REVIEW

The Rise of AI in Structural Biology

How recent algorithms will kick-start a new generation in biological research

Nicky Kromdijk

Abstract

In this review, I will explore how recent protein structure prediction methods that make use of deep learning will impact the field of structural biology. I will particularly search for purposes where these methods can be applied and what are their current pitfalls. After an intensive literature study, it became clear that accurate prediction algorithms are useful to solve experimental structures faster and help dividing proteins in functional domains. Structural predictions improve interpreting the molecular processes of (unknown) proteins and therefore provide insights into diseases and how to design potential treatments. That way, researchers can rapidly work out the structure of every protein in new and dangerous pathogens, speeding up the process of screening for drug targets. In addition to that, the open-source nature of these prediction methods enables scientists to continue the advances to create even more powerful software. Besides great advances, limitations of these deep learning structure prediction methods include that the predicted models only provide one conformational state of the protein and that no ligands are included in the model which is crucial data for inferring exact biological function and designing new drugs. Additionally, no folding information is present for the predicted structure. We should not state that these algorithms solved the folding problem, as they solved the prediction problem. Although scientists should remain critical of computational predictions, these technologies are still ground-breaking and will lead to more advances in the field of structural biology.

Keywords: CASP, Structure prediction, AI, AlphaFold2, RoseTTaFold

1 | LAYMAN'S SUMMARY

2021 was a great year for structural biology. Because in this year we saw the advances of recent protein structure prediction methods that make use of artificial intelligence (AI). It even became Nature Methods' Method of the Year. Researchers are so invested in predicting the three-dimensional (3D) structures of protein because the 3D structure is crucial for understanding their biological function, measuring the impact of human pathogenic mutations, and improving the design of new drugs. Thanks to The Critical Assessment of Techniques for Protein Structure Prediction (CASP), a structure prediction contest, computational predictions have been steadily improving over the past decades. A significant improvement in model accuracy was observed since recent contesters such AlphaFold, AlphaFold2, and RoseTTaFold started to make use of advanced deep learning. In 2020, for the first time in history, accuracies of predicted models nearly agreed with the experimentally solved structures, something that scientists were not sure to experience in their lifetime. In this review, I researched how these recent structure prediction methods will transform the field of structural biology. After an intensive literature study, it became clear that accurate prediction algorithms are useful to solve experimental structures faster because solving structures with conventional methods is time-consuming and expensive. The models may also improve dividing the protein into functional domains. Structural predictions help interpret the molecular processes of (unknown) proteins and therefore provide insights into how diseases arise and how to design potential treatments. That way researchers can rapidly work out the structure of every protein in new and dangerous pathogens, speeding up the process of testing for drug targets. In addition to that, the open-source nature of these prediction methods enables scientists to continue the advances to create even more powerful software. Besides great advances, limitations of these deep learning structure prediction methods include that the predicted models only provide one conformational state of the protein, even though a protein can have multiple conformations such as active or inactive. There are also no ligands (small molecules, cofactors, DNA, or metals) included in the model. This lacking interaction data is crucial for observing protein function in complexes and designing new drugs. Furthermore, for the predicted structure, no information is present of how the protein has folded leading to its end-state model. We should not state that these deep learning algorithms solved the folding problem. How exactly proteins fold is still unclear. They did however solve the prediction problem. Although scientists should remain critical of computational predictions, these technologies are still ground-breaking and will lead to more advances in the field of structural biology. We will have to see what these structure prediction methods are capable of in the coming years.

2 | INTRODUCTION

When the first atomic-resolution protein structures became available in 1960, an immediate question emerged that today's scientists are still trying to answer. How does the primary sequence (1D) of a protein determine how a protein folds into its three-dimensional (3D) structure (**Figure 1A**)? This problem is also known as the protein folding problem.[1] Research shows that understanding the 3D structure of proteins is crucial to comprehend their biological function, measure the impact of human pathogenic mutations, and improve the design of new drugs. One of the reasons why this question is so complex is that a protein can have countless possible protein conformations. With experimental work alone, this could never be solved. For that reason, researchers started using computational prediction methods to predict a protein's 3D structure from its amino acid sequence.[2]

To promote solving protein structures computationally, a community-wide contest named The Critical Assessment of Techniques for Protein Structure Prediction (CASP) was introduced, taking place biannually since 1994. In this competition, around 100 research groups worldwide try to predict the unknown 3D structures of proteins using their own structure prediction method. The structures of these proteins will be experimentally solved using current methods such as X-ray crystallography, cryogenic electron microscopy (cryo-EM), and nuclear magnetic resonance spectroscopy (NMR), but not made public yet. CASP predictions are evaluated by a GDT_TS score (Global Distance Test - Total Score) which measures the accuracy of a structural prediction compared to the experimentally solved structure with a value between 0 to 100, from inaccurate to perfect respectively. The team that submits the predictions with the highest accuracies wins the contest. Thanks to CASP, computational predictions have been steadily improving over the past decades. Especially during CASP13 (2018), there was a turning point. The first implementation of high-performance deep-learning systems by multiple participants under which AlphaFold[3], submitted by the London-based and Google-owned AI research group. Two years later during CASP14 (2020), DeepMind presented its new and improved AlphaFold2 with remarkable accuracy close to experimentally solved structures. At the same time, an academic team led by David Baker developed RoseTTaFold[4] which performs similarly to AlphaFold2. Recent technologies that make use of advanced deep learning are astonishing at predicting the 3D structures of a single protein using only the primary sequence which seemed to be a deemed problem a few years ago. It is expected that the accuracies of these advanced prediction methods will keep increasing.

However, today's structure prediction methods did not just fall out of the sky. Methods like AlphaFold, AlphaFold2, and RoseTTaFold were built on the foundation set by various researchers over the last decade. Recent prediction methods are co-evolution-dependent which means that they first create a multiple sequence alignment (MSA) from aligning the protein sequence of interest to homologs. Then, evolutionary connections are extracted from the MSA by looking at amino acids that co-evolved over time. Co-evolving amino acids are residues that mutated during evolution due to other mutations in the sequence, indicating that these residues are potentially physically close. Protein structure prediction methods use this co-evolution information to create a contact matrix (**Figure 1B**). From the contact data, the 3D coordinates of the protein can be obtained and then be fold according to physical laws.[5] In CASP11 (2014) these coevolution-based methods were surpassing other approaches, and in CASP12 (2016) these methods were brought to a higher level when coupled to machine learning. Besides that, CASP12 saw an increased performance as well due to the growing number of

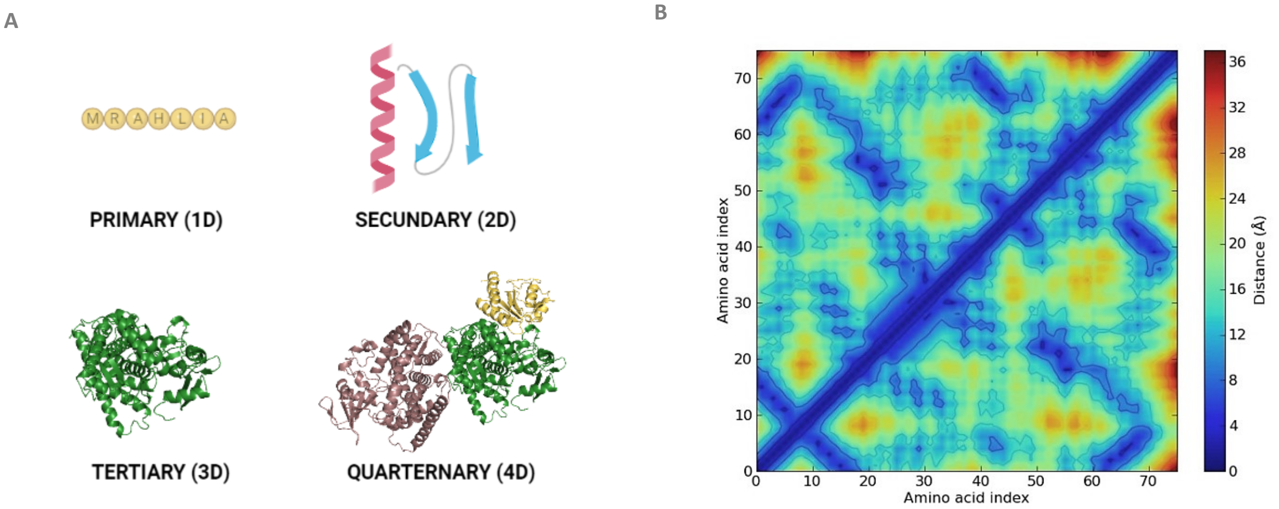

FIGURE 1: **Protein structure prediction.** (A) Proteins have four structural levels. The primary structure of a protein consists of its amino acid sequence, the secondary structure consists of repeated arrangements in a protein's amino acid sequence such as α-helices and β-sheets. The tertiary structure represents a 3D shape of a single protein. In protein structure prediction researchers try to build the tertiary structure of a protein from its primary structure. The quaternary structure represents a complex of two or more proteins. (B) An example of a contact matrix, also named distance map. The colors of the pixels correspond to the distance in Ångström (Å) between residues in a protein sequence. In this case, the bluer the pixel, the closer the pair, and the redder the pixel, the bigger the distance. The contact matrix is used to predict the tertiary structure of the protein. (Rafferty et al, 2020)[a]

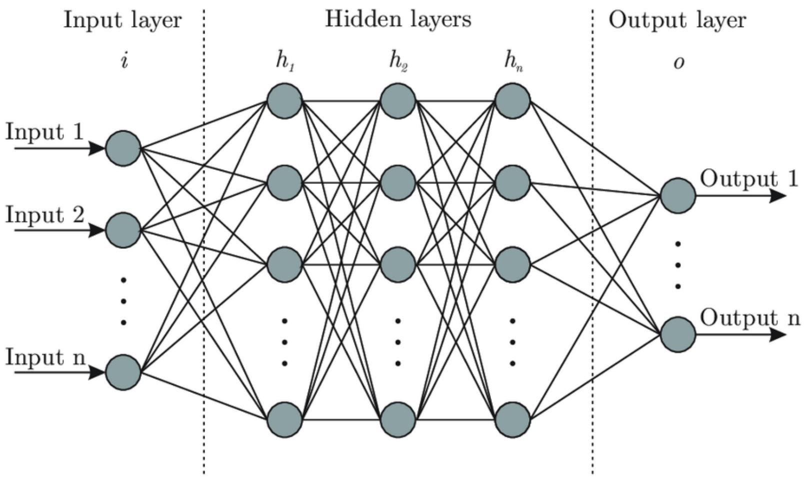

FIGURE 2: The architecture of a neural network. Inside the neural network, data flows from the input layer through the hidden layers to the output layer. All layers can have different functions on the transmitting data. The edges that connect the nodes, can contain a weight. (Bre et al, 2018)[b]

available sequences and therefore more evolutionary information could be obtained.[6] Then, in CASP13, incredible performance was observed in one particular category of CASP namely "template free" modeling, which was extremely difficult for years. As mentioned, this performance was motivated by the implementation of advanced deep learning techniques. The algorithms are able to predict distances between amino acids when no structural template is available for the target sequence.[7] In CASP14, predicted model accuracy kept increasing for the first time above a GDT_TS >90, an accuracy that is considered very close to the experimentally solved structure. Some scientists even say that in a sense the protein folding problem is solved.[8]

AlphaFold and RoseTTaFold make use of advanced deep learning methods. Deep learning models are a type of artificial neural networks (**Figure 2**). A neural network is a machine learning method that can be extremely useful for classification, clustering, pattern recognition, and making predictions. A neural network consists of connected nodes (artificial neurons) which can transmit a signal to other nodes. The part that connects nodes is termed an edge. A weight can be placed on these edges while the neural network is trained. The neurons are divided into multiple layers, starting with the input layer, and ending with the output layer, which can have a diverse function on the transmitting data. Deep learning models are neural networks with multiple complex layers and are useful for handling big data. Neural networks can work out non-linear challenges, which make them applicable to all sorts of problems, therefore also to protein structure prediction.[9]

We have seen that the accuracy of these deep learning structure prediction methods can touch the accuracy of experimentally solved structures (GDT_TS >90). Some believe that the protein folding problem is finally solved and therefore the field of structural biology will change drastically. In this report, I will review how these recent structure prediction methods will transform the field of structural biology and the future of its research. I will particularly search for purposes where these methods can be applied and what are their current pitfalls.

3 | RECENT DEEP LEARNING STRUCTURE PREDICTION METHODS

3.1 | Single-chain prediction algorithms

AlphaFold

In 2018, CASP13 saw the first-time implementation of high-performance deep-learning systems by multiple contesters, under which DeepMind's entry AlphaFold. DeepMind is a London-based and Google-owned company specialized in AI. Their AlphaFold framework consists of a neural network that can be trained to predict the distance between amino acids, which contains more structural information than contact interactions.[10] AlphaFold neural networks stand out due to their hundreds of layers (deep learning), they can model long-range interactions over the complete amino acid sequence of a protein, and they perform under reduced memory and processing requirements so that they can be faster trained. The neural network methods that AlphaFold uses are not novel; however, it is the first time they are used like this for biological purposes in protein structure prediction. One other main idea that DeepMind came up with it is that AlphaFold's neural network could be simply optimized using a minimization algorithm called a gradient descent.[5] A few predicted models by AlphaFold showed high accuracy (GDT_TS >80). The average GDT_TS score of AlphaFold's models was 64.4 and the algorithm placed first in CASP13.[11]

AlphaFold2

The first AlphaFold algorithm presented in CASP13 was able to predict nearly correct structures for some proteins by presenting 15 models with a GDT_TS score of >80. However, no entry, including AlphaFold was able to predict the exact atomic locations within the protein. With these challenges in mind, at CASP14, DeepMind presented their new and improved algorithm AlphaFold2 which differs completely from its predecessor. The new neural network processes the MSA and pairwise representations (such as templates) simultaneously using a particular block of the network named the Evoformer. The main function of the Evoformer is to handle a protein structure prediction as a graph interference problem. This makes sure that long-range residue interactions in the sequence are managed. A new neural network module was designed to build the protein structure instantly including the atomic information. Then, the complete system is executed several times to refine the predictions. At last, to relax the predicted structures, a gradient descent is used.[3,12] AlphaFold2 performed incredibly at CASP14 and placed first. The average GDT_TS score of AlphaFold2's models was 88. Some of the models even scored a GDT_TS of >90, a number that is competitive with the accuracy of an experimental structure, and was never observed in CASP before.[13] A problem that existed for many years seems solved.

RoseTTaFold

Around the time of CASP14, an academic team from the University of Washington led by David Baker developed an AI algorithm as well named RoseTTaFold, which performs similarly to AlphaFold2. RoseTTaFold also makes use of deep learning and can predict a protein's structure in under ten minutes while using limited information. This gives them a slight advantage (depending on the research goals) as AlphaFold2 has a longer prediction time. RoseTTaFold's system consists of a three-track neural network. It simultaneously considers information about sequence (patterns), distance (how residues interact), and coordinates (possible 3D structure). Using RoseTTaFold researchers can quickly study functions of proteins where no structure was available before.[4] The average GDT_TS score of RoseTTaFold's models was 66.9 and placed second in CASP14.[13]

3.2 | Protein Complex Prediction Algorithms

The majority of proteins do not work alone. Biological processes happen mostly due to the interaction between two or more proteins that form a complex. The binding partners might be crucial for forming the quaternary structure (4D) of the proteins. Besides the remarkable accuracy of Alphafold2 for predicting the structure of single chains, handling complexes consisting of multiple chains is still very difficult. The algorithm was initially not designed to predict protein complexes. However, very recent studies showed that a special-designed AlphaFold2 model, named AlphaFold-Multimer, is in fact able to predict the structures of complexes as well.[14,15] While this is an excellent side effect of the AlphaFold2 algorithm, other methodologies were already focused on predicting complexes, such RoseTTaFold. The co-evolutionary information that is used to predict residue distances within one sequence can also be used to predict protein-protein interactions. RoseTTaFold can besides single-chain proteins also be used for predicting complexes of proteins. This is very valuable because protein function does not only rely on structure but also protein-protein interactions.[16]

4 | APPLICATIONS

4.1 | The AlphaFold Protein Structure Database

To make all the predicted structures by AlphaFold widely and freely accessible, DeepMind collaborated with the European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI). This collaboration established the AlphaFold Protein Structure Database (AlphaFoldDB). After AlphaFold's first release of structures, 360.000 proteins of nearly the complete human proteome and the ones of other model organisms were published in the database. After the second release of AlphaFold, many manually build proteins from UniProt were published. The current AlphaFoldDB consists of over 800.000 protein structures, and the goal is to increase that number to over 100 million this year. To emphasize how large this amount is, the Protein Data Bank (PDB) currently consists of 186.000 protein structures. This large number of high-accuracy models will have a range of biological impacts, from speeding up experimental work to providing more data to study diseases and eventually developing structure-based therapies [17,18]

4.2 | Helping Structure Determination by Experimental Methods

The current methods for determining atomic structures include X-ray crystallography, nuclear magnetic resonance spectroscopy (NMR), and cryogenic electron microscopy (cryo-EM). Even though these methods contributed to solving many proteins, they consist of time-consuming meticulous laboratory work and require expensive equipment. How will recent structure prediction methods change the way scientists will use these current methods? Firstly, using deep learning tools, experimental structures can be solved faster and simpler. Not by replacing the current methods but as an addition e.g., when building a protein's structure with X-ray crystallography, all information about the diffracted X-ray waves is needed, which is the amplitude and the phase. While the amplitude can be easily measured, the phase cannot. This is called the phase problem. There are methods to overcome this problem e.g., by applying molecular replacement. This method requires a template for a protein that has over ~25% sequence identity and a backbone RMSD (Cα atoms) of less than ~2.0Å between the template model and the crystal structure that you are building. Using molecular fragment programs, the template model can be lined up with the crystal structure which provides the phase information that was missing before. Because of the expanding number of structures in the AlphaFoldDB, there are now many more available high-accuracy models to perform molecular replacement with. Even for low-quality or low-resolution data, structures can be solved, which would otherwise be very difficult (and therefore time-consuming) or even impossible.[19]

4.3 | Constructing Protein Domains

The residues in the predicted models are labeled with a certain confidence. High confidence resulted in models of higher accuracies. It was believed that low confidence meant the failure of the prediction tool. However, low confidence of a region can also mean that there is disorder in the protein.[12] These disordered protein domains are able to shift from a disordered to an ordered state by binding to other ligands and proteins and can be essential for intracellular signaling.[20]

An advantage of using the predicted models is that it would be easier to divide the protein into its domains, an issue that experimental scientists faced. Understanding the topology of the protein and its domains can lead to a better way to set up structure determination experiments. For some current methods such as cryo-EM, this may help overcome tiresome model-building steps.[21]

4.4 | Function and Disease

It is known that structural information of a protein is crucial for research, however, a 3D structure on its own does not automatically deliver its complete biological function. Structure prediction algorithms are not able to predict protein function yet. But, the 3D structure does assist to infer the biological function of unknown proteins.[22] Inferred functions from sequences and structures are determined by evolutionary relationships. With the increase of new accurately predicted structures researchers can look further back in time to find the farthest relative possible and might infer the function from those relatives. 3D structures can also help interpret the molecular processes of a protein and therefore not only provide insights into how mutation of a protein might result in disease but also how to design a potential treatment. For example, if researchers can accurately predict the spike proteins of viruses, much quicker structural analysis can be performed on rising viruses in potential future pandemics so that vaccines can be developed quicker.

Structural models can also be useful to help identify pathogenic variants in the human genome by computational mutagenesis. By entering a sequence with a mutation in the structure prediction software, it can be observed which mutations are more likely to result in a normal structure and which result in a mutated protein that can lead to disease. Additionally, with computational mutagenesis insights can be gained into the minimal residues that are needed for folding which can give researchers a better understanding of the relationship between the sequence and structure of a protein.[23,24]

4.5 | Drug discovery

Many researchers predicted recent AI applications would benefit drug design. Because of the expanding databases, there is an enormous increase in human proteins as well as protein structures from various pathogens. This way, more potential drug targets can be analyzed, and the chance of cross-reactivity can be better assessed. Researchers can rapidly work out the structure of every protein in new and dangerous pathogens, speeding up the process of screening for drug targets. Target identification is the first stage in drug design which could take multiple years.[17]

4.6 | Open source and open science

One of the great benefits of these recent tools like RoseTTaFold and AlphaFold2 as well as the database AlphaFoldDB, is that they are open source. Open source means that the software, models, or data are entirely transparent, of high quality, and free to use (no license).[25] This way, anyone can continue the prior work to build even better tools. Without open-source data, such as from the Protein Data Bank (PDB), recent AI tools would never been able to train their algorithm this good.[26]

A group of scientists developed an open-source tool for fast and easy protein structure- and complex prediction named ColabFold. This tool unites fast homology search with either the AlphaFold2 or RoseTTaFold algorithm. In collaboration with Google Colab, ColabFold is easy-to-use in the form of Jupyter Notebooks.[27] The open-source software of ColabFold is accessible at github.com/sokrypton/ColabFold.

Another important concept in science is “open science” which means that together as a community scientists develop concepts, methods, and applications. Additionally, there is invested in education for all, and everyone should be able to reproduce results. According to the EMBL, open science is crucial for reliable, transparent, and more inclusive research.[25] Even though the open-source nature of recent AI tools provides many benefits, it should be noted that they are not always open science. This is because DeepMind released the AlphaFold2 software in such a way that anyone could use it, but not all neural network-specific parameters were released for commercial use. David T. Jones and Janet M. Thornton published in a review that “it’s somewhat debatable whether the weights of a neural network can be protected as intellectual property, and it’s also worth noting that DeepMind is not alone in using this possible loophole to prevent commercial usage of supposedly open-source machine-learning software. The RoseTTaFold software from David Baker’s laboratory at the University of Washington has similar restrictions on the use of its neural network weights. As we say, open-source certainly, but not truly open science.”[28]

5 | LIMITATIONS

Besides great advances, deep learning structure predictions methods also have their limitations. An important one is that these methods only predict one protein state even though a protein can exist in multiple conformational states. The disadvantage of only one static protein structure is that it remains unclear what the function of other conformational states is, if the predicted state is in its active or inactive form, and what are all the protein's possible interaction partners.[22,29] Modelling flexible proteins can therefore be a problem. Also, large proteins consisting of many domains are not modeled very accurately.[23] The AlphaFoldDB does also not contain models for protein sequences that are longer than 2700 residues. This means that there are still no structures for 207 large and important human proteins.[20]

Another limitation is that the predicted models do not include ligands such as small molecules, cofactors, DNA, or metals. Additionally, protein-protein prediction is still very novel. Without any interaction data, it can be very hard to directly infer the biological function of a protein.[23] Additionally, one could wonder how accurate it would be to predict the structure of a single protein that is out of its complex or how accurate a predicted binding site is if there is no ligand information.[30]

It should be noted that even though recent algorithms can predict structures agreeing with the experimentally solved structures, they did not solve the problem of how a protein would fold in solution or inside a cell. Instead, it delivers a possible end-state that the algorithm has learned from the outcomes of folding at the amino acid residue contact level and can thus correctly predict proteins that would never exist without being in their complex.[2] Also, the folding process of this possible end-state of a protein remains unclear. There has been performed a study by Outeiral et al. where they examined if recent structure prediction tools (including AlphaFold2 and RoseTTaFold) provide an understanding of how proteins fold. Their results showed that their ability to predict the folding process is in most cases worse than a sequence-agnostic linear classifier which implies that recent structure prediction methods do not yet provide insights into the concepts that determine protein folding.[31]

6 | DISCUSSION & CONCLUSION

Although the great advances in structure prediction methods, experimental scientists do not have to worry about losing their jobs, instead it will help their work by overcoming tiresome model-building steps with the use of phasing and clearer domain separation.

Even though the accuracies of computational predictions do sometimes come near the accuracy of experimentally solved structures, experimental work is still crucial for understanding exact protein function and interaction with ligands and other proteins. Since the predicted models do not include ligands there is no interaction data which can make it difficult to use these predictions to infer function and improve the drug-designing process. However, there have been successful studies showing that AlphaFold2 is able to perfectly predict a single hemoglobin chain even though it was lacking crucial binding partners for folding,[2] and that it is also able to perfectly predict a zinc binding site in a peptidase domain without including the zinc ion in the structural prediction.[30] High-accuracy models will speed up the first stage of drug discovery and lead to more focused experimental work.

Methods as AlphaFold2 and RoseTTaFold did not solve the problem of how a protein would fold in solution or inside a cell. Instead, it delivers a possible end-state that the algorithm has learned and can thus correctly predict proteins that would never exist without being in a complex. Moreover, these prediction tools are not able to provide insights into the folding process. For that reason, we shouldn't state that these algorithms solved the folding problem because how exactly a protein folds remains an unanswered question in biology. Nevertheless, what these prediction tools did solve, especially AlphaFold2, is the prediction problem. Researchers are now able to predict a protein’s 3D structure from its amino acid sequence.

Even though the protein folding problem is not solved, it may help for future research on how to do it. Recent years saw incredible progress, but the exact link between protein sequence, folded structure, and its function is not completely understood yet. One of the main areas of interest for now is predicting the structure of multiple interacting proteins as a complex. This is important for inferring biological function and therefore drug development.[26]

To answer the research question of this review how recent structure prediction methods will transform the field of structural biology and the future of its research, I can state that for experimental scientists, deep learning structure prediction methods are able to speed up a lot of their work, but it will not be completely replaced by computational methods. For studying diseases and finding new drugs, some stages can be accelerated and there is an immense increase of structural data which will kick-start even more research. However, structural predictions alone are not enough to change the way we design drugs. Advanced deep learning algorithms are amazing at predicting protein’s 3D structures from their amino acid sequence, but a lot more must be done before we can see the greater applications of these advances. We are now at the point when we will have to wait for these methods to disentangle their full potential, e.g. protein complex prediction, and see what these structure prediction methods will be capable of in the coming years.

REFERENCES

1. Dill KA, Ozkan SB, Shell MS, Weikl TR. The Protein Folding Problem. *Annu Rev Biophys*. 2008;37:289-316. doi:10.1146/annurev.biophys.37.092707.153558

2. AI revolutions in biology. *EMBO reports*. 2021;22(11):e54046. doi:10.15252/embr.202154046

3. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021;596(7873):583-589. doi:10.1038/s41586-021-03819-2

4. Baek M, DiMaio F, Anishchenko I, et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*. 2021;373(6557):871-876. doi:10.1126/science.abj8754

5. AlQuraishi M. AlphaFold at CASP13. Valencia A, ed. *Bioinformatics*. 2019;35(22):4862-4865. doi:10.1093/bioinformatics/btz422

6. Schaarschmidt J, Monastyrskyy B, Kryshtafovych A, Bonvin AMJJ. Assessment of contact predictions in CASP12: Co-evolution and deep learning coming of age. *Proteins: Structure, Function, and Bioinformatics*. 2018;86(S1):51-66. doi:10.1002/prot.25407

7. Croll TI, Sammito MD, Kryshtafovych A, Read RJ. Evaluation of template-based modeling in CASP13. *Proteins*. 2019;87(12):1113-1127. doi:10.1002/prot.25800

8. Callaway E. ‘It will change everything’: DeepMind’s AI makes gigantic leap in solving protein structures. *Nature*. 2020;588(7837):203-204. doi:10.1038/d41586-020-03348-4

9. Abiodun OI, Jantan A, Omolara AE, Dada KV, Mohamed NA, Arshad H. State-of-the-art in artificial neural network applications: A survey. *Heliyon*. 2018;4(11):e00938. doi:10.1016/j.heliyon.2018.e00938

10. Senior AW, Evans R, Jumper J, et al. Improved protein structure prediction using potentials from deep learning. *Nature*. 2020;577(7792):706-710. doi:10.1038/s41586-019-1923-7

11. Groups Analysis - CASP13. Accessed February 6, 2022. https://predictioncenter.org/casp13/groups_analysis.cgi

12. Jumper J, Evans R, Pritzel A, et al. Applying and improving AlphaFold at CASP14. *Proteins: Structure, Function, and Bioinformatics.* 2021;89(12):1711-1721. doi:10.1002/prot.26257

13. Groups Analysis - CASP14. Accessed February 6, 2022. https://predictioncenter.org/casp14/groups_analysis.cgi

14. Akdel M, Pires DEV, Porta Pardo E, et al. *A Structural Biology Community Assessment of AlphaFold 2 Applications*. Biophysics; 2021. doi:10.1101/2021.09.26.461876

15. Evans R, O’Neill M, Pritzel A, et al. *Protein Complex Prediction with AlphaFold-Multimer*. Bioinformatics; 2021. doi:10.1101/2021.10.04.463034

16. Pennisi E. Protein structure prediction now easier, faster. *Science*. Published online July 16, 2021. doi:10.1126/science.373.6552.262

17. Subramaniam S, Kleywegt GJ. A paradigm shift in structural biology. *Nat Methods*. 2022;19(1):20-23. doi:10.1038/s41592-021-01361-7

18. 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 Research*. 2022;50(D1):D439-D444. doi:10.1093/nar/gkab1061

19. Taylor G. The phase problem. *Acta Crystallogr D Biol Crystallogr*. 2003;59(Pt 11):1881-1890. doi:10.1107/s0907444903017815

20. David A, Islam S, Tankhilevich E, Sternberg MJE. The AlphaFold Database of Protein Structures: A Biologist's Guide. *Journal of Molecular Biology*. 2022;434(2):167336. doi:10.1016/j.jmb.2021.167336

21. Jumper J, Hassabis D. Protein structure predictions to atomic accuracy with AlphaFold. *Nat Methods*. 2022;19(1):11-12. doi:10.1038/s41592-021-01362-6

22. Skolnick J, Gao M, Zhou H, Singh S. AlphaFold 2: Why It Works and Its Implications for Understanding the Relationships of Protein Sequence, Structure, and Function. *J Chem Inf Model*. 2021;61(10):4827-4831. doi:10.1021/acs.jcim.1c01114

23. Thornton JM, Laskowski RA, Borkakoti N. AlphaFold heralds a data-driven revolution in biology and medicine. *Nat Med*. 2021;27(10):1666-1669. doi:10.1038/s41591-021-01533-0

24. Zhang Y, Li P, Pan F, et al. Applications of AlphaFold beyond Protein Structure Prediction. Published online December 13, 2021:2021.11.03.467194. doi:10.1101/2021.11.03.467194

25. Lippert J, Burghaus R, Edginton A, et al. Open Systems Pharmacology Community—An Open Access, Open Source, Open Science Approach to Modeling and Simulation in Pharmaceutical Sciences. *CPT: Pharmacometrics & Systems Pharmacology*. 2019;8(12):878-882. doi:10.1002/psp4.12473

26. Callaway E. DeepMind's AI for protein structure is coming to the masses. *Nature*. Published online July 15, 2021. doi:10.1038/d41586-021-01968-y

27. Mirdita M, Ovchinnikov S, Steinegger M. *ColabFold - Making Protein Folding Accessible to All*.; 2021. doi:10.1101/2021.08.15.456425

28. Jones DT, Thornton JM. The impact of AlphaFold2 one year on. *Nat Methods*. 2022;19(1):15-20. doi:10.1038/s41592-021-01365-3

29. Bershtein S, Kleiner D, Mishmar D. Predicting 3D protein structures in light of evolution. *Nat Ecol Evol*. 2021;5(9):1195-1198. doi:10.1038/s41559-021-01519-8

30. Golinelli-Pimpaneau B. Prediction of the Iron–Sulfur Binding Sites in Proteins Using the Highly Accurate Three-Dimensional Models Calculated by AlphaFold and RoseTTAFold. *Inorganics*. 2022;10(1):2. doi:10.3390/inorganics10010002

31. Outeiral C, Nissley DA, Deane CM. Current protein structure predictors do not produce meaningful folding pathways. Published online September 20, 2021:2021.09.20.461137. doi:10.1101/2021.09.20.461137

FIGURES

a. Rafferty B, Flohr ZC, Martini A. Protein Contact Maps (2020) https://nanohub.org/resources/contactmaps. doi: 10.21981/02DQ-MT84

b. Bre F, Gimenez JM, Fachinotti VD. Prediction of Wind Pressure Coefficients on Building Surfaces Using Artificial Neural Networks. *Energy and Buildings* 158 (2018): 1429–41. https://doi.org/10.1016/j.enbuild.2017.11.045.