Importance of updated benchmark sets for statistically correct AlphaFold applications

Laszlo Dobson[1,2,*] Gábor E. Tusnády[1,2], Peter Tompa[1,3,4]

[1]*Institute of Molecular Life Sciences, Research Centre for Natural Sciences, Magyar Tudósok Körútja 2, Budapest, 1117, Hungary*
[2]*Department of Bioinformatics, Semmelweis University, Tűzoltó u. 7, Budapest, 1094, Hungary*
[3]*VIB-VUB Center for Structural Biology, Vlaams Instituut voor Biotechnologie (VIB), 1050 Brussels, Belgium*
[4]*Structural Biology Brussels (SBB), Vrije Universiteit Brussel (VUB), 1050 Brussels, Belgium*

*to whom correspondence should be addressed at: dobson.laszlo@ttk.hu

Abstract

AlphaFold2 changed structural biology by providing high-quality structure prediction for all possible proteins. Since that a plethora of applications were built on AlphaFold2, expediting discoveries on virtually all fields related to protein science. In many cases it seems like optimism made scientists forget about data leakage, a serious issue that needs to be addressed when evaluating machine learning methods. Here we provide a rigorous benchmark set that can be used in a broad range of applications built around AlphaFold2/3.

Introduction

The advent of AlphaFold2 (AF2) in 2020 revolutionized protein science, by significantly raising the standards of structure prediction at CASP XIV[1]. Within a year, the method was published[2], soon followed by a database of predicted all-atom structures for almost every known protein sequence[3]. Based on this spectacular feat, thousands of papers appeared that analyzed AF2 structures to predict and/or assess various structural and physicochemical features of proteins in specific areas, such as protein stability[4], conformational heterogeneity[5], protein function[6], structure[7] or interfaces[8] of protein complexes, domain definitions[9], or disordered regions that do not fold into unique structures[10]. A similar, or even more swepping campaign of applications can be expected now with the appearance of AlphaFold3 (AF3), with its

advanced architecture and features, probably more reliably predicting protein-ligand (another protein), protein-nucleic acid and protein-small molecule interactions and complex structures[11].

Results

The apparent insight provided by these noted, and many other original applications gave cause for optimism, having boosted a broad range of studies, with artificial intelligence promising to expand the proteome coverage enabled by laborious and time-consuming experiments. We noted, however, many of these studies did not take a stringent approach to selecting proteins that were not used to train AF2, which is a potential source of statistical inaccuracies (cf. Fig. 1). We used search terms "AlphaFold2"/"AlphaFold" and "benchmark"/"evaluation"/"performance"/"analysis"/"accuracy" in PubMed, which resulted in over a thousand hits. Reading all of them would have been a rather strenuous process, but after reviewing about a hundred papers (where such a step could have been expected), in about ~90% of the cases the correct filtering for the training/template set was arguably not performed, causing data leakage, an issue that must not be disregarded when applying machine learning methods. Most of these papers are still highly valuable and point to relevant applications and evaluations. It is of relevance that our perusal of literature and analysis of data could not be done right after AF2 publication, only after an intense period of diverse analyses based on its applications. We figure the same will happen after AF3 code and/or access will be made broadly available, thus our conclusion and suggested solution also apply across a wide range of future applications.

To give an example, although AF2 has not been trained on protein complexes, its performance in predicting their structures is remarkable[7], probably because protein-protein interactions have a similar biophysical background to protein folding, which AF2 has been trained to model[12] – these capabilities are promised to further improve with AF3. Intriguingly, the capabilities of AF2 on complexes is not limited to those built on domain-domain interactions, it can well handle more transient interactions driven by short linear motifs[8]. It should be noted, however, that although the evaluation of multimers appears to be not affected by the choice of benchmarks, the bound state of single chains incorporated in AF2 prediction does cause data leakage and compromises the resulting statistical inference.

To assess the magnitude of the resulting discrepancies and to place AF2 – and by the same logic, AF3 – based structure-function studies on a solid ground, we present here the Benchmarking Evaluation Test for AlphaFold2 (BETA), a dataset of structures and sequences that were not used in AF2 training (Suppl. Material S1, https://beta.pbrg.hu, https://github.com/brgenzim/Beta). As AF2 can use template

structures which can further improve prediction, for any benchmark it is necessary to take into consideration which version of the PDB was used as the template library. To generate the appropriate datasets, we used Foldseek[13] and PSI-Blast[14] to perform a structure and sequence based search, as well as Voronota[15] to detect interchain contacts on the most probable assembly displayed on PDBe[6] and BIOGRID[16], when looking for interactions between SwissProt entries. In BETA, we list (i) monomeric PDB chains with no homologs before the selected date; (ii) interacting PDB chain pairs where participants have no homologs prior the selected date; (iii) complete protein sequences from SwissProt without any homologs in the PDB before the selected date and (iv) complete interacting sequences from SwissProt where the interacting participants do not have homologs before the selected date.

Careful assembly of this benchmark set enables us to demonstrate the importance of properly filtering benchmark data for evaluation. As an illustration, our analysis on AF2 performance showed a huge difference between transmembrane protein structures with and without homologous PDB entries[17]. Assessment of AF2 accuracy on this newly generated benchmark set shows significant differences between entries with and without homologous associated structures for both monomeric and multimeric structures (Fig. 1A, B, Mann-Whitney p-values <0.001 for each comparison, Suppl. Material S1, Suppl. Table S1). We have carried out a similar demonstrative test with AF3, with the caveat that its limitation on daily use and a later cutoff date limits the number of predictable cases that have no homologues in the training set. Yet, the statistical difference between cases predicted with and without homologues in the training set is statistically significant (Fig. 1C, Mann-Whitney p-values <0.01 for each comparison, Suppl. Material S1, Suppl. Table S1). It is to be noted specifically that in the case of multimer structures one cannot rule out the possibility that the predicted monomer structures are more accurate, yet, in a similar manner developers of multimer applications should ask what if they aren't. Clearly, any further application built around AlphaFold should take the emerging limitations into consideration.

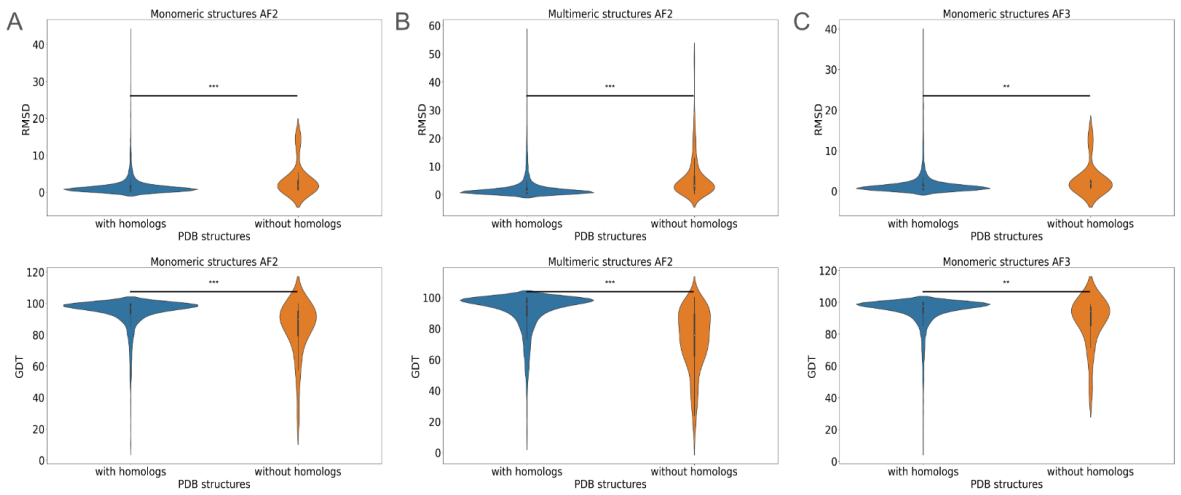

Figure 1: *AF2 and AF3 performance with and without homologs in training/template sets. Violin plots comparing Root Mean Square Deviation (RMSD) and Global Distance Test (GDT) values between structures of monomeric proteins (A) and complexes (B) determined experimentally or predicted by AF2 or monomers predicted by AF3 (C) (cf. Suppl. Material S1, Suppl. Table S1). Mann-Whitney tests were performed to test for significance (***: p-value < 0.001, **: p-value < 0.01). The mean values of the respective distributions (without vs with homologues) for AF2 are: monomeric RMSD: 3.47 vs 1.67, monomeric GDT: 84.08 vs 93.92, multimeric RMSD: 5.72 vs 2.31, multimeric GDT: 72.72 vs 90.56; for AF3: monomeric RMSD: 2.94 vs 1.64, monomeric GDT: 87.74 vs 94.11.*

BETA can be utilized in many different ways to help benchmark set construction for surveys or prediction methods, while also providing a convenient way for researchers without a computational background to browse data. We aimed to construct datasets for all users, including those accessing AlphaFold via the AlphaFold database[3], ColabFold[18] or running their predictions on their own computer.

For defining an appropriate benchmark set using BETA, one should first determine the appropriate cutoff date based on the input data (UniProt sequences or PDB codes) and the version of AlphaFold (AlphaFold database, ColabFold or running on one's own computer), and then select the first date after this cutoff to avoid using structures used during AF2 training or prediction. A detailed description and illustration of the protocol can be found in Online methods, „Protocol to define benchmark sets" (Suppl. Material S2), with five relevant Use cases (also at https://beta.pbrg.hu): (1) Building a protein disorder prediction method; (2) Predicting antibody epitopes; (3) Predicting phase-separating regions in proteins; (4) Assessing the pathological effects of missense mutations; and (5) Ranking interactions driven by short linear motifs. These examples explain how an appropriate benchmark dataset can be selected and used. It should be noted that it is difficult to construct a universal benchmark set, so additional steps may be required depending on the type of evaluation/analysis. In addition, to objectively evaluate AF2 and AF3 in the future, we also suggest further redundancy filtering on these datasets, depending on the exact use case.

As suggested above, during the preparation of this manuscript, AF3[11] was published, providing highly accurate prediction of multiple interacting molecules, including the bound state of proteins with DNA, RNA, ligands or other proteins. While AF2 is already available through multiple resources or as a standalone program, AF3 is currently available as a webserver, accepting a limited number of jobs each day, with a broad range of applications for predicting various features on the horizon. Our initial comparison on the monomer dataset (Fig. 1C) demonstrates that our conclusion and suggestion are equally useful for future applications of AF3, thus we plan to update the underlying BETA dataset on a regular basis, taking into account

how AF2/AF3 based resources will update their training and template libraries in the future.

References

1. Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K. & Moult, J. Critical assessment of methods of protein structure prediction (CASP)-Round XIV. *Proteins* **89**, 1607–1617 (2021).
2. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
3. Varadi, M. *et al.* AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–D444 (2022).
4. Pak, M. A. *et al.* Using AlphaFold to predict the impact of single mutations on protein stability and function. *PLoS One* **18**, e0282689 (2023).
5. Stein, R. A. & Mchaourab, H. S. SPEACH_AF: Sampling protein ensembles and conformational heterogeneity with Alphafold2. *PLoS Comput. Biol.* **18**, e1010483 (2022).
6. PDBe-KB consortium. PDBe-KB: a community-driven resource for structural and functional annotations. *Nucleic Acids Res.* **48**, D344–D353 (2020).
7. Bryant, P. *et al.* Predicting the structure of large protein complexes using AlphaFold and Monte Carlo tree search. *Nat. Commun.* **13**, 6028 (2022).
8. Lee, C. Y. *et al.* Systematic discovery of protein interaction interfaces using AlphaFold and experimental validation. *Mol. Syst. Biol.* **20**, 75–97 (2024).
9. Bordin, N. *et al.* AlphaFold2 reveals commonalities and novelties in protein structure space for 21 model organisms. *Commun Biol* **6**, 160 (2023).
10. Conte, A. D. *et al.* Critical assessment of protein intrinsic disorder prediction (CAID) - Results of round 2. *Proteins* **91**, 1925–1934 (2023).
11. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature.* https://doi.org/10.1038/s41586-024-07487-w (2024).
12. Roney, J. P. & Ovchinnikov, S. State-of-the-Art Estimation of Protein Model Accuracy Using AlphaFold. *Phys. Rev. Lett.* **129**, 238101 (2022).
13. van Kempen, M. *et al.* Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.* **42**, 243–246 (2024).
14. Altschul, S. F. & Koonin, E. V. Iterated profile searches with PSI-BLAST--a tool for discovery in protein databases. *Trends Biochem. Sci.* **23**, 444–447 (1998).
15. Olechnovič, K. & Venclovas, C. Voronota: A fast and reliable tool for computing the vertices of the Voronoi diagram of atomic balls. *J. Comput. Chem.* **35**, 672–681 (2014).
16. Oughtred, R. *et al.* The BioGRID interaction database: 2019 update. *Nucleic Acids Res.* **47**, D529–D541 (2019).
17. Jambrich, M. A., Tusnady, G. E. & Dobson, L. How AlphaFold2 shaped the structural coverage of the human transmembrane proteome. *Sci. Rep.* **13**, 20283 (2023).

18. Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nat. Methods* **19**, 679–682 (2022).