Engineering 12 (2022) 9–11

Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng

News & Highlights

Machine Learning Turbocharges Structural Biology

Sean O'Neill

Senior Technology Writer

On 28 January 2022, DeepMind Technologies announced the addition of the proteomes of 27 organisms to its AlphaFold Protein Structures Database (AlphaFold DB), a free online resource for scientists [1]. DeepMind, the London-based, artificial intelligence (AI)-focused subsidiary of Google's parent company, Alphabet, selected these proteomes in alignment with the priorities of the World Health Organization. That is, they published the predicted structures of proteins in organisms that cause neglected tropical diseases, such as leprosy and schistosomiasis, and others of great concern due to antimicrobial resistance. Announcing the additions, the DeepMind team said: "We hope this release can help accelerate research and support those already working tirelessly to eradicate these conditions" [1].

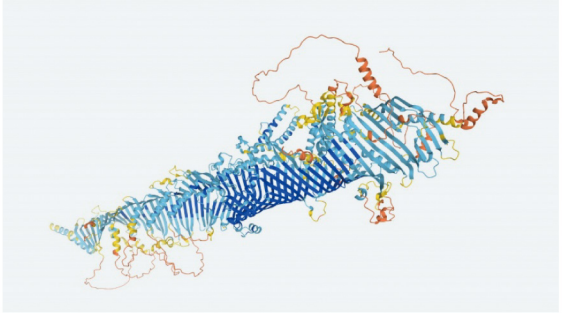
The announcement followed a slew of prior additions to AlphaFold DB since its launch in July 2021 [2]. DeepMind had initially made available structural predictions for proteins from 21 model organisms (Fig. 1) [3], including human, mouse, fruit fly, important crops such as maize, Asian rice, soybean and yeast, pathogens such as *Escherichia coli*, *Candida albicans*, and disease-causing parasites such as *Trypanosoma cruzi* (Chagas disease) and *Leishmania infantum* (leishmaniasis). More additions quickly followed.

These rapid developments indicate how the field of structural biology is being transformed by machine-learning tools that allow scientists to predict the shape of proteins with unprecedented accuracy, based purely on their genetic sequences. Predicting the structure of proteins from their genetic sequences had been a "grand challenge" in biology for five decades [4,5]. It is important because it is often the shape that a protein folds into, and not the genetic sequence itself, that reveals its function. Predicting structures with confidence opens possibilities from designing highly targeted drug molecules to creating crops more resistant to climate change.

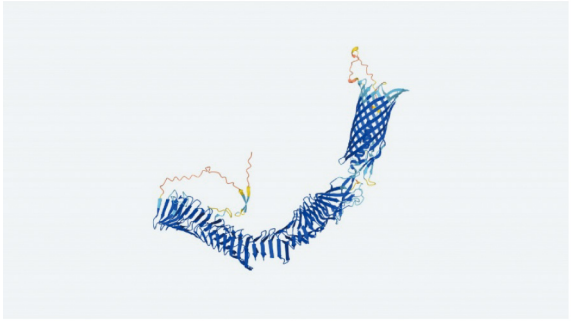
It was in late 2020 that DeepMind's AlphaFold system displayed stunning accuracy [4,5], winning an international biennial experiment called Critical Assessment of Protein Structure Prediction (CASP), in which teams compete to predict the structures of proteins. In many cases, AlphaFold's Protein structure predictions were indistinguishable from experimentally determined structures [4,5].

Back then, no one knew how much the AlphaFold team would make public about their system. That changed in July 2021, when DeepMind published two landmark papers in the journal *Nature*. The first, on 15 July, described in detail how AlphaFold "greatly improves the accuracy of structure prediction by incorporating novel neural network architectures and training procedures based on the evolutionary, physical and geometric constraints of protein structures" [6]. The publication coincided with the open-source release of the AlphaFold code [2], which enabled scientists all over the world to use the system.

The second paper, on 22 July, announced that DeepMind had made available structural predictions for 98.5% of proteins in the



(a)



(b)

Fig. 1. These AlphaFold-generated schematics show the predicted structures of (a) protein Q9VZS7 from the fruit fly, *Drosophila melanogaster* and (b) protein P39180 from the bacterium *Escherichia coli*. Both the fly and bacterium are widely used model organisms for basic research. These examples are among those for the proteomes of 21 such organisms whose predicted structures were initially included in AlphaFold DB. The coloring represents AlphaFold's confidence measure for the predicted positions of the amino acids that make up the protein, from dark blue (high confidence) through light blue (medium), yellow (low), and orange (very low). Credit: DeepMind/AlphaFold (public domain).

https://doi.org/10.1016/j.eng.2022.03.007

entire human proteome. Less than one-fifth of human protein structures have been discovered through experimental determination [7]. That same day, DeepMind—in partnership with the European Bioinformatics Institute (EBI), part of the European Molecular Biology Laboratory (EMBL) group—announced the launch of AlphaFold DB [8]. The database initially contained over 360 000 predicted protein structures, from the 21 model organism proteomes previously mentioned.

Then in December 2021, DeepMind and EMBL-EBI announced that they had expanded AlphaFold DB to cover protein sequences held in the UniProtKB/Swiss-Prot database—a high-quality database of manually-annotated records drawn from the scientific literature [9]. This took AlphaFold DB to over 800 000 predicted protein structures. A planned further update in 2022 will take AlphaFold DB to over 100 million protein structures [3].

The impact of all this in the structural biology world has been "really extraordinary," said CASP co-founder and organizer John Moult, professor and fellow at the University of Maryland's Institute for Bioscience and Biotechnology Research in Rockville, MD, USA. "I have never seen such a rapid uptake of a piece of software. It is not much of an exaggeration to say that all structural biologists are now using either AlphaFold DB or their own installed versions of the software."

AlphaFold is "enormously convenient" for structural biologists, said Jinbo Xu, professor of computational biology at the University of Chicago's Toyota Technological Institute in Chicago, IL, USA. "However, the AlphaFold software tool itself is more important and represents a breakthrough," said Xu, who developed RaptorX, another protein structure predictor and former CASP winner [10].

Richard Wheeler, a principal investigator at the University of Oxford, UK, agrees. Wheeler's lab explores the fundamental cell biology of the *Leishmania* and *Trypanosoma* parasites, both single-celled organisms from the Discoba supergroup of eukaryotes. "I have been hoping for something like AlphaFold for a very long time," he said. "I was super excited because, working with neglected tropical pathogens, we do not have the amazing databases of experimentally determined data that exist for humans, or for model organisms like yeast."

However, Wheeler was immediately concerned that the sparsity of genetic data and knowledge of less well studied organisms like *Trypanosoma cruzi* (Fig. 2) would be a problem for the AlphaFold database. "The protein-sequence databases they were using to do the predictions were likely not very good for Discoba," he said. By comprehensively gathering the latest available protein sequence data for Discoba species and feeding it to an implementation of AlphaFold himself, Wheeler obtained significant improvements on many structural predictions in AlphaFold DB and subsequently made his enhancements freely available to the parasitology community [11].

Such work has important implications for other structural biologists using AlphaFold DB to study neglected organisms, Wheeler said. "For about one-third of Discoba proteins, I saw no improvement over AlphaFold DB, but for about two-thirds I saw anything from noticeable improvements to first high-confidence structural predictions," he said. "AlphaFold is profoundly amazing."

AlphaFold is not the only open-source prediction tool newly available, however. RoseTTaFold, developed by professor of biochemistry David Baker and colleagues at the Institute for Protein Design at the University of Washington in Seattle, WA, USA, was made available shortly after AlphaFold. RosseTTaFold produces predictions approaching the accuracy of AlphaFold, but requires markedly less computer power to run, so is faster [12].

While such tools are transforming protein structure prediction, headway is also being made in the less-crowded field of ribonucleic acid (RNA) structure prediction. A nucleic acid like deoxyribonucleic acid (DNA), but single stranded and with differing functions, RNA also plays a key role in cellular physiology. Various types of RNA perform myriad biological tasks, including messenger RNA (mRNA) that translates the information from DNA into proteins (Fig. 3).

Predicting the structure that a strand of RNA will fold into, based only on its genetic sequence, is a machine-learning challenge like the protein folding challenge, but with far fewer experimentally determined RNA structures to train machine learning models on; the tally of confirmed RNA structures available to science is less than one-hundredth that of proteins [14]. Nevertheless, researchers at Stanford University in Stanford, CA, USA, reported substantial progress in this area last year with their Atomic Rotationally Equivariant Scorer (ARES) system.

The Stanford researchers trained ARES using a machine learning approach with data comprised of the structural configurations of just 18 RNA molecules. Unlike AlphaFold's training on proteins, the ARES training incorporated no domain-specific information about how RNA molecules fold or behave but used merely the relative coordinates of the atoms in the RNA molecules. When given the genetic sequence of an RNA molecule with an unknown (to ARES) structure, the system uses an open-source RNA-

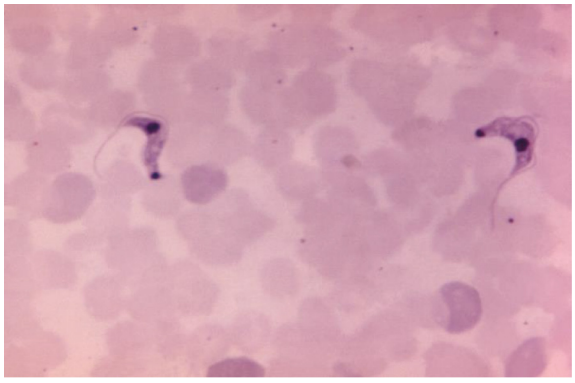


Fig. 2. This photomicrograph of a blood smear specimen reveals two *Trypanosoma cruzi* trypomastigotes, single-celled flagellated parasites that cause Chagas disease. DeepMind has released predicted structures of proteins in the proteomes of this parasite and other pathogens that cause neglected tropical diseases, which could help accelerate the development of more effective therapies. Credit: CDC/Myron G. Schultz (public domain).

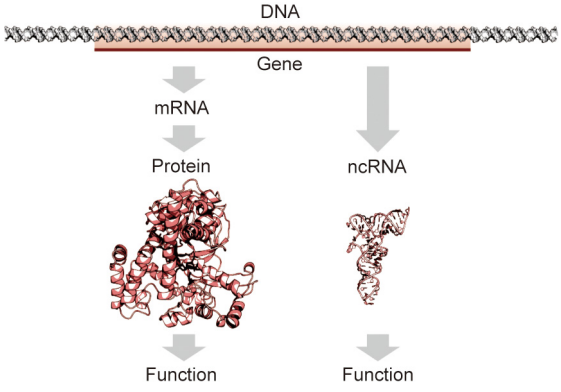


Fig. 3. While the DNA sequences of protein-coding genes are transcribed to mRNA, which is then translated into functional proteins, RNA-coding genes are directly transcribed into functional non-coding RNA (ncRNA). Understanding the folded structures of RNA, like understanding those of proteins, is important to understanding how these molecules function in both health and disease. Credit: Thomas Shafee (CC BY 4.0).

modelling tool called Rosetta FARFAR2 [13] to generate more than 1500 candidate structures for that RNA molecule. Based on its training, it then picks the candidate it deems closest to reality. ARES outperformed competing structure-prediction methods, according to a benchmarking of the predicted models in RNA-Puzzles, a CASP-like, blind RNA structure prediction challenge. The team published their results in *Science* in August 2021 [14].

"In structural biology, you can think of the atom as a fundamental machine-learning data type," said first author on the paper Raphael Townshend, who left Stanford University to become founder and chief executive officer of Atomic AI, a San Francisco-based biotech startup focused on using machine learning approaches to design new molecules and medicines. "We adapted machine learning models that we had successfully used in the protein space and applied them in the RNA space. And it worked beautifully," Townshend said. "It was a nice proof of the generalizability of machine learning."

ARES represents an improvement on existing RNA-structure prediction systems, but as professor of chemistry at the University of North Carolina at Chapel Hill. Weeks noted in a Perspective piece accompanying the *Science* article: "ARES is still short of the level consistent with atomic resolution or sufficient to guide identification of key functional sites or drug discovery efforts" [15].

Townshend, who had previously worked at DeepMind on the AlphaFold team, acknowledged this point. "The ARES network is the most accurate in the world, but it is only the first step on the road to rational drug discovery," he said. "However, it can immediately be used as a powerful screening tool, in conjunction with experiments." Townshend said he wants to do for RNA what has been done for proteins—provide a dramatic increase in accuracy over just a few years, powered by AI. It remains to be seen, however, whether the models can achieve such accuracy without incorporating domain-specific information about how RNA molecules behave.

Regardless, the success in protein structure prediction—and the growing arsenal of open-source tools—has been a boon for the RNA folding challenge. CASP15, which begins in May 2022, will expand its focus to include more structure prediction for RNA molecules. "We are adapting in accordance with the new excitements, as it were, and working with the RNA Puzzles team to bring in a bigger audience," said Moult. "Protein people are interested in moving into the RNA arena as well."

CASP15 will also increase its emphasis on predicting the structures of protein complexes. This is where the field is heading, said Xu, because proteins do not exist in isolation. "Proteins fashion themselves by interacting with other proteins and molecules, and I would say this is even more important than predicting the structure of single proteins. It is a fundamental problem with tremendous application in industry, particularly in drug design."

References

[1] DeepMind [Internet]. San Francisco: Twitter; 2022 Jan 28 [cited 2022 Jan 26]. Available from: https://twitter.com/DeepMind/status/1487021347565940738.

[2] Hassabis D. Putting the power of AlphaFold into the world's hands [Internet]. London: DeepMind; 2021 Jul 22 [cited 2022 Jan 26]. Available from: https://deepmind.com/blog/article/putting-the-power-of-alphafold-into-the-worlds-hands.

[3] Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, et al. AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res 2022;50(D1):D439–44.

[4] AlphaFold: a solution to a 50-year-old grand challenge in biology [Internet]. London: DeepMind; 2020 Nov 30 [cited 2022 Feb 4]. Available from: https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology.

[5] O'Neill S. Artificial intelligence cracks a 50-year-old grand challenge in biology. Engineering 2021;7(6):706–8.

[6] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature 2021;596:583–9.

[7] Tunyasuvunakool K, Adler J, Wu Z, Green T, Zielinski M, Žídek A, et al. Highly accurate protein structure prediction for the human proteome. Nature 2021;596:590–6.

[8] Hatch V. DeepMind and EMBL release the most complete database of predicted 3D structures of human proteins [Internet]. Hinxton: European Bioinformatics Institute; 2022 Jul 22 [cited 2022 Jan 26]. Available from: https://ebi.ac.uk/about/news/announcements/alphafold-database-launch/.

[9] DeepMind [Internet]. San Francisco: Twitter; 2021 Dec 9 [cited 2022 Jan 26]. Available from: https://twitter.com/DeepMind/status/1468945984378056707.

[10] Xu J, Wang S. Analysis of distance-based protein structure prediction by deep learning in CASP13. Proteins Struct Funct Bioinform 2019;87:1069–81.

[11] Wheeler RJ. A resource for improved predictions of *Trypanosoma* and *Leishmania* protein three-dimensional structure. PLoS ONE 2021;16(1):e0259871.

[12] Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science 2021;373(6557):871–6.

[13] Watkins AM, Rangan R, Das R. FARFAR2: improved *de novo* Rosetta prediction of complex global RNA folds. Structure 2020;28:963–76.

[14] Townshend RJL, Eismann S, Watkins AM, Rangan R, Karelina M, Das R, et al. Geometric deep learning of RNA structure. Science 2021;373(6558):1047–51.

[15] Weeks KM. Piercing the fog of the RNA structure-ome. Science 2021;373(6558):964–5.