https://doi.org/10.1038/s41589-024-01570-z

Big data and benchmarking initiatives to bridge the gap from AlphaFold to drug design

Matthieu Schapira, Levon Halabelian, Cheryl H. Arrowsmith & Rachel J. Harding

AlphaFold is a breakthrough in protein structure prediction, but limitations in its application to computation- and structure-guided drug discovery remain. As with structure prediction, public-domain data and benchmarking initiatives will be essential to advance the field of computational drug design.

Computational modeling of ligands bound to protein drug targets can have a significant role in the early phases of drug-discovery programs. However, before the release of AlphaFold and its peers, many targets lacked a structural model, hindering this line of enquiry. The recent advent of large-scale protein structure prediction technologies has led to a paradigm shift across many fields in how scientists from diverse disciplines consider, interrogate and exploit 3D protein structures. Since its original release[1], AlphaFold has been hailed for its far-reaching utility and potential, and it has inspired a cohort of experts who are actively building upon its foundational technology. The application of AlphaFold to structure-based drug discovery and related computational chemistry approaches has been the subject of intense scrutiny by drug hunters, revealing benefits and weaknesses in the utility of predicted structures for these efforts. These recent studies, and the growing potential of artificial intelligence (AI) in computational drug discovery, underscore the need for more experimental data to train machine learning models and continued benchmarking challenges to accelerate the field.

Open data enabled structure prediction breakthroughs

The breakthrough in structure prediction algorithms was enabled by the confluence of two key fields. Recent advances in AI and deep learning methods exploited the vast amount of experimental protein structure data *publicly* available under open license in the Protein Data Bank (PDB), generated over 50 years by structural biologists around the world. This rich database includes thousands of structures from structural genomics endeavors such as the NIH Protein Structure Initiative, which specifically sought to increase the number of structures of 'novel folds', serving as a critical training dataset to enable methods development for the computational prediction of protein structures.

The superiority of AI-based computational structure prediction methods was demonstrated via the benchmarking exercise CASP (Critical Assessment of Protein Structure Prediction), now in its fifteenth iteration[2]. CASP is a global competition that assesses the state of the art in the prediction of protein structures for which experimentally determined structures are solved, but not yet disclosed. Each prediction team's method is evaluated by assessing the accuracy and reliability of the predicted structure relative to the reference experimental structure. This mechanism for independent validation of different computational strategies helped monitor progress and enabled continuous improvement of protein structure prediction methods over the years. Despite these advances, limitations remain for the use of predicted structures, especially in drug-discovery efforts. Here we provide an overview of some challenges in this field and highlight the need for benchmarking exercises in small-molecule computational drug design using predicted protein structures, and in computational macromolecule–ligand interaction analyses in general.

Challenges in using predicted structures in virtual screening

The ability to predict the globular portions of protein structures readily and accurately can help accelerate the experimental structural biology component of a drug-discovery program in several ways. For example, predicted structures can provide critical information, such as domain boundaries, for designing expression constructs, thereby enabling or accelerating protein production or crystallization. Predicted models can also be used in X-ray crystallography to obtain initial phases by molecular replacement, and for building larger multiprotein complexes in cryogenic electron microscopy maps.

A key issue is that many predicted structures, or domains and regions within larger structures, are predicted at low confidence. These low-confidence regions tend to be in areas that require binding of partner proteins, cofactors, peptides or substrates to take on their 3D folded structure and can constitute critical regions of druggable binding pockets. So far, most prediction methods are limited by their consideration of proteins in isolation as monomeric, solely proteinogenic entities, with their physiologically active homo- or hetero-complex form not presented. This limitation was exemplified in an analysis of predictions for fold-switching proteins[3], in which conformational changes from the ground to excited state are induced upon binding of a substrate, ligand or other protein. AlphaFold2 (AF2) failed to predict the ground-state conformation in 30% of cases of this class. To overcome these challenges, advances are rapidly being made to 'add in' missing cofactors[4] and to model protein–protein complexes[5] and protein–nucleic acid complexes[6] so as to improve the protein models.

Another current limitation of prediction technologies is the omission of post-translational modifications (PTMs) of amino acids. Many proteins, especially multidomain proteins, perform their functions via conformational changes regulated by post-translational modification. Such conformational transitions are an attractive process to target through allosteric modulators, as recently illustrated for Casitas B lymphoma-b (Cbl-b), a multidomain E3 ubiquitin ligase that regulates tyrosine kinase signaling. Cbl-b has an N-terminal region comprising a tyrosine kinase-binding domain (TKBD), a short linker helix region (LHR) and a RING finger domain. A recent co-crystal structure of Cbl-b in complex with an inhibitor revealed that the inhibitor

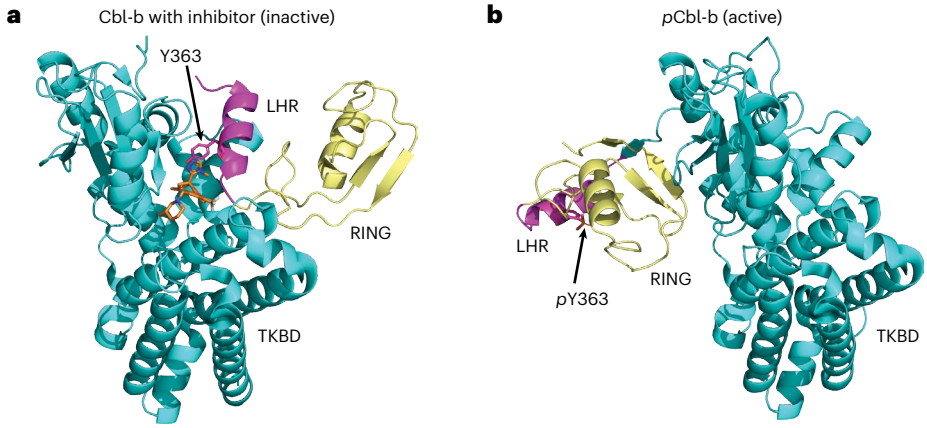

Fig. 1 | Comparison of inactive and active conformations of Cbl-b, the latter induced by phosphorylation. a, Crystal structure of Cbl-b bound to an inhibitor, adopting inactive (closed) conformation (PDB 8GCY). The inhibitor is shown in orange sticks. **b**, Crystal structure of Y363-phosphorylated Cbl-b adopting active conformation (PDB 3ZNI). TKBD, LHR and RING domains in both structures are shown in cyan, magenta and yellow, respectively. *p*, phospho.

interacts with both the TKBD and LHR, causing Cbl-b to adopt a closed (inactive) conformation (Fig. 1). By contrast, in its active form, a critical tyrosine in the LHR is phosphorylated, which triggers large conformational changes resulting in an open protein conformation. In such cases, it is difficult to predict a priori the conformational changes that a protein undergoes upon phosphorylation or substrate binding to create distinct druggable pockets for inhibitor design. Moreover, in most cases, the true differences between conformational states mediated by PTMs are not known experimentally, and thus the functional state of a single AlphaFold-predicted model may not be known.

Benchmarking ligand discovery using predicted structures

Even in cases in which a druggable pocket is identified in a predicted structure, the side chains lining the pocket must adopt a precise conformational arrangement to accommodate a given small-molecule ligand, which is generally not correctly predicted. As such, the utility of protein structure predictions for drug discovery has been the subject of intense investigation in the past few years. Below we highlight several informative studies that have benchmarked the utility of predicted protein structures for use in computational ligand discovery, evaluating their performance in structure-based drug design as compared to experimental structures.

In one study, known active small molecules and decoys from the Directory of Useful Decoys Enhanced (DUD-E) benchmark dataset, commonly used for evaluating virtual screening methods, were virtually screened against crystal structures and AF2 models of 27 diverse protein targets using Glide (Schrödinger, NY), a reputable virtual screening tool[7]. When using crystal structures in which the binding pocket is occupied by a known ligand and is therefore captured in a favorable state for ligand binding, the top 1% of predicted hits were enriched 24-fold in active molecules as compared with a random selection. The enrichment factor dropped to 13-fold when using AF2 structures, comparable to the value obtained with apo crystal structures (11-fold), in which binding pockets are not occupied by ligands. Importantly, this indicates that the most critical factor making a 3D protein model optimal for virtual screening is not the source of the model (AF2 versus X-ray diffraction) but the availability of a previously known ligand, making it possible to capture the binding pocket in a conformational state that best accommodates drug-like molecules. Further supporting this notion, the enrichment factor obtained with AF2 models in the study increased to 18-fold when the virtual screen was preceded by a step in which the structure of the pocket was refined to better accommodate known ligands[7]. As an example, the ligand-binding pocket of the chromatin factor EED is occluded by side chains in both the AF2 model and the experimental structure determined in the absence of ligand (Fig. 2, top).

A related study focused on comparing the docking accuracy observed across 2,474 ligand–receptor protein pairs from the PDBbind database. Although 41% of compounds were re-docked within 2-Å root mean square deviation of the experimental structure, when the latter orginally contained the docked compound (an oversimplified prediction task that rarely reflects the needs of prospective structure-based design), the percentage dropped to just 17% when using AF2 models. Crystal structures of the ligand-free protein led to an even lower success rate of 10%, suggesting that crystal structures of empty binding pockets were on average even more distant from ligand-occupied conformations than AF2 models[8]. Again, this suggests that experimental structures of ligand-bound proteins are the best option for virtual screening efforts. This is the case, for instance, for the PWWP domain of the methyltransferase NSD2, in which the ligand-binding pocket is severely occluded in the crystal structure without ligand, but not in the AF2 model (Fig. 2, bottom).

Another benchmarking exercise did not reveal a significant advantage when using AF2 as compared with protein structure prediction tools predating AlphaFold, such as SWISS-Model, to derive structural models of the target protein from experimental structures of homologous proteins bound to drug-like molecules. Forgoing the protein model–building step and docking compounds directly to the crystallized protein homolog (in complex with ligand) was as efficient in predicting active molecules, questioning the merit of protein structure prediction in rational drug design. Only when the sequence identity of the crystallized homolog was <30% did AF2 models significantly outperform other methods, but results were highly variable depending on the protein and virtual screening tool, and enrichment in active molecules was often low[9].

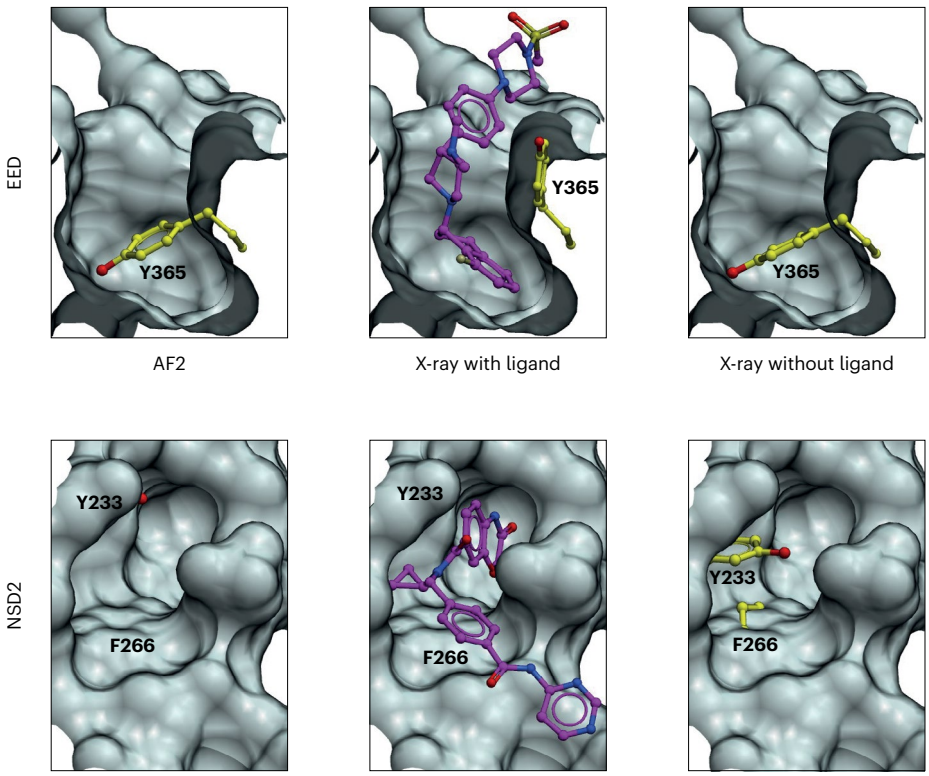

Fig. 2 | Comparison of ligand-free and ligand-bound pockets. Pockets are often occluded by amino acid side chains (yellow) in ligand-free crystal structures (right; EED, PDB 3JZN; NSD2, PDB 5VC8) or AF2 models (left), even when the protein backbones match perfectly. Crystallized ligands bound to EED (top; PDB 5K0M) and NSD2 (bottom; PDB 6XCG) are shown in pink.

Outlook for the use of predictive technologies in drug discovery

Collectively, these benchmarking studies indicate that AF2 models on their own are generally insufficient for virtual screening. In some cases, AF2 models may be used productively for structure-based drug design, but only when combined with approaches such as molecular dynamics to refine the structure of the binding pocket and predict induced-fit conformations elicited by the ligands. This finding poses a significant hurdle for virtual screening efforts, especially for novel drug target classes, as only ~20% of the structures currently available in the PDB are for proteins bound to drug-like molecules. Thus, for AI-based predictions to reach their potential in accurately aiding structure-based drug discovery, far more public-domain experimental data encompassing a wider variety of protein–ligand structures will be required to better train the models. This is the goal of the global Target 2035 initiative[10], which seeks to identify pharmacological modulators of most human proteins by 2035.

In the interim, it may become possible to complement the limited experimental data with robust but slow computational approaches, as has been done in a related field using ANI, a deep neural network for organic molecules trained on quantum mechanical density function theory calculations[11]. It would also be of interest to systematically evaluate whether docking to a conformational ensemble of a binding pocket generated with AF2 leads to significant improvement. As an alternative, incorporation of receptor flexibility in ligand binding simulations using AF2 structures should be evaluated. For instance, low-throughput free energy perturbation calculations, in which the binding pocket is conformationally dynamic, to calculate the relative affinities of compounds from a given chemical series achieved similar accuracy when using AF2 or crystal structures[12], whereas the same group found that high-throughput virtual screening with a rigid receptor produced poor results[13]. A recent study corroborated the conclusion that AF2 structures are poor for retrospective docking studies with known ligands but, interestingly, found that they yield much better results with prospective virtual screening efforts. The authors suggest that AF2 finds valid structures for drug design that are incompatible with previous ligands but can be used to find new ones[14].

Finally, as described above, unbiased benchmarking of different computational methods for protein–ligand prediction will be essential to advance the field. For example, CACHE (Critical Assessment of Computational Hit-finding Experiments)[15] provides a forum for the computational community to prospectively test and advance small-molecule hit-finding algorithms by comparing and improving their performance through prediction and experimental testing cycles, in a format similar to CASP. Importantly, community engagement in recurring benchmarking exercises such as CACHE is essential to define the state of the art and inform future development. As was the case for 3D protein structure prediction, we believe that the combination of robust community-endorsed benchmarking challenges with increasing data availability in the public domain will guide development in AI-based methods in coming years and contribute to breakthroughs in drug design on a scale analogous to that seen for protein fold prediction.

Matthieu Schapira[1,2], **Levon Halabelian**[1,2], **Cheryl H. Arrowsmith**[1,3,4] **& Rachel J. Harding**[1,2] ✉
[1]Structural Genomics Consortium, University of Toronto, Toronto, Ontario, Canada. [2]Department of Pharmacology & Toxicology, University of Toronto, Toronto, Ontario, Canada. [3]Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada. [4]Princess Margaret Cancer Centre, Toronto, Ontario, Canada.
✉e-mail: rachel.harding@utoronto.ca

Published online: 8 March 2024

References

1. Jumper, J. et al. *Nature* **596**, 583–589 (2021).
2. Elofsson, A. *Curr. Opin. Struct. Biol.* **80**, 102594 (2023).
3. Chakravarty, D. & Porter, L. L. *Protein Sci.* **31**, e4353 (2022).
4. Hekkelman, M. L., de Vries, I., Joosten, R. P. & Perrakis, A. *Nat. Methods* **20**, 205–213 (2023).
5. Humphreys, I. R. et al. *Science* **374**, eabm4805 (2021).
6. Baek, M. et al. *Nat. Methods* **21**, 117–121 (2024).
7. Zhang, Y. et al. *J. Chem. Inf. Model.* **63**, 1656–1667 (2023).
8. Holcomb, M., Chang, Y.-T., Goodsell, D. S. & Forli, S. *Protein Sci.* **32**, e4530 (2023).
9. Kersten, C., Clower, S. & Barthels, F. *J. Chem. Inf. Model.* **63**, 2218–2225 (2023).
10. Carter, A. J. et al. *Drug Discov. Today* **24**, 2111–2115 (2019).
11. Smith, J. S., Isayev, O. & Roitberg, A. E. *Chem. Sci.* **8**, 3192–3203 (2017).
12. Beuming, T. et al. *J. Chem. Inf. Model.* **62**, 4351–4360 (2022).
13. Díaz-Rovira, A. M. et al. *J. Chem. Inf. Model.* **63**, 1668–1674 (2023).
14. Lyu, J. et al. Preprint at *bioRxiv* https://doi.org/10.1101/2023.12.20.572662 (2023).
15. Ackloo, S. et al. *Nat. Rev. Chem.* **6**, 287–295 (2022).

Acknowledgements

The Structural Genomics Consortium is a registered charity (UK charities commission no. 1097737) that receives funds from Bayer AG, Boehringer Ingelheim, Bristol Myers Squibb, Genentech, Genome Canada through the Ontario Genomics Institute (OGI-196), EU/EFPIA/OICR/McGill/KTH/Diamond Innovative Medicines Initiative 2 Joint Undertaking (EUbOPEN grant 875510), Janssen, Merck (known as EMD in Canada and the United States), Pfizer and Takeda.

Author contributions

M.S. and R.J.H. created the figures. M.S., R.J.H., L.H. and C.H.A. conceived and wrote this Commentary.

Competing interests

The authors declare no competing interests.