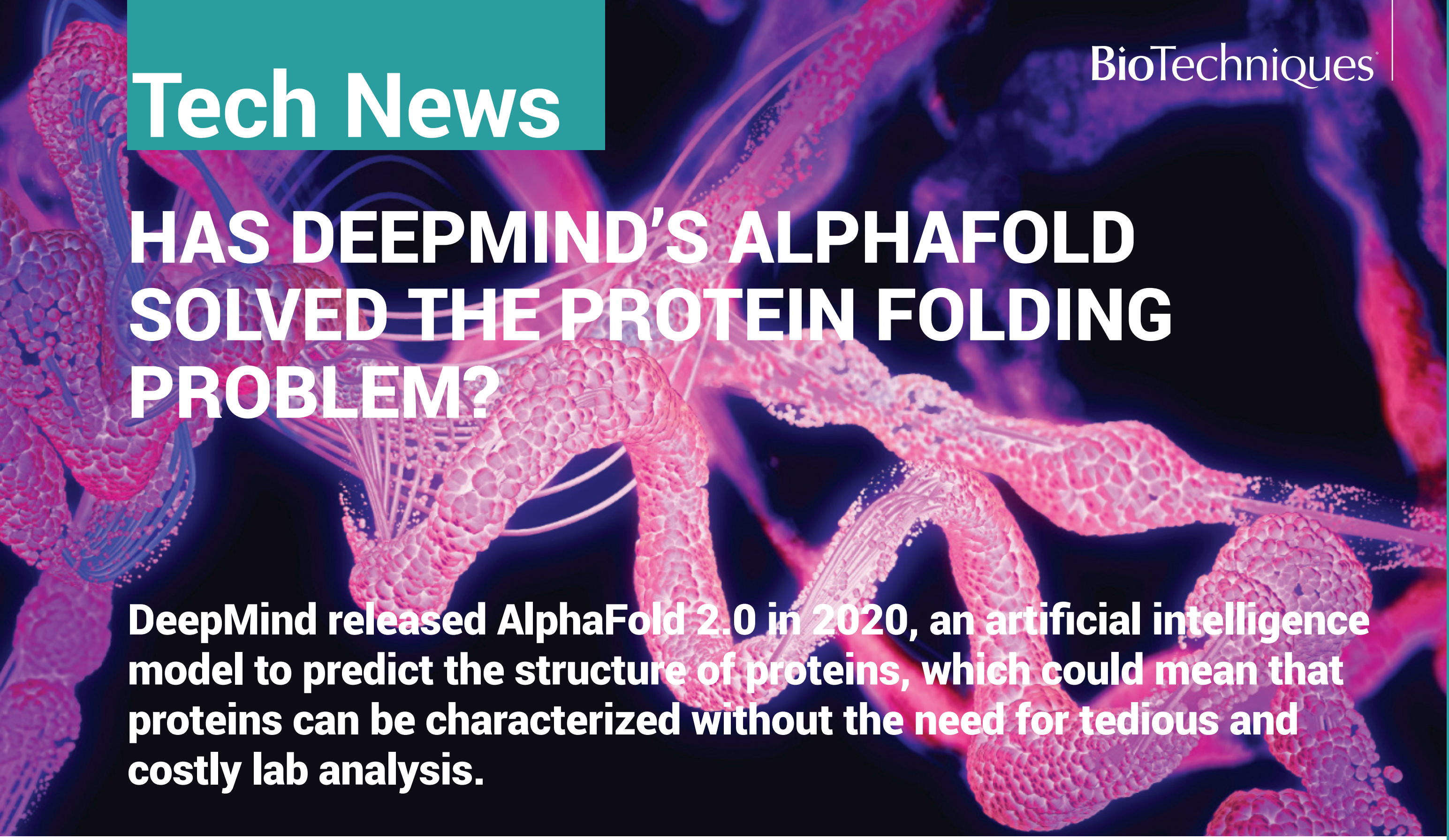


HAS DEEPMIND'S ALPHAFOLD SOLVED THE PROTEIN FOLDING PROBLEM?

DeepMind released AlphaFold 2.0 in 2020, an artificial intelligence model to predict the structure of proteins, which could mean that proteins can be characterized without the need for tedious and costly lab analysis.

In 1972 biochemist Christian Anfinsen won the Nobel Prize for his work on the relationship between an amino acid sequence and its biologically active three-dimensional (3D) conformation. Anfinsen ended his Nobel Prize lecture paper by saying: *"Empirical considerations of the large amount of data now available on the correlations between sequence and the three-dimensional structure, together with an increasing sophistication in the theoretical treatment of the energetics of polypeptide chain folding are beginning to make more realistic the idea of the a priori prediction of protein conformation."* [1]

Now, 50 years later, DeepMind has developed an artificial intelligence (AI) model that takes us one step closer to predicting the 3D shape of a protein from only its one-dimensional (1D) amino acid sequence.

THE GREAT PROTEIN FOLDING PROBLEM

Proteins are the building blocks of life, each with a unique shape that determines their function, such as catalyzing biochemical reactions or enabling muscles to contract [2].

The ability to accurately predict or determine the spatial arrangement of amino acids would provide a better insight into the role of proteins and the mechanisms through which these essential macromolecules function [3]. For instance, protein misfolding is known to contribute to the pathogenesis of diseases such as Alzheimer's; therefore, having accessible information on the structure of different proteins would be highly beneficial in the study of neurodegenerative diseases [4]. Additionally, the shape of a protein will affect which molecules can bind to it; therefore, understanding the folded shape of protein targets could have implications in drug design [5].

Currently the Universal Protein Resource (UniProt) has around 220 million unique protein sequences on record, obtained with methods such as Edman degradation or mass spectrometry. In comparison, the Protein Data Bank only has 180,000 3D protein structures from 55,000 different protein in its archives [3]. Despite advances in experimental methods of structure determination, it continues to be a time-consuming process based on trial and error using x-ray crystallography, nuclear magnetic resonance spectroscopy or cryogenic electron microscopy, for example [6].

This is partly because of the large number of degrees of freedom in an amino acid chain before folding, meaning that there are many possible 3D conformations. In 1969 Cyrus Levinthal, a molecular biologist, estimated that each protein could fold into 10^{300} conformations, and yet, proteins achieve the correct conformation in a fraction of a second [7]. This means it is mathematically unfeasible for proteins to randomly fold until they achieve their functional conformation, a phenomenon referred to as Levinthal's Paradox.

Additionally, techniques used to analyze a protein's structure require crystallized proteins, which is not ideal for hydrophobic membrane proteins that aggregate in aqueous solutions and are notoriously difficult to crystalize [3].

Therefore, computational models for accurately predicting and visualizing a protein's structures is appealing. It is thought that AI could play a role in expanding these databases of protein structures and close the gap between our understanding of 1D and 3D protein structures [4].

A GLOBAL PROTEIN PREDICTION EXPERIMENT

To encourage development in computational methods to predict protein structures, Krzysztof Fidelis (University of California,

▶

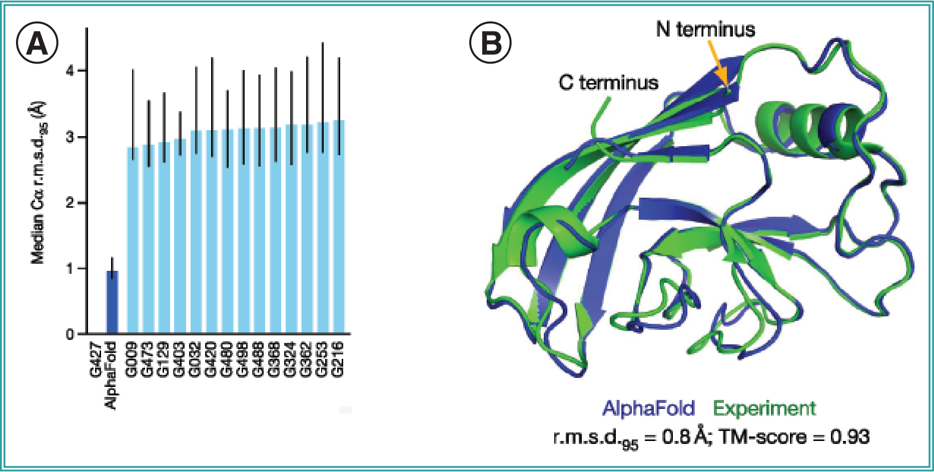


Figure 1. (A) AlphaFold's performance compared to the top-15 entries in CASP14. (B) AlphaFold's prediction of CASP14 target T1049 (blue) compared to the experimental structure (green). This figure has been reused following the terms of the CC-BY licence from Figure 1 in [9].

CA, USA) and John Moult (University of Maryland, MD, USA) founded the Critical Assessment of Structure Prediction (CASP) in 1994, which is a biennial event testing these computational models [2].

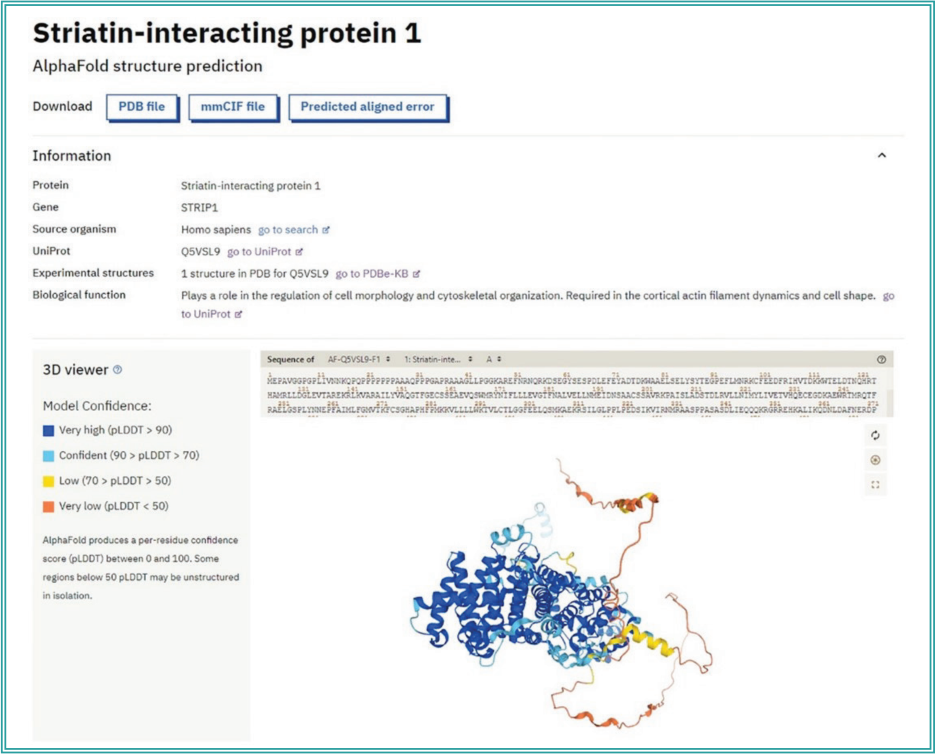
The founders prefer to think of this as an experiment, instead of a competition, where around 100 teams use their models to predict the 3D structure of proteins from their given amino acid sequence. These protein structures have been experimentally determined but not released to the public, and will be compared to the computationally predicted conformation [8].

The Global Distance Test (GDT) is the main metric used to assess the success of the computational models and ranges from 0 to 100. This can be thought of as the percentage of amino acid residues that are within a certain distance from the correct position, where the experimental structures are used as the 'ground truth'. Getting a GDT score of around 90 is considered comparable to experimental methods, according to Moult.

There are two paths of developing computational methods for the protein folding problem. These can be based on physical interactions by applying our understanding of molecular driving forces, or our evolutionary understanding, relying on bioinformatic analysis of the evolutionary history of proteins [9].

INTRODUCING ALPHAFOLD: THE GOLD STANDARD FOR PREDICTING PROTEINS

In previous years at CASP, the GDTs have only reached around 60 [2]. However, in 2020 Google's DeepMind introduced AlphaFold 2.0, which achieved an average GDT of 90 at CASP14 [7].



Striatin-interacting protein 1

AlphaFold structure prediction

Information

Protein	Striatin-interacting protein 1
Gene	STRIP1
Source organism	Homo sapiens go to search
UniProt	Q5VSL9 go to UniProt
Experimental structures	1 structure in PDB for Q5VSL9 go to PDBe-KB
Biological function	Plays a role in the regulation of cell morphology and cytoskeletal organization. Required in the cortical actin filament dynamics and cell shape. go to UniProt

3D viewer

Model Confidence:

- 🟦 Very high (pLDDT > 90)
- 🟦 Confident (90 > pLDDT > 70)
- 🟨 Low (70 > pLDDT > 50)
- 🟧 Very low (pLDDT < 50)

AlphaFold produces a per-residue confidence score (pLDDT) between 0 and 100. Some regions below 50 pLDDT may be unstructured in isolation.

```
MEPAVGGPGPLIVNNKQPQPPPPPPPAAAQPPPGAPRAAAGLLPGGKAREFNRNQRKDSEGYSESPDLEFEYADTDKWAAELSELYSYTEGPEFLMNRKCFEEDFRIHVTDKKWTELDTNQHRT
HAMRLLDGLEVTAREKRLKVARAILYVAQGTFGECSSEAEVQSWMRYNIFLLLEVGTFNALVELLNMEIDNSAACSSAVRKPAISLADSTDLRVLLNIMYLIVETVHQECEGDKAEWRTMRQTF
RAELGSPLYNNEPFAIMLFGMVTKFCSGHAPHFPMKKVLLLLWKTVLCTLGGFEELQSMKAEKRSILGLPPLPEDSIKVIRNMRAASPPASASDLIEQQQKRGRREHKALIKQDNLDAFNERDP
```

Figure 2. An AlphaFold predicted structure for striatin-interacting protein 1 alongside meta-information. This copyright is held by Oxford University press and the figure has been reused following the terms of the CC-BY licence from Figure 2 in [3].

"We have been stuck on this one problem – how do proteins fold up – for nearly 50 years. To see DeepMind produce a solution for this, having worked personally on this problem for so long and after so many stops and starts, wondering if we'd ever get there, is a very special moment," said Moult [7]. AlphaFold 2.0 is the first computational method that can predict protein structures with near experimental accuracy and is being called the gold standard for this type of technology [9].

DeepMind also achieved one of the highest GDT scores in CASP13 in 2018 and a median score of around 70–75 with the first version of AlphaFold [2]. After this, DeepMind went back to the drawing board and created a completely new model for predicting structures from amino acid sequences resulting in AlphaFold 2.0, a model that uses both physical and evolutionary constraints in its predictions [3]. The original AlphaFold combined local physics and pattern recognition, and would often overestimate the effect of interactions between nearby residues [10].

Instead, AlphaFold 2.0 relies exclusively on pattern recognition and is an attention-based neural network architecture combined with a deep learning framework [9]. An attention-based algorithm behaves the way one might approach a jigsaw puzzle, by connecting smaller sections of the jigsaw – in this case, amino acids – before these sections are joined together to complete the jigsaw puzzle picture, or 3D protein structure [11].

The neural network is initially trained on the 170,000 protein structures publicly available in the Protein Data Bank (PDB) [7]. This trained network was then used to make structural predictions from a further 350,000 sequences from UniClust, a database of annotated protein sequences and alignments. High-confidence predictions from the UniClust dataset were selected and combined with the PBD data to create a new dataset. This new dataset is then used to retrain the AlphaFold architecture anew, enhancing the accuracy of predictions. Further training with added alterations to protein sequences challenge AlphaFold's capabilities to produce the same structures as it previously predicted [9].

AlphaFold uses an iterating process to improve its predictions and has an internal measure called the predicted Local Distances Difference Test, pLDDT, to assess the reliability of its predictions. This is based on an existing metric in protein structure prediction, the Local Distance Difference Test or LDDT, which compares the local distances of atoms in computational models to the experimentally determined structures [12]. This assigns a high score to regions with a high local accuracy, regardless of the accuracy of the whole predicted protein [3]. This type of system allows AlphaFold to refine its predictions, resulting in a more accurate structure.

AlphaFold's median score in CASP14 was 92.4 GDT, with an average error of 1.6Å calculated using the root-mean-square deviation (RMSD) from the correct atomic positions. This means AlphaFold predicted structures that were accurate to within the width of one atom [7]. The accuracy of the protein backbone had a median accuracy of 0.96Å RMSD, compared to the next best performing model with an accuracy of 2.8Å RMSD [9].

The assessors of CASP14 gave AlphaFold an additional protein sequence to predict, a membrane protein from an ancient group of microbes that have been studied using experimental techniques for 10 years without successfully obtaining the 3D structure. AlphaFold successfully produced a 3D image of the protein, which the research team could then use to interpret their x-ray crystallography data. They reported fitting the predicted model to the x-ray data within half an hour, saying it was an almost perfect prediction [11]. This highlights that AI may not completely replace existing experimental techniques but can be used as a tool alongside them to interpret low-resolution data [13].

Whilst it may be considered the gold standard of protein prediction, there is still room for improvement as AlphaFold only provides one prediction of a stable structure for each protein; however, proteins are dynamic and can change shape throughout the body, for example under different pH conditions [5]. Additionally, AlphaFold is not able to determine the shape of multiprotein complexes and does not include any ligands such as cofactors or metals, meaning no data are available for such interactions [13]. Despite these shortcomings, AlphaFold is the first step in protein prediction technology, and it is likely that solving these challenges will also be done so using deep learning and AI.

AlphaFold was awarded Nature's Method of the Year 2021 as it significantly advances protein prediction technology and presents a paradigm shift in this type of technology [14]. In addition to this, the success of AlphaFold in CASP14 led the organizers to declare that the protein folding problem had been solved, although there is some disagreement on this in the wider community [5]. Lior Pachter, a computational biologist at Caltech (CA, USA), took to Twitter to highlight that AlphaFold wasn't the winner across all the protein sequences in CASP14. One of the most notable contenders was another AI model called RoseTTaFold, developed by an academic group lead by David Barker at The University of Washington (WA, USA) [15]. Unlike AlphaFold, RoseTTaFold can predict the structures of protein complexes, whilst AlphaFold remains limited to single protein structures [16].

ALPHAFOLD BEYOND CASP14

Since DeepMind's success at CASP14 in 2020, the source code for the computer model has been published in Nature, meaning structural biologists can independently validate AlphaFold's predicted structures as well as use it to predict other protein structures in their own research [2].

DeepMind has also published an open-source database with over 350,000 protein structures from 21 model organisms that have been predicted using AlphaFold in a collaboration with the European Molecular Biology Laboratory (EMBL, Heidelberg, Germany) [14]. This database includes the structures for 98.5% of the human proteome, where 36% are predicted to a very high accuracy and 22% with a high accuracy [5]. This presents a vast expansion in data, as previously only 17% of the human proteome had reported 3D structures, which were obtained through experimental efforts [2].

"The AlphaFold database is the perfect example of the virtuous circle of open science," said Edith Heard, the Director General at EMBL [8]. *"AlphaFold was trained using data from public resource and built by the scientific community, so it makes sense for its predictions to be public. Sharing AlphaFold predictions openly and freely empowers researchers everywhere to gain new insights and drive discovery."*

Other than realizing Anfinsen's vision from his Nobel Laureate essay of predicting the 3D structure of proteins, technology like AlphaFold has the potential to impact several areas of scientific research. This could enable scientists to make *de novo* proteins, those that don't exist in nature, with tuneable properties depending on the desired application. Being able to computationally predict how these proteins will fold, and what their stable conformation is, could open up a whole new area of biological research [5]. For example, proteins could be designed to recycle single-use plastic, an application that is currently being studied at the Centre for Enzyme Innovation (University of Portsmouth, UK) [8].

Accurately predicting protein structures is often thought to be beneficial in the process of designing therapeutics, enabling researchers to visualize the target proteins' shape. However, the current limitations of AlphaFold mean we are yet to see a significant change in the drug design game just yet. Predicting the shape of larger multidomain protein complexes and knowing the locations of all the amino acid side chains is important for designing a drug molecule: these are areas that AlphaFold currently struggles to predict [5]. A recent paper also highlighted that whilst this extra structural data may accelerate the early stages of research, it is unlikely to revolutionize the speed at which new drugs go from the lab bench to patients [13].

However, this type of structural data is useful for identifying pathogenic variants and the spike proteins on the surface of microbes like coronaviruses. Researchers at The University of California, San Francisco (CA, USA) used AlphaFold and cryogenic electron microscopy to analyze Nsp2, a protein in SARS-CoV-2 viruses. The structure and function of this particular protein are not known, but the results using AlphaFold show it has a zinc ion-binding site, indicating that this protein plays a role in RNA binding, which could have implications for further research [17].

Another promising application of open-source computational methods is in neglected disease research where funding is often limited, meaning access to these freely available resources is hugely beneficial. DeepMind is currently partnered with the Drugs for Neglected Disease Initiative (DNDi; Geneva, Switzerland), who hope to use AlphaFold to investigate treatments diseases such as Chagas disease, a life-threatening illness caused by the Trypanosoma cruzi parasite [5]. Recently, researchers have identified a molecule that is able to bind to a protein in the parasite, eventually killing it. The team at DNDi hope to use AlphaFold to characterize this protein's structure and use this knowledge to develop new treatments for Chagas disease [18].

Whilst AlphaFold dazzled the judges at CASP14 and is already being used in a variety of research areas, it is only the start of this type of computational technology. For example, we are yet to explain how proteins manage to spontaneously fold into the correct shape in the blink of an eye, despite the potential 10^{300} conformations so, for now, the Leventhal Paradox remains unanswered.

As both AlphaFold and RoseTTafold are freely available, it will be interesting to compare their performance outside of a competition setting. It may be the case that different AI models will be used depending on the analysis carried out; for example, RoseTTaFold may be the ideal choice for characterizing multiprotein complexes, whilst AlphaFold will the take home the trophy for research needing accurate single-protein analysis.

Whilst AlphaFold currently represents the gold standard for AI protein prediction, this benchmark will continue to be raised as this technology develops and evolves – who knows what will be in store for CASP15 in 2022. Paul Nurse, the chair of the EMBL science advisory committee reflected on the impact of AphaFold: *"With this resource freely and openly available, the scientific community will be able to draw on collective knowledge to accelerate discovery, ushering in a new era for AI-enabled biology."* [8].

Written by Aisha Al-Janabi

REFERENCES

1. Anfinsen C. Principles that govern the folding of protein chains. *Science* 181(4096), 223–230 (1973).
2. The Institute of Cancer Research. Reflecting on DeepMind's AlphaFold artificial intelligence success – what's the real significance for protein folding research and drug discovery? www.icr.ac.uk/blogs/the-drug-discoverer/page-details/reflecting-on-deepmind-s-alphafold-artificial-intelligence-success-what-s-the-real-significance-for-protein-folding-research-and-drug-discovery
3. Varadi M, Anyango S, Deshpande M *et al.* AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* 50(D1), D439–D444 (2021).
4. The Conversation. AI makes huge progress predicting how proteins fold - one of biology's greatest challenges - promising rapid drug development. https://theconversation.com/ai-makes-huge-progress-predicting-how-proteins-fold-one-of-biologys-greatest-challenges-promising-rapid-drug-development-151181
5. Forbes. AlphaFold is the most important achievement in AI – ever. www.forbes.com/sites/robtoews/2021/10/03/alphafold-is-the-most-important-achievement-in-ai-ever/?sh=8d9be86e0af3
6. *Nature*. "It will change everything": DeepMind's AI makes gigantic leap in solving protein structures. www.nature.com/articles/d41586-020-03348-4
7. DeepMind. AlphaFold: a solution to a 50-year-old grand challenge in biology. https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology
8. EMBL-EBI. DeepMind and EMBL release the most complete database of predicted 3D structures of human proteins. www.ebi.ac.uk/about/news/press-releases/alphafold-database-launch
9. Jumper J, Evans R, Pritzel A *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589 (2021).
10. Senior AW, Evans R, Jumper J *et al.* Improved protein structure prediction using potentials from deep learning. *Nature* 577(7792), 706–710 (2020).
11. *Science*. "The game has changed." AI triumphs at solving protein structures. www.science.org/content/article/game-has-changed-ai-triumphs-solving-protein-structures
12. Mariani V, Biasini M, Barbato A, Schwede T. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics* 29(21), 2722–2728 (2013).
13. Thornton JM, Laskowski RA, Borkakoti N. AlphaFold heralds a data-driven revolution in biology and medicine. *Nat. Med.* 27, 1666–1669 (2021).
14. Method of the Year 2021: protein structure prediction. *Nat. Methods* 19(1), 1–1 (2022).
15. *Nature*. DeepMind's AI for protein structure is coming to the masses. www.nature.com/articles/d41586-021-01968-y
16. C&EN. Accurate protein structure prediction AI made openly available. https://cen.acs.org/analytical-chemistry/structural-biology/Accurate-protein-structure-prediction-AI/99/i26
17. Gupta M, Azumaya CM, Moritz M *et al.* CryoEM and AI reveal a structure of SARS-CoV-2 Nsp2, a multifunctional protein involved in key host processes. *bioRxiv* doi:10.1101/2021.05.10.443524 (2021) (Epub ahead of print).
18. Wired. DeepMind wants to use its AI to cure neglected diseases. www.wired.co.uk/article/deepmind-alphafold-protein-diseases