HHS Public Access

Author manuscript

Nat Struct Mol Biol. Author manuscript; available in PMC 2023 August 01.

Published in final edited form as:

Nat Struct Mol Biol. 2023 February ; 30(2): 129–130. doi:10.1038/s41594-023-00924-w.

Structural biology at the scale of proteomes

Nazim Bouatta[a,*], **Mohammed AlQuraishi**[b,*]

[a]Laboratory of Systems Pharmacology, Program in Therapeutic Science, Harvard Medical School, Boston, MA, USA.

[b]Department of Systems Biology, Columbia University, New York, NY, USA.

Abstract

AlphaFold2 has already changed structural biology but its true power may lie in how it changes the way we think about cells and organisms. Two studies broadly analyze and assess the performance of AlphaFold2 to outline the extent of its utility and limitations in providing structural models that shed light on biological questions, including mutations, post-translational modifications, and protein-protein complex interactions.

The Human Genome Project and its accompanying technological advances paved the way for the genomics revolution, yielding an abundance of sequenced genes and genomes, and enabling biologists to study cells, organisms, and their evolution using the firm molecular basis of genetic data. Combined with physically-inspired mathematical methods such as maximum entropy models[1] and, more recently, machine learning, this wealth of sequence data has formed the foundation for the recent revolution in protein structure prediction[2], epitomized by DeepMind's and the European Bioinformatics Institute's recent release of 200 million AlphaFold2-predicted protein structures[3]. As one revolution begets another, the widespread availability of structural information posits a new possibility: a *structural systems biology* in which biological phenomena across the varied scales of life are studied through a structural and mechanistic prism. In this and a previous issue of *Nature Structural & Molecular Biology*, Akdel *et al*.[4] and Burke *et al*.[5] take some of the first steps toward this goal. Akdel *et al*. assess AlphaFold2 across multiple tasks, including its structural coverage of multiple proteomes, the extent of fold space it models, and its ability to predict ligand binding sites, finding that AlphaFold2 systematically outperforms existing state-of-the-art tools. In a complementary paper, using the wealth of AlphaFold2-predicted structures, Burke *et al*. assess AlphaFold2's abilities, and limitations, to structurally model the human protein-protein interaction network (less than 5% of which is estimated to have been structurally characterized). They find that AlphaFold2 increases the number of high accuracy predictions of human protein-protein interactions and use these predictions to provide structural insights into pathogenic mutations and the phosphorylation of protein interaction interfaces.

*Corresponding authors: nazim_bouatta@hms.harvard.edu and ma4129@cumc.columbia.edu.

Conflict of interest

M.A. is a member of the Scientific Advisory Boards of Cyrus Biotechnology, Deep Forest Sciences, Nabla Bio, Oracle Therapeutics, and FL2021-002, a Foresite Labs company.

To evaluate the breadth of applicability of AlphaFold2, *Akdel et al.* assess the quantity and quality of new structural knowledge made possible by AlphaFold2. They start by quantifying for different organisms the amount of additional coverage AlphaFold2 provides over experimentally determined structures. On average, 25% of all residues in any given proteome are predicted with very high accuracy by AlphaFold2 (accuracy is predicted using AlphaFold2's own calibrated self-assessment, which has been shown to be reliable in other studies[6]). This amount of coverage far exceeds what was previously possible using traditional homology modeling techniques. The breadth and quality of coverage for any given species/protein depends on the number and diversity of available homologous protein sequences, which are correlated with AlphaFold2 accuracy. A large fraction of low-confidence AlphaFold2 predictions also likely correspond to intrinsically disordered protein regions.

Another way to assess structural coverage is not by proteomes but by the extent of fold space modeled. Existing experimental coverage of fold space is difficult to quantify due to over sampling of species and proteins most pertinent to contemporary research interests, which likely leaves out swaths of archaeal and prokaryotic proteomes. Past studies have however suggested that the Protein Data Bank may already encompass near complete coverage of single domain space[7]. *Akdel et al.* shed new light on this question by showing that among structures predicted with high accuracy by AlphaFold2, a large fraction contains either novel folds or folds that have yet to be functionally characterized. *Akdel et al.* conduct their analysis using the original release of the AlphaFold2 database with ~360,000 structures focused on key model organisms; the newest release with nearly complete coverage of the UniProt[8] protein database may thus increase this fraction. Studying the emergence and diversification of new folds may further provide a complementary approach to evolutionary analyses, which have traditionally been driven by sequence-based methods.

Having addressed the question of quantity, *Akdel et al.* turn their attention to the quality of AlphaFold2-predicted structures. Experimental protein structures can provide biologists with molecular insights into protein function and dysfunction, and the same may be expected of predicted structures. *Akdel et al.* first assess the utility of AlphaFold2 structures in predicting the impact of missense mutations on protein function. When comparing predictions of change in stability using experimental structures, traditional homology modeling, and AlphaFold2, they find that when high-confidence predictions are available, AlphaFold2 is nearly as informative as experimental structures. Another task for function prediction is the identification of protein pockets and binding sites from structure. Here too, the authors find that when using a set of proteins with known binding sites, high-confidence regions of predicted structures are as informative as experimental structures in pinpointing binding sites. This suggests the possibility of discovering previously unknown pockets for potential binding by small molecule drugs.

In a parallel study, Burke *et al.* apply various augmentations of AlphaFold2, which was originally trained to predict structures of individual proteins, to tackle the prediction of multi-protein complexes. In prior work by the same team[9], a clever trick was developed to goad AlphaFold2 into predicting protein-protein complexes by concatenating sequences of proteins belonging to the same complex and computationally inserting a flexible linker

between them (an idea inspired by RoseTTAFold[10]). Remarkably, this trick resulted in more accurately predicted protein complexes than ones inferred by dedicated protein docking software. Taking advantage of this approach, *Burke et al.* tackle the human interactome by first predicting the structures of ~65,000 pairs of interacting human proteins. Among these are ~3,000 high-confidence pairs with a pDockQ score greater than 0.5 (the team previously[9] introduced pDockQ as a metric to assess the accuracy of predicted protein complexes, with scores greater than 0.5 corresponding to accurate predictions based on comparisons with experimental structures). Interestingly, ~1,400 of the newly predicted high-confidence complexes lack homology to existing structures, suggesting that the model is capable of substantially enlarging our corpus of structural knowledge.

Burke *et al.* next consider higher-order assemblies, a more challenging problem as naïve one-step prediction of whole assemblies is beyond reach due to intractable computational cost. Instead, the team tackle the problem using a multi-step pairwise approach. Their algorithm starts by first selecting the highest-ranked dimers in a complex then iteratively adds additional dimers that share one non-overlapping subunit with the complex. The order by which dimers are added is based on their pDockQ scores. This approach works for some assemblies but fails for others, such as the 20S proteasome complex. Another method, based on Monte Carlo tree search, is also considered and appears to perform better[11].

With these tools at their disposal, Burke *et al.* turn their attention to the systems biology problem of how these protein-protein interactions are regulated. Specifically, they look at phosphorylation at protein interfaces as a regulatory mark. Of the over 100,000 known human phosphorylation sites, few have been functionally characterized. By analyzing the structures of protein-protein interaction interfaces and performing a Gene Ontology (GO) enrichment analysis, the authors observe that clusters of interface phosphosites are involved in different GO processes. They furthermore analyze these interfaces for coordinated regulation by combining their analysis with experimental measurements of changes in phosphorylation levels across hundreds of experimental conditions. They find that in certain conditions, one or a set of kinases appear to phosphorylate the same set of sites across multiple proteins, suggesting that coordinated regulation is indeed taking place.

Together, the papers by Akdel *et al.* and Burke *et al.* offer a glimpse of the power of structural systems biology to gain biological insights through large-scale structural analyses, building on prior experimentally- and computationally-driven work in this space[12]. These remain early days however, and we expect rapid progress on multiple fronts. First, prediction speed and memory efficiency remain a bottleneck, particularly when predicting multimeric protein complexes, which must be applied at scale to assemble a comprehensive picture of all molecular machines. Advances in more efficient neural networks[13], especially the widely used attention architecture[14], have resulted in a flurry of highly optimized AlphaFold2 reimplementations[15–17], including our own OpenFold system[18]. These methods predict structures at rates up to two times faster than the original AlphaFold2 and utilize substantially less memory, making it possible to tackle larger assemblies. Another promising approach is the use of protein language models[19], which implicitly encode a rich representation of protein sequence space. When coupled to protein structure prediction, language models obviate the need for the computationally costly search for homologous

protein sequences. This too is a rapidly evolving area with new methods appearing regularly (RGN2[20], trRosettaX-Single[21], OmegaFold[22], and ESMFold[23] to name a few) that are already providing notable gains in efficiency.

Second, proteins and the molecular machines they comprise are dynamic objects whose function is often driven by changes in conformational state. For the time being, the ability of AlphaFold2, and more generally machine learning based methods, to provide an accounting of the conformational landscape of proteins remains very limited. Physics-based methods, including molecular dynamics, have historically played a key role in understanding protein motion and we expect this to continue. The outstanding challenge for structural systems biology is scaling such approaches to the proteome scale, which remains out of reach.

The current state of computational molecular biology is perhaps reminiscent of early 20th-century physics and the remarkable developments that transpired then, where novel theoretical ideas, driven mainly by quantum mechanics, radically changed our understanding of physical phenomena. AlphaFold2 and related deep learning techniques may similarly change the way we model and understand biological phenomena. Experiments, although ultimately the final arbiters of truth, cannot comprehensively characterize the remarkable diversity that is life; the combinatorics of biology, from molecules to cells to organisms, are simply too overwhelming. Machine learning in combination with simulation may tame this combinatorial complexity, as AlphaFold2 has arguably done for the space of protein folds. One can only hope that if this successful, this approach will reveal general principles about the organization and behavior of biological phenomena at all scales.

Acknowledgements

N.B. is supported by DARPA PANACEA program grant HR0011-19-2-0022 and NCI grant U54-CA225088.

References

1. Mora T & Bialek W Are Biological Systems Poised at Criticality? Journal of Statistical Physics 144, 268–302 (2011).
2. Jumper J et al. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589 (2021). [PubMed: 34265844]
3. Varadi M et al. AlphaFold Protein Structure Database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Research 50, D439–D444 (2022). [PubMed: 34791371]
4. Akdel M et al. A structural biology community assessment of AlphaFold2 applications. Nature Structural and Molecular Biology 29, 1056–1067 (2022).
5. Burke DF et al. Towards a structurally resolved human protein interaction network. bioRxiv 2021.11.08.467664 (2021).
6. Tunyasuvunakool K et al. Highly accurate protein structure prediction for the human proteome. Nature 596, 590–596 (2021). [PubMed: 34293799]
7. Zhang Y & Skolnick J The protein structure prediction problem could be solved using the current PDB library. Proceedings of the National Academy of Sciences of the United States of America 102, 1029–1034 (2005). [PubMed: 15653774]
8. Bateman A et al. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Research 49, D480–D489 (2021). [PubMed: 33237286]
9. Bryant P, Pozzati G & Elofsson A Improved prediction of protein-protein interactions using AlphaFold2. Nature Communications 13, 1–11 (2022).

10. Baek M et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science 373, 871–876 (2021). [PubMed: 34282049]
11. Bryant P et al. Predicting the structure of large protein complexes using AlphaFold and Monte Carlo tree search. bioRxiv 2022.03.12.484089 (2022) doi:10.1038/s41467-022-33729-4.
12. Maritan M et al. Building Structural Models of a Whole Mycoplasma Cell. Journal of Molecular Biology 434, 167351 (2022). [PubMed: 34774566]
13. Dao T, Fu DY, Ermon S, Rudra A & Ré C FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness. https://arxiv.org/abs/2205.14135 (2022).
14. Vaswani A et al. Attention is all you need. Advances in Neural Information Processing Systems 2017-Decem, 5999–6009 (2017).
15. Wang G et al. HelixFold: An Efficient Implementation of AlphaFold2 using PaddlePaddle. https://arxiv.org/abs/2207.05477 (2022).
16. Cheng S et al. FastFold: Reducing AlphaFold Training Time from 11 Days to 67 Hours. https://arxiv.org/abs/2203.00854 (2022).
17. Li Z et al. Uni-Fold: An Open-Source Platform for Developing Protein Folding Models beyond AlphaFold. bioRxiv 2022.08.04.502811 (2022).
18. Ahdritz G, et al. OpenFold: Retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization. bioRxiv 2022.11.20.517210 (2022).
19. Alley EC, Khimulya G, Biswas S, AlQuraishi M & Church GM Unified rational protein engineering with sequence-based deep representation learning. Nature Methods 1315–1322 (2019) doi:10.1038/s41592-019-0598-1. [PubMed: 31636460]
20. Chowdhury R et al. Single-sequence protein structure prediction using a language model and deep learning. Nature Biotechnology 2022 1–7 (2022) doi:10.1038/s41587-022-01432-w.
21. Wang W, Peng Z & Yang J Single-sequence protein structure prediction using supervised transformer protein language models. bioRxiv 2022.01.15.476476 (2022).
22. Wua R & Dinga Fan, Wanga Rui, et al. High-resolution de novo structure prediction from primary sequence. bioRxiv 2022.07.21.500999 (2022).
23. Lin Z et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. bioRxiv 2022.07.20.500902 (2022).