ANNUAL REVIEWS

Annual Review of Biomedical Data Science

AlphaFold and Protein Folding: Not Dead Yet! The Frontier Is Conformational Ensembles

Gregory R. Bowman

Departments of Biochemistry and Biophysics and Bioengineering, University of Pennsylvania, Philadelphia, Pennsylvania, USA; email: grbowman@seas.upenn.edu

Annu. Rev. Biomed. Data Sci. 2024. 7:51–57

First published as a Review in Advance on April 11, 2024

The *Annual Review of Biomedical Data Science* is online at biodatasci.annualreviews.org

https://doi.org/10.1146/annurev-biodatasci-102423-011435

Keywords

structure prediction, molecular dynamics, machine learning

Abstract

Like the black knight in the classic Monty Python movie, grand scientific challenges such as protein folding are hard to finish off. Notably, AlphaFold is revolutionizing structural biology by bringing highly accurate structure prediction to the masses and opening up innumerable new avenues of research. Despite this enormous success, calling structure prediction, much less protein folding and related problems, "solved" is dangerous, as doing so could stymie further progress. Imagine what the world would be like if we had declared flight solved after the first commercial airlines opened and stopped investing in further research and development. Likewise, there are still important limitations to structure prediction that we would benefit from addressing. Moreover, we are limited in our understanding of the enormous diversity of different structures a single protein can adopt (called a conformational ensemble) and the dynamics by which a protein explores this space. What is clear is that conformational ensembles are critical to protein function, and understanding this aspect of protein dynamics will advance our ability to design new proteins and drugs.

INTRODUCTION

The protein folding problem is a grand challenge that has long been recognized to have multiple parts (1). Historically, two major questions have motivated the field. First, how can one predict the structure of a protein from its amino acid sequence? Second, how does a protein get to this final structure? Or, put another way, what is the mechanism of protein folding?

Work to understand protein folding mechanisms led to an even broader question: What does the energy landscape of a protein look like? The core idea is that a protein does not have a single structure. Rather, a protein stochastically hops between an enormous set of alternative structures, often called an ensemble. This biased random walk is often referred to as protein dynamics. The probability that a protein adopts any one structure is related to the energy of that structure, with proteins spending exponentially more time in lower-energy structures with more energetically favorable interactions (e.g., hydrogen bonding in an α-helix) than in structures with less favorable interactions (e.g., strained torsions). A protein's structural ensemble can then be conceptualized as a rugged energy landscape, full of low-energy minima separated by ridges and mountains made up of structures with higher energies (**Figure 1**). From this perspective, predicting a protein's structure is a matter of finding the lowest-energy minima in the energy landscape, whereas the folding mechanism is the dominant path(s) from the unfolded state to the folded state. The energy

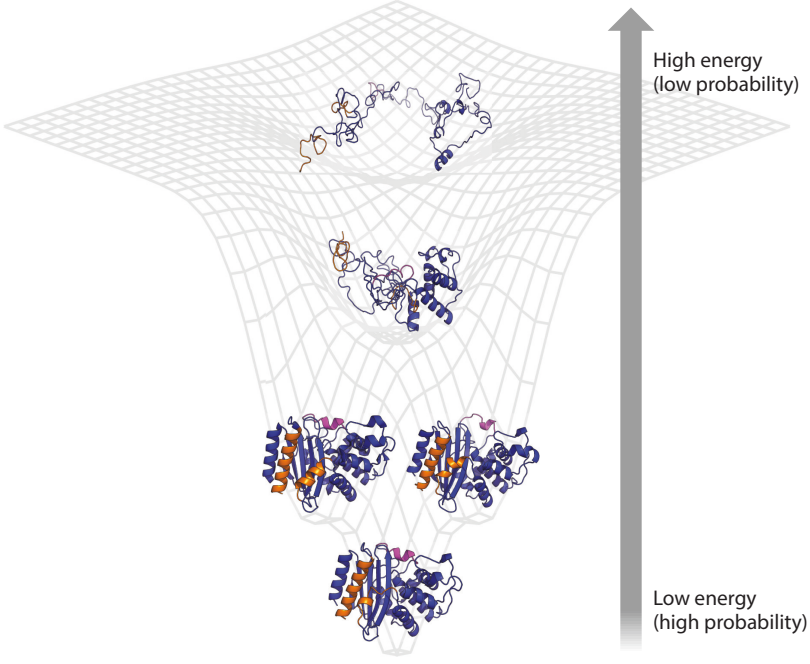

Figure 1

A schematic of an energy landscape representing the conformational ensemble of the enzyme TEM β-lactamase. The structure at the bottom has the lowest energy (highest probability) and is what one would expect to see in a crystal structure or AlphaFold prediction. The structures moving toward the top of the figure have progressively higher energy (lower probability). For example, the next two structures up have cryptic pockets (highlighted by the *orange* and *magenta* residues) that are absent in the lowest-energy structure but that have been shown to exist experimentally and provide new opportunities for drug design. The other structures represent partially and fully unfolded structures that are known to be important for the folding mechanism of this protein. Figure adapted with permission from Knoverek et al. (3).

landscape also holds a wealth of other information. For example, there is growing recognition that protein dynamics are often critical for function (2, 3). Protein dynamics also promise new opportunities for drug discovery. For example, there is growing interest in targeting cryptic pockets that are absent in experimentally derived structures but open due to protein motion (4). In a recent study, we found that at least half of a set of 5,000 proteins that were previously thought to lack a druggable pocket have cryptic pockets that present new opportunities for drug design (5). Results like these give a sense of the broad relevance of energy landscapes and protein dynamics.

AlphaFold represents a huge advance in the protein folding field, most notably for structure prediction (6). Briefly, AlphaFold is a deep learning algorithm that takes the primary amino acid sequence of a protein as input and predicts the structure of the protein as would be observed with an experimental technique like X-ray crystallography or cryo-electron microscopy (cryoEM). The algorithm was trained on the Protein Data Bank (PDB) (7), which is a publicly available repository of over 200,000 protein structures that have been accumulated over decades through a requirement that structural biologists deposit their structures during peer review of their work. Prior to AlphaFold, other algorithms had been developed to predict protein structures using a combination of physics and machine learning based on available structures. For decades, the performance of these methods was regularly tested through blind predictions via the critical assessment of protein structure prediction (CASP) competition (8). While the field made great progress over time, it had hit somewhat of a plateau in recent years. AlphaFold broke this trend, making a substantial improvement in accuracy. Its predictive power is one of the most compelling examples of the enormous power that computational methods have to offer biomedical research.

This review covers the implications of AlphaFold for understanding protein folding, including the original questions about structure prediction and folding mechanisms as well as the broader question about energy landscapes. A major theme is the need to go beyond single structures and understand energy landscapes, as others have also pointed out (9).

FIELDS ARE GENERALLY ADVANCED, NOT SOLVED

Solved is a dangerous word when it comes to grand scientific challenges. Such problems tend to evolve rather than to be totally completed. While calling a problem solved can be an appropriate homage to a tremendous advance, it can also backfire by impeding investment in the next major advance(s).

AlphaFold is a tremendous advance that has solved the structure prediction problem akin to the way the first commercial airline flight in 1914 solved the flight problem. That first commercial flight carried a single passenger from St. Petersburg to Tampa, Florida. While quaint by today's standards, that flight marked the beginning of routine flight for the masses, regardless of how much or little they may know about flight. It would not have been possible if we had called flight solved after the Wright brothers' first flight in 1903. Furthermore, my recent transatlantic flight with ~230 other passengers would not have been possible if we had stopped investing in research on flight after that first commercial flight in 1914.

Like that first commercial flight, one of the most exciting things about AlphaFold is the new opportunities it creates by making structure prediction a routine process that is available to the masses rather than just expert users. For example, AlphaFold's predictions have enabled molecular replacement solutions to previously unsolvable crystallographic datasets (8, 10, 11). Others are devising new structure-based hypotheses about biological problems using AlphaFold structures (12), much as scientists have done with crystal structures since they first became available. Those of us studying protein dynamics with molecular dynamics simulations are using AlphaFold structures as starting points in cases where experimentally derived structures are not available (13). Similarly,

AlphaFold structures are enabling structure-based drug discovery for targets nobody has solved experimental structures for yet (14). One can even go to new scales, scanning entire genomes for interesting features, like new proteins and folds (15–17).

In addition to AlphaFold's use for different applications, there are many natural extensions of its success in structure prediction. For example, there are still limitations when it comes to predicting the structures of large assemblies or incorporating nonprotein elements into predicted structures, so improved methods for such applications are needed despite recent progress (18, 19). Efforts to make open source versions of powerful algorithms akin to AlphaFold are also valuable catalysts that empower the broader community to contribute new ideas (20, 21).

THE FRONTIER OF CONFORMATIONAL ENSEMBLES

In addition to the new opportunities that AlphaFold creates, the advance also highlights the frontiers where there are still large gaps in our understanding.

In the case of AlphaFold, the method's greatest strength is also its greatest weakness: AlphaFold predicts a single structure even for proteins that are known to switch between different conformations as part of their function (22) and is subject to the same limitations as experimental techniques that resolve a single structure. AlphaFold does not give us the mechanism of protein folding. Further, AlphaFold often predicts the same structure for a sequence with a point mutation as it did for the wild-type sequence (23), just as experimentally derived structures of protein variants are often indistinguishable (3). Structure-based drug discovery remains difficult (24). Disordered regions that are left unresolved in experimentally derived structures appear as unrealistic swirls in AlphaFold-predicted structures (25).

One critical frontier that AlphaFold helps highlight, then, is understanding proteins' conformational ensembles and energy landscapes. The ability to predict such ensembles should reveal the mechanisms of protein folding and the different structural states that are important to a protein's function, whether they are the on/off states of a molecular switch or the different conformations of a catalytic cycle. One could examine how point mutations or small molecules shift the relative probabilities of different structures and how these shifts alter the affinities of a protein for different binding partners (26, 27). Structure-based drug design could become more rational and routine (28–30), and one could sift through the myriad structures adopted by disordered proteins to find patterns tied to function.

MACHINE LEARNING AND CONFORMATIONAL ENSEMBLES

In light of AlphaFold's success at structure prediction, a natural question is what role machine learning will play in our evolving understanding of conformational ensembles.

A major challenge here is the availability of large amounts of high-quality data, or the lack thereof. One of the key ingredients of AlphaFold's success was the availability of the PDB (7). The PDB is dominated by atomically detailed structures solved by X-ray crystallography, as well as structures provided by practitioners of nuclear magnetic resonance (NMR) spectroscopy and cryoEM in the same format. While protein conformational ensembles have been of interest for many years, there is no analogous repository of high-resolution ensembles in a consistent format. No doubt many in the community would be happy to contribute to a repository of ensemble data if there were a clear path to doing so. However, there are a multitude of methods for studying conformational ensembles and they vary so greatly (e.g., in their spatial and temporal resolution) that they cannot be represented in a single common format. Hypothetically, one could learn proteins' conformational ensembles by drawing on disparate data types. However, doing so would be

far more complicated due to both logistical reasons (e.g., how to amass and organize such data) and theoretical considerations (e.g., how much weight to put on different types of data).

One possibility is that an algorithm like AlphaFold can succeed in modeling conformational ensembles given existing resources like the PDB. There are many examples where researchers have captured different conformations of a single protein and deposited these structures to the PDB. In principle, one could design a machine learning algorithm that can leverage these examples of conformational diversity to predict conformational ensembles. Del Alamo et al. (31) took an early step in this direction by showing that one can trick AlphaFold into generating diverse predicted structures by asking it to generate structures for different subsets of a large sequence alignment instead of giving it all the sequence information and asking it for a single structure. While this approach can be useful, later work showed that it often fails to recapitulate alternative structures of a protein that are known to exist (32). An added layer of complexity arises because the number of structures of a given protein in a given conformation that have been deposited to the PDB does not necessarily reflect the relative probabilities that the protein adopts those different structures in solution. Therefore, it is unclear how one would train a machine learning algorithm to capture these physical relationships (33). Difficult does not mean impossible, though, and new algorithms are being developed to try and capture proteins' conformational ensembles (32, 34, 35).

Another possibility is that a combination of machine learning and physics-based approaches will provide a route to predictive models of proteins' conformational ensembles. One way to achieve this is by training machine learning algorithms based on large sets of atomically detailed computer simulations of protein dynamics. The Folding@home project that I lead is providing a wealth of data to explore this possibility by enlisting citizen scientists from around the world to help generate atomically detailed molecular dynamics simulations of proteins (36, 37). For example, the PocketMiner algorithm for predicting the locations of cryptic pockets was trained on these data (5). PocketMiner was trained to take a structure from a simulation as input and predict the probability that a given residue becomes more exposed to solvent due to formation of a cryptic pocket in a fixed amount of simulation time after that snapshot. We demonstrated that PocketMiner does an excellent job of predicting if and where cryptic pockets are likely to form in experimentally derived structures and is likely to work equally well on AlphaFold-predicted structures. Other labs are training generative models of conformational ensembles by drawing on simulation data (38–40).

CONCLUSIONS AND FUTURE OUTLOOK

AlphaFold is revolutionizing structural biology by providing routine access to highly accurate predictions of protein structures. This structural information opens up innumerable avenues of research. However, important limitations remain, necessitating continued investment in structure prediction. Moreover, our understanding of protein folding mechanisms and protein dynamics more broadly remains limited. Further insight into conformational ensembles will advance our understanding of protein function and dysfunction and enhance our ability to design new proteins and drugs. As with structure prediction, machine learning is likely to play a prominent role in advances on these important frontiers.

DISCLOSURE STATEMENT

G.R.B. holds grants from the National Institutes for Health and the National Science Foundation, has equity in Decrypt Bio, and serves as the director of the Folding@home distributed computing project.

ACKNOWLEDGMENTS

This work was funded by National Institutes for Health grants R01GM124007, RF1AG067194, and U19AG069701 and National Science Foundation grant MCB-2218156. G.R.B. holds a Packard Fellowship from the David and Lucile Packard Foundation.

LITERATURE CITED

1. Dill KA, MacCallum JL. 2012. The protein-folding problem, 50 years on. *Science* 338:1042–46
2. Henzler-Wildman K, Kern D. 2007. Dynamic personalities of proteins. *Nature* 450:964–72
3. Knoverek CR, Amarasinghe GK, Bowman GR. 2019. Advanced methods for accessing protein shape-shifting present new therapeutic opportunities. *Trends Biochem. Sci.* 44:351–64
4. Kuzmanic A, Bowman GR, Juarez-Jimenez J, Michel J, Gervasio FL. 2020. Investigating cryptic binding sites by molecular dynamics simulations. *Acc. Chem. Res.* 53:654–61
5. Meller A, Ward M, Borowsky J, Kshirsagar M, Lotthammer JM, et al. 2023. Predicting locations of cryptic pockets from single protein structures using the PocketMiner graph neural network. *Nat. Commun.* 14:1177
6. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, et al. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* 596:583–89
7. Burley SK, Berman HM, Duarte JM, Feng Z, Flatt JW, et al. 2022. Protein Data Bank: a comprehensive review of 3D structure holdings and worldwide utilization by researchers, educators, and students. *Biomolecules* 12:1425
8. Millan C, Keegan RM, Pereira J, Sammito MD, Simpkin AJ, et al. 2021. Assessing the utility of CASP14 models for molecular replacement. *Proteins* 89:1752–69
9. Lane TJ. 2023. Protein structure prediction has reached the single-structure frontier. *Nat. Methods* 20:170–73
10. Barbarin-Bocahu I, Graille M. 2022. The X-ray crystallography phase problem solved thanks to AlphaFold and RoseTTAFold models: a case-study report. *Acta Crystallogr. D Struct. Biol.* 78:517–31
11. McCoy AJ, Sammito MD, Read RJ. 2022. Implications of AlphaFold2 for crystallographic phasing by molecular replacement. *Acta Crystallogr. D Struct. Biol.* 78:1–13
12. Oliver MR, Toon K, Lewis CB, Devlin S, Gifford RJ, Grove J. 2023. Structures of the Hepaci-, Pegi-, and Pestiviruses envelope proteins suggest a novel membrane fusion mechanism. *PLOS Biol.* 21:e3002174
13. Meller A, De Oliveira S, Davtyan A, Abramyan T, Bowman GR, van den Bedem H. 2023. Discovery of a cryptic pocket in the AI-predicted structure of PPM1D phosphatase explains the binding site and potency of its allosteric inhibitors. *Front. Mol. Biosci.* 10:1171143
14. Ren F, Ding X, Zheng M, Korzinkin M, Cai X, et al. 2023. AlphaFold accelerates artificial intelligence powered drug discovery: efficient discovery of a novel CDK20 small molecule inhibitor. *Chem. Sci.* 14:1443–52
15. Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, et al. 2022. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* 50:D439–44
16. Bouatta N, AlQuraishi M. 2023. Structural biology at the scale of proteomes. *Nat. Struct. Mol. Biol.* 30:129–30
17. Bayly-Jones C, Whisstock JC. 2022. Mining folded proteomes in the era of accurate structure prediction. *PLOS Comput. Biol.* 18:e1009930
18. Evans R, O'Neill M, Pritzel A, Antropova N, Senior A, et al. 2021. Protein complex prediction with AlphaFold-Multimer. bioRxiv 2021.10.04.463034. **https://doi.org/10.1101/2021.10.04.463034**
19. Hekkelman ML, de Vries I, Joosten RP, Perrakis A. 2023. AlphaFill: enriching AlphaFold models with ligands and cofactors. *Nat. Methods* 20:205–13
20. Ahdritz G, Bouatta N, Kadyan S, Xia Q, Gerecke W, et al. 2022. OpenFold: Retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization. bioRxiv 2022.11.20.517210. **https://doi.org/10.1101/2022.11.20.517210**

21. Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, et al. 2021. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* 373:871–76
22. Chakravarty D, Porter LL. 2022. AlphaFold2 fails to predict protein fold switching. *Protein Sci.* 31:e4353
23. Cheng J, Novati G, Pan J, Bycroft C, Zemgulyte A, et al. 2023. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* 381:eadg7492
24. Karelina M, Noh JJ, Dror RO. 2023. How accurately can one predict drug binding modes using AlphaFold models? *eLife* 12:RP89386
25. Ruff KM, Pappu RV. 2021. AlphaFold and implications for intrinsically disordered proteins. *J. Mol. Biol.* 433:167208
26. Meller A, Lotthammer JM, Smith LG, Novak B, Lee LA, et al. 2023. Drug specificity and affinity are encoded in the probability of cryptic pocket opening in myosin motor domains. *eLife* 12:e83602
27. Porter JR, Meller A, Zimmerman MI, Greenberg MJ, Bowman GR. 2020. Conformational distributions of isolated myosin motor domains encode their mechanochemical properties. *eLife* 9:e55132
28. Hart KM, Moeder KE, Ho CMW, Zimmerman MI, Frederick TE, Bowman GR. 2017. Designing small molecules to target cryptic pockets yields both positive and negative allosteric modulators. *PLOS ONE* 12:e0178678
29. Smith LG, Novak B, Osato M, Mobley DL, Bowman GR. 2023. PopShift: a thermodynamically sound approach to estimate binding free energies by accounting for ligand-induced population shifts from a ligand-free Markov state model. *J. Chem. Theory Comput.* 20:1036–50
30. Fischer M, Coleman RG, Fraser JS, Shoichet BK. 2014. Incorporation of protein flexibility and conformational energy penalties in docking screens to improve ligand discovery. *Nat. Chem.* 6:575–83
31. Del Alamo D, Sala D, Mchaourab HS, Meiler J. 2022. Sampling alternative conformational states of transporters and receptors with AlphaFold2. *eLife* 11:e75751
32. Meller A, Bhakat S, Solieva S, Bowman GR. 2023. Accelerating cryptic pocket discovery using AlphaFold. *J. Chem. Theory Comput.* 19:4355–63
33. Vani BP, Aranganathan A, Tiwary P. 2023. Exploring kinase DFG loop conformational stability with AlphaFold2-RAVE. arXiv:2309.03649 [physics.bio-ph]
34. Mansoor S, Baek M, Park H, Lee GR, Baker D. 2023. Protein ensemble generation through variational autoencoder latent space sampling. bioRxiv 2023.08.01.551540. **https://doi.org/10.1101/2023.08.01.551540**
35. Vani BP, Aranganathan A, Wang D, Tiwary P. 2023. AlphaFold2-RAVE: from sequence to Boltzmann ranking. *J. Chem. Theory Comput.* 19:4351–54
36. Voelz VA, Pande VS, Bowman GR. 2023. Folding@home: achievements from over 20 years of citizen science herald the exascale era. *Biophys. J.* 122:2852–63
37. Zimmerman M, Porter J, Ward M, Singh S, Vithani N, et al. 2021. SARS-CoV-2 simulations go exascale to predict dramatic spike opening and cryptic pockets across the proteome. *Nat. Chem.* 13:651–59
38. Noé F, Olsson S, Köhler J, Wu H. 2019. Boltzmann generators: sampling equilibrium states of many-body systems with deep learning. *Science* 365:eaaw1147
39. Mehdi S, Smith Z, Herron L, Zou Z, Tiwary P. 2023. Enhanced sampling with machine learning: a review. arXiv:2306.09111 [cond-mat.stat-mech]
40. Jones MS, McDargh ZA, Wiewiora RP, Izaguirre JA, Xu H, Ferguson AL. 2023. Molecular latent space simulators for distributed and multimolecular trajectories. *J. Phys. Chem. A* 127:5470–90