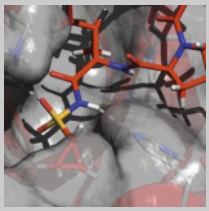
doi:10.2533/chimia.2022.364 Chimia 76 (2022) 364–366



Medicinal Chemistry and Chemical Biology Highlights

Division of Medicinal Chemistry and Chemical Biology
A Division of the Swiss Chemical Society

AlphaFold: Deep Learning, Drug Discovery and the Protein Structure Revolution

Christopher M. Baker* and Alessio Atzori

*Correspondence: Dr. C. Baker, E-mail: chris.baker@syngenta.com
Syngenta, Jealott's Hill International Research Centre, Bracknell, RG42 6EY, UK,

Keywords: Artificial intelligence · Deep learning · Protein structure prediction · Structure-based drug design

The protein folding problem began with Christian Anfinsen, and his observation that a protein's structure is determined entirely by its amino acid sequence.[1] Aside from earning Anfinsen a Nobel Prize, this realization had obvious and hugely significant implications: if we could understand how sequence encoded structure, we would know the three-dimensional structure of any protein as soon as we determined its sequence. The challenge, though, was that Anfinsen's simple statement masked a problem of almost unimaginable complexity. A typical protein may have 10^{300} possible conformations: how could we possibly identify which of those is the true folded conformation? Simply calculating every conformation of one protein and evaluating their relative energies, for example, would take longer than the age of the universe. And that is before we even ask whether we can accurately calculate those relative energies.

The difficulty of the problem did not, of course, dissuade researchers from trying to solve it, and in 1994 the community founded the Critical Assessment of protein Structure Prediction (CASP) experiment.[2] This biennial challenge allowed scientists to test their protein structure prediction methods against approximately 100 experimentally determined but as yet unpublished protein structures. For 20 years, CASP recorded a steady, incremental improvement in the performance of protein structure prediction methods: by 2016, the best performing method scored 41 out of 100 in a measure assessing similarity between predicted and experimental structures.[3] In 2020, AlphaFold, a machine-learning based method from Google-owned DeepMind, predicted 2/3 of structures with a score of more than 90, indicating accuracy equivalent to the experimental structures, with an overall average score of 87.[3] For practical purposes, DeepMind had solved the protein folding problem. Forbes magazine described it as "the most important achievement in AI – ever"[4] and many scientists saw huge and far reaching implications.

"This will change medicine." One CASP judge was quoted as saying, "It will change research… It will change everything."[3]

AlphaFold's success – and the rapid improvement in protein structure prediction after a period of incremental advances – is built on the use of deep learning methods. Where machine learning refers to any approach in which a computer attempts to learn from data without a reliance on human-coded rules, deep learning is a subset of machine learning that employs methods inspired by the architecture of a human brain. Neural networks with many layers allow models to extract features from the data in one layer, and then combine those features into higher order features in subsequent layers, steadily building a more complex description of the information contained in the data. Though powerful, deep learning methods are also demanding: they require large data sets for training, and depend on high performance computing to such an extent that they were not practical until around 2010.

The 2020 version of AlphaFold (also known as AlphaFold2,[5] to differentiate it from the original method used for the 2018 CASP[6]), takes as a central notion the idea that, in a protein, neighboring residues can be treated as nodes that are connected by edges to define a 'spatial graph'. In this version of AlphaFold, a neural network system trained on publicly available protein structures and protein sequences attempts to interpret the spatial graph of a protein.

To perform this task, AlphaFold uses three specific inputs: multiple sequence alignments (MSA); evolutionarily related sequences, and a representation of all the amino acid residue pairs in the input primary sequence. The main body of the network then employs transformers, models that identify the most important parts of the input data, to build a picture of the interrelationships between the protein sequences and template structures. It then iteratively improves this picture by updating the information from both the evolutionarily related sequences and the amino acid residue pairs, before finally producing a 3D model.

The deep learning methods employed by AlphaFold are state of the art, and the code itself is a fantastic feat of software engineering, but it is important to realize that artificial intelligence alone, no matter how sophisticated, could never have solved the protein folding problem. AlphaFold only works because its developers had access to the PDB: a source of high-quality data that they could use to train their algorithms. In fact, as early as 2005 the PDB contained an essentially complete library of single domain protein structures.[7] AlphaFold's success comes from the ability of the deep learning algorithms to robustly connect input sequences to the most appropriate template structures, and then also to learn how individual residues pack together in space.

Beyond the scientific achievement of AlphaFold, a key issue in determining the likely impact in drug discovery is the accessibility of the method. And for the user, running AlphaFold is incredibly simple. The only input required is a sequence of amino acid residues. It is far simpler to run than traditional homology modelling, and takes computational protein structure prediction from being an expert activity to something that anyone can access (though the interpretation of the resulting structures still benefits from expert input). Beyond this, DeepMind have released the AlphaFold code under an open source license, meaning that anyone can modify and improve it, and the community has already used this to produce improved versions for the prediction of multimeric complexes. Initially, there was some uncertainty about whether the model parameters could be used for commercial applications,[8] but a recent relaxing of the license conditions appears to have rendered the method truly open for all.[9]

Can you show us your Medicinal Chemistry and Chemical Biology Highlight?
Please contact: Prof. Dr. Kathrin Lang, Dept. of Chemistry and Applied Biosciences, ETH Zurich, Vladimir-Prelog-Weg 1-5/10, CH-8093 Zurich, E-mail: katrin.lang@org.chem.ethz.ch

DeepMind have also partnered with the European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI) to build a freely available database of predicted structures. The initial release included structures for the entire human proteome,[10] as well as the proteomes of model organisms, totaling 350,000 proteins – more than double the number in the PDB. By the middle of 2022, they aim to expand this database to cover 130 million protein structures: the structure of practically every protein already sequenced.[11,12] There are some concerns about the practicalities of this approach: are all structures updated every time a new version of AlphaFold is released? Do we create new databases for the new structure-prediction methods that will follow AlphaFold? What is the environmental cost of running so many calculations? But there are also advantages, especially to researchers working without access to high performance computing, and researchers working in industries, such as agrochemistry or animal health, that consider a wide range of different (and often poorly-studied) organisms.

As this discussion indicates, AlphaFold represents a huge jump forward in protein structure prediction, often producing structures comparable to those obtained experimentally, and in only a year has led to an explosion in the number of protein structures available. It is important to acknowledge though, that the method does come with limitations, and that there are a number of important challenges that remain to be overcome. Perhaps the most significant for those of us wishing to use AlphaFold structures for drug discovery, is that they are often incomplete. They lack any non-amino acid components, including cofactors or ions that can play significant roles in the protein's function and inhibition. The method as originally published also applied only to single chains, though has now been expanded to cover multimeric proteins,[13] and cannot model binding partners such as DNA, RNA or ligands.

A further limitation comes from the fact that AlphaFold has been trained using the data present in the PDB, which is biased towards well-ordered, folded proteins. AlphaFold cannot be used to say anything meaningful about the structure of intrinsically disordered proteins or disordered regions – amino acid sequences that do not adopt a single well-defined secondary or tertiary structure in their native state. This is a significant limitation given that 50% of eukaryotic proteins are estimated to contain disordered regions (but could be converted into an advantage, if AlphaFold can be used as a tool to reliably predict regions that do not fold). Similarly, AlphaFold is unable to predict the impact of single mutations or post-translational modifications that change the protein structure.

The final limitation of significance for drug discovery, is that only a single protein structure is predicted per sequence (in effect, the one most likely to be found in the PDB). In reality, proteins are dynamic objects that occupy an ensemble of different conformational states, and in many cases two (or more) significantly different structures of the same protein may play important functional roles, and the correct choice of structure may be important to understand the binding of drug molecules.

To mitigate the impact of some of the limitations described above, AlphaFold does provide metrics on the reliability of the predicted model: the predicted Local Distance Difference Test (pLLDT) and the predicted aligned error (PAE). The former is used to identify domains and possible disordered regions, and to assess confidence within a domain, with the latter used to measure confidence in the relative positions of residue pairs, and to assess relative domain positions in a multidomain protein.

Taking all this information together allows us to formulate an answer to the key question: what, overall, does AlphaFold mean for drug discovery? Structure-based drug design is a well-established component of drug discovery programs. In this approach, the structure of a protein target in complex with a ligand is used to rationalize and optimize observed binding affinity. It follows, logically, that the more, earlier, or better structural information that you have about a given target, the more effectively you can apply structure-based design. One would expect, therefore, that AlphaFold would have a significant impact on this part of a project. A key barrier to this impact, though, is the fact that AlphaFold provides you with no information about how a ligand binds to a protein. For less challenging cases, where the binding site can be identified and understood by analogy to known structures, using AlphaFold structures for drug design becomes more realistic, and there is at least one report of a novel inhibitor being designed using an AI-driven workflow that combined an AlphaFold structure with a generative chemistry engine.[8] But for a truly novel protein it is not always possible to identify, without additional information, where the binding site is located, let alone the binding mode of the compound(s) of interest. In these cases, experimental structure determination will still be essential, and AlphaFold will be of greatest utility in supporting that experimental work, by helping to evaluate proposed protein constructs or to interpret experimental data.

If DeepMind, or anyone else, could develop an algorithm that could accurately predict the structure of a protein in complex with any arbitrary ligand, that would be the point at which computational structure prediction might genuinely challenge experimental structure determination as the primary tool for use in drug discovery projects. But this is a much more difficult problem. Ligand chemical space is far more heterogenous than amino acid space; there are far fewer ligands than amino acids in the PDB, and the quality of ligand structures is often more uncertain than that of the protein itself. Almost certainly, the majority of high quality experimental protein–ligand crystal structures are not even publicly accessible: they are in proprietary databases. If this problem were to be solved in the near future, it would almost certainly require some arrangement that would permit those structures to be used in the training of the AI algorithm, a task that has likely been made more difficult by the formation of DeepMind's own spin-out drug discovery company.

In spite of this, there are areas of drug discovery where AlphaFold is likely to become the method of choice for obtaining structural information. Researchers targeting neglected diseases – diseases that overwhelmingly affect poor communities in the developing world – often do not have access to expensive experimental facilities, and are targeting organisms that are not well covered by PDB structures: DeepMind has already announced a partnership with the Drugs for Neglected Diseases Initiative.

Beyond small molecules, AlphaFold might, in the short term, have a more significant role to play in the development of novel therapies based on proteins, including in antibody and vaccine development.

It is also worth considering what AlphaFold teaches us about the impact that artificial intelligence, more generally, will have on drug discovery. AlphaFold demonstrates, emphatically, that artificial intelligence has the potential to solve longstanding problems in drug discovery. But it also makes clear that the realization of that potential relies on the underlying data. AlphaFold only works because the protein structure community has, over 50 years, committed to a well-defined standard of data collection and sharing. Artificial intelligence is a powerful tool for extracting information from data, but it cannot create new information, and it cannot create order from chaos. Communities and organizations that have not invested in maintaining high quality, standardized and easily accessible data will not, in the short term, reap the same benefits from artificial intelligence. They will have to invest time and effort in curating historic data sources and creating structures to enable the sharing of data, but even this on its own may not be enough. The data will also need to be the right data: data that is appropriate for modelling. The organizations that benefit most from artificial intelligence, therefore, will be those that close the gaps

between experimental and computational scientists; that foster close collaboration and mutual understanding.

AlphaFold will not, on its own, fundamentally change drug discovery. It will not significantly decrease the cost or time to bring a drug to market. But that does not mean that it is not hugely significant. AlphaFold proves, beyond any doubt, that artificial intelligence is not just hype, that it will impact on drug discovery. AlphaFold is not the end of the story, it is the beginning. The beginning of a revolution.

Received: February 24, 2022

[1] C. B. Anfinsen, *Science* **1973**, *181*, 223, https://doi.org/10.1126/science.181.4096.223.
[2] J. Moult, J. T. Pedersen, R. Judson, K. Fidelis, *Proteins: Struct. Funct. Genet.* **1995**, *23*, ii, https://doi.org/10.1002/prot.340230303.
[3] E. Callaway, *Nature* **2020**, *588*, 203.
[4] https://www.forbes.com/sites/robtoews/2021/10/03/alphafold-is-the-most-important-achievement-in-ai-ever/?sh=4767e0e36e0a
[5] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, *Nature* **2021**, *596*, 583, https://doi.org/10.1038/s41586-021-03819-2.
[6] A. W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Žídek, A. W. R. Nelson, A. Bridgland, H. Penedones, S. Petersen, K. Simonyan, S. Crossan, P. Kohli, D. T. Jones, D. Silver, K. Kavukcuoglu, D. Hassabis, *Nature* **2020**, *577*, 706, https://doi.org/10.1038/s41586-019-1923-7.
[7] J. Skolnick, M. Gao, H. Zhou, S. Singh, *J. Chem. Inf. Model.* **2021**, *61*, 4827, https://doi.org/10.1021/acs.jcim.1c01114.
[8] F. Ren, X. Ding, M. Zheng, M. Korzinkin, X. Cai, W. Zhu, A. Mantsyzov, A. Aliper, V. Aladinskiy, Z. Cao, S. Kong, X. Long, B. Hei Man Liu, Y. Liu, V. Naumov, A. Shneyderman, I. V. Ozerov, J. Wang, F. W. Pun, A. Aspuru-Guzik, M. Levitt, A. Zhavoronkov, *arXiv* **2022**, 2201.09647.
[9] https://github.com/deepmind/alphafold/commit/8173117130e6df8749ab7349722ec465666df548
[10] K. Tunyasuvunakool, J. Adler, Z. Wu, T. Green, M. Zielinski, A. Žídek, A. Bridgland, A. Cowie, C. Meyer, A. Laydon, S. Velankar, G. J. Kleywegt, A. Bateman, R. Evans, A. Pritzel, M. Figurnov, O. Ronneberger, R. Bates, S. A. A. Kohl, A. Potapenko, A. J. Ballard, B. Romera-Paredes, S. Nikolov, R. Jain, E. Clancy, D. Reiman, S. Petersen, A. W. Senior, K. Kavukcuoglu, E. Birney, P. Kohli, J. Jumper, D. Hassabis, *Nature* **2021**, *596*, 590, https://doi.org/10.1038/s41586-021-03828-1.
[11] D. T. Jones, J. M. Thornton, *Nat. Methods* **2022**, *19*, 15, https://doi.org/10.1038/s41592-021-01365-3.
[12] M. Varadi, S. Anyango, M. Deshpande, S. Nair, C. Natassia, G. Yordanova, D. Yuan, O. Stroe, G. Wood, A. Laydon, A. Žídek, T. Green, K. Tunyasuvunakool, s. Petersen, J. Jumper, E. Clancy, R. Green, A. Vora, M. Lutfi, M. Figurnov, A. Cowie, N. Hobbs, P. Kohli, G. Kleywegt, E. Birney, D. Hassabis, S. Velankar, *Nucleic Acids Res.* **2022**, 50, 439, https://doi.org/10.1093/nar/gkab1061.
[13] R. Evans, M. O'Neill, A. Pritzel, N. Antropova, A. Senior, T. Green, A. Žídek, R. Bates, S. Blackwell, J. Yim, O. Ronneberger, S. Bodenstein, M. Zielinski, A. Bridgland, A. Potapenko, A. Cowie, K. Tunyasuvunakool, R. Jain, E. Clancy, P. Kohli, J. Jumper, D. Hassabis, *bioRxiv* **2021**, https://doi.org/10.04.463034.