Commentary

Structure is beauty, but not always truth

James S. Fraser[1],* and **Mark A. Murcko[2],***
[1]Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, CA, USA
[2]Disruptive Biomedical LLC, Holliston, MA, USA
*Correspondence: jfraser@fraserlab.com (J.S.F.), mark.murcko@gmail.com (M.A.M.)
https://doi.org/10.1016/j.cell.2024.01.003

Structural biology, as powerful as it is, can be misleading. We highlight four fundamental challenges: interpreting raw experimental data; accounting for motion; addressing the misleading nature of *in vitro* structures; and unraveling interactions between drugs and "anti-targets." Overcoming these challenges will amplify the impact of structural biology on drug discovery.

Introduction

Amidst the many uncertainties that complicate drug discovery, structural biology anchors the process in beautiful and concrete images of drugs interacting with receptors. Structure can enable the tackling of many of the key challenges of drug design. Atomistic models emerging from cryo-electron microscopy (cryo-EM), X-ray crystallography, and nuclear magnetic resonance (NMR) provide strong starting points for thinking broadly and creatively about how to modulate protein function by identifying binding pockets and potential allosteric sites. Ligand-bound structures greatly focus the search of chemical space to molecules that maintain key interactions with the receptor.

By offering a "ground truth," structural biology is clear, quantifiable, and interpretable. In the best cases, the precise location of every atom is clearly defined. For example, measuring the distances between atoms on the ligand and receptor allow us to infer that "this hydrogen bond is better than that one." By comparison, the biology of the target is often quite complex and difficult to model quantitatively. Cellular assays and animal models are approximations that do not fully recapitulate the human disease process or the potential of small molecules to induce toxicities. In addition, medicinal chemistry is full of uncertainty. Even with a structure, in the course of inhibitor optimization, it is difficult to know what molecules to make next or how to synthesize them. Finally, we struggle to understand why some molecules are more potent or more bioavailable than others, making it next to impossible to optimize the pharmacokinetic and safety profile of a drug candidate.

However, the "truth" of structural biology raises a legitimate concern: does the availability of structural information, such as a protein crystal structure, irreparably *constrain* the creative process? This is a potential risk for those who fail to recognize the inherent limitations in the structures and the new predictions (e.g., AlphaFold2[1] trained on the corpus of the Protein DataBank (PDB). We suggest it is useful to consider four kinds of limitations and offer ways that the field can address each of them to optimize the value we derive from structural biology and further improve the quality of predictive modeling.

Four harsh truths about structural biology and drug discovery

(1) A structure is a model, not experimental reality.
(2) Representing wiggling and jiggling is hard.
(3) *In vitro* can be deceiving.
(4) Drugs mingle with many different receptors.

A structure is a model, not experimental reality

Undoubtedly, AlphaFold2[1] shook up the field of structural biology by "solving" the protein structure prediction problem. "Solving" means that the predicted models are highly similar to "ground truth" experimentally determined structures by the metrics used by the CASP (Critical Assessment of Structure Prediction) competition, a community-wide experiment to determine and advance the state of the art in modeling protein structure from amino acid sequence. It is

important to note that ground truth structures contain inaccuracies beyond the signal-to-noise of the experiments that generate them. For example, in X-ray crystallography, the experimental data is measured very precisely (usually to less than 5% error), but the structures refined against that data have large residual errors compared to the experimental data (generally $>\sim20\%$).[2] Adding in prior knowledge, such as geometry restraints, is especially important as the resolution of the experimental data gets worse.[2] "Truth" therefore may not lie only in comparison to the atomic coordinates, especially when the structure is based on low-resolution data. Rather, comparisons to density maps (or even raw diffraction images or micrographs) may reveal a deeper form of truth (Figure 1).

Recent work has started the important task of comparing AlphaFold2 models directly to experimental crystallographic density maps.[3] In many cases, predictions closely matched experimental maps. Refinement of the AlphaFold2 models against experimental data can resolve some global scale distortion and issues of domain orientation. Refinement also improves local backbone and side-chain conformations. However, most very high-confidence predictions differed from experimental maps to a greater extent than independently determined experimental structures.

Beyond suggesting that direct agreement with experimental data, not "structures," could be a new benchmark of "ground truth," these findings prompt us to ask how we can maximize the utility of computationally predicted models in drug discovery. Some differences
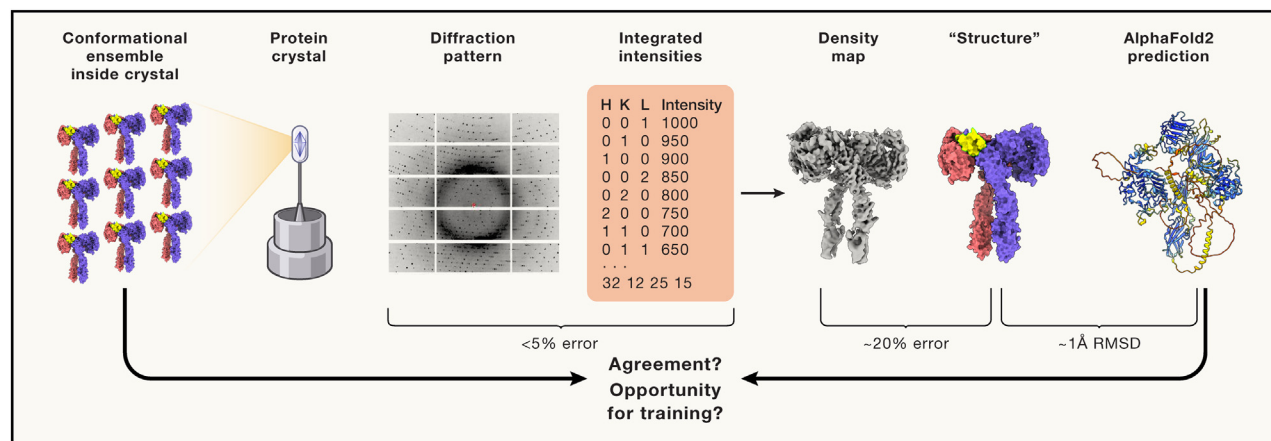
Figure 1. The loss of information along the dataflow of structural biology
While computationally predicted models are currently trained against "structure" from the PDB, there is potential to increase the quality of models by looking at the agreement between earlier data transformations. In X-ray crystallography, the diffraction from a crystal represents contributions from many molecules that adopt distinct compositions inside the crystal. The estimated precision of integrated experimental intensities is typically much higher than the agreement between the density map and the model, indicating that the "structure" can still be improved. Agreement between AlphaFold2 predictions and the structure are typically in the range of what would be expected of independently determined low-resolution (4–5 Å) experimental structures. The agreement may be improved in the future by looking earlier in the dataflow for training the models. Analogies to these rawer forms of truth exist in cryo-EM (e.g., raw micrographs, particle stacks, 3D volumes) and NMR (e.g., Nuclear Overhauser Effect [NOE], Residual Dipolar Coupling [RDC], chemical shifts).

between predicted models and the experimental structure (and even underlying data) may reflect a bias toward a less-explored part of the energy landscape of conformations populated by the protein.[4] Moreover, the value of exposing predicted models to orthogonal computational techniques, like long molecular dynamics simulations, is currently unclear. Despite these concerns, AlphaFold2 and related approaches are already having a huge impact in drug discovery, ranging from areas often considered mundane (e.g., DNA construct design) to those widely considered to be exciting (e.g., generative artificial intelligence [AI] modeling of ligands into predicted binding pockets).

Computationally predicted models therefore have great potential to reduce some of the early-stage uncertainty in drug discovery that occurs prior to structure enablement. The release of the AlphaFold code base spurred a Cambrian explosion of structural bioinformatics and unanticipated findings (e.g., prediction of protein complexes using AlphaFold Multimer). Disturbingly, though, the next stage of development of AlphaFold is clouded with uncertainty, as the disclosure of methods has moved from preprints, GitHub, and journals to a company blogpost without accompanying methods.[5] Without open methods, it is difficult to

tell whether we are approaching a plateau in structure prediction accuracy. When such a plateau is reached, we will need to know how much of it is due to a faulty definition of "ground truth." It is likely that more direct training against experimental data, not refined structures, will be required for further improvement in structure prediction accuracy. Moreover, the recognition that the structure is a model of experimental data and that the experimental data actually represents the average of many (moving) molecules may unlock new capabilities.

Representing wiggling and jiggling is hard

The profound wisdom in Feyman's statement, "Everything that living things do can be understood in terms of the jigglings and wigglings of atoms," suggests the need for a wholesale redefinition of ground truth. We can account for the macromolecular movements that are crucial for drug discovery and reshape our perspective to account for the dynamic nature of biomolecules and the existence of ensembles.

A few proteins are so simple that they can largely be considered static for structure-based drug design. But even in the paradigmatic example, carbonic anhydrase, an active site residue, His 64, can undergo a side chain $\chi 1$ rotation and change the shape of the binding pocket.

Recognizing this rotation was essential to optimize the properties of the glaucoma drug dorzolamide.[6] Even this type of simple side chain motion is currently difficult to predict, revealing an important reason to get co-complexes quickly, with any kinds of ligands, whether considered "drug-like" or not, in a drug discovery campaign. Banging on the walls of the protein surface is also an effective way to find alternate binding sites and cryptic pockets. Such strategies could lead a team to make compounds that simply should not fit in a static binding pocket, but against all odds do bind, and thereby reveal the intrinsic dynamics of the receptor.

Even with the recognition that a single structure can be misleading, making current AI pipelines aware of the multiple truths and generating a probabilistic ensemble remains a frontier challenge. Current generative models can produce structures from a latent space that may be related to the underlying energy landscape of the system. Much like the change from classical to quantum mechanics a century ago in physics, a more probabilistic view of protein conformational landscapes will likely explain properties that cannot be explained from single structures alone.

However, current simulation methods are hard to apply because the resulting states are often rare and interconvert
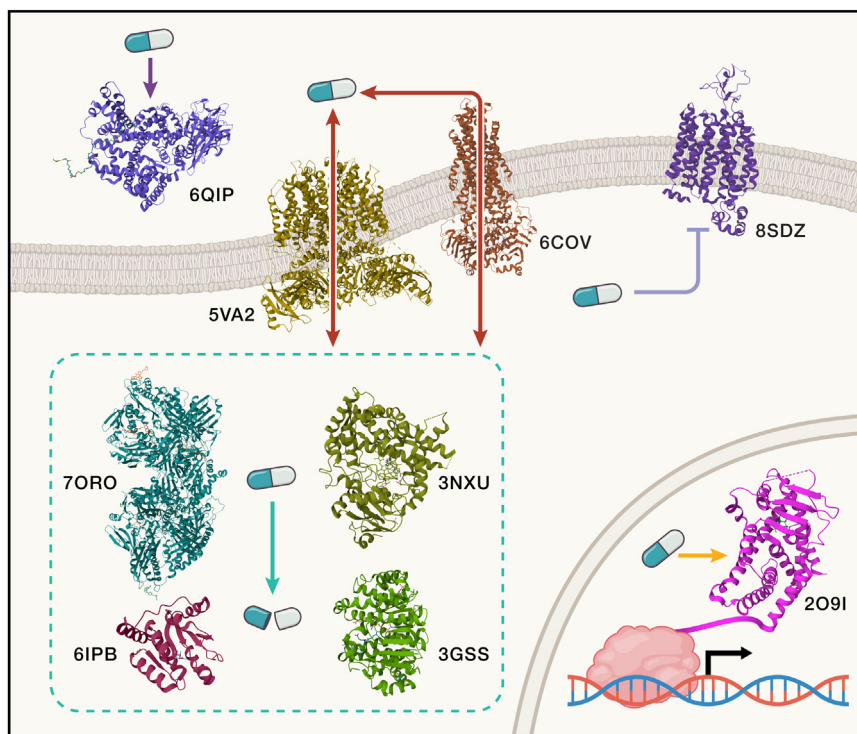
Figure 2. Structural representations of selected proteins associated with the drug metabolism and pharmacokinetics (DMPK)-related "avoidome"

Proteins in the extracellular milieu, like human serum albumin (PDB: 6QIP), greatly affect distribution by binding to drugs (purple arrow). Membrane proteins are involved in transport of drugs into and out of cells (red arrows), including: P-glycoprotein 1 (PGP, MDR1, ABCB1) (PDB: 6C0V) and the Organic anion transporter 1 (OAT1) (PDB: 8SDZ). Enzymes involved in metabolism alter the chemical structures of drugs (green arrow), including: Glutathione S-Transferase (PDB: 3GSS), UDP glucuronosyltransferase (PDB: 6IPB), P450 CYP3A4 (PDB: 3NXU), and Aldehyde Oxidase (PDB: 7ORC). The xenobiotic transcriptional response (yellow arrow) is mediated by direct binding to transcription factors, including the Pregnane X Receptor (PXR) (PDB: 2O9I). Finally, toxicology can emerge due to promiscuous binding (blue arrow) to anti-targets, including the human Ether-à-go-go-Related Gene (hERG) potassium channel (PDB: 5VA2). We recognize that occasionally it may be desirable to target certain proteins in the "avoidome"[12,13]; for example, the COVID-19 medicine paxlovid contains two active ingredients, nirmatrelvir (the actual antiviral agent) and ritonavir, which blocks cytochrome P450 3A4. Ritonavir reduces the metabolism of nirmatrelvir, increasing its effectiveness. These special cases notwithstanding, *in general*, the goal of a drug discovery team is to avoid interacting with the avoidome class of proteins. These structures exemplify the molecular diversity and the intricate interplay of protein-ligand interactions within the avoidome.

slowly. Refining ensembles with greater agreement to experimental data may provide the substrates for the next breakthrough in both single-structure and ensemble prediction. Analogous to how structure-based drug design is great for optimizing "surface complementarity" and electrostatics, future protein modeling approaches will unlock ensemble-based drug design with an ability to predictably tune new and important aspects of design, including entropic contributions[7] and residence times[8] of bound ligands.

In vitro *can be deceiving*

While purifying a protein out of its cellular context can be enabling for *in vitro* drug discovery, it can also provide a false impression. Recombinant expression can lead to missing post-translational modifications (e.g., phosphorylation or glycosylation) that are critical to understanding the function of a protein. One of the most exciting realizations of Alpha-Fold2 predictions was that the model was somehow "aware" of parts of the native environment that a purely physics-based prediction would miss. Predicted structures are so poised to be filled with prosthetic groups (e.g., heme), metals, and metabolites that they can be "transplanted" into the models with minimal refinement.[9]

Isolated structures of proteins become more and more misleading as the focus of drug discovery shifts to complex biological systems that include multi-protein complexes, protein-RNA interactions, and cellular condensates enriched with intrinsically disordered proteins.[10] Emerging techniques, especially cryo-electron tomography (cryo-ET), have great potential to deliver atomistic insights directly from observations in cells. An early example of cryo-ET has revealed how ribosomes bound to the antibiotic chloramphenicol are enriched in elongation states that lead to collisions.[11] These techniques will eventually answer questions about the residual structure in "disordered regions" that cannot be addressed without considering the local cellular environment. In doing so, the applicability and relevance of structural biology to drug discovery will undoubtedly increase.

Drugs mingle with many different receptors

The sad reality that all drug discoverers must face is that however well designed we may believe our compounds to be, they will find ways to interact with many other proteins or nucleic acids in the body and interfere with the normal functions of those biomolecules. While occasionally, the ability of a medicine to bind to multiple biomolecules will increase a drug's efficacy, such polypharmacology is far more likely to produce undesirable effects. These undesirable outcomes take two forms. Obviously, the direct binding to an anti-target can lead to a bewildering range of toxicities, many of which render the drug too hazardous for any use. More subtly, the binding to anti-targets reduces the ability of the drug to reach the *desired* target. A drug that largely avoids binding to anti-targets will partition more effectively through the body, enabling it to accumulate at high enough concentrations in the disease-relevant tissue to effectively modulate the function of the target.

A particular challenge results from the interaction of drugs with the enzymes, transporters, channels, and receptors that are largely responsible for controlling the metabolism and pharmacokinetic properties (DMPK) of those drugs—their absorption, distribution, metabolism, and elimination. Drugs often bind to plasma proteins, preventing them from reaching the intended tissues; they can block or be substrates for all manner

of pumps and transporters, changing their distribution through the body; they occasionally interfere with xenobiotic sensors such as PXR that turn on transcriptional programs recognizing foreign substances; and they often block enzymes like cytochrome P450s, thereby changing their own metabolism and that of other medicines. They are themselves substrates for P450s and other metabolizing enzymes and, once altered, can no longer carry out their assigned, life-saving function.

Taken together, we refer to these DMPK-related proteins, somewhat tongue-in-cheek, as the "avoidome" (Figure 2). Unfortunately, the structures of the vast majority of avoidome targets have not yet been determined. Further, many of these proteins are complex machines that contain multiple domains and exhibit considerable structural dynamism. Their binding pockets can be quite large and promiscuous, favoring distinct binding modes for even closely related compounds. As a consequence, multiple structures spanning a range of bound ligands and protein conformational states will be required to fully understand how best to prevent drugs from engaging these problematic anti-targets.

We believe the structural biology community should "embrace the avoidome" with the same enthusiasm that structure-based design has been applied to intended targets. The structures of these proteins will shed considerable light on human biology and represent exciting opportunities to demonstrate the power of cutting-edge structural techniques. Crucially, a detailed understanding of the ways that drugs engage with avoidome targets would significantly expedite drug discovery. This information holds the potential to achieve a profound impact on the discovery of new and enhanced medicines.

Conclusion

In drug discovery, truth is a molecule that transforms the practice of medicine. A drug prevents, ameliorates, or cures a disease. It is well tolerated and practical to use in the real world. Sadly, few important new medicines are created each year. Despite the limitations imposed by the four harsh truths we have described, structural information, thoughtfully applied, has consistently demonstrated its utility for drug discovery. Indeed, 2024 will mark the 30th anniversary of the FDA approval of dorzolamide, the first drug that benefited from structure-based design.[6] The coming decade will witness exciting progress at addressing these limitations, unlocking new efficiencies in the drug discovery process and contributing to an ever-increasing extent to the discovery of future medicines. We suggest that focusing machine learning efforts on these four challenges will complement and enhance the coming improvements in experimental disciplines to further accelerate our progress.

DECLARATION OF GENERATIVE AI USED IN THE WRITING PROCESS

During the preparation of this work, the authors used ChatGPT to summarize some notes from our back-and-forth emails planning the article. It was helpful, but honestly less than we had hoped for when agreeing to write this perspective. The authors take full blame/responsibility for the content of the publication.

REFERENCES

1. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589.

2. Holton, J.M., Classen, S., Frankel, K.A., and Tainer, J.A. (2014). The R-factor gap in macromolecular crystallography: an untapped potential for insights on accurate structures. FEBS J. 281, 4046–4060.

3. Terwilliger, T.C., Liebschner, D., Croll, T.I., Williams, C.J., McCoy, A.J., Poon, B.K., Afonine, P.V., Oeffner, R.D., Richardson, J.S., Read, R.J., et al. (2023). AlphaFold predictions are valuable hypotheses and accelerate but do not replace experimental structure determination. Nat. Methods. https://doi.org/10.1038/s41592-023-02087-4.

4. Tyka, M.D., Keedy, D.A., André, I., Dimaio, F., Song, Y., Richardson, D.C., Richardson, J.S., and Baker, D. (2011). Alternate states of proteins revealed by detailed energy landscape mapping. J. Mol. Biol. 405, 607–618.

5. Google DeepMind AlphaFold team; Isomorphic Labs team (2023). A glimpse of the next generation of AlphaFold. Isomorphic Labs. https://www.isomorphiclabs.com/articles/a-glimpse-of-the-next-generation-of-alphafold.

6. Baldwin, J.J., Ponticello, G.S., Anderson, P.S., Christy, M.E., Murcko, M.A., Randall, W.C., Schwam, H., Sugrue, M.F., Springer, J.P., and Gautheron, P. (1989). Thienothiopyran-2-sulfonamides: novel topically active carbonic anhydrase inhibitors for the treatment of glaucoma. J. Med. Chem. 32, 2510–2513.

7. Wankowicz, S.A., de Oliveira, S.H., Hogan, D.W., van den Bedem, H., and Fraser, J.S. (2022). Ligand binding remodels protein side-chain conformational heterogeneity. Elife 11, e74114. https://doi.org/10.7554/eLife.74114.

8. Copeland, R.A. (2016). The drug-target residence time model: a 10-year retrospective. Nat. Rev. Drug Discov. 15, 87–95.

9. Hekkelman, M.L., de Vries, I., Joosten, R.P., and Perrakis, A. (2023). AlphaFill: enriching AlphaFold models with ligands and cofactors. Nat. Methods 20, 205–213.

10. Mitrea, D.M., Mittasch, M., Gomes, B.F., Klein, I.A., and Murcko, M.A. (2022). Modulating biomolecular condensates: a novel approach to drug discovery. Nat. Rev. Drug Discov. 21, 841–862.

11. Xue, L., Spahn, C.M.T., Schacherl, M., and Mahamid, J. (2023). Structural insights into context-dependent inhibitory mechanisms of chloramphenicol in cells. bioRxiv. https://doi.org/10.1101/2023.06.07.544107.

12. Foti, R.S. (2023). Cytochrome P450 and Other Drug-Metabolizing Enzymes As Therapeutic Targets. Drug Metab. Dispos. 51, 936–949.

13. Guengerich, F.P. (2023). Cytochrome P450 enzymes as drug targets in human disease. Drug Metab. Dispos. DMD-MR-2023-001431. https://doi.org/10.1124/dmd.123.001431.