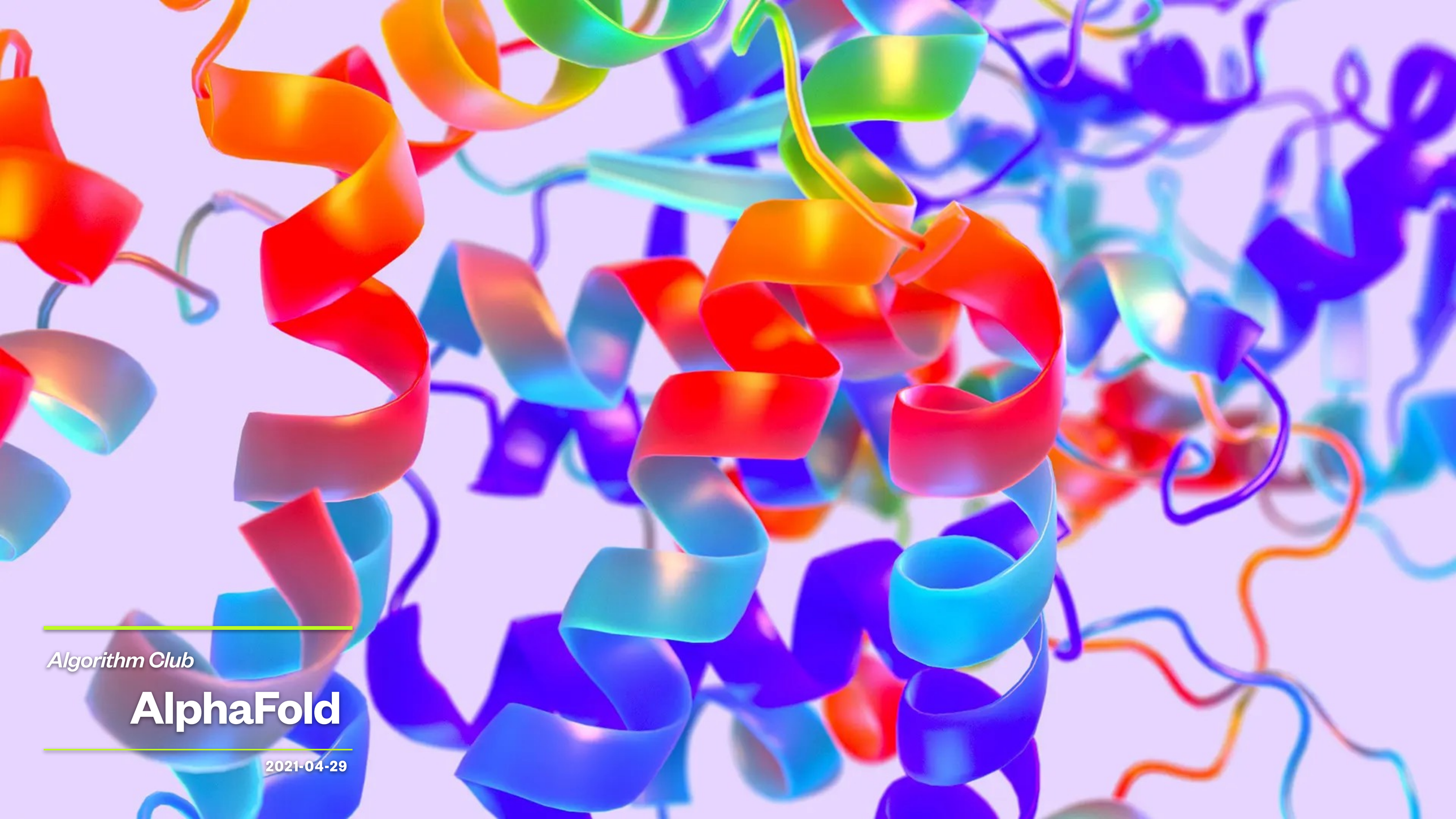
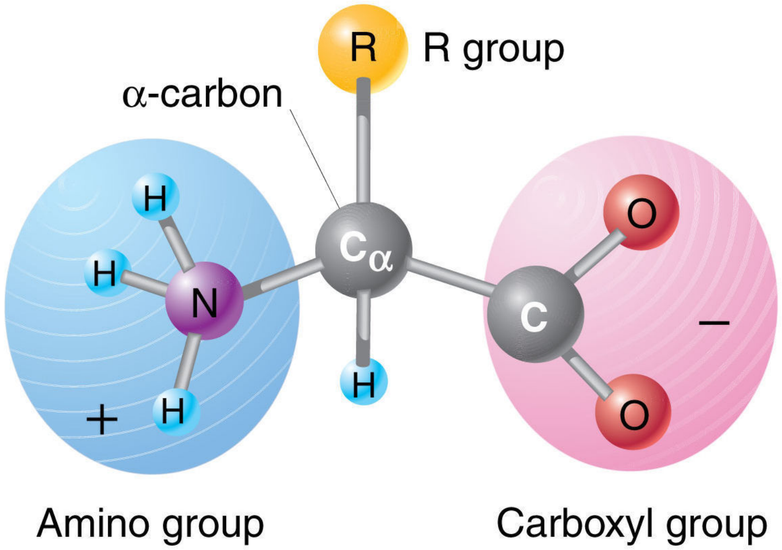


Algorithm Club

AlphaFold

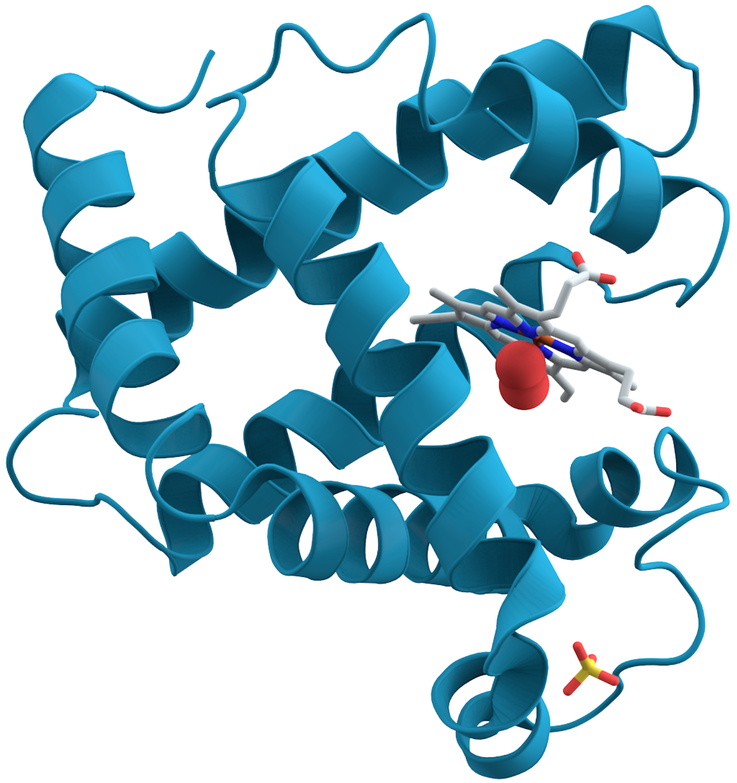
2021-04-29

Protein Folding Problem



Amino Acid

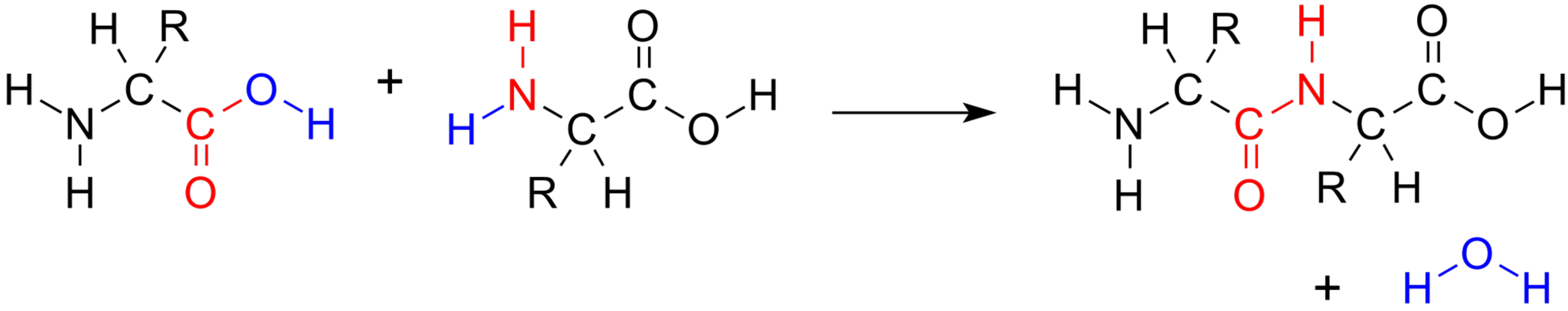
Building Units of Proteins



Protein

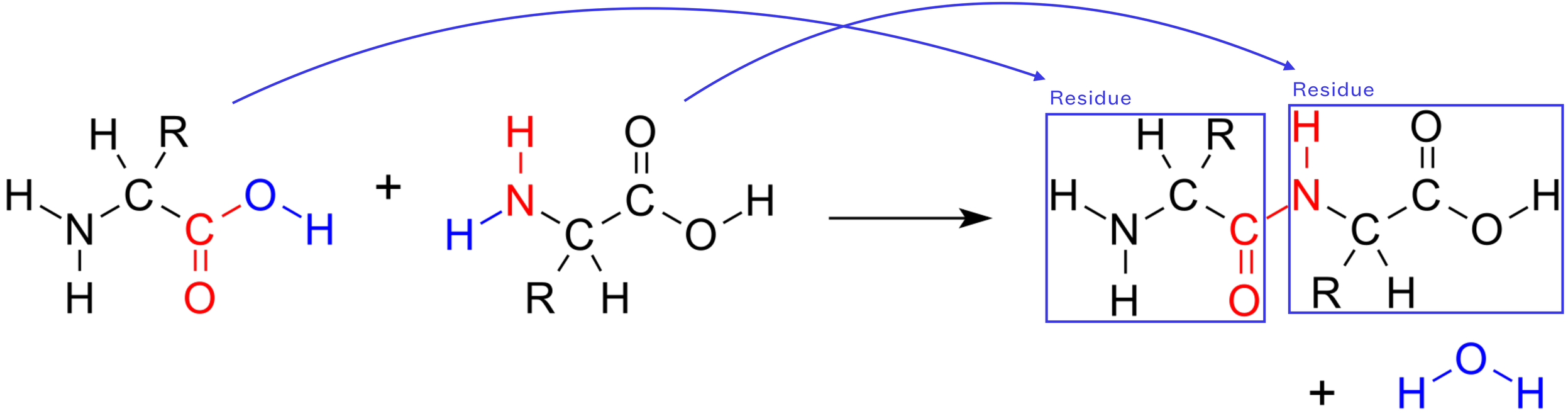
Comprised of a Chain of Amino Acid Residues

Protein Folding Problem



Condensation Reaction of Amino Acids

Protein Folding Problem



Condensation Reaction of Amino Acids

Protein Folding Problem

Christian Anfinsen
1972 Nobel Laureate in Chemistry

Hypothesis

“A protein’s amino acid sequence should fully determine its structure.”

Protein Folding Problem

Christian Anfinsen
1972 Nobel Laureate in Chemistry

Hypothesis

"A protein's amino acid sequence should fully determine its structure."

Learning Task

To computationally predict a protein's 3D structure based solely on its 1D amino acid sequence

AlphaFold - Performance

Article

Improved protein structure prediction using potentials from deep learning

https://doi.org/10.1038/s41586-019-1923-7
Received: 2 April 2019
Accepted: 10 December 2019
Published online: 15 January 2020

Andrew W. Senior[1,4*], Richard Evans[1,4], John Jumper[1,4], James Kirkpatrick[1,4], Laurent Sifre[1,4], Tim Green[1], Chongli Qin[1], Augustin Žídek[1], Alexander W. R. Nelson[1], Alex Bridgland[1], Hugo Penedones[1], Stig Petersen[1], Karen Simonyan[1], Steve Crossan[1], Pushmeet Kohli[1], David T. Jones[2,3], David Silver[1], Koray Kavukcuoglu[1] & Demis Hassabis[1]

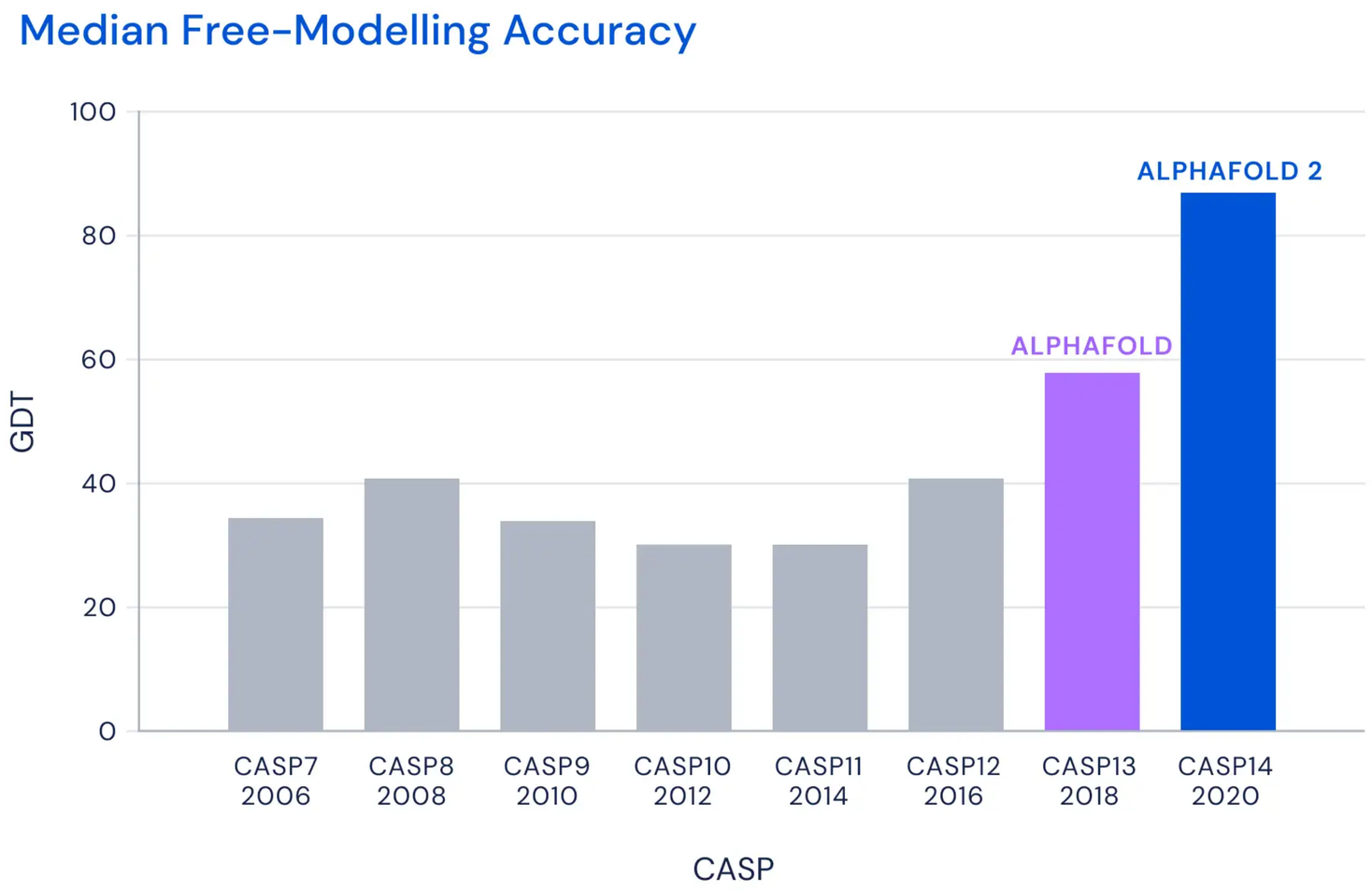
Protein structure prediction can be used to determine the three-dimensional shape of a protein from its amino acid sequence[1]. This problem is of fundamental importance as the structure of a protein largely determines its function[2]; however, protein structures can be difficult to determine experimentally. Considerable progress has recently been made by leveraging genetic information. It is possible to infer which amino acid residues are in contact by analysing covariation in homologous sequences, which aids in the prediction of protein structures[3]. Here we show that we can train a neural network to make accurate predictions of the distances between pairs of residues, which convey more information about the structure than contact predictions. Using this information, we construct a potential of mean force[4] that can accurately describe the shape of a protein. We find that the resulting potential can be optimized by a simple gradient descent algorithm to generate structures without complex sampling procedures. The resulting system, named AlphaFold, achieves high accuracy, even for sequences with fewer homologous sequences. In the recent Critical Assessment of Protein Structure Prediction[5] (CASP13)—a blind assessment of the state of the field—AlphaFold created high-accuracy structures (with template modelling (TM) scores[6] of 0.7 or higher) for 24 out of 43 free modelling domains, whereas the next best method, which used sampling and contact information, achieved such accuracy for only 14 out of 43 domains. AlphaFold represents a considerable advance in protein-structure prediction. We expect this increased accuracy to enable insights into the function and malfunction of proteins, especially in cases for which no structures for homologous proteins have been experimentally determined[7].

Proteins are at the core of most biological processes. As the function of a protein is dependent on its structure, understanding protein structures has been a grand challenge in biology for decades. Although several experimental structure determination techniques have been developed and improved in accuracy, they remain difficult and time-consuming[2]. As a result, decades of theoretical work has attempted to predict protein structures from amino acid sequences.

CASP[5] is a biennial blind protein structure prediction assessment run by the structure prediction community to benchmark progress in accuracy. In 2018, AlphaFold joined 97 groups from around the world in entering CASP13[5]. Each group submitted up to 5 structure predictions for each of 84 protein sequences for which experimentally determined structures were sequestered. Assessors divided the proteins into 104 domains for scoring and classified each as being amenable to template-based modelling (TBM, in which a protein with a similar sequence has a known structure, and that homologous structure is modified in accordance with the sequence differences) or requiring free modelling (FM, in cases in which no homologous structure is available), with an intermediate (FM/TBM) category. Figure 1a shows that AlphaFold predicts more FM domains with high accuracy than any other system, particularly in the 0.6–0.7 TM-score range. The TM score—ranging between 0 and 1—measures the degree of match of the overall (backbone) shape of a proposed structure to a native structure. The assessors ranked the 98 participating groups by the summed, capped z-scores of the structures, separated according to category. AlphaFold achieved a summed z-score of 52.8 in the FM category (best-of-five) compared with 36.6 for the next closest group (322). Combining FM and TBM/FM categories, AlphaFold scored 68.3 compared with 48.2. AlphaFold is able to predict previously unknown folds to high accuracy (Fig. 1b). Despite using only FM techniques and not using templates, AlphaFold also scored well in the TBM category according to the assessors' formula 0-capped z-score, ranking fourth for the top-one model or first for the best-of-five models. Much of the accuracy of AlphaFold is due to the accuracy of the distance predictions, which is evident from the high precision of the corresponding contact predictions (Fig. 1c and Extended Data Fig. 2a).

[1]DeepMind, London, UK. [2]The Francis Crick Institute, London, UK. [3]University College London, London, UK. [4]These authors contributed equally: Andrew W. Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre. *e-mail: andrewsenior@google.com

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



AlphaFold - Performance

Article

Improved protein structure prediction using potentials from deep learning

https://doi.org/10.1038/s41586-019-1923-7
Received: 2 April 2019
Accepted: 10 December 2019
Published online: 15 January 2020

Andrew W. Senior[1,4], Richard Evans[1,4], John Jumper[1,4], James Kirkpatrick[1,4], Laurent Sifre[1,4], Tim Green[1], Chongli Qin[1], Augustin Žídek[1], Alexander W. R. Nelson[1], Alex Bridgland[1], Hugo Penedones[1], Stig Petersen[1], Karen Simonyan[1], Steve Crossan[1], Pushmeet Kohli[1], David T. Jones[2,3], David Silver[1], Koray Kavukcuoglu[1] & Demis Hassabis[1]

Protein structure prediction can be used to determine the three-dimensional shape of a protein from its amino acid sequence[1]. This problem is of fundamental importance as the structure of a protein largely determines its function[2]; however, protein structures can be difficult to determine experimentally. Considerable progress has recently been made by leveraging genetic information. It is possible to infer which amino acid residues are in contact by analysing covariation in homologous sequences, which aids in the prediction of protein structures[3]. Here we show that we can train a neural network to make accurate predictions of the distances between pairs of residues, which convey more information about the structure than contact predictions. Using this information, we construct a potential of mean force[4] that can accurately describe the shape of a protein. We find that the resulting potential can be optimized by a simple gradient descent algorithm to generate structures without complex sampling procedures. The resulting system, named AlphaFold, achieves high accuracy, even for sequences with fewer homologous sequences. In the recent Critical Assessment of Protein Structure Prediction[5] (CASP13)—a blind assessment of the state of the field—AlphaFold created high-accuracy structures (with template modelling (TM) scores[6] of 0.7 or higher) for 24 out of 43 free modelling domains, whereas the next best method, which used sampling and contact information, achieved such accuracy for only 14 out of 43 domains. AlphaFold represents a considerable advance in protein-structure prediction. We expect this increased accuracy to enable insights into the function and malfunction of proteins, especially in cases for which no structures for homologous proteins have been experimentally determined[7].

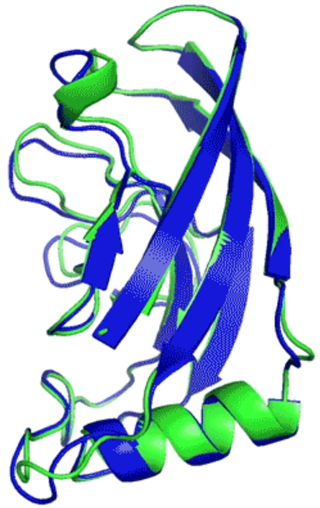
Proteins are at the core of most biological processes. As the function of a protein is dependent on its structure, understanding protein structures has been a grand challenge in biology for decades. Although several experimental structure determination techniques have been developed and improved in accuracy, they remain difficult and time-consuming[2]. As a result, decades of theoretical work has attempted to predict protein structures from amino acid sequences.

CASP[5] is a biennial blind protein structure prediction assessment run by the structure prediction community to benchmark progress in accuracy. In 2018, AlphaFold joined 97 groups from around the world in entering CASP13[5]. Each group submitted up to 5 structure predictions for each of 84 protein sequences for which experimentally determined structures were sequestered. Assessors divided the proteins into 104 domains for scoring and classified each as being amenable to template-based modelling (TBM, in which a protein with a similar sequence has a known structure, and that homologous structure is modified in accordance with the sequence differences) or requiring free modelling (FM, in cases in which no homologous structure is available), with an intermediate (FM/TBM) category. Figure 1a shows that AlphaFold predicts more FM domains with high accuracy than any other system, particularly in the 0.6–0.7 TM-score range. The TM score—ranging between 0 and 1—measures the degree of match of the overall (backbone) shape of a proposed structure to a native structure. The assessors ranked the 98 participating groups by the summed, capped z-scores of the structures, separated according to category. AlphaFold achieved a summed z-score of 52.8 in the FM category (best-of-five) compared with 36.6 for the next closest group (322). Combining FM and TBM/FM categories, AlphaFold scored 68.3 compared with 48.2. AlphaFold is able to predict previously unknown folds to high accuracy (Fig. 1b). Despite using only FM techniques and not using templates, AlphaFold also scored well in the TBM category according to the assessors' formula 0-capped z-score, ranking fourth for the top-one model or first for the best-of-five models. Much of the accuracy of AlphaFold is due to the accuracy of the distance predictions, which is evident from the high precision of the corresponding contact predictions (Fig. 1c and Extended Data Fig. 2a).

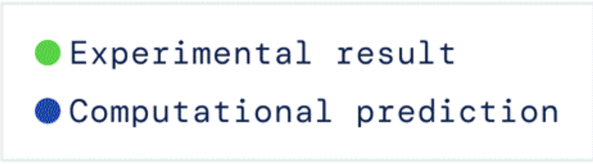
[1]DeepMind, London, UK. [2]The Francis Crick Institute, London, UK. [3]University College London, London, UK. [4]These authors contributed equally: Andrew W. Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre. *e-mail: andrewsenior@google.com

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

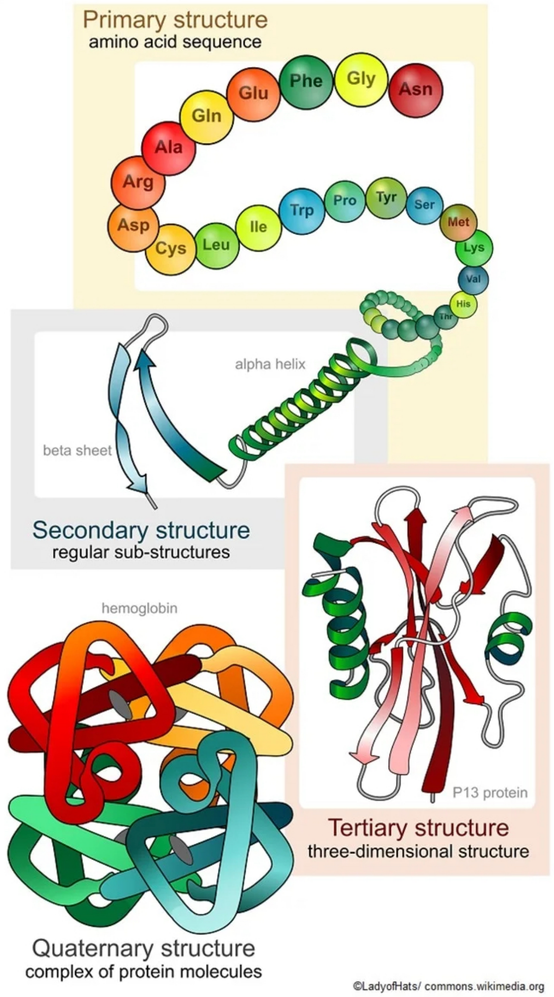
T1037 / 6vr4
90.7 GDT
(RNA polymerase domain)

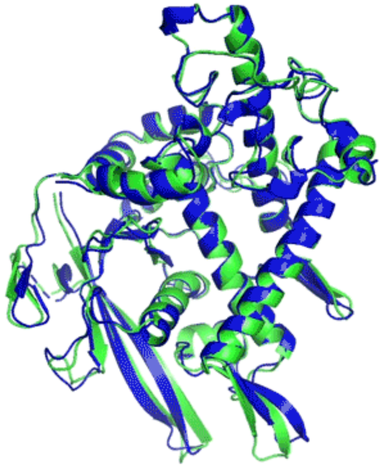


T1049 / 6y4f
93.3 GDT
(adhesin tip)

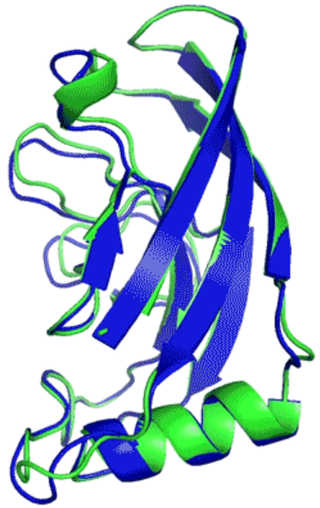


AlphaFold - Performance





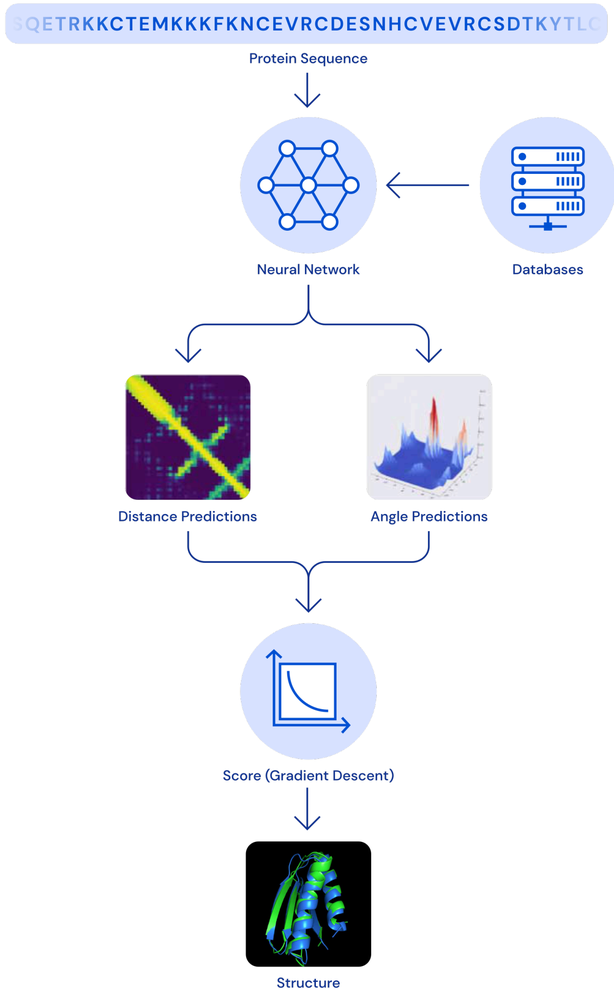
T1037 / 6vr4
90.7 GDT
(RNA polymerase domain)

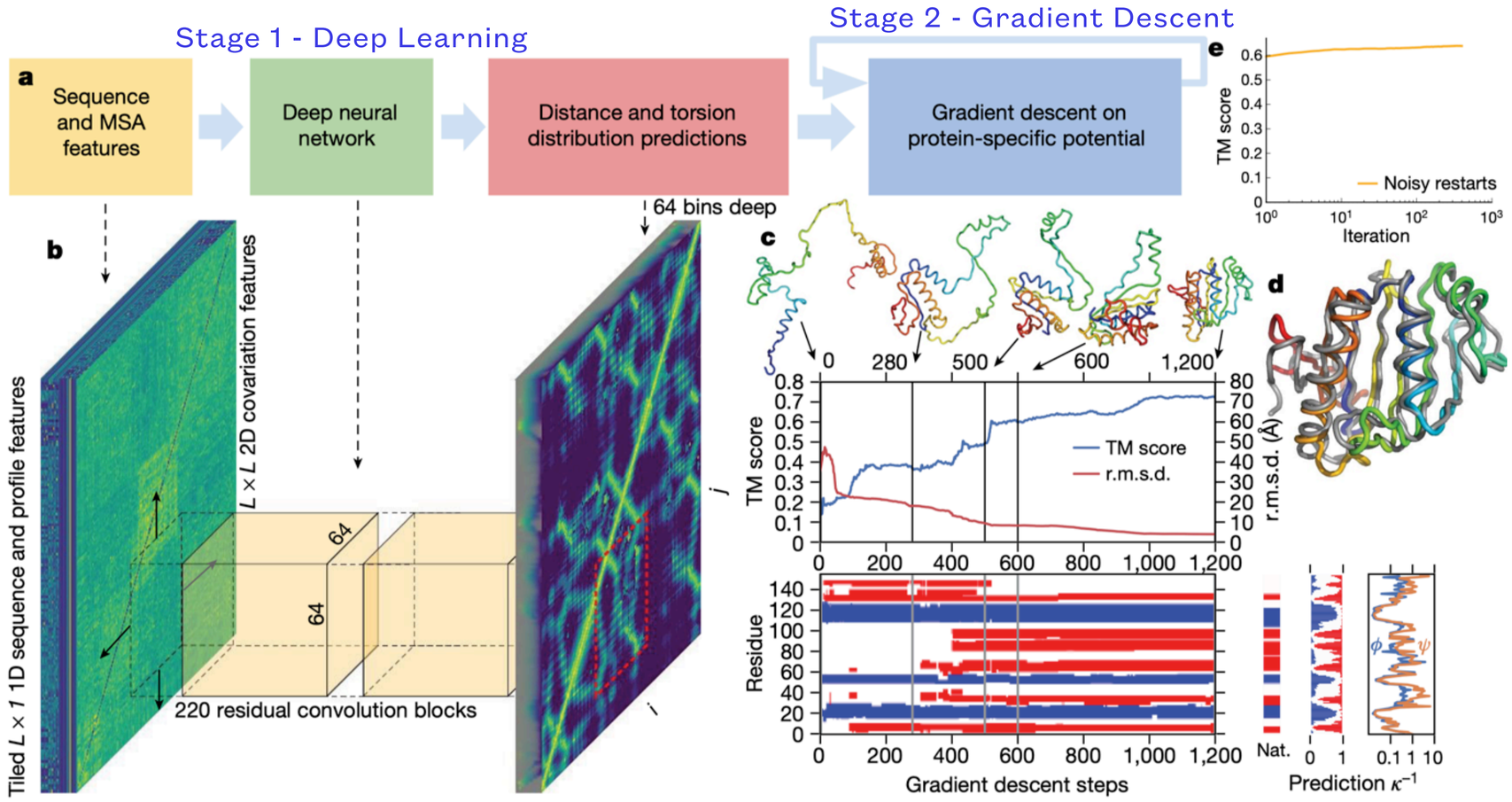


T1049 / 6y4f
93.3 GDT
(adhesin tip)

Experimental result
Computational prediction

AlphaFold - Model Overview

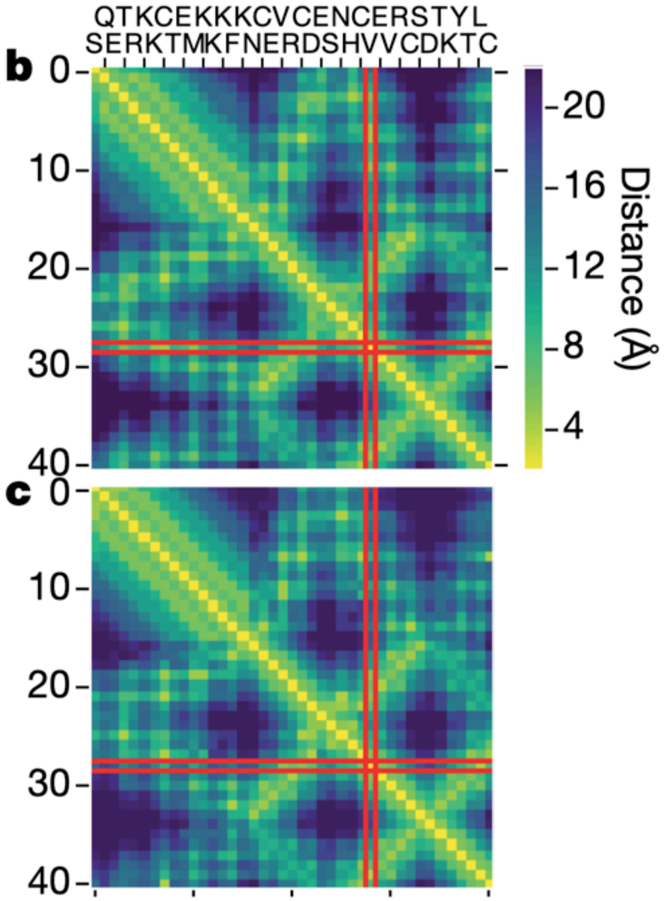




AlphaFold - Model Overview

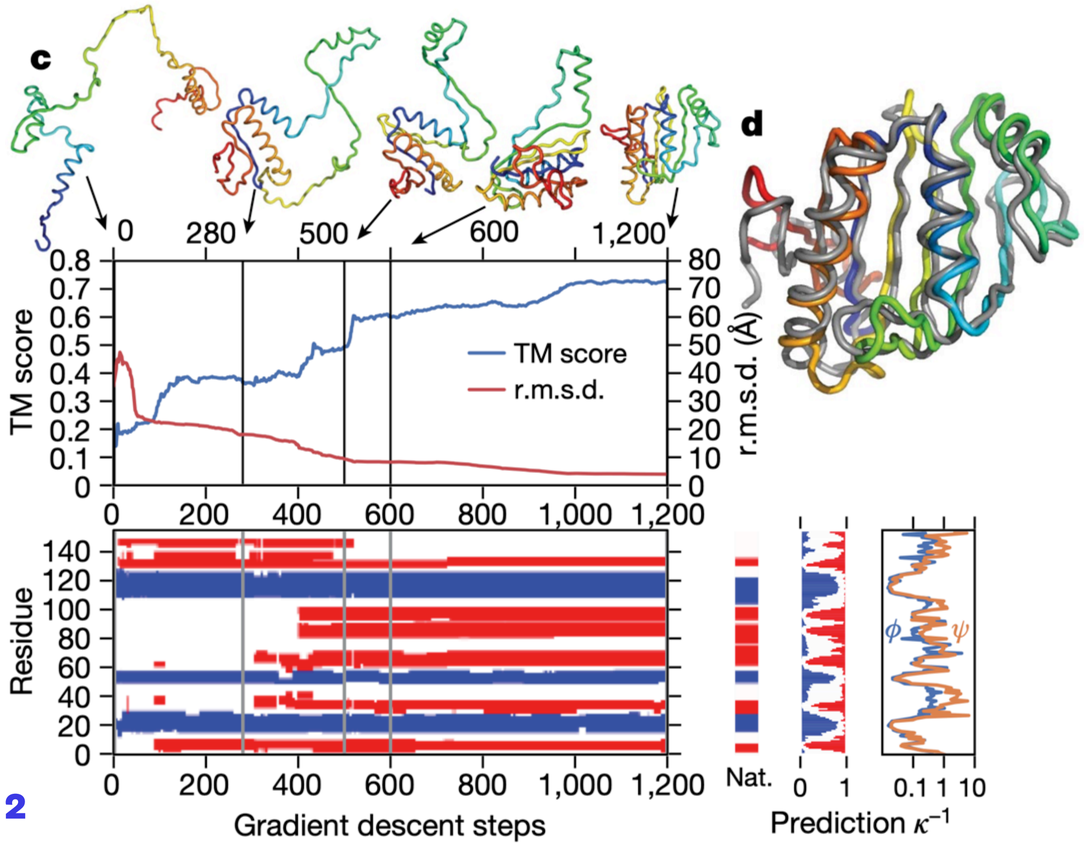
Stage 1

Given a sequence of amino acids, predict pair-wise distances.



b. Groundtruth Distance Matrix

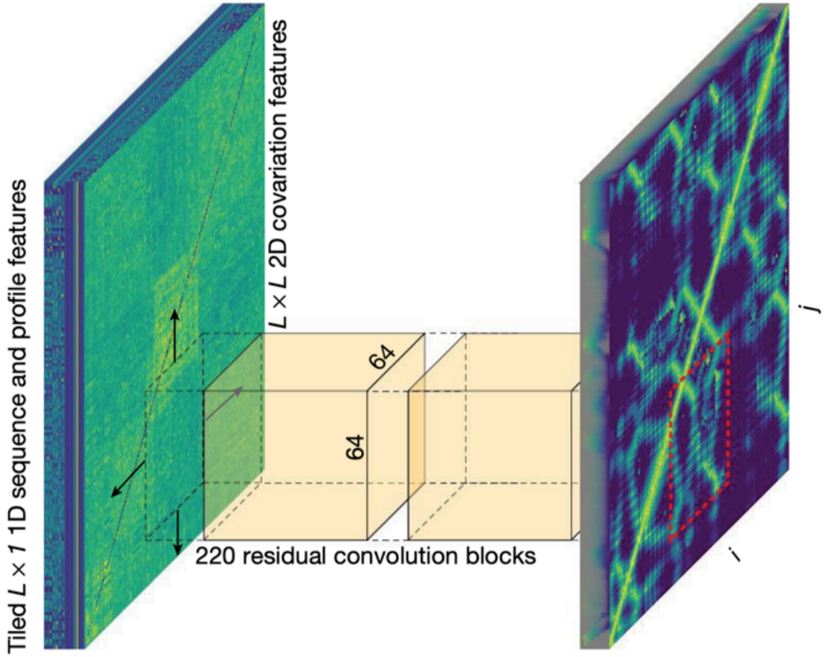
c. Predicted Distance Matrix



Stage 2

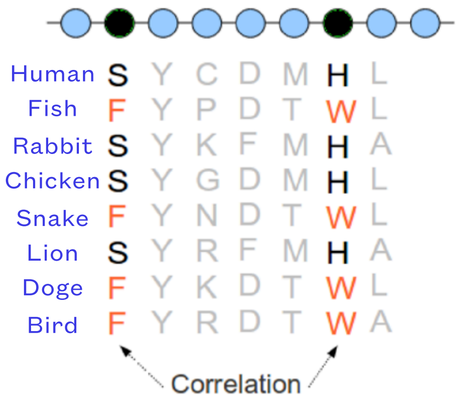
Based on the output of stage 1, build differentiable 3D geometric structures.

AlphaFold - Stage 1



“Image-to-Image Translation” CNN

- Input feature map has the size of sequence length by sequence length;
- Pair-wise or tiled bio-chemical features are stacked together;



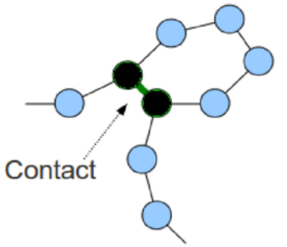


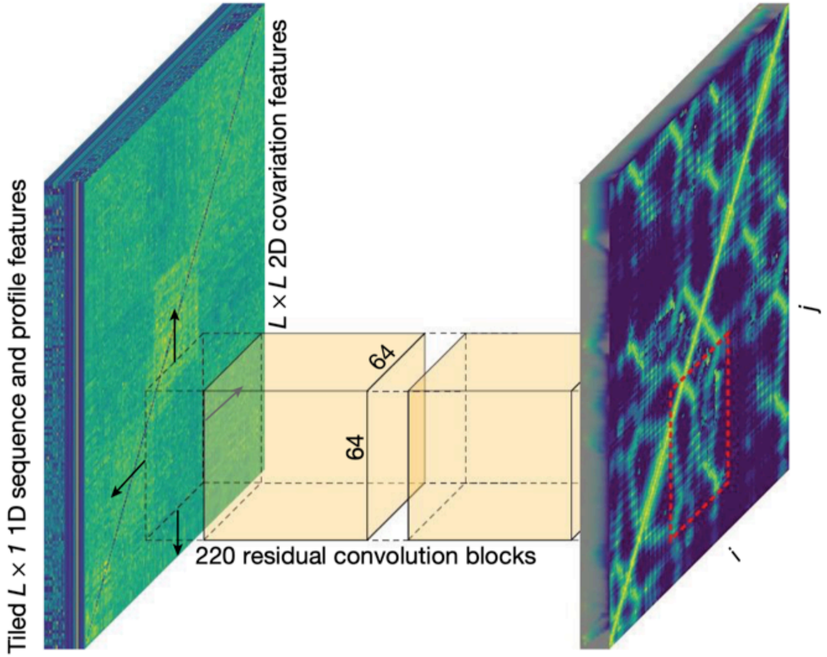
FIG. 1. (Color online) Left panel: small MSA with two positions of correlated amino-acid occupancy. Right panel: hypothetical corresponding spatial conformation, bringing the two correlated positions into direct contact.

Multiple Sequence Alignment

a table of the amino acid sequences of all the protein domains in the family lined up to be as similar as possible.

- different amino acids can still form proteins with very similar structure and functionality;
- if amino acids in these proteins are in contact, there’s explicit correlation between them.

AlphaFold - Stage 1

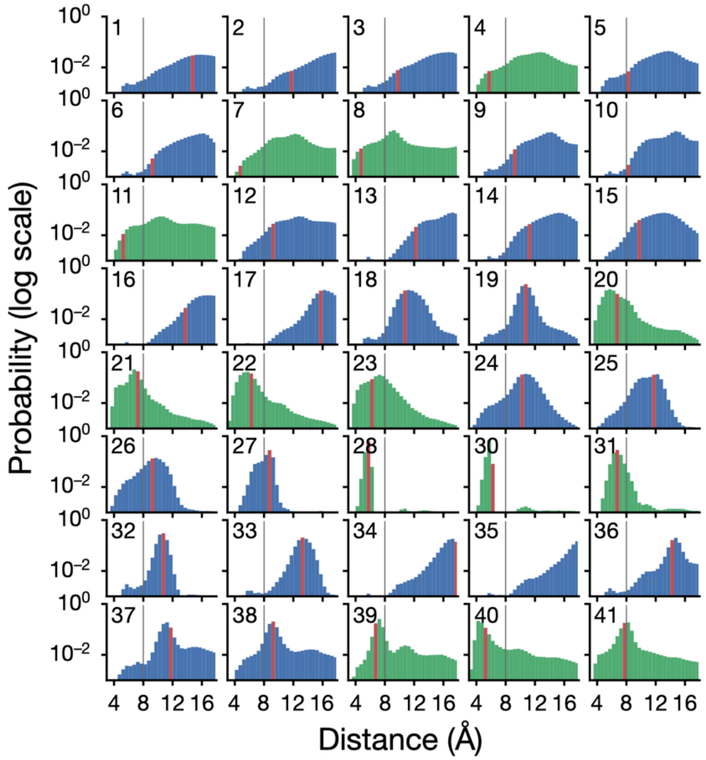


“Image-to-Image Translation” CNN

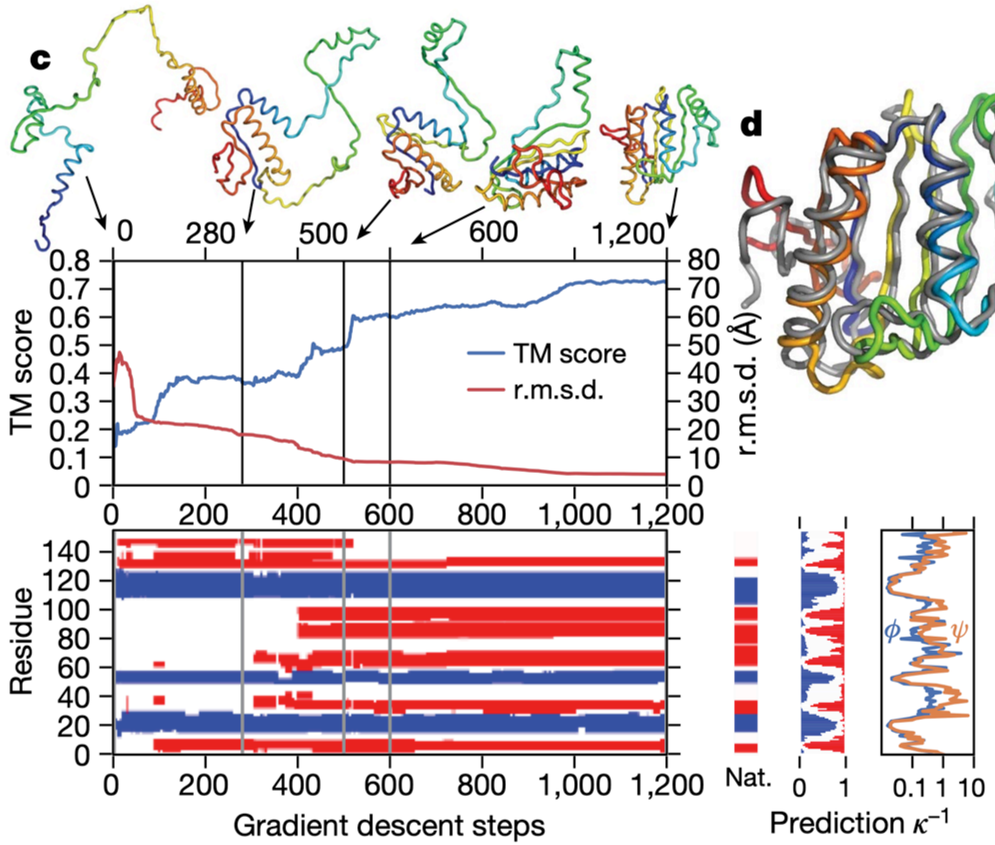
- Input feature map has the size of sequence length by sequence length;
- Pair-wise or tiled bio-chemical features are stacked together;
- Predict pair-wise distance.

Distribution of Distance

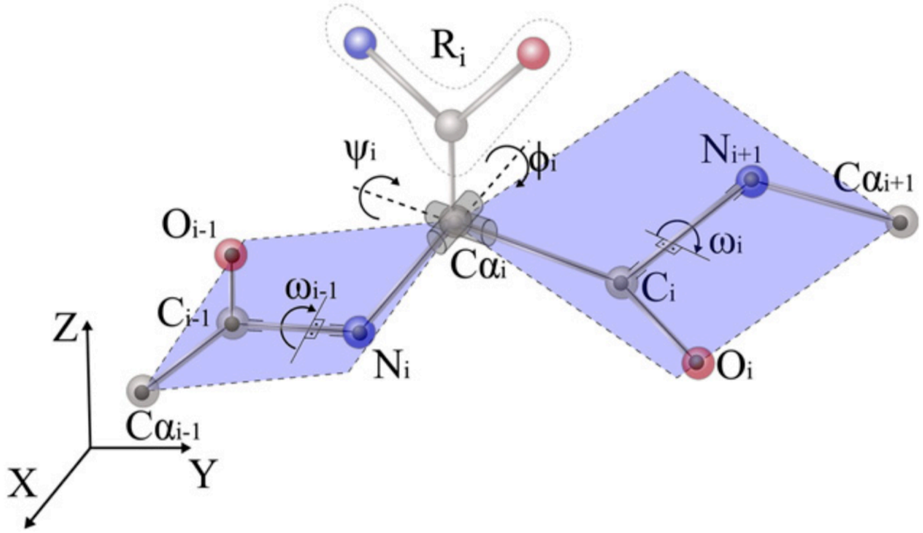
Red: groundtruth distance
Green: exists chemical bond
Blue: no chemical bond



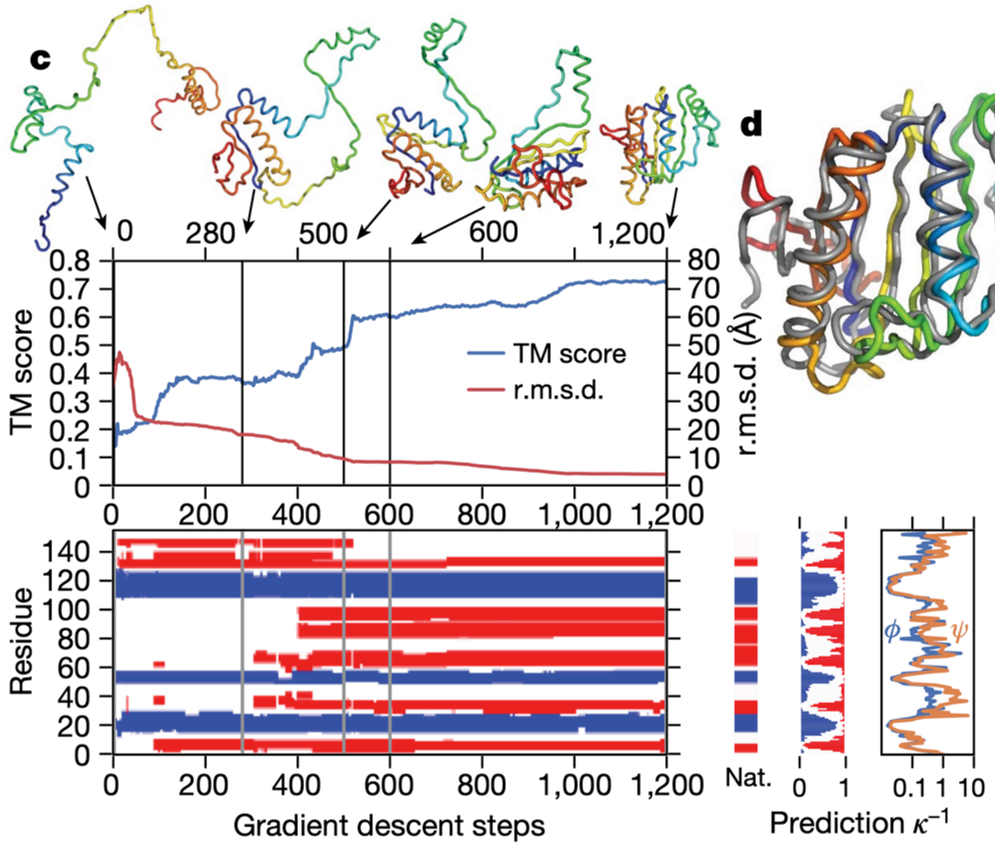
AlphaFold - Stage 2



(see Methods). We parameterized protein structures by the backbone torsion angles (φ, ψ) of all residues and build a differentiable model of protein geometry $\mathbf{x} = G(\varphi, \psi)$ to compute the C_β coordinates, $\mathbf{x}_i$ for all residues i and thus the inter-residue distances, $d_{ij} = ||\mathbf{x}_i - \mathbf{x}_j||$, for each



AlphaFold - Stage 2

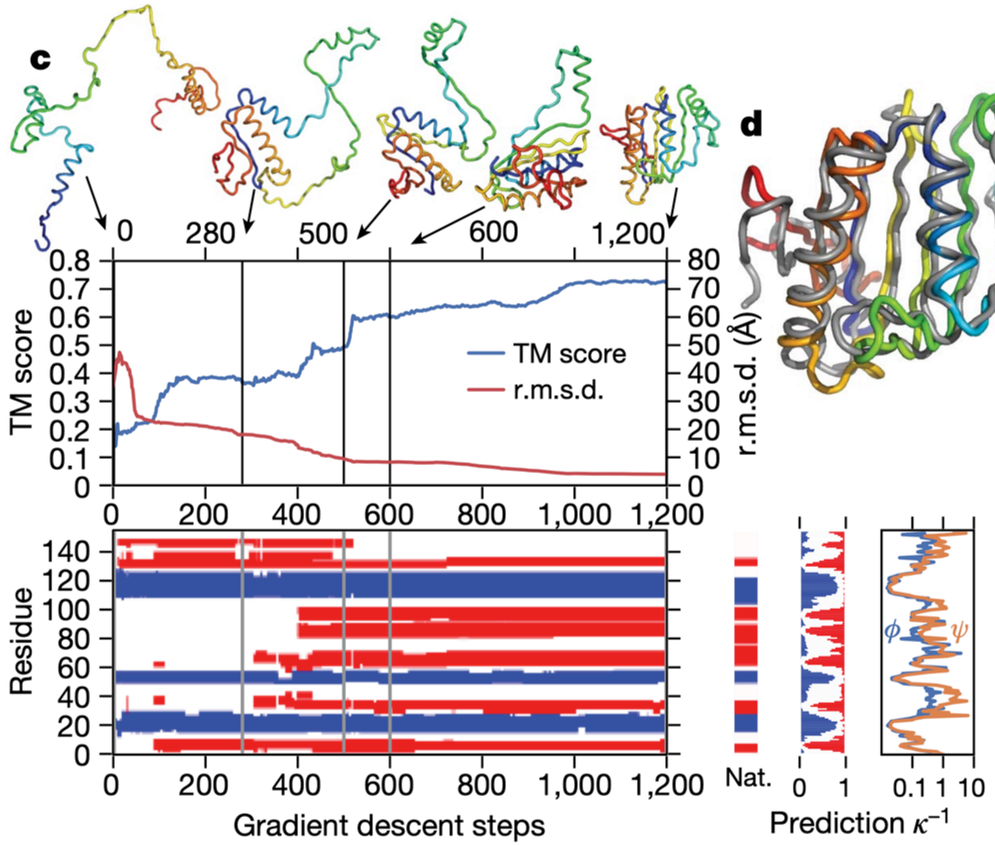


(see Methods). We parameterized protein structures by the backbone torsion angles (φ, ψ) of all residues and build a differentiable model of protein geometry $\mathbf{x} = G(\varphi, \psi)$ to compute the C_β coordinates, $\mathbf{x}_i$ for all residues i and thus the inter-residue distances, $d_{ij} = ||\mathbf{x}_i - \mathbf{x}_j||$, for each

Simplified Version

1. Take predicted distances in the previous stage as input, predict torsion angles;
2. Run gradient decent on L2 loss such that the 3D structure derived from torsion angles fulfill the constraints of predicted pair-wise amino acid distance.

AlphaFold - Stage 2

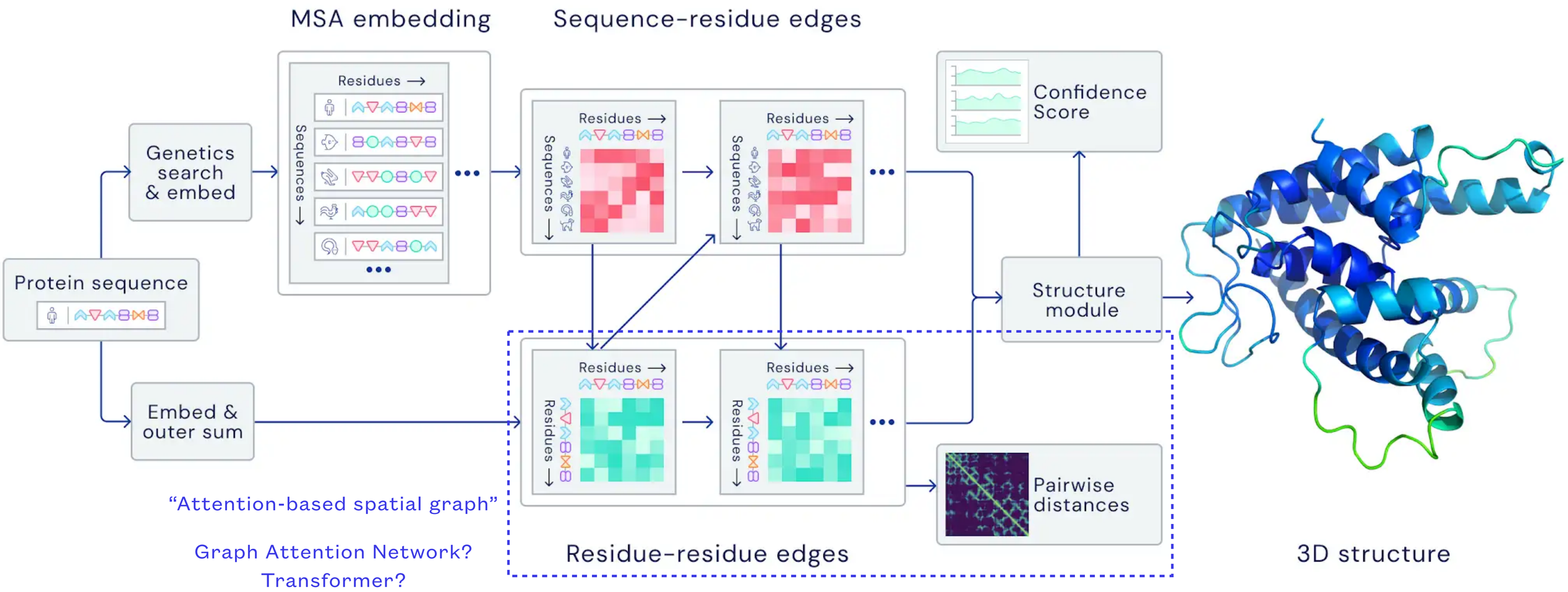


(see Methods). We parameterized protein structures by the backbone torsion angles (φ, ψ) of all residues and build a differentiable model of protein geometry $\mathbf{x} = G(\varphi, \psi)$ to compute the C_β coordinates, $\mathbf{x}_i$ for all residues i and thus the inter-residue distances, $d_{ij} = ||\mathbf{x}_i - \mathbf{x}_j||$, for each

As the training goes on:

- TM score rises, root mean square deviation decreases;
- The predicted secondary protein structure (helix in blue, strand in red) becomes more similar to the native structure, i.e. the groundtruth.

AlphaFold 2



"Attention-based spatial graph"

Graph Attention Network?
Transformer?