Highly accurate protein structure prediction with AlphaFold2

Loghman Samani
March 2024

Table of Contents

Introducing AlphaFold

What is AlphaFold?

AlphaFold is a cutting-edge computational algorithm developed by DeepMind
AlphaFold1 (2018)
AlphaFold2 (2020)
Critical Assessment of Structure Prediction (CASP)

Purpose

It's designed to predict the 3D structure of proteins

Introducing AlphaFold

Methodology

Employes a deep learning architecture
Attention mechanisms
Gradient-based optimization

Key Features

High accuracy
Scalability
Accessibility

Applications

Drug discovery
Biotechnology
Basic research

Machine Learning Model Overview

Model Representation

Model = $W \cdot X^T + b$

- W: Weight matrix
- X : Input features
- B : Bias term

Training Process

Initialization : randomly initialize W and b
Repeat until Convergence:

- Calculate Gradient:
  - dJ/dW and dJ/db
- Update Parameters:
  - $W = W - \alpha \cdot dJ/dW$
  - $B = b - \alpha \cdot dJ/db$

α (alpha) : learning rate

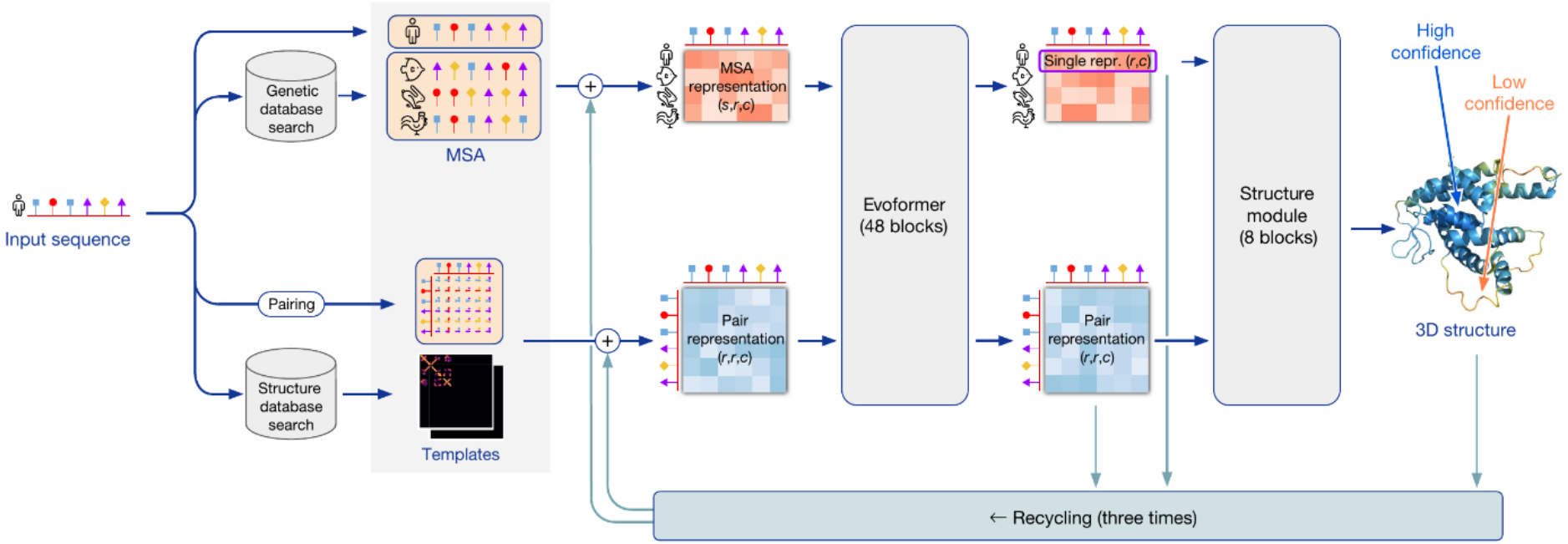
Cost Function

$J(W, b) = 1/2m \sum (\text{predicted} - \text{actual})^2$

- m: number of training examples

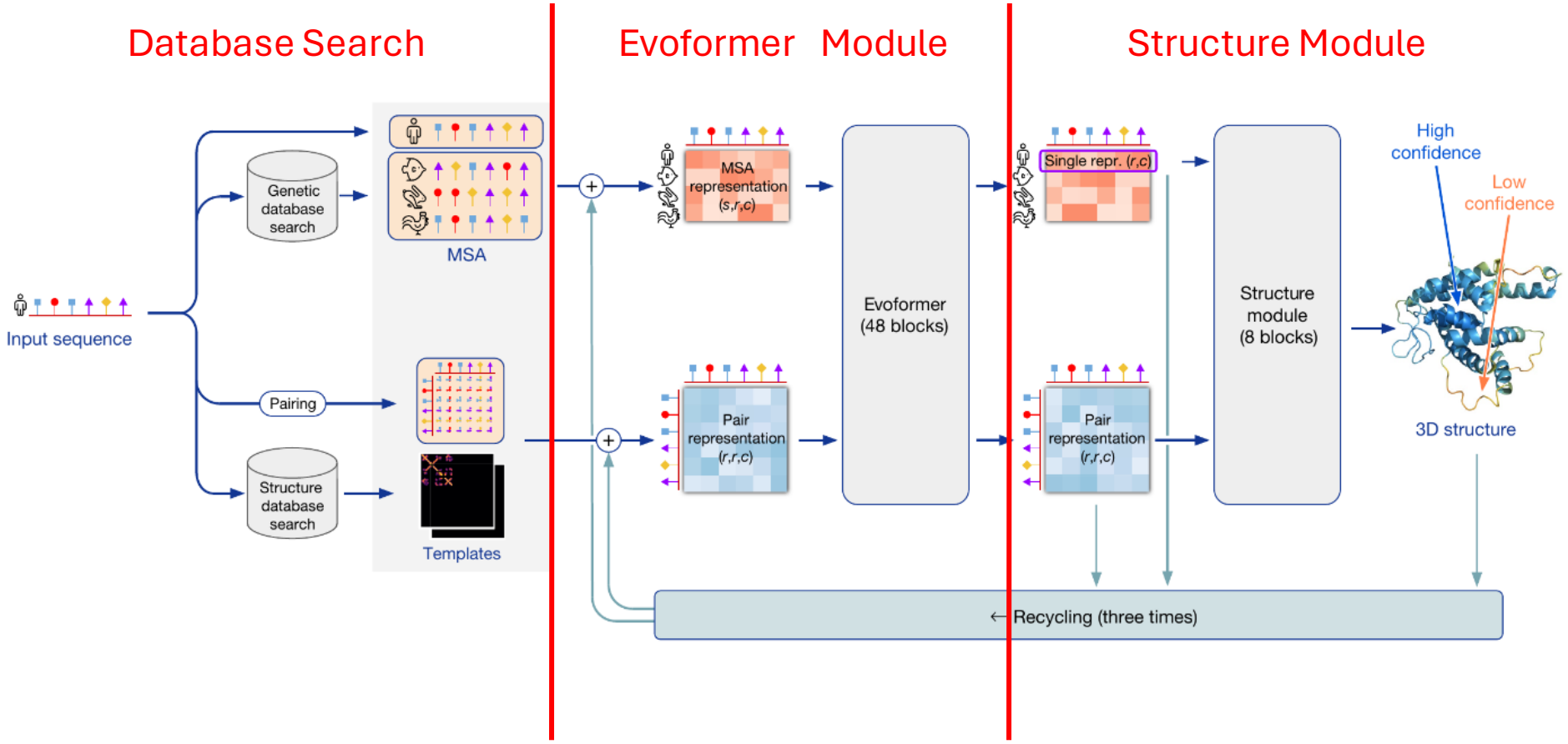
Objective: minimize the cost function J(W,b) by updating parameters W and b iteratively

AlphaFold Structure



John Jumper et al. 15 July 2021. https://www.nature.com/articles/s41586-021-03819-2

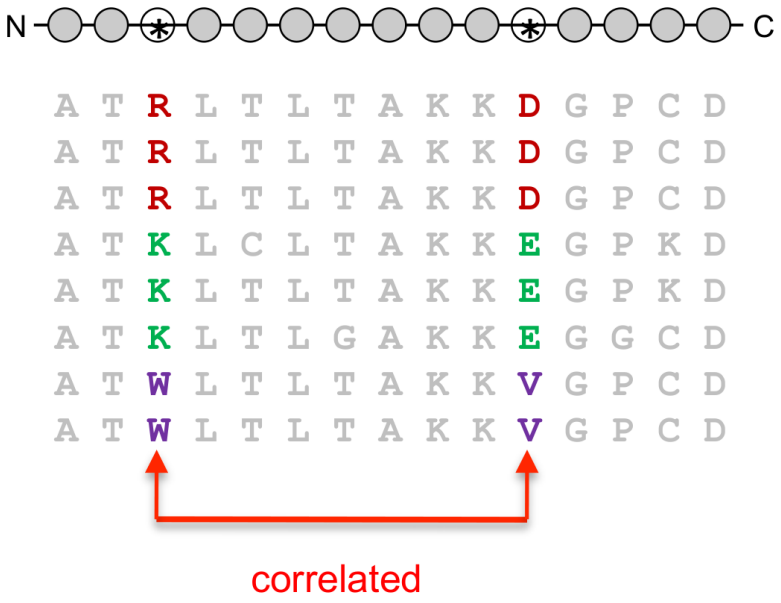
AlphaFold Structure



Database Search

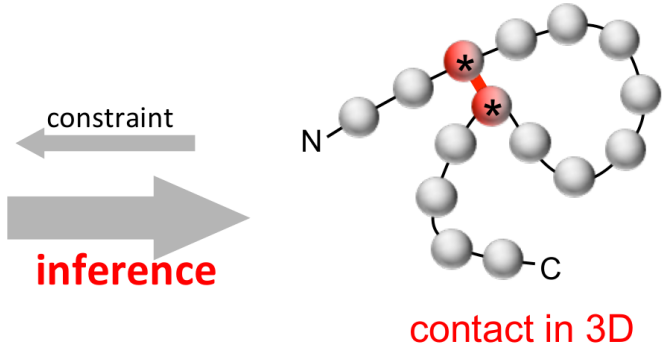
Genetic Database Search

- Big Fantastic Database (BFD)
- Multiple Sequence Alignment (MSA)



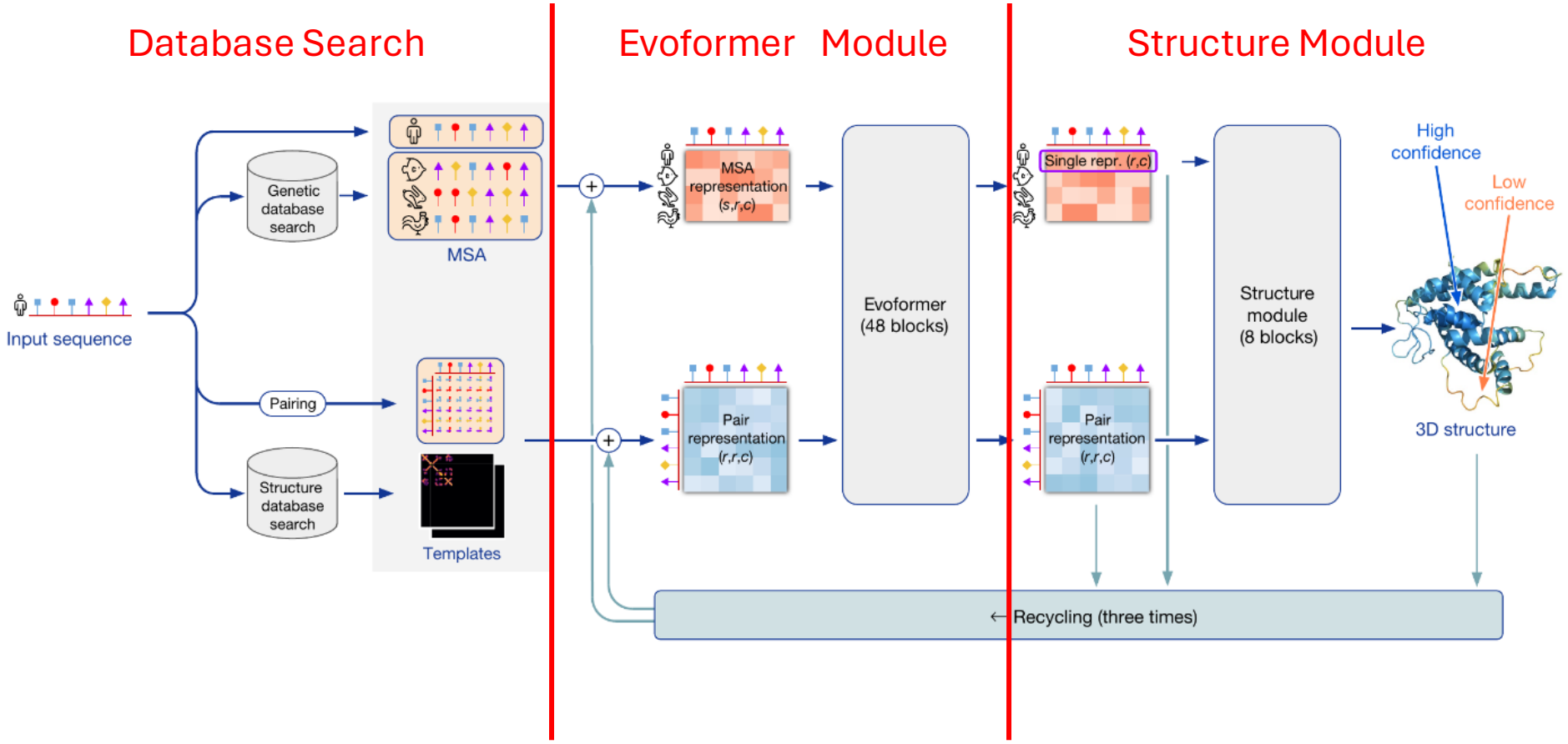
Structure Database Search

- Protein Data Bank (PDB)
- The 3D Structure of Proteins
- X-ray crystallography



Debora S. Marks et al. December 7, 2011. https://doi.org/10.1371/journal.pone.0028766

AlphaFold Structure



Evoformer (Evolutionary Transformer ?)

Introduction to Transformers

Attention is all you need, Google Brain (2017)

Application Area

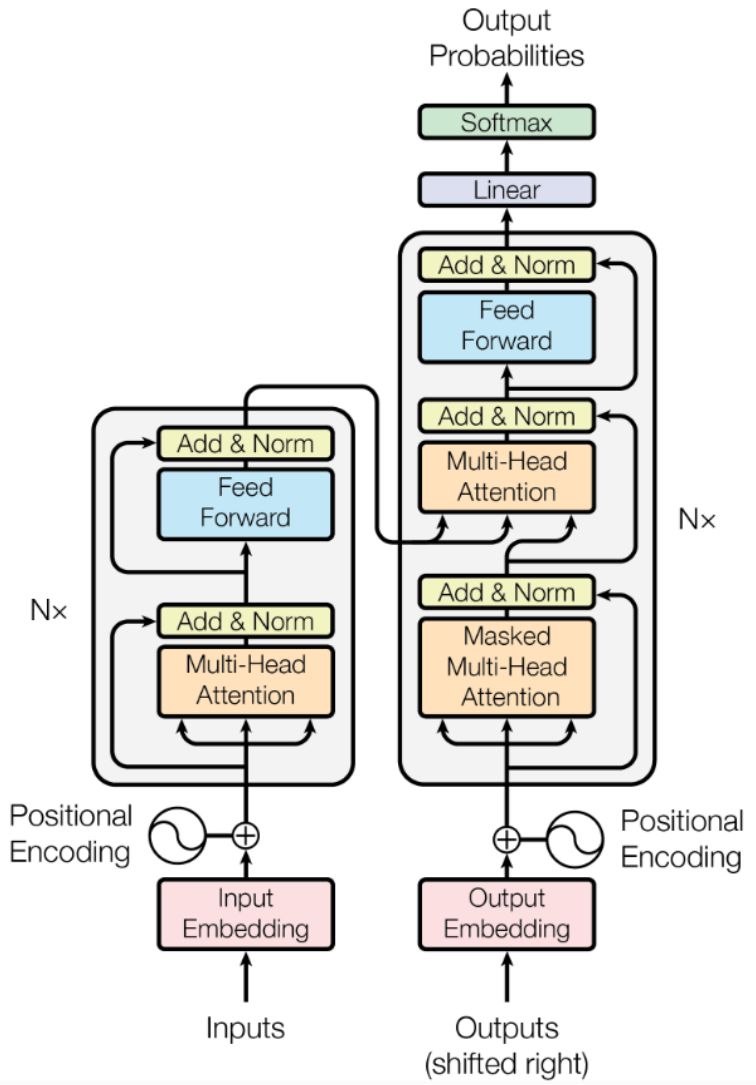
Natural Language Processing

Notable Models

GPT-2, GPT-3, Gemini, Google BERT, AlphaFold2

Transformer Components

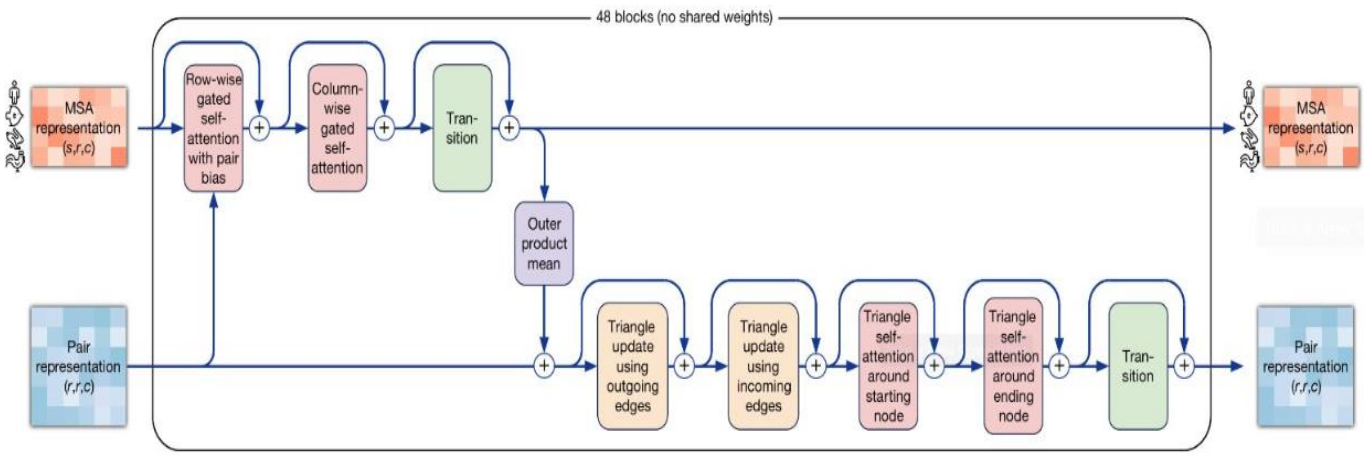
Embedding Input
Multi-Head Attention
Feed Forward Neural Network



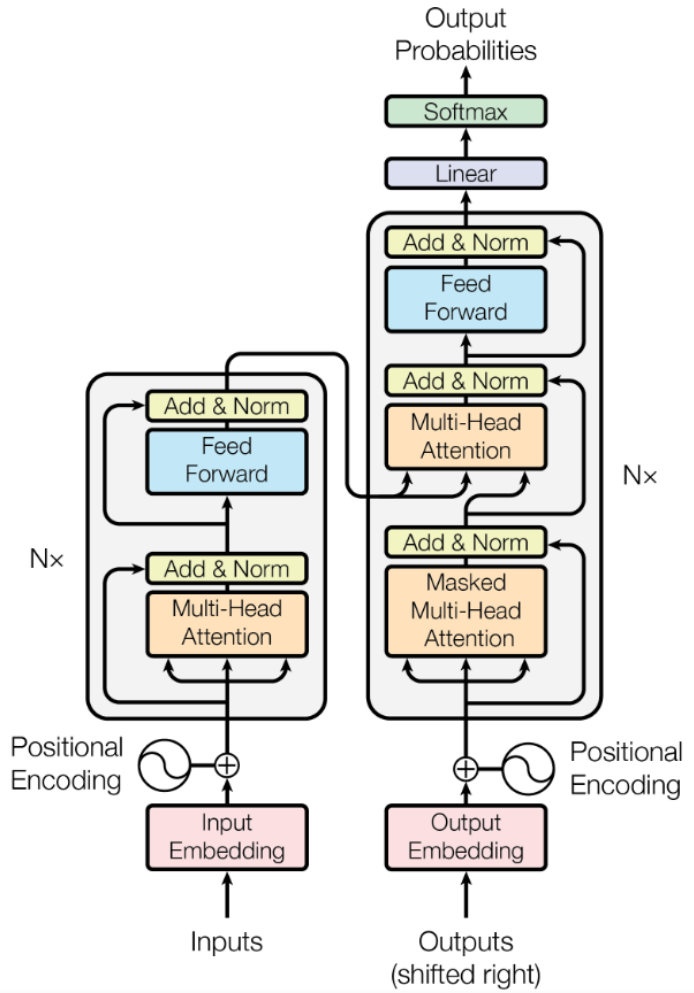
Ashish Vaswani et al. 12 Jun 2017. https://arxiv.org/abs/1706.03762

Evoformer (Evolutionary Transformer ?)

- MSA Representation Transformer
- Pair Representation Transformer



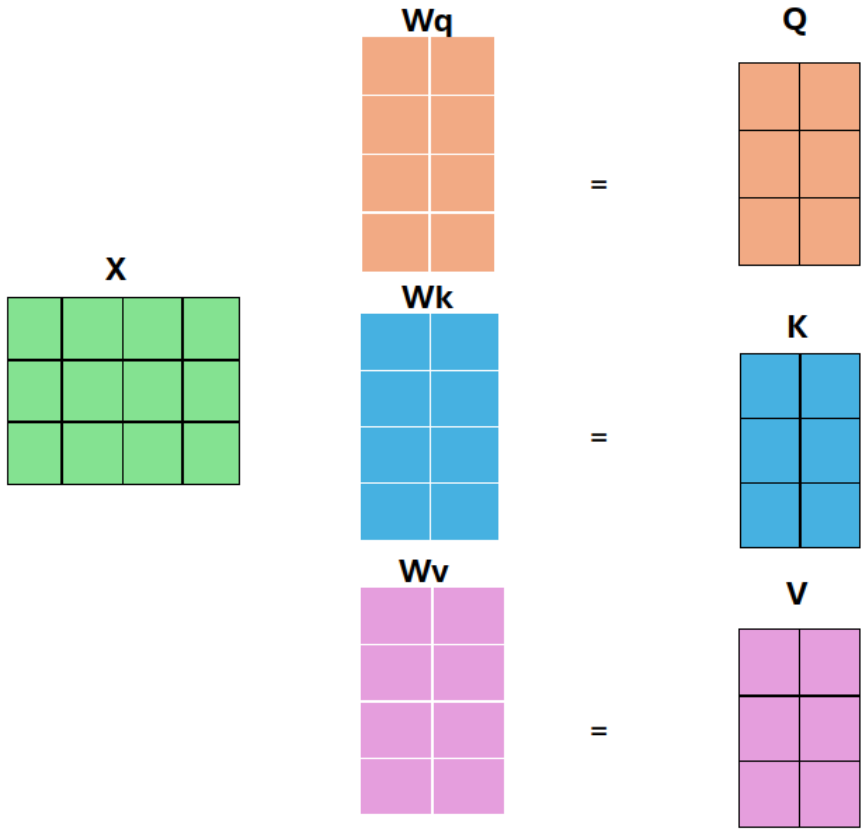
John Jumper et al. 15 July 2021. https://www.nature.com/articles/s41586-021-03819-2



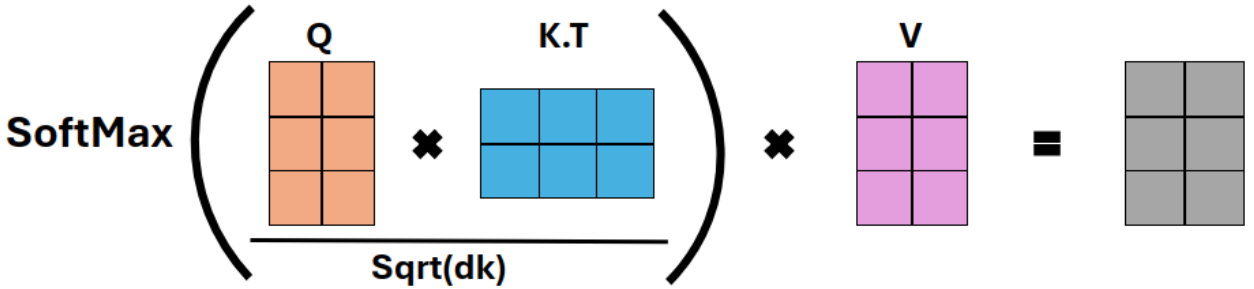
Evoformer (Evolutionary Transformer ?)

Attention Mechanism

Capture Dependencies and Relationships within Input Sequence



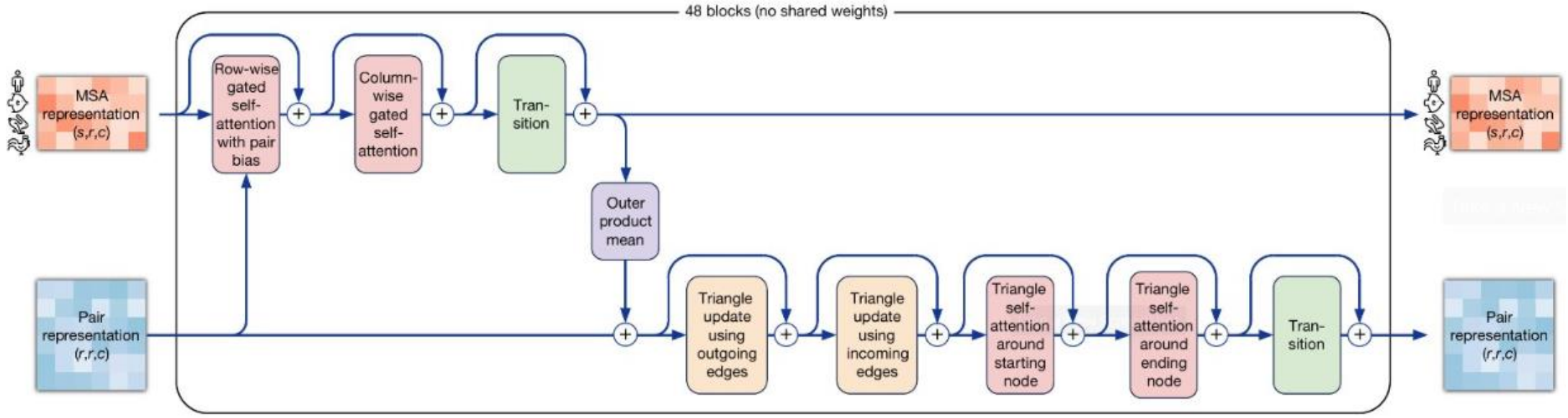
$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$



Evoformer (Evolutionary Transformer ?)

MSA Representation Transformer

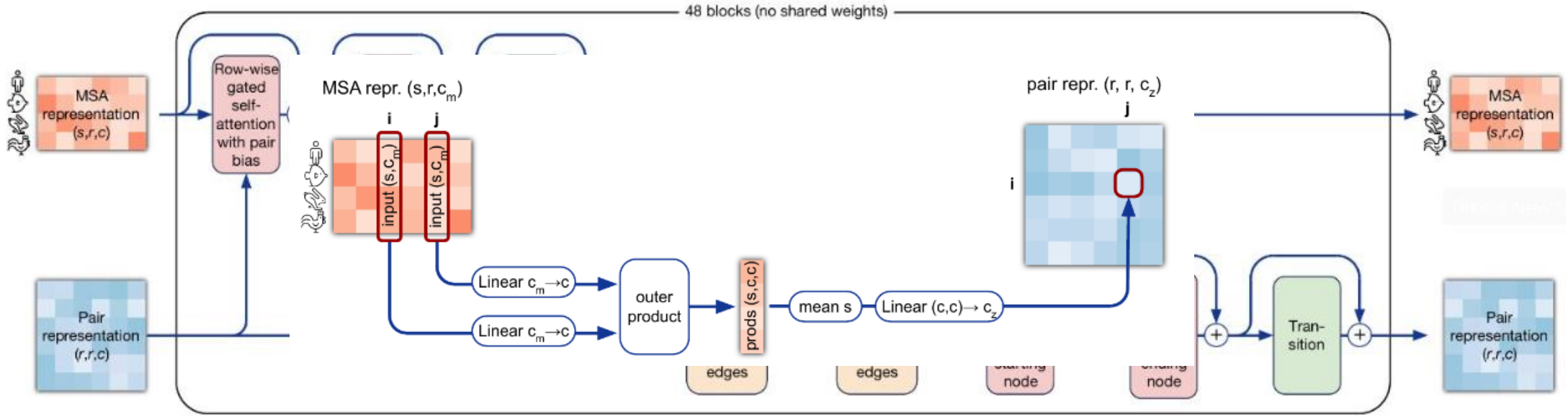
- Row-wise and Column-wise Attention ($\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$)
- Feed Forward Neural Network ($\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2$)



Evoformer (Evolutionary Transformer ?)

MSA Representation Transformer

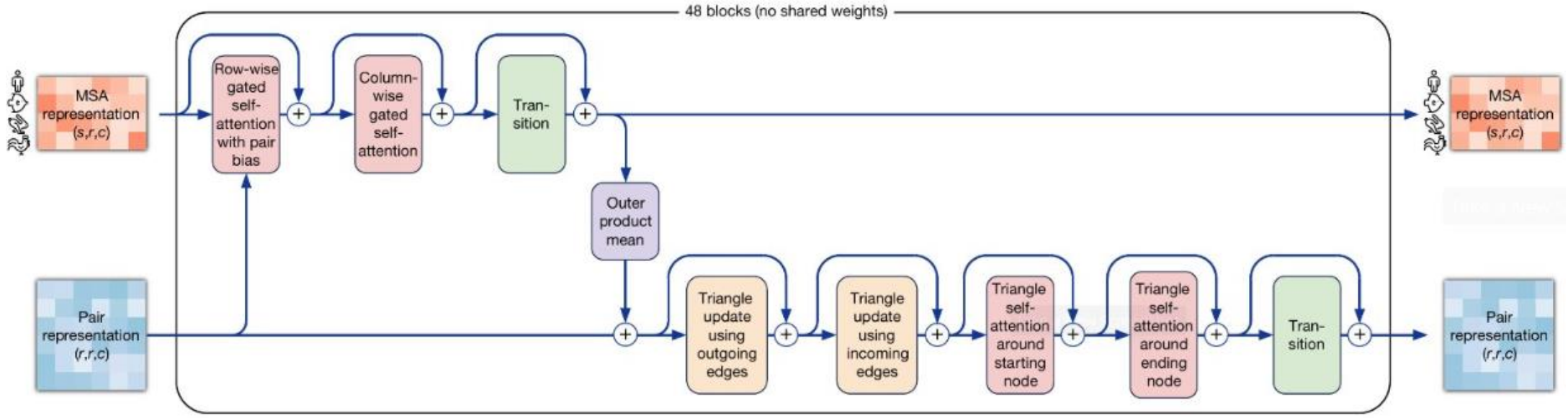
- Row-wise and Column-wise Attention ($\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$)
- Feed Forward Neural Network ($\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2$)



Evoformer (Evolutionary Transformer ?)

MSA Representation Transformer

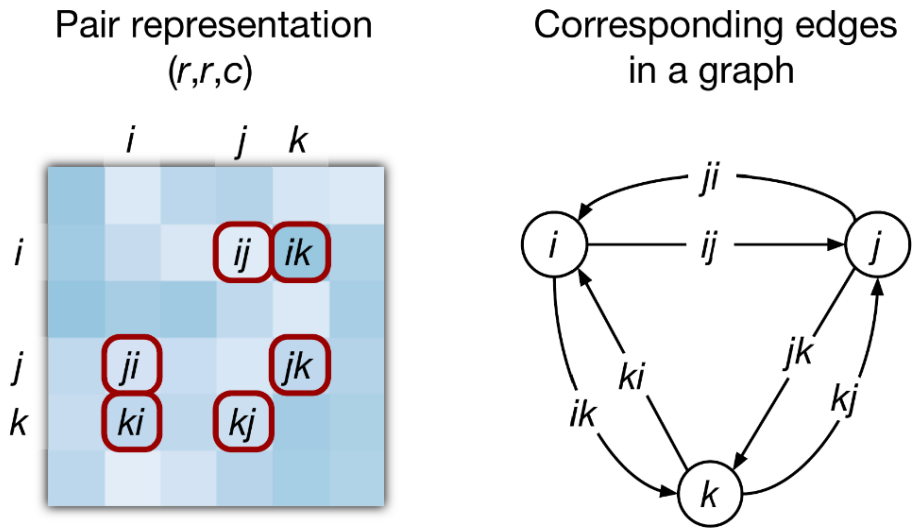
- Row-wise and Column-wise Attention ($\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$)
- Feed Forward Neural Network ($\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2$)

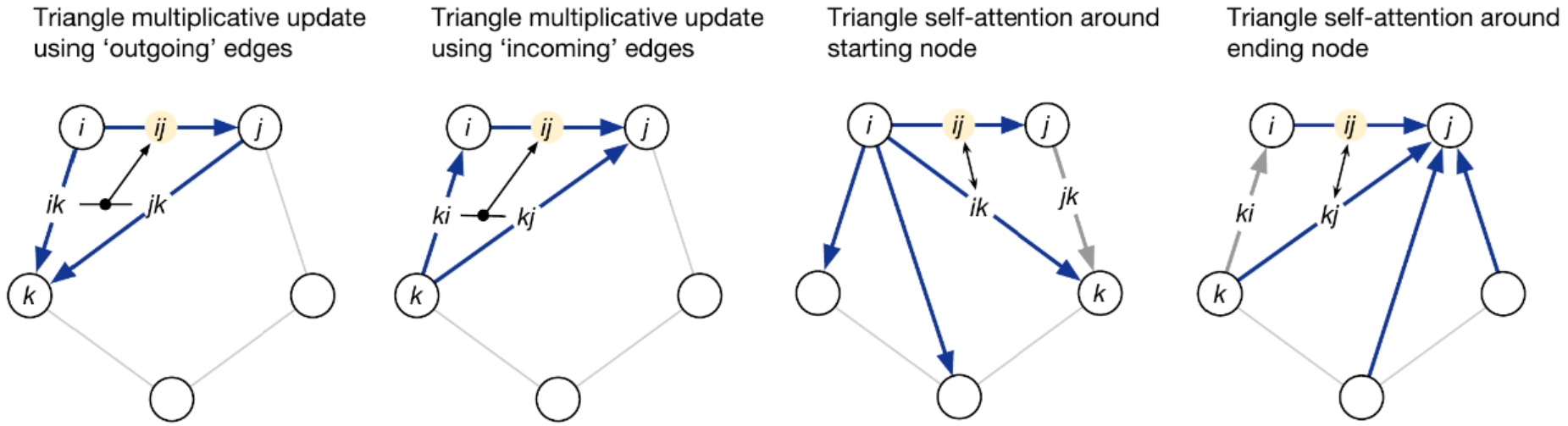


Evoformer (Evolutionary Transformer ?)

Pair Representation Transformer

- Triangle Updates
- Self-Attention
- Feed Forward Neural Network



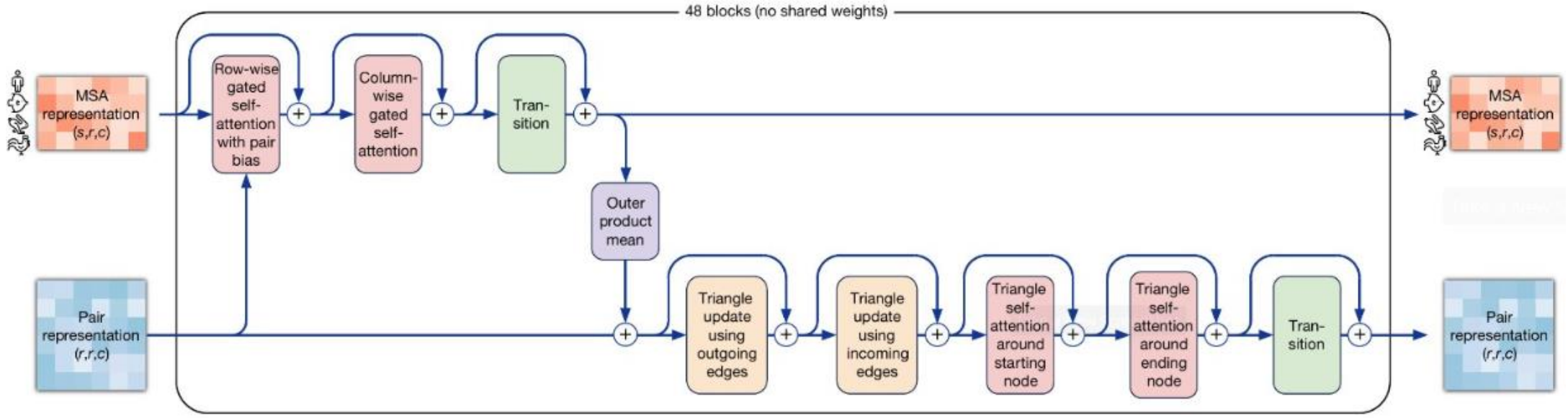


John Jumper et al. 15 July 2021. https://www.nature.com/articles/s41586-021-03819-2

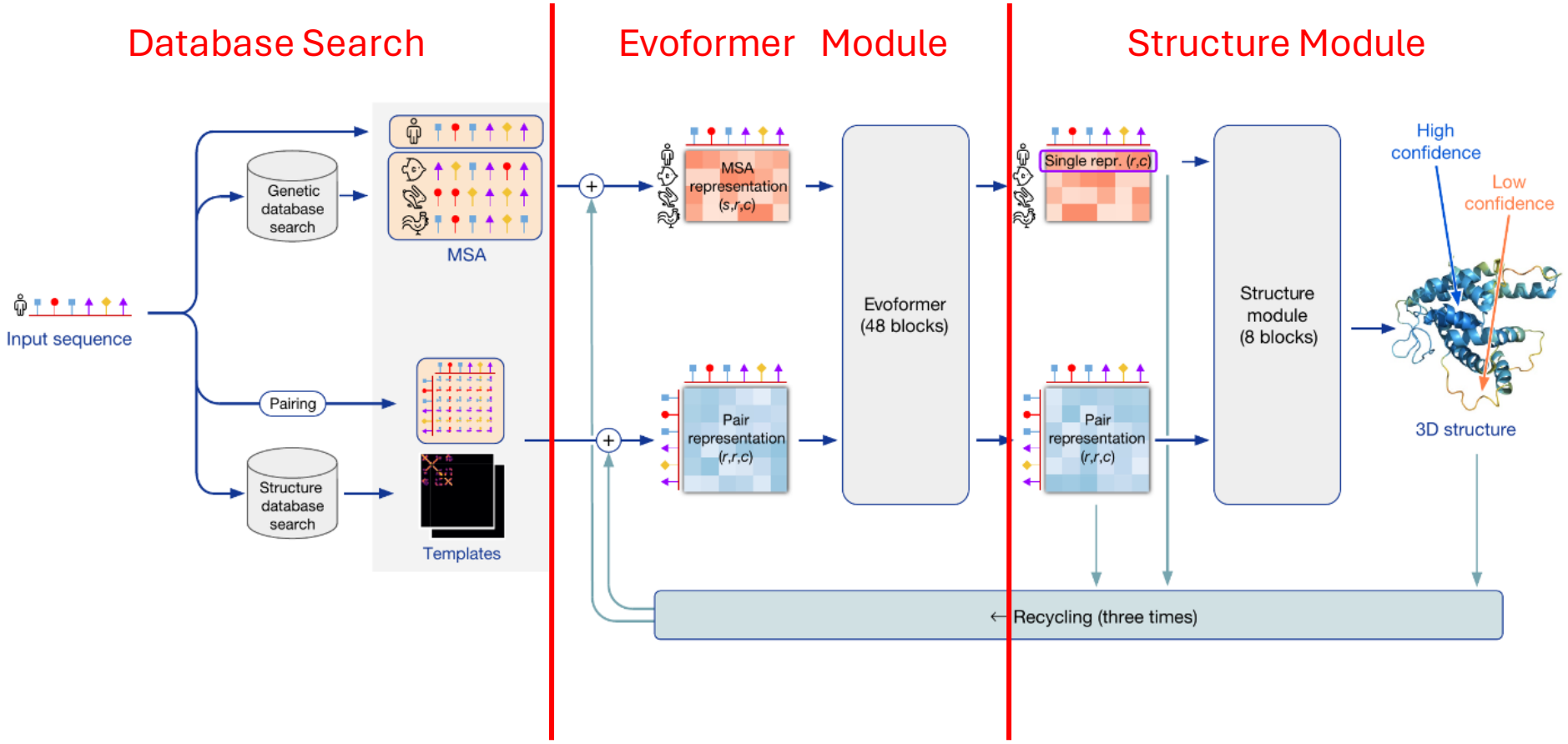
Evoformer (Evolutionary Transformer ?)

MSA Representation Transformer

- Row-wise and Column-wise Attention ($\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$)
- Feed Forward Neural Network ($\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2$)



AlphaFold Structure



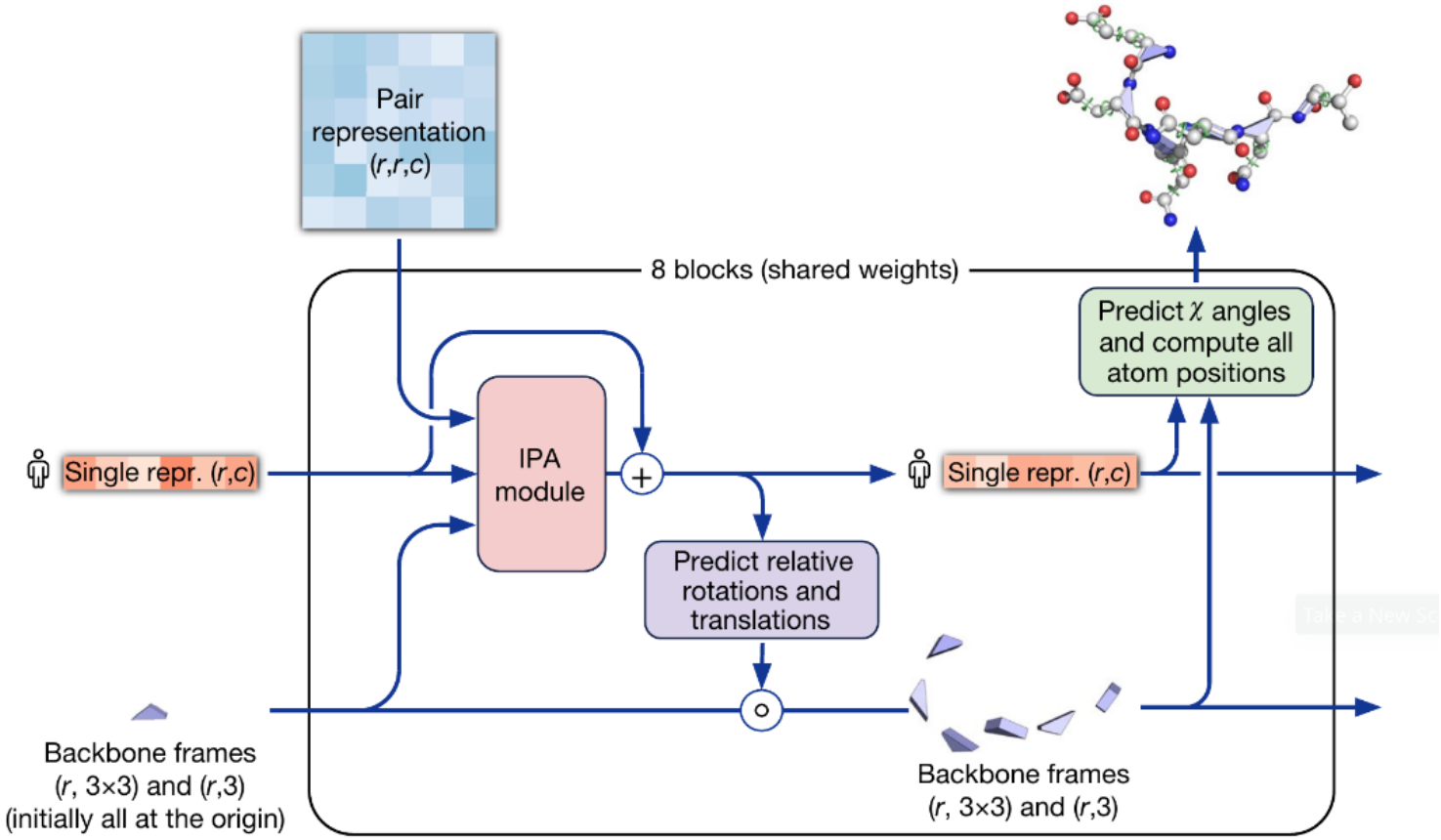
Structure Module

Input

- Single Representation
- Pair Representation
- Backbone Frames

Output

- 3D Model of the Protein
- Cartesian Coordinates in Protein Data Bank (PDB) Format



John Jumper et al. 15 July 2021. https://www.nature.com/articles/s41586-021-03819-2

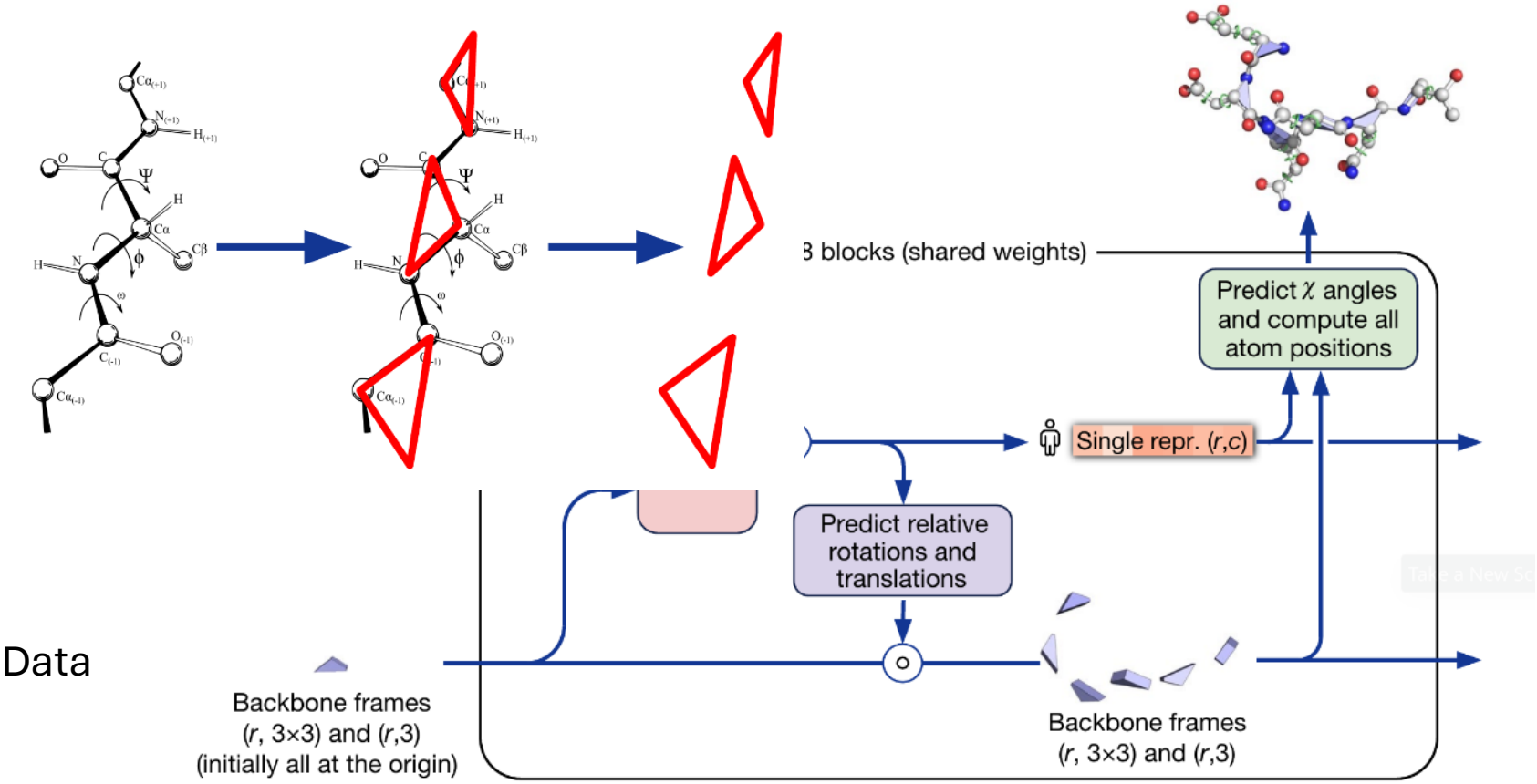
Structure Module

Input

- Single Representation
- Pair Representation
- Backbone Frames

Output

- 3D Model of the Protein
- Cartesian Coordinates in Protein Data Bank (PDB) Format



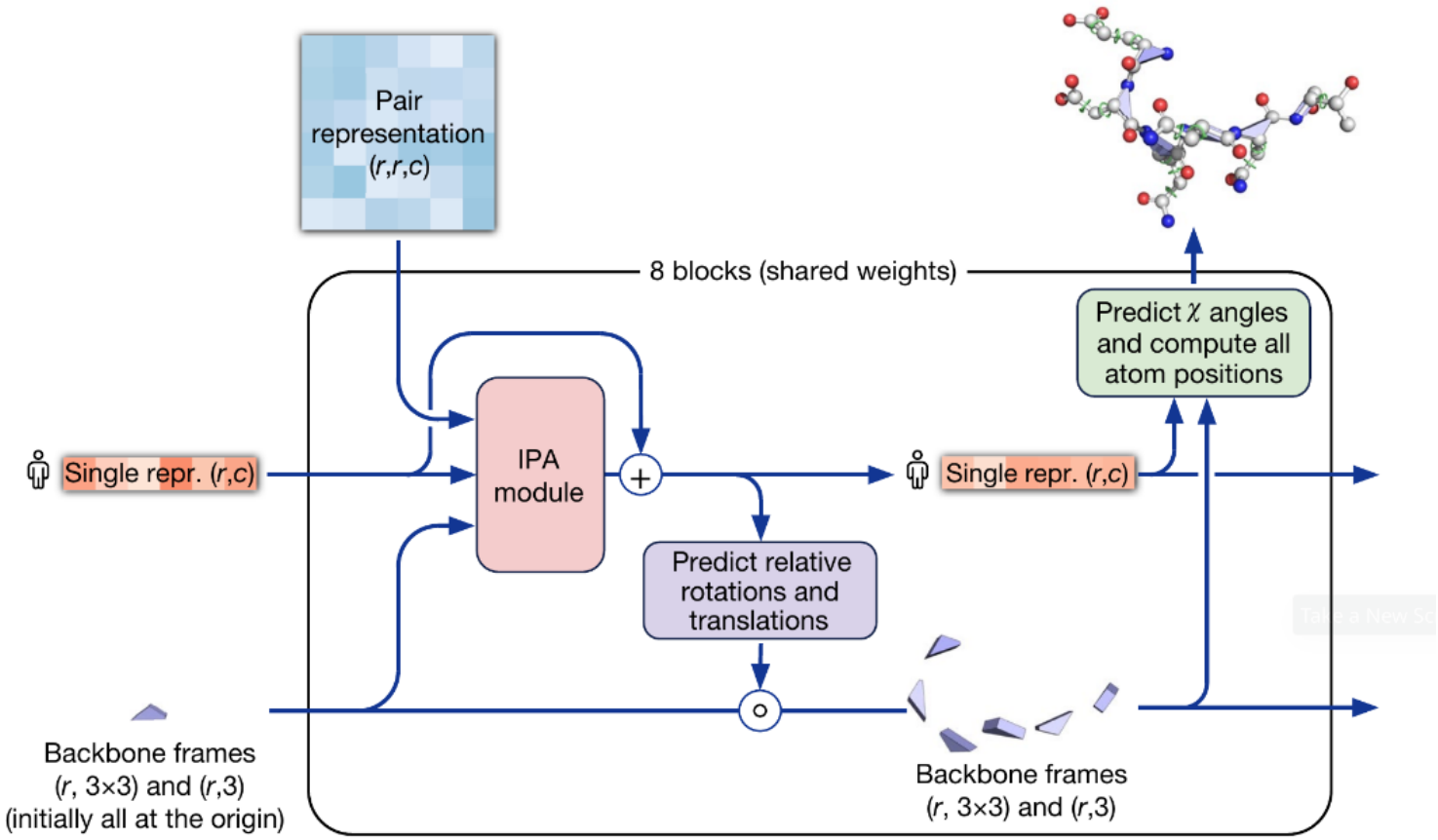
Structure Module

Input

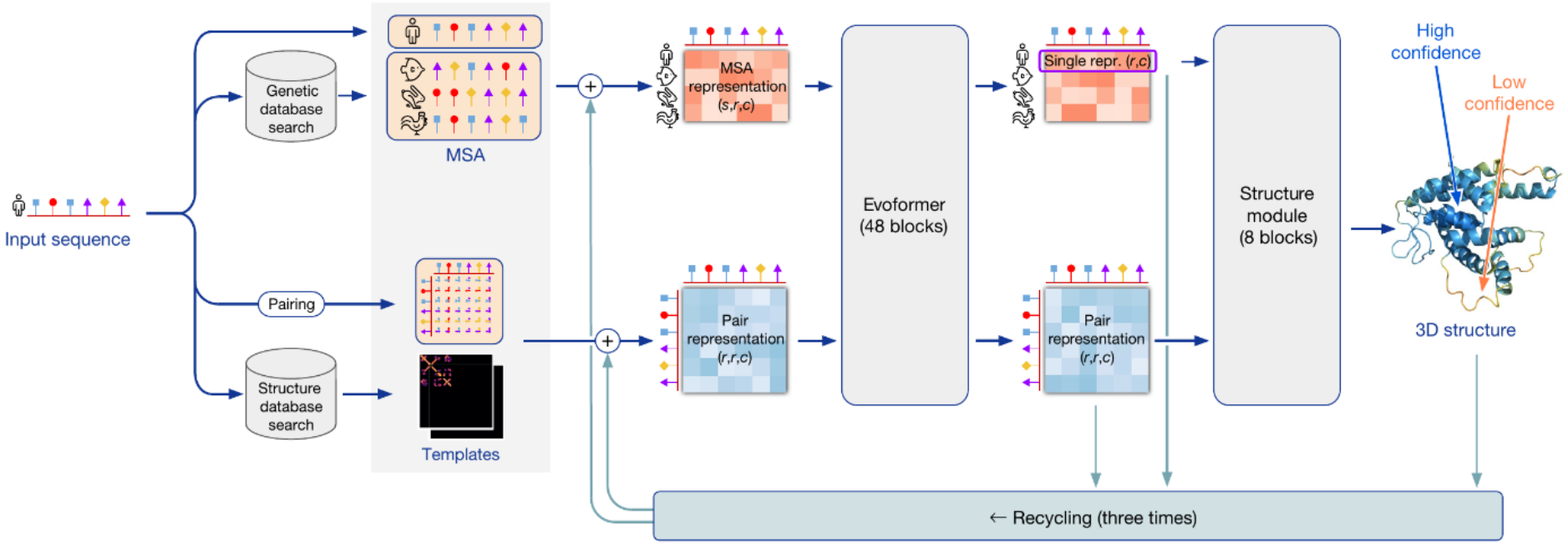
- Single Representation
- Pair Representation
- Backbone Frames

Output

- 3D Model of the Protein
- Cartesian Coordinates in Protein Data Bank (PDB) Format



AlphaFold Structure



John Jumper et al. 15 July 2021. https://www.nature.com/articles/s41586-021-03819-2

Machine Learning Model Overview

Model Representation

Model = $W \cdot X^T + b$

- W: Weight matrix
- X : Input features
- B : Bias term

Training Process

Initialization : randomly initialize W and b
Repeat until Convergence:

- Calculate Gradient:
  - dJ/dW and dJ/db
- Update Parameters:
  - $W = W - \alpha \cdot dJ/dW$
  - $B = b - \alpha \cdot dJ/db$

α (alpha) : learning rate

Cost Function

$J(W, b) = 1/2m \sum (\text{predicted} - \text{actual})^2$

- m: number of training examples

Objective: minimize the cost function J(W,b) by updating parameters W and b iteratively

AlphaFold Cost Function

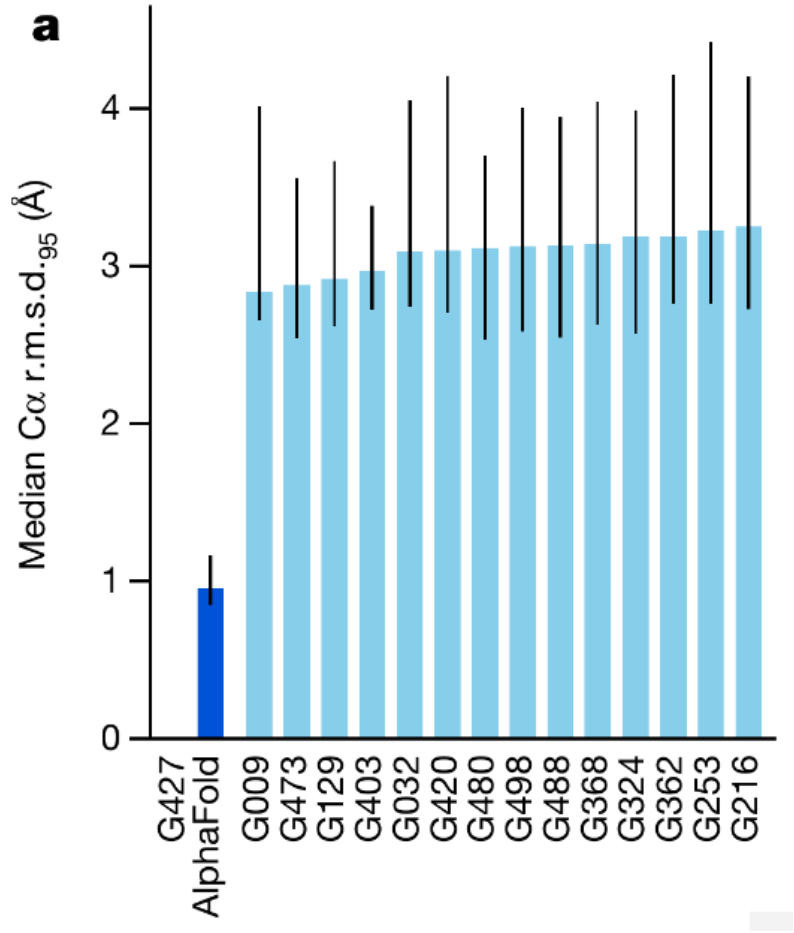
$$\mathcal{L} = \begin{cases} 0.5\mathcal{L}_{\text{FAPE}} + 0.5\mathcal{L}_{\text{aux}} + 0.3\mathcal{L}_{\text{dist}} + 2.0\mathcal{L}_{\text{msa}} + 0.01\mathcal{L}_{\text{conf}} & \text{training} \\ 0.5\mathcal{L}_{\text{FAPE}} + 0.5\mathcal{L}_{\text{aux}} + 0.3\mathcal{L}_{\text{dist}} + 2.0\mathcal{L}_{\text{msa}}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

John Jumper et al. 15 July 2021. https://www.nature.com/articles/s41586-021-03819-2

L (FAPE) : Frame Alined Point Error (FAPE) loss
L (aux) : auxiliary loss from the Structure Module
L (dist) : averaged cross-entropy loss for distogram prediction
L (msa) : averaged cross-entropy loss for masked MSA prediction
L (conf) : model confidence loss
L (exp) : experimentally resolved loss
L (viol) : violation loss

Achievements

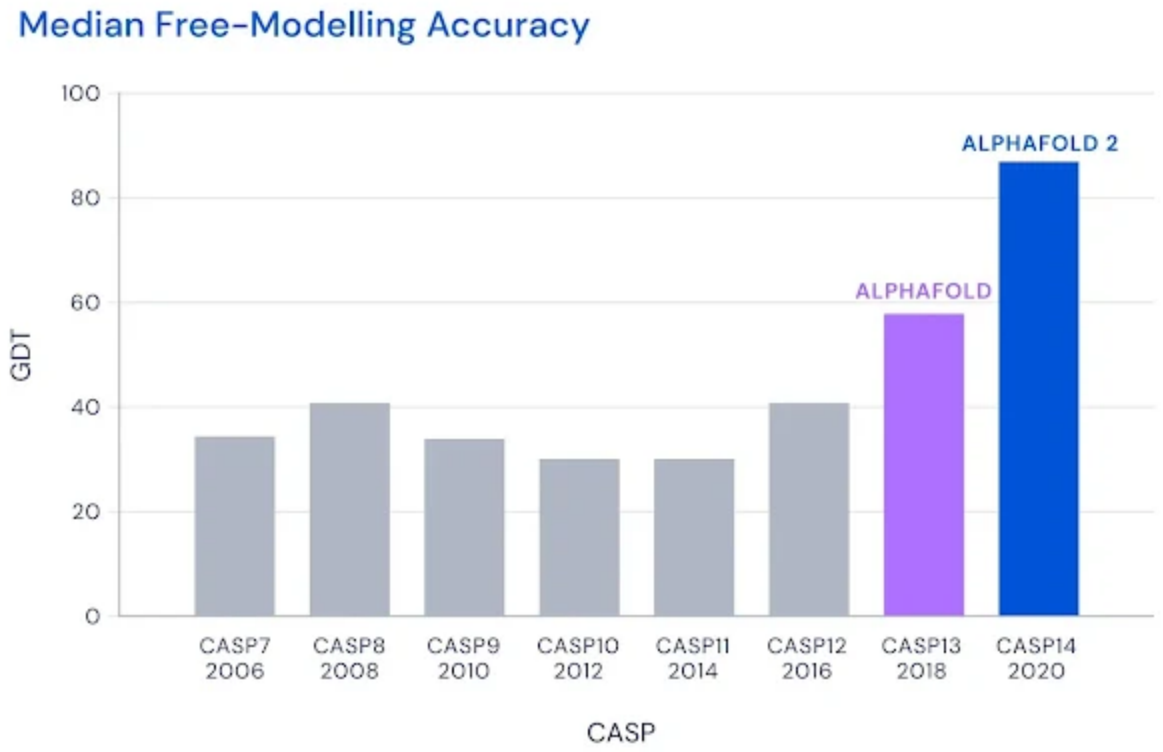
- First Place in the 14th Critical Assessment of Protein Structure Prediction (CASP14)
- Median Backbone Accuracy: 0.96 Å
- All-Atom Accuracy of AlphaFold: 1.5 Å
- Predicted 98.5% of the Human Proteome
- Database with 214 Million Protein Structure Predictions



John Jumper et al. 15 July 2021. https://www.nature.com/articles/s41586-021-03819-2

Achievements

- First Place in the 14th Critical Assessment of Protein Structure Prediction (CASP14)
- Median Backbone Accuracy: 0.96 Å
- All-Atom Accuracy of AlphaFold: 1.5 Å
- Predicted 98.5% of the Human Proteome
- Database with 214 Million Protein Structure Predictions



References

[1] Jumper, J., Evans, R., Pritzel, A. et al. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589 (2021)

[2] Tunyasuvunakool, K., Adler, J., Wu, Z. et al. Highly accurate protein structure prediction for the human proteome. Nature 596, 590–596 (2021).

[3] Senior, A.W., Evans, R., Jumper, J. et al. Improved protein structure prediction using potentials from deep learning. Nature 577, 706–710 (2020).

[4] Big Fantastic Database. Accessed 14 January 2024.

[5] Wang, S., Zhou, W. & Jiang, C. A survey of word embeddings based on deep learning. Computing 102, 717–740 (2020). https://doi.org/10.1007/s00607-019-00768-7.

[6] Chunyan Xu, Zhen Cui, Xiaobin Hong, Tong Zhang, Jian Yang, Wei Liu. Graph Inference Learning for Semi-supervised Classification.

[7] Protein Data Bank: the single global archive for 3D macromolecular structure data. Nucleic acids research 47, no. D1 (2019): D520-D528.

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin. Attention Is All You Need.

[9] Richard E. Turner. An Introduction to Transformers.

[10] G. Bebis and M. Georgiopoulos, Feed-forward neural networks, in IEEE Potentials, vol. 13, no. 4, pp. 27-31, Oct.-Nov. 1994, doi: 10.1109/45.329294.

Thank you for your attention