Protein structure predictions to atomic accuracy with AlphaFold

AlphaFold is a neural-network-based approach to predicting protein structures with high accuracy. We describe how it works in general terms and discuss some anticipated impacts on the field of structural biology.

John Jumper and Demis Hassabis

In the 2020 Critical Assessment of protein Structure Prediction (CASP14), the AlphaFold system[1] predicted almost two-thirds of the target protein structures at an accuracy that the assessors considered competitive with that of experimental methods (~1 Å typical deviation on the backbone)[2]. This advance, which builds on decades of work to create comprehensive databases of protein sequences and structures[3–5], has enabled large increases in the structural coverage of model organism proteomes, including a doubling of the fraction of the human proteome whose structure is known to high accuracy[6].

At the core of AlphaFold is a new kind of neural network whose building blocks are adapted specifically to the problem of predicting protein structure. Neural networks are a large class of machine-learning algorithms, consisting of pipelines of alternating linear and nonlinear components, called layers, that are typically 'trained' (the process of optimizing parameters) using gradient descent on the error of the final predictions. The accuracy and generality of the trained neural network is highly dependent on the design of the network architecture (the layers used and how they are connected) and its training. To develop AlphaFold's neural network, we set out to create new network architectures and training procedures that are aligned with our understanding of protein biology.

A key to AlphaFold's success is the establishment of communication patterns within and between components of the network that are sympathetic to the concepts of protein physics and biology. For example, wherever there is an interaction that can be interpreted as a communication between different sequence positions, we add a special connection to our 'pair representation' that enables the network to modulate these interactions on the basis of its understanding of pairwise residue interactions. In practice, this means that the network learns rapidly in training to enforce communication between sequence-distant positions in the protein that are spatially local in the folded structure, without requiring hard-coding of a specific geometric algorithm. Similarly, the training is adapted so that the neural network can make effective use of protein sequence data even when the structures are unknown and is encouraged to learn generalized coevolution patterns. The combined effect of these and many other ideas on the network are dramatic: AlphaFold can be trained to produce vastly more accurate structure

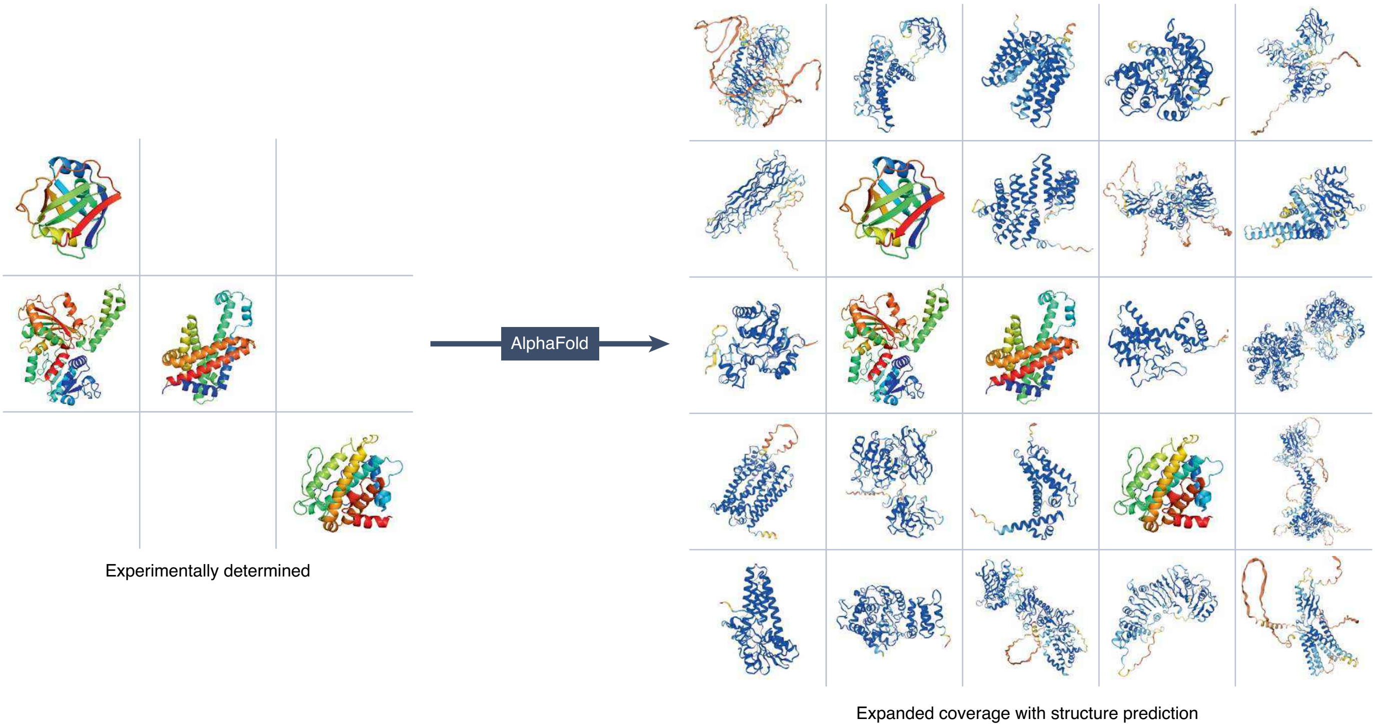
AlphaFold
Experimentally determined
Expanded coverage with structure prediction

Fig. 1 | AlphaFold as an amplifier of sparse experimental data. Schematic illustration of the role of machine learning, which converts a smaller amount experimentally determined data into a comprehensive set of experimental predictions.

predictions using the same Protein Data Bank (PDB) training data as earlier, less accurate methods.

Despite being trained on only single protein chains that appear in the PDB, the AlphaFold network shows significant generalization to other protein structure tasks due to its very high accuracy. AlphaFold predicts the structure of proteins with novel folds at approximately the same accuracy as that of proteins with known folds, implying that the network is prioritizing local interactions over recognition of global patterns. Similarly, the network is tolerant of sequences containing large disordered segments, and the low confidence of AlphaFold in these regions can be repurposed as a reliable predictor of intrinsically disordered regions[6,7].

As a more stringent test of generalization, AlphaFold is able to accurately predict the structure of artificial constructs that include multiple proteins joined with flexible linkers or artificial sequence gaps[8]. Note that this is a situation that is rarely, if ever, observed in AlphaFold's training data, but obeys the same physical and geometric principles as regular protein structure. Using these artificial linker or gap sequences, the AlphaFold network can be used to predict protein–protein interactions at an accuracy that exceeds even that of specialized protein interaction predictors despite being trained only on single protein chains[9]. Prediction of heteromeric interactions can be substantially increased, however, by including protein–protein interactions in the training of the neural network[10]. Despite these advances, the prediction of protein interactions still requires development, and current issues including false negatives and difficulties with antibody binding can likely be greatly reduced with further research.

In the immediate term, the availability of an accurate method of predicting protein structure will allow many functional studies to proceed based on structural hypotheses developed from the predicted models that would previously have required experimental models. This will be especially useful for understudied organisms and in metagenomics, where structural coverage is often very sparse and slow to extend but genomic coverage can be increased much more rapidly. In effect, AlphaFold amplifies the combined output of the experimental protein structure community to create a vastly larger universe of reliable protein structures (Fig. 1). The structural prediction of entire proteomes also creates opportunities to interpret protein structure at scale and to add geometric and biophysical context to protein-coding variants in the genome. Just as with experimental models, however, care will need to be taken to interpret the confidence and limitations of the computational models to make sure that conclusions drawn from them are well founded and possible errors in the models are understood.

These computational models can also be expected to accelerate progress in experimental structure determination. A large fraction of X-ray structures can be phased via molecular replacement using AlphaFold-predicted structures[11], and the network's predictions are excellent starting points for building models into experimental electron densities[12]. Additionally, knowledge of the protein topology and domain structures will enable better design of experimental constructs for structure determination. Recent work on the nuclear pore complex[13] has shown that monomer and pairwise heteromer modeling can be very effectively combined with low-resolution cryo-electron microscopy data to provide atomic-scale models of enormous molecular machines.

AlphaFold and related technologies will make it possible to build atomistic models of many more cellular processes using abundant pairwise connections made by protein interaction models. Early work in this direction has already uncovered many new eukaryotic interactions[14,15], and further advances in heteromer modeling will greatly expand our coverage of protein interaction networks. This will create the need for new computational methods to interpret the structural biology of molecular pathways at scale and is likely to create new opportunities for deep learning systems to interpret these data in conjunction with large-scale, low-resolution experimental techniques such as cryo-electron tomography.

Although these and related developments will ultimately bring us quite a bit further toward modeling the geometry of well-structured protein components, there is still much more to achieve in understanding the dynamical and functional behavior of these components, as well as understanding the vast disordered regions of the proteome. It is quite possible, though, that other areas of cell biology can replicate what occurred for structure prediction: the impact of carefully collected, diverse biological resources like the PDB can be amplified by many orders of magnitude through the development of the right machine-learning tools. ❐

John Jumper ✉ and Demis Hassabis ✉
DeepMind, London, UK.
✉*e-mail: jumper@deepmind.com; dhcontact@deepmind.com*

Published online: 11 January 2022
https://doi.org/10.1038/s41592-021-01362-6

References

1. Jumper, J. et al. *Nature* **596**, 583–589 (2021).
2. Pereira, J. et al. *Proteins* **89**, 1687–1699 (2021).
3. wwPDB Consortium. *Nucleic Acids Res.* **47**, D520–D528 (2018).
4. Bateman, A. et al. *Nucleic Acids Res.* **49**, D480–D489 (2021).
5. Mitchell, A. L. et al. *Nucleic Acids Res.* **48**(D1), D570–D578 (2020).
6. Tunyasuvunakool, K. et al. *Nature* **596**, 590–596 (2021).
7. Akdel, M., Pires, D. E. V., Pardo, E. P., Jänes, J. & Zalevsky, A. O. Preprint at *bioRxiv* https://doi.org/10.1101/2021.09.26.461876 (2021).
8. Yin, R., Feng, B. Y., Varshney, A. & Pierce, B. G. Preprint at *bioRxiv* https://doi.org/10.1101/2021.10.23.465575 (2021).
9. Bryant, P., Pozzati, G. & Elofsson, A. Preprint at *bioRxiv* https://doi.org/10.1101/2021.09.15.460468 (2021).
10. Evans, R., O'Neill, M., Pritzel, A., Antropova, N. & Senior, A.W. *bioRxiv* (2021).
11. Millán, C. et al. *Proteins* **89**, 1752–1769 (2021).
12. Kryshtafovych, A. et al. *Proteins* **89**, 1633–1646 (2021).
13. Mosalaganti, S. et al. Preprint at *bioRxiv* https://doi.org/10.1101/2021.10.26.465776 (2021).
14. Humphreys, I. R. et al. *Science* https://doi.org/10.1126/science.abm4805 (2021).
15. Burke, D. F. et al. Preprint at *bioRxiv* https://doi.org/10.1101/2021.11.08.467664 (2021).

Acknowledgements

We thank K. Tunyasuvunakool for helping with the figure; R. Bates, M. Figurnov, T. Green and Z. Wu for their suggestions and comments; and C. Meyer for helping to prepare the manuscript.

Competing interests

The authors have filed patent applications in the name of DeepMind Technologies Limited relating to machine learning for protein structure prediction.