Title: AlphaFold

URL: https://en.wikipedia.org/wiki/AlphaFold

PageID: 59766171

Categories: Category:Applied machine learning, Category:Bioinformatics software, Category:Deep learning software applications, Category:Google DeepMind, Category:Molecular modelling software, Category:Protein folding

-----

Artificial general intelligence

Intelligent agent

Recursive self-improvement

Planning

Computer vision

General game playing

Knowledge representation

Natural language processing

Robotics

AI safety

Machine learning

Symbolic

Deep learning

Bayesian networks

Evolutionary algorithms

Hybrid intelligent systems

Systems integration

Open-source

Bioinformatics

Deepfake

Earth sciences

Finance

Generative AI Art Audio Music

Art

Audio

Music

Government

Healthcare Mental health

Mental health

Industry

Software development

Translation

Military

Physics

Projects

AI alignment

Artificial consciousness

The bitter lesson

Chinese room

Friendly AI

Ethics

Existential risk

Turing test

Uncanny valley

Timeline

Progress

AI winter

AI boom

AI bubble

Glossary

v

t

e

AlphaFold is an artificial intelligence (AI) program developed by DeepMind , a subsidiary of Alphabet , which performs predictions of protein structure . [1] It is designed using deep learning techniques. [2]

AlphaFold 1 (2018) placed first in the overall rankings of the 13th Critical Assessment of Structure Prediction (CASP) in December 2018. It was particularly successful at predicting the most accurate structures for targets rated as most difficult by the competition organizers, where no existing template structures were available from proteins with partially similar sequences.

AlphaFold 2 (2020) repeated this placement in the CASP14 competition in November 2020. [3] It achieved a level of accuracy much higher than any other entry. [2] [4] It scored above 90 on CASP's global distance test (GDT) for approximately two-thirds of the proteins, a test measuring the similarity between a computationally predicted structure and the experimentally determined structure, where 100 represents a complete match. [2] [5] The inclusion of metagenomic data has improved the quality of the prediction of MSAs . One of the biggest sources of the training data was the custom-built Big Fantastic Database (BFD) of 65,983,866 protein families, represented as MSAs and hidden Markov models (HMMs), covering 2,204,359,010 protein sequences from reference databases, metagenomes, and metatranscriptomes. [6]

AlphaFold 2's results at CASP14 were described as "astounding" [7] and "transformational". [8] However, some researchers noted that the accuracy was insufficient for a third of its predictions, and that it did not reveal the underlying mechanism or rules of protein folding for the protein folding problem , which remains unsolved. [9] [10]

Despite this, the technical achievement was widely recognized. On 15 July 2021, the AlphaFold 2 paper was published in Nature as an advance access publication alongside open source software and a searchable database of species proteomes . [6] [11] [12] As of February 2025, the paper had been cited nearly 35,000 times. [13]

AlphaFold 3 was announced on 8 May 2024. It can predict the structure of complexes created by proteins with DNA , RNA , various ligands , and ions . [14] [15] The new prediction method shows a minimum 50% improvement in accuracy for protein interactions with other molecules compared to existing methods. Moreover, for certain key categories of interactions, the prediction accuracy has effectively doubled. [16]

Demis Hassabis and John Jumper of Google DeepMind shared one half of the 2024 Nobel Prize in Chemistry , awarded "for protein structure prediction," while the other half went to David Baker "for computational protein design." [17] Hassabis and Jumper had previously won the Breakthrough Prize in Life Sciences and the Albert Lasker Award for Basic Medical Research in 2023 for their leadership of the AlphaFold project. [18] [19]

Background

Proteins consist of chains of amino acids which spontaneously fold to form the three dimensional (3-D) structures of the proteins. The 3-D structure is crucial to understanding the biological function of the protein.

Protein structures can be determined experimentally through techniques such as X-ray crystallography , cryo-electron microscopy and nuclear magnetic resonance (NMR), which are all expensive and time-consuming. [20] Such efforts, using the experimental methods, have identified the structures of about 170,000 proteins over the last 60 years, while there are over 200 million known proteins across all life forms. [5]

Over the years, researchers have applied numerous computational methods to predict the 3D structures of proteins from their amino acid sequences, accuracy of such methods in best possible scenario is close to experimental techniques (NMR) by the use of homology modeling based on molecular evolution. CASP , which was launched in 1994 to challenge the scientific community to produce their best protein structure predictions, found that GDT scores of only about 40 out of 100 can be achieved for the most difficult proteins by 2016. [5] AlphaFold started competing in the 2018 CASP using an artificial intelligence (AI) deep learning technique. [20]

Algorithm

DeepMind is known to have trained the program on over 170,000 proteins from the Protein Data Bank , a public repository of protein sequences and structures. The program uses a form of attention network , a deep learning technique that focuses on having the AI identify parts of a larger problem, then piece it together to obtain the overall solution. [2] The overall training was conducted on processing power between 100 and 200 GPUs . [2]

AlphaFold 1 (2018)

AlphaFold 1 (2018) was built on work developed by various teams in the 2010s, work that looked at the large databanks of related DNA sequences now available from many different organisms (most without known 3D structures), to try to find changes at different residues (peptides) that appeared to be correlated, even though the residues were not consecutive in the main chain. Such correlations suggest that the residues may be close to each other physically, even though not close in the sequence, allowing a contact map to be estimated. Building on recent work prior to 2018, AlphaFold 1 extended this by estimating a probability distribution for the distances between residues, effectively transforming the contact map into a distance map. It also used more advanced learning methods than previously to develop the inference. [21] [22]

AlphaFold 2 (2020)

The 2020 version of the program (AlphaFold 2 , 2020) is significantly different from the original version that won CASP 13 in 2018, according to the team at DeepMind. [24] [25]

AlphaFold 1 used a number of separately trained modules to produce a guide potential, which was then combined with a physics-based energy potential. AlphaFold 2 replaced this with a system of interconnected sub-networks, forming a single, differentiable, end-to-end model based on pattern recognition. This model was trained in an integrated manner. [25] [26] After the neural network's prediction converges, a final refinement step applies local physical constraints using energy minimization based on the AMBER force field. This step only slightly adjusts the predicted structure. [27]

A key part of the 2020 system are two modules, believed to be based on a transformer design, which are used to progressively refine a vector of information for each relationship (or " edge " in graph-theory terminology) between an amino acid residue of the protein and another amino acid residue (these relationships are represented by the array shown in green); and between each amino acid position and each different sequences in the input sequence alignment (these relationships are represented by the array shown in red). [26] Internally these refinement transformations contain layers that have the effect of bringing relevant data together and filtering out irrelevant data (the "attention mechanism") for these relationships, in a context-dependent way, learnt from training data. These transformations are iterated, the updated information output by one step becoming the input of the next, with the sharpened residue/residue information feeding into the update of the residue/sequence information, and then the improved residue/sequence information feeding into the update of the residue/residue information. [26] As the iteration progresses, according to one report, the "attention algorithm ... mimics the way a person might assemble a jigsaw puzzle: first connecting pieces in small clumps—in this case clusters of amino acids—and then searching for ways to join the clumps in a larger whole." [5] [needs update]

The output of these iterations then informs the final structure prediction module, [26] which also uses transformers, [28] and is itself then iterated. In an example presented by DeepMind, the structure prediction module achieved a correct topology for the target protein on its first iteration, scored as having a GDT_TS of 78, but with a large number (90%) of stereochemical violations – i.e. unphysical bond angles or lengths. With subsequent iterations the number of stereochemical violations fell. By the third iteration the GDT_TS of the prediction was approaching 90, and by the eighth iteration the number of stereochemical violations was approaching zero. [29]

The training data was originally restricted to single peptide chains. However, the October 2021 update, named AlphaFold-Multimer, included protein complexes in its training data. DeepMind stated this update succeeded about 70% of the time at accurately predicting protein-protein interactions. [30]

AlphaFold 3 (2024)

Announced on 8 May 2024, AlphaFold 3 was co-developed by Google DeepMind and Isomorphic Labs , both subsidiaries of Alphabet . AlphaFold 3 is not limited to single-chain proteins, as it can also predict the structures of protein complexes with DNA , RNA , post-translational modifications and selected ligands and ions . [31] [14]

AlphaFold 3 introduces the "Pairformer," a deep learning architecture inspired by the transformer, which is considered similar to, but simpler than, the Evoformer used in AlphaFold 2. [15] [32] The Pairformer module's initial predictions are refined by a diffusion model . This model begins with a cloud of atoms and iteratively refines their positions, guided by the Pairformer's output, to generate a 3D representation of the molecular structure. [14]

The AlphaFold server was created to provide free access to AlphaFold 3 for non-commercial research. [33] As of May 2025, the AlphaFold 3 research paper has been directly cited more than 4000 times. [34]

Competitions

CASP13

In December 2018, DeepMind's AlphaFold placed first in the overall rankings of the 13th Critical Assessment of Techniques for Protein Structure Prediction (CASP). [35] [36]

The program was particularly successfully predicting the most accurate structure for targets rated as the most difficult by the competition organisers, where no existing template structures were available from proteins with a partially similar sequence. AlphaFold gave the best prediction for 25 out of 43 protein targets in this class, [36] [37] [38] achieving a median score of 58.9 on the CASP's global distance test (GDT) score, ahead of 52.5 and 52.4 by the two next best-placed teams, [39] who were also using deep learning to estimate contact distances. [40] [41] Overall, across all targets, AlphaFold 1 achieved a GDT score of 68.5. [42]

In January 2020, implementations and illustrative code of AlphaFold 1 was released open-source on GitHub . [43] [20] but, as stated in the "Read Me" file on that website: "This code can't be used to predict structure of an arbitrary protein sequence. It can be used to predict structure only on the CASP13 dataset (links below). The feature generation code is tightly coupled to our internal infrastructure as well as external tools, hence we are unable to open-source it." Therefore, in essence, the code deposited is not suitable for general use but only for the CASP13 proteins. The company has not announced plans to make their code publicly available as of 5 March 2021.

CASP14

In November 2020, DeepMind's new version, AlphaFold 2, won CASP14. [44] [45] Overall, AlphaFold 2 made the best prediction for 88 out of the 97 targets. [7]

On the competition's preferred global distance test (GDT) measure of accuracy, the program achieved a median score of 92.4 (out of 100), meaning that more than half of its predictions were scored at better than 92.4% for having their atoms in more-or-less the right place, [46] [47] a level of accuracy reported to be comparable to experimental techniques like X-ray crystallography . [24] [8] [42] In 2018 AlphaFold 1 had only reached this level of accuracy in two of all of its predictions. [7] 88% of predictions in the 2020 competition had a GDT_TS score of more than 80. On the group of targets classed as the most difficult, AlphaFold 2 achieved a median score of 87. [citation needed]

Measured by the root-mean-square deviation (RMS-D) of the placement of the alpha-carbon atoms of the protein backbone chain, which tends to be dominated by the performance of the worst-fitted outliers, 88% of AlphaFold 2's predictions had an RMS deviation of less than 4 Å for the set of overlapped C-alpha atoms. [7] 76% of predictions achieved better than 3 Å, and 46% had a C-alpha atom RMS accuracy better than 2 Å, [7] with a median RMS deviation in its predictions of 2.1 Å for a set of overlapped CA atoms. [7] AlphaFold 2 also achieved an accuracy in modelling surface side chains described as "really really extraordinary".

To further validate AlphaFold 2, the conference organizers approached four leading experimental groups working on structures they found particularly challenging and had been unable to determine. In all four cases the three-dimensional models produced by AlphaFold 2 were sufficiently accurate to determine structures of these proteins by molecular replacement . These included target T1100 (Af1503), a small membrane protein studied by experimentalists for ten years. [5]

Of the three structures that AlphaFold 2 had the least success in predicting, two had been obtained by protein NMR methods, which define protein structure directly in aqueous solution, whereas AlphaFold was mostly trained on protein structures in crystals . The third exists in nature as a multidomain complex consisting of 52 identical copies of the same domain , a situation AlphaFold was not programmed to consider. For all targets with a single domain, excluding only one very large protein and the two structures determined by NMR, AlphaFold 2 achieved a GDT_TS score of over 80.

CASP15

In 2022, DeepMind did not enter CASP15, but most of the entrants used AlphaFold or tools incorporating AlphaFold. [48]

Reception

AlphaFold 2 scoring more than 90 in CASP 's global distance test (GDT) is a great achievement in computational biology [5] . [8] Nobel Prize winner and structural biologist Venki Ramakrishnan called the result "a stunning advance on the protein folding problem", [5] adding that "It has occurred decades before many people in the field would have predicted. It will be exciting to see the many ways in which it will fundamentally change biological research." [44]

AlphaFold 2's success received wide media attention. [49] [44] [50] News pieces appeared in the science press, such as Nature , [8] Science , [5] MIT Technology Review , [2] and New Scientist , [51] [52] and the story was covered by national newspapers. [53] [54] [55] [56] A frequent theme was the ability to predict protein structures based on the constituent amino acid sequence, expected to have benefits in the life sciences--accelerating drug discovery and enabling better understanding of diseases. [8] [57] Some have noted that even a perfect answer to the protein prediction problem still leaves questions about the protein folding problem (and thus protein dynamics)—understanding in detail how the folding process actually occurs in nature (and how sometimes they can also misfold). [58]

In 2023, Demis Hassabis and John Jumper won the Breakthrough Prize in Life Sciences [19] as well as the Albert Lasker Award for Basic Medical Research for their management of the AlphaFold project. [59] Hassabis and Jumper proceeded to win the Nobel Prize in Chemistry in 2024 for their work on “protein structure prediction” with David Baker of the University of Washington. [18] [60]

Source code

Open access to source code of several AlphaFold versions (excluding AlphaFold 3) has been provided by DeepMind after requests from the scientific community. [61] [62] [63] The source code of AlphaFold 3 [64] was made available for non-commercial use to the scientific community upon request in November 2024.

Database of protein models generated by AlphaFold

The AlphaFold Protein Structure Database , a joint project between AlphaFold and EMBL-EBI , was launched on July 22, 2021. At launch, the database contained AlphaFold-predicted models for nearly the complete UniProt proteome of humans and 20 model organisms , totaling over 365,000 proteins. The database does not include proteins with fewer than 16 or more than 2700 amino acid residues , [65] but for humans they are available in the whole batch file. [66] AlphaFold's initial goal (as of early 2022) was to expand the database to cover most of the UniRef90 set, which contains over 100 million proteins. As of May 15, 2022, the database contained 992,316 predictions. [67]

In July 2021, UniProt-KB and InterPro [68] has been updated to show AlphaFold predictions when available. [69]

On July 28, 2022, the team uploaded to the database the structures of around 200 million proteins from 1 million species, covering nearly every known protein on the planet. [70]

Performance, validations and limitations

Despite its impressive success, AlphaFold has also shown certain limitations:

AlphaFold DB provides models of individual protein chains (monomers), rather than their biologically relevant complexes. [71]

Many protein regions are predicted with low confidence score, including the intrinsically disordered protein regions. [72]

Alphafold-2 was validated for predicting effects of point mutations on structure [73] and free energy [74] , with a partial success.

The model relies, to some extent, on co-evolutionary information from similar proteins. Therefore, it may not perform as well on synthetic proteins or proteins with very low homology to those in the training database. [75]

The model's ability to predict multiple native conformations of proteins is limited.

AlphaFold 3 version can predict structures of protein complexes with a very limited set of selected cofactors and co- and post-translational modifications . [76] Between 50% and 70% of the structures of the human proteome are incomplete without covalently-attached glycans. [77] [71] AlphaFill, a derived database, adds cofactors to AlphaFold models where appropriate. [78]

In the algorithm, the residues are moved freely, without any restraints. Therefore, during modeling the integrity of the chain is not maintained. As a result, AlphaFold may produce topologically wrong results, like structures with an arbitrary number of knots. [79]

Applications

AlphaFold has been used to predict structures of proteins of SARS-CoV-2 , the causative agent of COVID-19 . The structures of these proteins were pending experimental detection in early 2020. [80] [8] Results were reviewed by scientists at the Francis Crick Institute in the United Kingdom before being released to the broader research community. The team also confirmed accurate prediction against the experimentally determined SARS-CoV-2 spike protein that was shared in the Protein Data Bank , an international open-access database, before releasing the computationally determined structures of the under-studied protein molecules. [81] The team acknowledged that although these protein structures might not be the subject of ongoing therapeutical research efforts, they will add to the community's understanding of the SARS-CoV-2 virus. [81] Specifically, AlphaFold 2's prediction of the structure of the ORF3a protein was very similar to the structure determined by researchers at University of California, Berkeley using cryo-electron microscopy . This specific protein is believed to assist the virus in breaking out of the host cell once it replicates. This protein is also believed to play a role in triggering the inflammatory response to the infection. [82]

Published works

Andrew W. Senior et al. (December 2019), "Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13)" , Proteins: Structure, Function, Bioinformatics 87 (12) 1141–1148 doi : 10.1002/prot.25834

Andrew W. Senior et al. (15 January 2020), "Improved protein structure prediction using potentials from deep learning" , Nature 577 706–710 doi : 10.1038/s41586-019-1923-7

John Jumper et al. (December 2020), "High Accuracy Protein Structure Prediction Using Deep Learning", in Fourteenth Critical Assessment of Techniques for Protein Structure Prediction (Abstract Book) , pp. 22–24

John Jumper et al. (December 2020), " AlphaFold 2 ". Presentation given at CASP 14.

Abramson, J., Adler, J., Dunger, J. et al. (May 2024), " Accurate structure prediction of biomolecular interactions with AlphaFold 3 ", Nature 630, 493–500 (2024)

See also

Folding@home

IBM Blue Gene

Foldit

Rosetta@home

Human Proteome Folding Project

AlphaZero

AlphaGo

AlphaGeometry

Predicted Aligned Error

References

Further reading

Carlos Outeiral, CASP14: what Google DeepMind's AlphaFold 2 really achieved, and what it means for protein folding, biology and bioinformatics , Oxford Protein Informatics Group. (3 December)

Mohammed AlQuraishi, AlphaFold2 @ CASP14: "It feels like one's child has left home." (blog), 8 December 2020

Mohammed AlQuraishi, The AlphaFold2 Method Paper: A Fount of Good Ideas (blog), 25 July 2021

External links

AlphaFold-3 web server

AlphaFold v2.1 code and links to model on GitHub

Open access to protein structure predictions for the human proteome and 20 other key organisms at European Bioinformatics Institute (AlphaFold Protein Structure Database)

CASP 14 website

AlphaFold: The making of a scientific breakthrough , DeepMind, via YouTube.

ColabFold (Mirdita, Milot; Schütze, Konstantin; Moriwaki, Yoshitaka; Heo, Lim; Ovchinnikov, Sergey; Steinegger, Martin (2022-05-30). "ColabFold: Making protein folding accessible to all" . Nature Methods . 19 (6): 679– 682. doi : 10.1038/s41592-022-01488-1 . PMC 9184281 . PMID 35637307 .), version for homooligomeric prediction and complexes

v

t

e

Google

Google Brain

Google DeepMind

AlphaGo (2015)

Master (2016)

AlphaGo Zero (2017)

AlphaZero (2017)

MuZero (2019)

Fan Hui (2015)

Lee Sedol (2016)

Ke Jie (2017)

AlphaGo (2017)

The MANIAC (2023)

AlphaFold (2018)

AlphaStar (2019)

AlphaDev (2023)

AlphaGeometry (2024)

AlphaGenome (2025)

Inception (2014)

WaveNet (2016)

MobileNet (2017)

Transformer (2017)

EfficientNet (2019)

Gato (2022)

Quantum Artificial Intelligence Lab

TensorFlow

Tensor Processing Unit

Assistant (2016)

Sparrow (2022)

Gemini (2023)

BERT (2018)

XLNet (2019)

T5 (2019)

LaMDA (2021)

Chinchilla (2022)

PaLM (2022)

Imagen (2023)

Gemini (2023)

VideoPoet (2024)

Gemma (2024)

Veo (2024)

DreamBooth (2022)

NotebookLM (2023)

Vids (2024)

Gemini Robotics (2025)

" Attention Is All You Need "

Future of Go Summit

Generative pre-trained transformer

Google Labs

Google Pixel

Google Workspace

Robot Constitution

Category

Commons

v

t

e

History timeline

timeline

Companies

Projects

Parameter Hyperparameter

Hyperparameter

Loss functions

Regression Bias–variance tradeoff Double descent Overfitting

Bias–variance tradeoff

Double descent

Overfitting

Clustering

Gradient descent SGD Quasi-Newton method Conjugate gradient method

SGD

Quasi-Newton method

Conjugate gradient method

Backpropagation

Attention

Convolution

Normalization Batchnorm

Batchnorm

Activation Softmax Sigmoid Rectifier

Softmax

Sigmoid

Rectifier

Gating

Weight initialization

Regularization

Datasets Augmentation

Augmentation

Prompt engineering

Reinforcement learning Q-learning SARSA Imitation Policy gradient

Q-learning

SARSA

Imitation

Policy gradient

Diffusion

Latent diffusion model

Autoregression

Adversary

RAG

Uncanny valley

RLHF

Self-supervised learning

Reflection

Recursive self-improvement

Hallucination

Word embedding

Vibe coding

Machine learning In-context learning

In-context learning

Artificial neural network Deep learning

Deep learning

Language model Large language model NMT

Large language model

NMT

Reasoning language model

Model Context Protocol

Intelligent agent

Artificial human companion

Humanity's Last Exam

Artificial general intelligence (AGI)

AlexNet

WaveNet

Human image synthesis

HWR

OCR

Computer vision

Speech synthesis 15.ai ElevenLabs

15.ai

ElevenLabs

Speech recognition Whisper

Whisper

Facial recognition

AlphaFold

Text-to-image models Aurora DALL-E Firefly Flux Ideogram Imagen Midjourney Recraft Stable Diffusion

Aurora

DALL-E

Firefly

Flux

Ideogram

Imagen

Midjourney

Recraft

Stable Diffusion

Text-to-video models Dream Machine Runway Gen Hailuo AI Kling Sora Veo

Dream Machine

Runway Gen

Hailuo AI

Kling

Sora

Veo

Music generation Riffusion Suno AI Udio

Riffusion

Suno AI

Udio

Word2vec

Seq2seq

GloVe

BERT

T5

Llama

Chinchilla AI

PaLM

GPT 1 2 3 J ChatGPT 4 4o o1 o3 4.5 4.1 o4-mini 5

1

2

3

J

ChatGPT

4

4o

o1

o3

4.5

4.1

o4-mini

5

Claude

Gemini Gemini (language model) Gemma

Gemini (language model)

Gemma

Grok

LaMDA

BLOOM

DBRX

Project Debater

IBM Watson

IBM Watsonx

Granite

PanGu-Σ

DeepSeek

Qwen

AlphaGo

AlphaZero

OpenAI Five

Self-driving car

MuZero

Action selection AutoGPT

AutoGPT

Robot control

Alan Turing

Warren Sturgis McCulloch

Walter Pitts

John von Neumann

Claude Shannon

Shun'ichi Amari

Kunihiko Fukushima

Takeo Kanade

Marvin Minsky

John McCarthy

Nathaniel Rochester

Allen Newell

Cliff Shaw

Herbert A. Simon

Oliver Selfridge

Frank Rosenblatt

Bernard Widrow

Joseph Weizenbaum

Seymour Papert

Seppo Linnainmaa

Paul Werbos

Geoffrey Hinton

John Hopfield

Jürgen Schmidhuber

Yann LeCun

Yoshua Bengio

Lotfi A. Zadeh

Stephen Grossberg

Alex Graves

James Goodnight

Andrew Ng

Fei-Fei Li

Alex Krizhevsky

Ilya Sutskever

Oriol Vinyals

Quoc V. Le

Ian Goodfellow

Demis Hassabis

David Silver

Andrej Karpathy

Ashish Vaswani

Noam Shazeer

Aidan Gomez

John Schulman

Mustafa Suleyman

Jan Leike

Daniel Kokotajlo

François Chollet

Neural Turing machine

Differentiable neural computer

Transformer Vision transformer (ViT)

Vision transformer (ViT)

Recurrent neural network (RNN)

Long short-term memory (LSTM)

Gated recurrent unit (GRU)

Echo state network

Multilayer perceptron (MLP)

Convolutional neural network (CNN)

Residual neural network (RNN)

Highway network

Mamba

Autoencoder

Variational autoencoder (VAE)

Generative adversarial network (GAN)

Graph neural network (GNN)

Category