Protein structure prediction has reached the single-structure frontier

Thomas J. Lane
Center for Free Electron Laser Science, Deutsches Elektronen-Synchrotron DESY
Notkestraße 85, 22607 Hamburg, Germany
<thomas.lane@desy.de>

Dramatic advances in protein structure prediction have sparked debate as to whether the problem of predicting structure from sequence is solved or not. Here, I argue that AlphaFold2 and its peers are currently limited by the fact that they predict only a single structure, instead of a structural distribution, and that this realization is crucial for the next generation of structure prediction algorithms.

The latest structure prediction methods, most prominently AlphaFold2 and RoseTTAFold,[1,2] have reduced the effort needed to obtain a structural model from months or years of laboratory work to a few keystrokes. While the predicted models do not reproduce experimental results in every case, this leap is dramatic enough to have provoked a series of self-reflective commentaries in the structural community,[3–5] some directly debating whether structural biology is "solved" or not.[6,7]

AlphaFold2 and its contemporaries aim to predict a *single* structure per sequence, yet proteins do not adopt a unique structural state. They move, and while not all motion is biologically meaningful, numerous lines of evidence have shown specific motions are necessary for protein function.[8,9] NMR, crystallography, and cryoEM have each been used to measure such motions,[10–12] which are often conceptualized as an energy landscape that describes the distribution of conformations a protein adopts and the rates at which those conformations interconvert.[13]

While the dynamic nature of proteins is widely accepted, the fact that structural heterogeneity manifests regularly in experimental structures – those in the protein databank (PDB) today – appears underappreciated. Indeed, it's common to hear talk of *the* structure of a particular protein, reflecting our biased thinking. This may be a byproduct of the expense of determining even a single structure for a given sequence. The PDB is strongly skewed towards single structures per sequence (supplemental figure 1). Nonetheless, whether a structure is derived from crystallography, cryoEM, or NMR data, it would be better to speak of it as *a* structure, one of many possible conformations.

Experimentally determined structures are not unique given a sequence

A distribution of structures can be determined from experimental data in one of two ways. First, multiple conformations are often required to model diffraction data, cryoEM images, or NMR NOEs from a single dataset. All three techniques average over an ensemble of many molecules in order to produce a signal. These averages are blurred by the structural differences from molecule to molecule, requiring models that incorporate structural heterogeneity to explain the data. This is accomplished by B-factors, multiple classes, alternative conformers, ensemble models, *etc*. Irrespective of the technique, it is essential to model the conformational heterogeneity to obtain satisfactory agreement with the measured experimental data.

A second level of variability exists. Repeated structural measurements of the same protein can yield different structures. While this may seem concerning, often these variations are intentional. For instance, different solution conditions can frequently be used to produce crystals with different lattices. The changes in packing can generate different structures.[14] While the structural changes are sometimes subtle, they exceed the coordinate error of the data. Alternatively, some structures may be determined in the presence of specific ligands, cofactors, ions, or other perturbations that change the protein structure from one experiment to the next.[15] Studies over the last decade show the temperature of data collection can alter the observed protein structure.[16] Additional variability arises from the experiments themselves: in crystallography, for instance, radiation damage incurred during data collection can do the same.[17] Even when these factors are controlled, variability persists. For instance, protein crystals are typically manually handled and plunge-frozen in liquid nitrogen before data collection, inducing idiosyncratic mechanical strain on the crystals lattice and frequently altering the determined structure.

The fact that a single sequence may give rise to many valid protein structures has implications for structure prediction. Because experimental structure determination can yield multiple outcomes, the accuracy of any single experiment is not an appropriate benchmark of success for structure prediction algorithms. Since no experimental data or model is perfect, every structure in the PDB will have some errors in the reported coordinate positions. There is some set of acceptable models that are "within error" for any given experiment. We might feel reasonable in saying the prediction has achieved experimental accuracy if a predicted structure's atomic coordinates are within this error, and that it has failed if it is outside this error limit. This would be too strict, however. Determining the structure of the same sequence but under different experimental conditions can produce appreciably different structures. These different structures would not be considered within error by any reasonable method of quantifying that error. Instead, if a predicted structure is contained within the *set* of possible experimental outcomes, it should be considered to have achieved experimental accuracy, even if no documented experiment precisely matches that result.

AlphaFold2 has reached the frontier of the single-structure approximation

The main protease (M^{pro}) from SARS-CoV-2 provides an excellent case study, highlighting why a single experiment cannot be used to assess the experimental accuracy of a prediction. During viral replication, much of the viral genome is synthesized into long polyproteins. M^{pro} processes these into individual proteins, and is essential for viral function. Intense interest in this system as a drug discovery target since late 2019 has resulted in a large number of high-quality structures of this system. As of June 2022, 452 structures of full-length, wild type M^{pro} were registered in the PDB, spanning many different crystallization conditions, bound ligands, and data collection temperatures. They show a correspondingly rich structural distribution (Fig. 1).

An AlphaFold2 model of M^{pro} compares favorably to this distribution. The smallest RMSD between the AlphaFold2 model and any PDB entry is 1.2 Å (PDB ID 7VLP, RMSD of non-hydrogen atoms modeled in all structures). The largest is 2.0 Å (PDB ID 7T46). Compare this to 1.75 Å, the largest RMSD between any pair of M^{pro} PDB depositions. Based on this simple metric, AlphaFold2 provides – in a few cases – a more accurate prediction of an experimental structure than would be provided by a different experimental structure!

The AlphaFold2 model is not in the geometric center of the experimental distribution (Fig. 1). On average, two randomly selected experimental structures will be more similar to each other than to the structure prediction. Still, the experimental-to-experimental distribution of RMSDs is large enough that it overlaps with the experimental-to-AlphaFold2 distribution. This result holds even when the experimental set is restricted to structures with no specific ligands bound (Fig. 1). AlphaFold2 has reached the frontier of the set of experimentally determined structures.

If a predicted structure is of sufficient quality to be contained in the set of experimental outcomes, this has direct implications for applications in which one might substitute a prediction for an experimental structure. Consider structure-based drug discovery, where atomically resolved structures are desirable. For targets with high-quality structure predictions and structural variability in the ligand-binding site, *in silico* ligand screening procedures that employ a rigid protein receptor will be more limited by a lack of protein flexibility than the accuracy of the structure prediction. M^{pro} is a prominent drug target that appears to fit this paradigm.

Putative drug-discovery applications highlight why the set of possible experimental outcomes is a more useful definition of "experimental accuracy" than coordinate error. For each individual experimental structure, the all-atom RMSDs between the AlphaFold2 model are at least twice as large as the reported coordinate error for these structures (supplemental figure 2). For any *single* M^{pro} experimental dataset, the AlphaFold2 structure would not be an acceptable model to explain the data. Still, this does not mean it is less informative in terms of scientific insight than an experimental structure of M^{pro} randomly selected from the PDB.

A single structure cannot capture functional motion

Structural heterogeneity underpins protein function. An archetypical example is hemoglobin, the protein that transports oxygen in humans and all other known vertebrates except icefish. In the process, hemoglobin toggles between a "tense" oxygen-free state (T) and a "relaxed" oxygen-bound state (R).[9] A survey of 16 human hemoglobin structures deposited in the PDB reveals that the oxygen-free T-state structures are highly similar to one another, reflecting the rigid, low-entropy nature of this state. In contrast, the R-state structures, bound to either O_2 or CO, are considerably more diverse. The atomic coordinates of AlphaFold2's model of hemoglobin lie geometrically halfway between these two structural extremes (Fig. 2). This is, perhaps, expected – with no information about the presence or absence of oxygen, a prediction intermediate between the R and T structures seems ideal. This case demonstrates both how good current structure prediction algorithms are and the opportunity to better understand biology through protein dynamics. The latter cannot be captured by a single structure.

In fact, AlphaFold2 may be capable of self-reporting when its single-structure approximation breaks down. ABL, a tyrosine kinase, provides an example. ABL is involved in cell differentiation, division, adhesion, and DNA repair, and is an oncology drug target. Like other kinases, ABL exhibits a flexible "activation" loop containing a three-amino acid D-F-G motif, which toggles between an active DFG-in and inactive DFG-out state. This structural flexibility is essential for the protein's regulation and function. In the 25 human ABL kinase structures deposited in the PDB, the activation loop adopts a variety of conformations. Each conformation is generated by the experimental conditions, notably ligands and crystallization reagents. In each crystal dataset, the various loop positions are unambiguous but, on average, less well-ordered than the surrounding atoms (judged by B-factors and map correlation, supplemental figure 4).

AlphaFold2 produces a per-residue prediction of confidence, the predicted Local Distance Difference Test (pLDDT).[1,18] In ABL's regions of high structural variability, the confidence is correspondingly lower (Fig. 3). Notably, AlphaFold2 correctly reports reduced confidence in the functional activation loop, where the ensemble is not well captured by a single structure. In this case, the algorithm simply predicts a common conformation and down-weights the prediction confidence in regions of high variability, highlighting the limit of the single-structure approximation. Structural variability is only one reason structure prediction confidence might be lowered. For instance, the AlphaFold authors note that regions of low sequence coverage also exhibit reduced pLDDT.[1] A precise accounting of the contributions to AlphaFold2's errors is beyond this work, but the ABL case study suggests structural heterogeneity should be considered a significant factor.

Supporting the notion that AlphaFold2 might already contain information about the distribution of protein structure is work on fold-switching proteins. Chakravarty and Porter have reported that for these systems, AlphaFold2 typically succeeds in predicting the structure of one fold, but not the other.[19] Predictions of fold-switchers have moderately reduced pLDDT *vs.* single-fold proteins, but the predicted LDDTs are still substantially higher than for disordered proteins or disordered regions of specific proteins. Even more strikingly, Wayment-Steele and colleagues have shown that by clustering the sequences used by AlphaFold2, the algorithm can in fact predict both folds in specific fold-switching systems.[20] This exciting result suggests that predictions of structural distributions may not be far off, as AlphaFold2 can already produce multiple correct structural outputs for the same protein. Further, these results reinforce the hypothesis that current models are limited by a single-structure output space.

Distributions of conformations are the future of structural biology

Structure prediction has advanced significantly, right up to the frontier where the single sequence, single structure approximation has broken down. While AlphaFold2 and its peers enable breakthroughs today, tomorrow's challenge is already clear: modeling protein structural distributions. If we could model and predict distributions, M^{pro}, hemoglobin and ABL kinase each provide a strong argument for what we might learn, and why a single-structure view cannot capture all of protein function.

Consider hemoglobin. A single structure cannot describe both the T and R states, cannot capture how oxygen is bound and released, and therefore cannot capture the protein's biological role. It is not difficult, however, to imagine an extension to the current structure prediction model, where the output space is a distribution of protein conformations, not just a single structure. The scientific impact of such models would be greatly enhanced if the distribution could be modeled as a function of relevant conditions: ligands, pH, binding partners, oxygen concentration, temperature, *etc.* How to build such models, manage the potentially endless list of possible model inputs, benchmark them against experimental data, and ensure humans can enjoy learning from such models remains hard work. But while the path is not known, the direction is clear. Machine learning algorithms capable of modeling continuous distributions of protein structure are already emerging, and we can expect significant progress in the coming years.[21–24]

The mission of learning protein distributions will require new abstractions and representations of protein structure, but also new experiments to train and guide those algorithms. Time-resolved crystallography,[25] modeling continuous structural distributions from cryoEM data,[21,22] and even the analysis of large, statistically significant sets of crystals of the same protein are already pushing the frontier of our knowledge. These efforts must continue and expand to support the training of models of protein structure distributions.

Understanding how protein structure changes upon ligand binding, interface formation, or changes in the surrounding environment represents an opportunity to get to the heart of protein function. In light of this, we might look back on this time not as the age in which structural biology was solved, but as its golden age.

Acknowledgements

The author is supported by a Young Investigator Group award from the Helmholtz Association, and would like to thank K. Dalton, A. Fadini, J. Greisman, R. Townshend, D. Hekstra, A. Peck, S. Günther, A. Mancuso and L. Aithani for discussion and suggestions. An anonymous reviewer provided suggestions that improved supplemental figure 2, and I want to sincerely thank T. Terwilliger for help implementing the map-model correlations reported in that figure. Further, K. Diederichs kindly provided insight into crystallographic coordinate errors, reported in supplemental figure 2. Finally, I apologize to the authors whose work I could not cite due to space limitations.

Competing Financial Interests

The author declares no competing financial interests at the time of writing and submission.
After submission but before final acceptance, the author became an employee and shareholder of CHARM Therapeutics Inc.

References

1. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

2. Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).

3. Cramer, P. AlphaFold2 and the future of structural biology. *Nat. Struct. Mol. Biol.* **28**, 704–705 (2021).

4. Jones, D. T. & Thornton, J. M. The impact of AlphaFold2 one year on. *Nat. Methods* **19**, 15–20 (2022).

5. Kleywegt, G. J. & Velankar, S. Whither structural biologists? *IUCrJ* **9**, 399–400 (2022).

6. Ourmazd, A., Moffat, K. & Lattman, E. E. Structural biology is solved — now what? *Nat. Methods* **19**, 24–26 (2022).

7. Moore, P. B., Hendrickson, W. A., Henderson, R. & Brunger, A. T. The protein-folding problem: Not yet solved. *Science* **375**, 507–507 (2022).

8. Schnell, J. R., Dyson, H. J. & Wright, P. E. Structure, dynamics, and catalytic function of dihydrofolate reductase. *Annu. Rev. Biophys. Biomol. Struct.* **33**, 119–140 (2004).

9. Yuan, Y., Tam, M. F., Simplaceanu, V. & Ho, C. New look at hemoglobin allostery. *Chem. Rev.* **115**, 1702–1724 (2015).

10. Eisenmesser, E. Z. *et al.* Intrinsic dynamics of an enzyme underlies catalysis. *Nature* **438**, 117–121 (2005).

11. Fraser, J. S. *et al.* Hidden alternative structures of proline isomerase essential for catalysis. *Nature* **462**, 669–673 (2009).

12. Dashti, A. *et al.* Retrieving functional pathways of biomolecules from single-particle snapshots. *Nat. Commun.* **11**, 4734 (2020).

13. Frauenfelder, H., Sligar, S. G. & Wolynes, P. G. The Energy Landscapes and Motions of Proteins. *Science* **254**, 1598–1603 (1991).

14. Kondrashov, D. A., Zhang, W., Aranda IV, R., Stec, B. & Phillips, G. N. Sampling of the native conformational ensemble of myoglobin via structures in different crystalline environments. *Proteins Struct. Funct. Genet.* **70**, 353–362 (2008).

15. Wankowicz, S. A., de Oliveira, S. H., Hogan, D. W., van den Bedem, H. & Fraser, J. S. Ligand binding remodels protein side-chain conformational heterogeneity. *eLife* **11**, e74114 (2022).

16. Keedy, D. A. *et al.* Mapping the conformational landscape of a dynamic enzyme by multitemperature and XFEL crystallography. *eLife* (2015) doi:10.7554/elife.07574.

17. Garman, E. F. Radiation damage in macromolecular crystallography: what is it and why should we care? *Acta Crystallogr. D Biol. Crystallogr.* **66**, 339–351 (2010).

18. Mariani, V., Biasini, M., Barbato, A. & Schwede, T. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics* **29**, 2722–2728 (2013).

19. Chakravarty, D. & Porter, L. L. AlphaFold2 fails to predict protein fold switching. 11.

20. Wayment-Steele, H. K., Ovchinnikov, S., Colwell, L. & Kern, D. *Prediction of multiple conformational states by combining sequence clustering with AlphaFold2*. http://biorxiv.org/lookup/doi/10.1101/2022.10.17.512570 (2022) doi:10.1101/2022.10.17.512570.

21. Zhong, E. D., Bepler, T., Berger, B. & Davis, J. H. CryoDRGN: reconstruction of heterogeneous cryo-EM structures using neural networks. *Nat. Methods* **18**, 176–185 (2021).

22. Rosenbaum, D. *et al.* Inferring a Continuous Distribution of Atom Coordinates from Cryo-EM Images using VAEs. Preprint at http://arxiv.org/abs/2106.14108 (2021).

23. Gupta, H., McCann, M. T., Donati, L. & Unser, M. CryoGAN: A New Reconstruction Paradigm for Single-particle Cryo-EM Via Deep Adversarial Learning. 2020.03.20.001016 Preprint at https://doi.org/10.1101/2020.03.20.001016 (2020).

24. Anand, N. & Achim, T. Protein Structure and Sequence Generation with Equivariant Denoising Diffusion Probabilistic Models. Preprint at https://doi.org/10.48550/arXiv.2205.15019 (2022).

25. Brändén, G. & Neutze, R. Advances and challenges in time-resolved macromolecular crystallography. *Science* **373**, eaba0954 (2021).

Figure Legends

Figure 1. AlphaFold2's model of M^{pro} is on the frontier of the set of experimentally determined structures. Panel A: the M^{pro} functional unit, a dimer (grey surface), is shown with the backbone traces of 64 ligand-free structures deposited in the PDB (blue) aligned to AlphaFold2's model (magenta). Panel B: RMSD between all pairs of M^{pro} PDB entries (blue) and from AlphaFold2's model to each PDB entry (yellow). Panel B, top: filtered for M^{pro} structures without specific ligands bound ("ligands" excludes crystallization reagents, buffers, salts). Panel B, bottom: all M^{pro} PDB entries. RMSDs are computed for all protein, non-hydrogen atoms modeled in all structures. PDB IDs included are those with 100% sequence identity match to PDBID 7ar6; a list is included in the supplementary information.

Figure 2. AlphaFold2's prediction of hemoglobin lies between the R and T states. Shown are experimental structures of O_2- or CO-bound (R, orange) and ligand-free (T, blue) hemoglobin, superimposed with the AlphaFold2 prediction (magenta). Right insert: detail of the structural superimposition. Left insert: PCA analysis of the dihedral angles shows the atomic coordinates in a space of reduced dimensionality. Plotted are the first two principal components, which explain 45% of the total variance (supplemental figure 3). Note: included T-state structure 1HGC is partially O2-bound, with the two α-subunits binding and the β-subunits ligand free. Structures selected from those highlighted in ref. [9].

Figure 3. AlphaFold2 uncertainty correlates with structural variability in ABL kinase. Left: for 25 structures of ABL kinase, the LDDT computed across the experimental ensemble plotted against the AlphaFold2 model's certainty metric, predicted LDDT (Spearman correlation coefficient 0.47). Color of markers shows position in sequence. The dynamic N-terminus and activation loop exhibit notably low confidence. Right: structural superimposition of the same experimental structures from the PDB, colored by AlphaFold2's confidence (pLDDT; AlphaFold2 model not shown). In regions of high variability between structures, such as the activation loop, the pLDDT reports lower confidence. PDB IDs listed in the supplemental information.