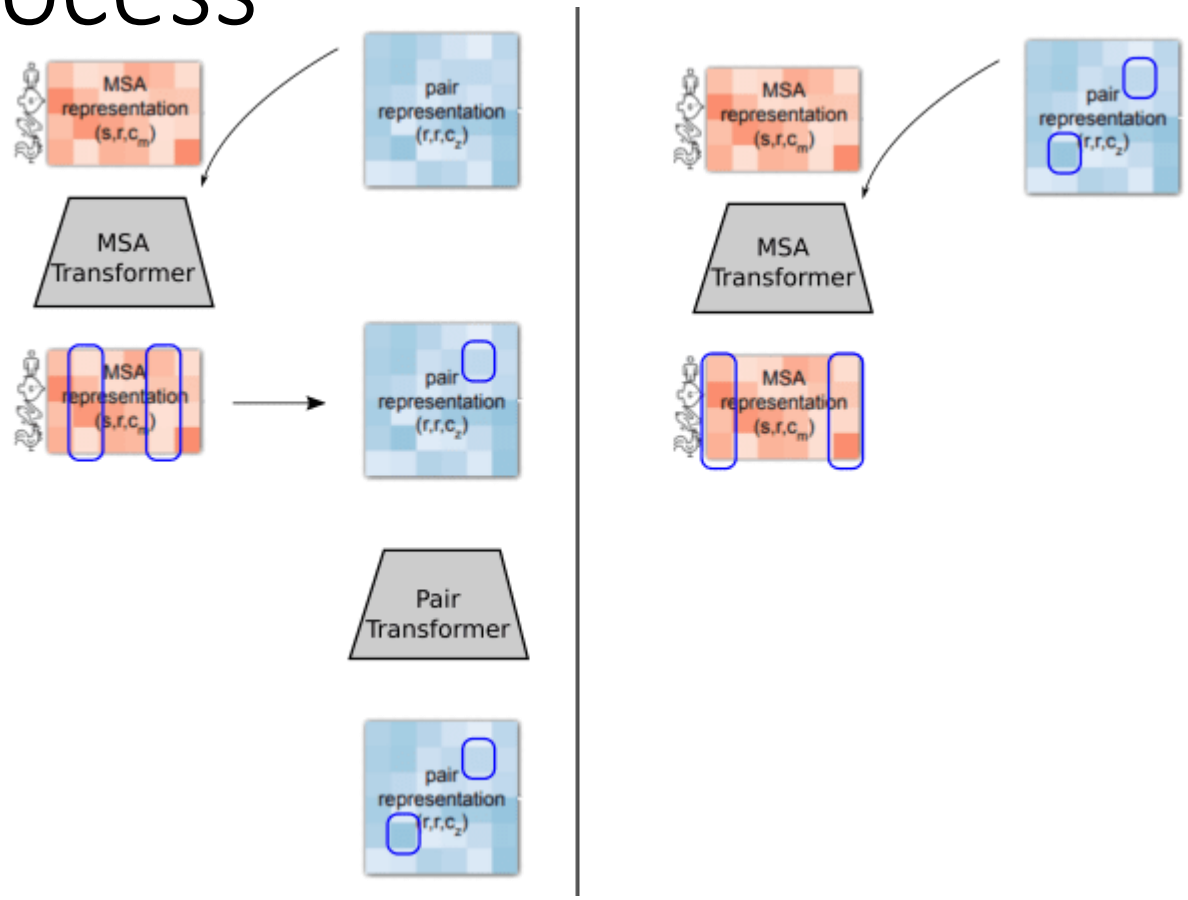




https://www.blopig.com/blog/2021/07/alphafold-2-is-here-whats-behind-the-structure-prediction-miracle/

The Evoformer Process

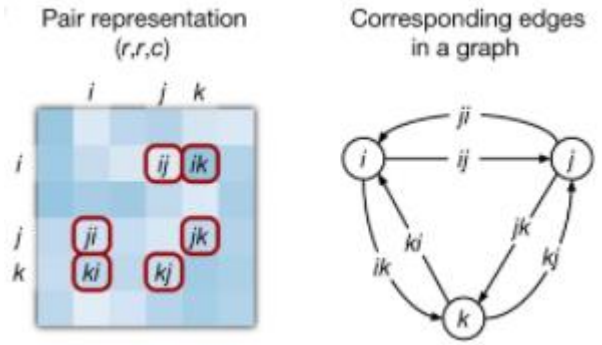
- Takes the MSA representation and the pair representation
- Uses the pair representation to limit the attention of the MSA transformer
- Model then determines two residues are close
- Given this information, the Pair Transformer notes that another two residues could be close
- Process is iterated until a possible structure is resolved

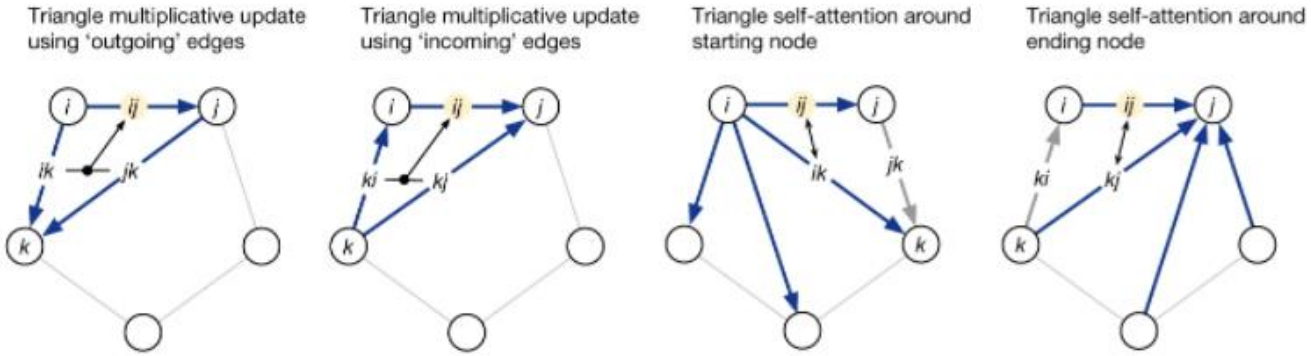


https://www.blopig.com/blog/2021/07/alphafold-2-is-here-whats-behind-the-structure-prediction-miracle/

A Closer Look at the Pair Transformer

- The pair transformer works on the principle of Triangle inequality, where the sum of two sides must be greater than or equal to the third side.
- Using this theorem, we can determine the likely distance residues have from one another because the distance between three points can never break that theorem

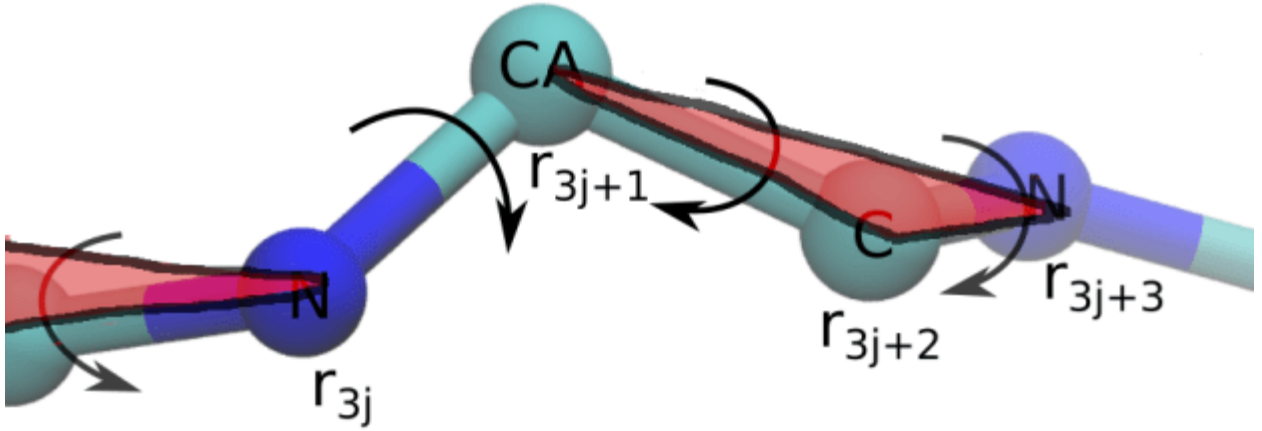




https://www.blopig.com/blog/2021/07/alphafold-2-is-here-whats-behind-the-structure-prediction-miracle/

The Structure Module

- Begins with each amino acid as a residue gas, or triangle with points at Nitrogen, R group Carbon and the Alpha Carbon
- These “gases” start at an origin point and are moved by the model using the pair distances and the information from the pairwise distance matric and the MSA



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

The Output

features.pkl	unrelaxed_model_*.pdb	relaxed_model_*.pdb	ranked_*.pdb	ranking_debug.json	timings.json	msas/	result_model_*.pkl
A pickle file w/ input feature NumPy arrays	A PDB file w/ predicted structure, exactly as outputted by the model	A PDB file w/ predicted structure, after performing an Amber relaxation procedure on the unrelaxed structure prediction	A PDB file w/ relaxed predicted structures, after reordering by model confidence (using predicted LDDT (pLDDT) scores). ranked_0.pdb = highest confidence ranked_4.pdb = lowest confidence	A JSON file w/ pLDDT values used to perform the model ranking, and a mapping back to the original model names.	A JSON file w/ times taken to run each section of the AlphaFold pipeline.	A directory containing the files describing the various genetic tool hits that were used to construct the input MSA.	– A pickle file w/ a nested dictionary of the various NumPy arrays directly produced by the model: • Structure Module Output • Distograms • Per-residue pLDDT scores • predicted TM-score • predicted pairwise aligned errors

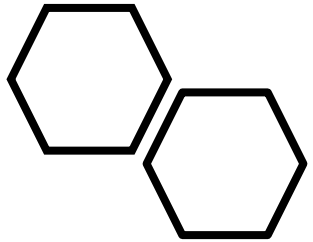
https://github.com/deepmind/alphafold

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



Next: Preparing the Batch Script