Deep learning and protein structure modeling

Deep learning has transformed protein structure modeling. Here we relate AlphaFold and RoseTTAFold to classical physically based approaches to protein structure prediction, and discuss the many areas of structural biology that are likely to be affected by further advances in deep learning.

Minkyung Baek and David Baker

Up until recently, computational structural biology—the prediction and design of biomolecular structures, dynamics and interactions—was based almost entirely on physically based models. Such models use force fields and energy functions that describe atomic interactions in biomolecules as the sum of terms representing non-covalent van der Waals, electrostatic and hydrogen bonding interactions along with covalent interactions between bonded atoms. Solvation interactions are modeled through either the explicit incorporation of water molecules or implicit models that average over their possible positions. The hundreds of parameters of these models cannot be collectively obtained from first-principles quantum mechanics (QM)-based calculations. Instead, various approaches have been developed over the years to obtain them from small-molecule experimental or QM data and/or protein data[1–4]. These force fields have been used to simulate macromolecular motion using molecular dynamics (MD) simulation and to predict and design protein structures using biomolecular modeling software such as Rosetta[5].

A major challenge for these methods has been the very large size of protein conformational space. Molecular dynamics approaches are typically limited to simulation times of less than a millisecond, and hence, for all but the smallest proteins, sample only the region around the starting structure. Monte Carlo (MC)-based structure prediction methods such as Rosetta that seek to identify the lowest energy state of the protein chain struggle with larger proteins, for which the conformational space becomes extremely large. A second challenge has been the accuracy of the force fields; the simulation of dynamics and prediction of structure are only as accurate as the description of the physics embedded in the force field. Supplementation with structural constraints derived from amino acid sequence covariation during evolution has increased the size and complexity of the structures that can be predicted using Rosetta and related approaches.

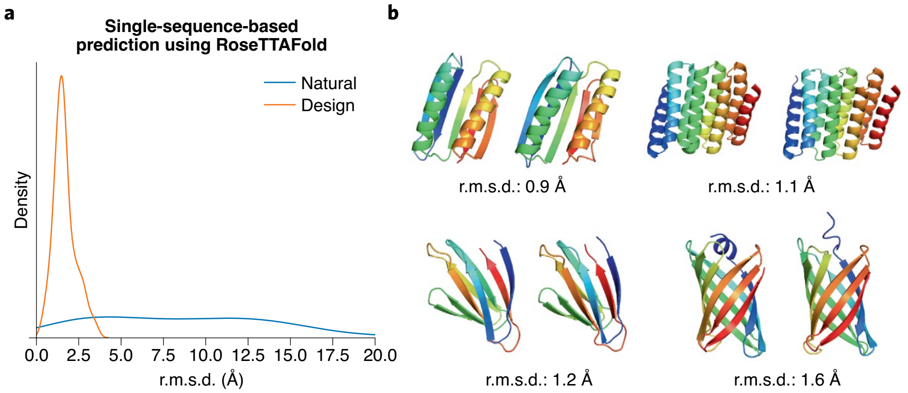

Fig. 1 | RoseTTAFold accurately predicts structures of de-novo-designed proteins from their amino acid sequences. a, The structures of designed proteins (orange) are accurately predicted using single-sequence information. The structures of naturally occurring proteins (blue) are less ideal, and are not accurately predicted by either AlphaFold or RoseTTAFold from single sequences (at least 30 homologous sequences are generally required). **b**, Examples of RoseTTAFold models for de-novo-designed proteins: α/β topology (PDB ID 1QYS), all-α topology (PDB ID 5CWF) and all-β topology (PDB IDs 6E5C and 6CZI). Experimental structures are on the left and the RoseTTAFold models are on the right. r.m.s.d., root mean squared deviation.

Very recently, deep learning methods such as RoseTTAFold[6] and AlphaFold[7] have achieved structure prediction accuracies far beyond that obtained with classical force-field-based models. These methods have millions of parameters, in contrast to the hundreds of parameters of classical approaches, and make no assumptions about the functional form of the interactions between atoms (such as Coulomb's law for electrostatic interactions). Unlike the energy-function-based classical approaches, the new methods learn millions of parameters directly by training the networks to generate correct three-dimensional structures from input amino acid sequences over sets of tens of thousands of experimentally determined protein structures. Despite these differences, the new methods do have an interesting resemblance to classical physical simulation (more than the first generation of less accurate convolutional-network-based deep learning methods for structure prediction): they iteratively update a representation of the structure, generating a trajectory that, in favorable cases, converges on the correct structure. These updates can be viewed as very sophisticated 'moves' analogous to those in a molecular simulation, but involving more concerted structural changes with magnitude adapting to the likely distance from the correct structure. The updates are based on the current representation of the structure, and unlike the moves in MD or MC trajectories, are directly optimized by the deep learning training procedure such that repeated updates result in accurate final structures. These smart structure updates help overcome the two challenges of classic molecular simulation described in the previous paragraph: global optimization is possible even in very large spaces if moves are almost always in the direction of the optimum (not the case for a classic MD or MC trajectory) and become very small when this is reached.

RoseTTAFold and AlphaFold are trained to predict structure not from single amino acid sequences, but from alignments of many homologous sequences, and they learn to extract rich structural information from these evolutionary data. However, the training of the models using extensive evolutionary information does not mean that such information is absolutely required for structure inference. RoseTTAFold very accurately predicts the structures of de-novo-designed proteins from single amino acid sequences (Fig. 1), indicating that it contains a sufficiently rich understanding of protein sequence–structure relationships to make evolutionary information unnecessary for such simple systems.

Although the recent advances in protein structure prediction are quite notable, this is just the beginning of the impact of deep learning on structural biology. The areas likely to be most immediately affected are protein interaction and assembly modeling, protein design and small-molecule drug discovery. Although the RoseTTAFold and AlphaFold papers were published only a few months ago, these methods have already started to have an impact on research in this field[8–13].

The combination of RoseTTAFold and AlphaFold can predict the structure of protein–protein complexes more accurately than either method alone, and we have used this approach to carry out full proteome-scale prediction of protein–protein interactions, which has resulted in models of core eukaryotic protein complexes that provide rich insights into biological function[11]. AlphaFold has been recently optimized for complex prediction[12]. Efforts are underway to develop deep learning approaches related to AlphaFold and RoseTTAFold that take as input not only the sequences of the proteins to be modeled, but also cryoelectron microscopy (cryo-EM) data to enable accurate modeling of large complexes from lower-resolution data. More generally, deep learning should greatly enhance the quality of models that can be built from limited experimental data.

On the protein design side, encouraged by the high accuracy of RoseTTAFold for predicting structures of de-novo-designed proteins (Fig. 1), we have inverted deep learning structure prediction networks to "hallucinate" a wide range of new proteins whose structures have been confirmed by X-ray crystallography and NMR[14]. This approach has been extended to design proteins scaffolding functional sites for catalysis and binding[15]. A second approach to protein function design uses an extended version of RoseTTAFold trained to recover sequences from structures in addition to structures from sequences in a manner analogous to the use of language models to complete sentences when given only the first few words: starting from the sequence and structure of a minimalist functional site, the model generates the sequence and structure of a full protein containing that site[13]. De novo protein design with the classical Rosetta method has resulted in the creation of new therapeutics[16] and vaccines[17] that have shown promise in animal trials and are currently in human clinical trials. We anticipate that incorporating deep learning approaches will enhance the already very rapid rate of progress in this field, considerably increasing the complexity of the proteins that can be designed.

The potential for deep learning in structural biology is enormous. What are the challenges ahead? First and foremost, deep learning methods require large and information-rich datasets for accurate model training. The success of RoseTTAFold and AlphaFold derives from the many tens of thousands of protein structures, each containing information on the atomic coordinates of thousands of atoms and millions of atom–atom pairwise interactions, solved by structural biologists over the past 50 years. For some of the most exciting areas of application, such as drug discovery, the available datasets (protein–small molecule complexes) are much smaller, and many are not publicly available. There are similarly far fewer data for designs with unnatural amino acids and non-protein backbones. In these areas, the most powerful approaches may combine deep learning with physically based models such as Rosetta which, at least at present, are more readily generalizable to problems for which limited training data exist. If we have learned anything from the rapid advances in the past few years, however, it is that predictions of the rate of progress in computational structural biology are much less accurate than the predictions of the models themselves; the next few years should be exciting indeed! ❐

Minkyung Baek[1,2] and David Baker[1,2,3] ✉

[1]Department of Biochemistry, University of Washington, Seattle, WA, USA. [2]Institute for Protein Design, University of Washington, Seattle, WA, USA. [3]Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA.

✉*e-mail: dabaker@uw.edu*

Published online: 11 January 2022
https://doi.org/10.1038/s41592-021-01360-8

References

1. Park, H. et al. *J. Chem. Theory Comput.* **12**, 6201–6212 (2016).
2. MacKerell, A. D. et al. *J. Phys. Chem. B* **102**, 3586–3616 (1998).
3. Ponder, J. W. & Case, D. A. *Adv. Protein Chem.* **66**, 27–85 (2003).
4. O'Meara, M. J. et al. *J. Chem. Theory Comput.* **11**, 609–622 (2015).
5. Leaver-Fay, A. et al. *Methods Enzymol.* **487**, 545–574 (2011).
6. Baek, M. et al. *Science* **373**, 871–876 (2021).
7. Jumper, J. et al. *Nature* **596**, 583–589 (2021).
8. Jendrusch, M., Korbel, J. O. & Sadiq, S. K. Preprint at *bioRxiv* https://doi.org/10.1101/2021.10.11.463937 (2021).
9. Moffat, L., Greener, J. G. & Jones, D. T. Preprint at *bioRxiv* https://doi.org/10.1101/2021.08.24.457549 (2021).
10. Burke, D. F. et al. Preprint at *bioRxiv* https://doi.org/10.1101/2021.11.08.467664 (2021).
11. Humphreys, I. R. et al. *Science* **374**, eabm4805 (2021).
12. Evans, R. et al. Preprint at *bioRxiv* https://doi.org/10.1101/2021.10.04.463034 (2021).
13. Wang, J. et al. Preprint at *bioRxiv* https://doi.org/10.1101/2021.11.10.468128 (2021).
14. Anishchenko, I. et al. *Nature* https://doi.org/10.1038/s41586-021-04184-w (2021).
15. Tischer, D. et al. Preprint at *bioRxiv* https://doi.org/10.1101/2020.11.29.402743 (2020).
16. Silva, D.-A. et al. *Nature* **565**, 186–191 (2019).
17. Walls, A. C. et al. *Cell* **183**, 1367–1382.e17 (2020).

Acknowledgements

This work was supported by Microsoft (M.B., D.B.), Open Philanthropy and HHMI (D.B.) and the Washington Research Foundation (M.B.).

Competing interests

The authors declare no competing interests.