Regularly updated benchmark sets for statistically correct evaluations of AlphaFold applications

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

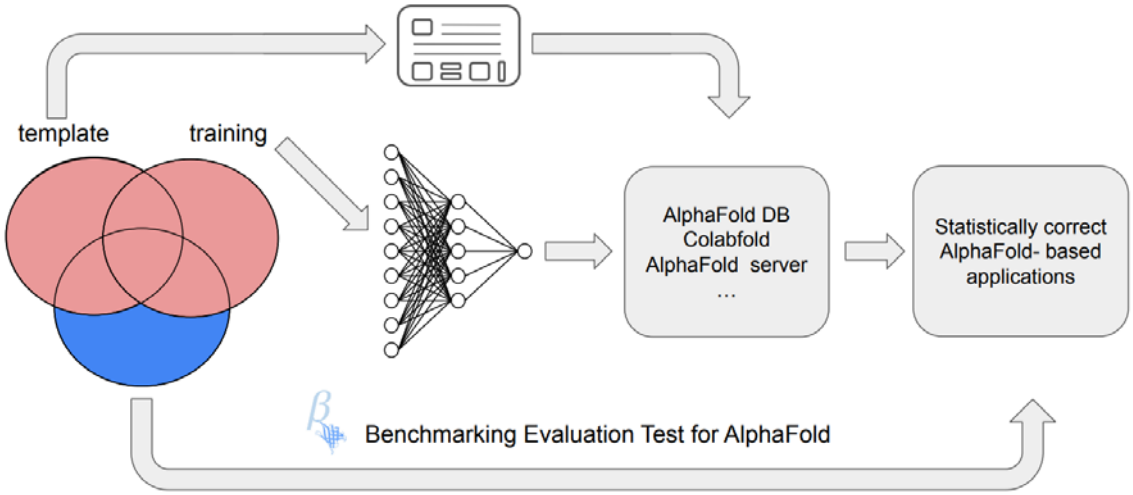
[1]*Institute of Molecular Life Sciences, Research Centre for Natural Sciences, Magyar Tudósok Körútja 2, Budapest, 1117, Hungary*
[2]*Department of Bioinformatics, Semmelweis University, Tűzoltó u. 7, Budapest, 1094, Hungary*
[3]*VIB-VUB Center for Structural Biology, Vlaams Instituut voor Biotechnologie (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 predictions for all possible proteins. Since its inception, a plethora of applications were built on AlphaFold2, expediting discoveries in virtually all areas related to protein science. In many cases, however, optimism seems to have 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.

Graphical abstract



Introduction

The creation of AlphaFold2 (AF2) in 2020 was bound to revolutionize 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 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].

The apparent insight provided by these noted, and many other original applications gave cause for optimism, having reinvigorated a broad range of structure-related studies, with artificial intelligence promising to expand the proteome coverage enabled by laborious and time-consuming experiments. We noted, however, that many of these studies did not take a stringent approach to selecting proteins that had not been used to train AF2, leading to data leakage, which is inevitably leads to statistical inaccuracies 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 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.

AF2 has not been trained on protein complexes, its performance in predicting their structures is remarkable, probably because protein-protein interactions have a similar biophysical background to protein folding, which AF2 has been trained to model – these capabilities are promised to further improve with AF3. Intriguingly, the capabilities of AF2 to model complexes is not limited to those built on domain-domain interactions [7], it can also well handle more transient interactions driven by short linear motifs [9]. 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 the respective single chains used for the training of AF2 does cause data leakage and compromises the resulting statistical inference.

To overcome this limitation we suggest a deep homology search for proteins in the training/template sets of AF. Although the solution may seem trivial, considering how

distant homologous structures can be still captured by AF and the many different versions of AF each with a different training cutoff, we prepared benchmark sets to facilitate this process. Here we provide Benchmarking Evaluation Test for AlphaFold (BETA), containing predefined datasets for AF applications, and demonstrate through a range of cases studies that it resolves the issue of data leakage-derived statistical uncertainties in AF-based analyses.

Methods

Defining benchmark sets

BETA was created as follows. PDB [6], SwissProt [12] and BIOGRID [13] were downloaded on the 21th of May, 2024. Considering the first day of the following month of each cutoff date (30.04.2018; 31.05.2020; 15.02.2021; 30.09.2021; 15.07.2022, 01.11.2022, 01.01.2023, 01.01.2024) we performed homology searches based on both sequence and structure. AF can be also used without incorporating template structures, and for training AF the models NMR structures were not used, therefore for cutoff dates related to trainining (30.04.2018,30.09.2021) NMR structures were not included in the background database. Homology searches were carrie dout as follows: I) Sequence search (PSI-BLAST [14]) of all structures after cutoff date against structures before cutoff date; II) Structure search (Foldseek [15]) of all structures after the cutoff date against structures before cutoff date; III) Sequence search (PSI-BLAST) of SwissProt proteins against structures before cutoff date. In the case of PSI-BLAST (e-value 0.0001, 3 iterations, maximum target sequences 50 000), all queries having any hit longer than 10 amino acids with at least 20% sequence identity were filtered In Foldseek searches (maximum target structures 50 000), all queries having any hit longer than 10 amino acids with at least 0.25 TM-Score were filtered. We used Voronota [16] to detect all interchain interaction in PDB structures, considering the most probable oligomerization state stored on PDBe. We used BIOGRID (only ‘direct interactions’) to search for interactions in SwissProt sequences. The pipeline and relevant resulting collections of different BETA benchmark sets are shown on Figure 1.

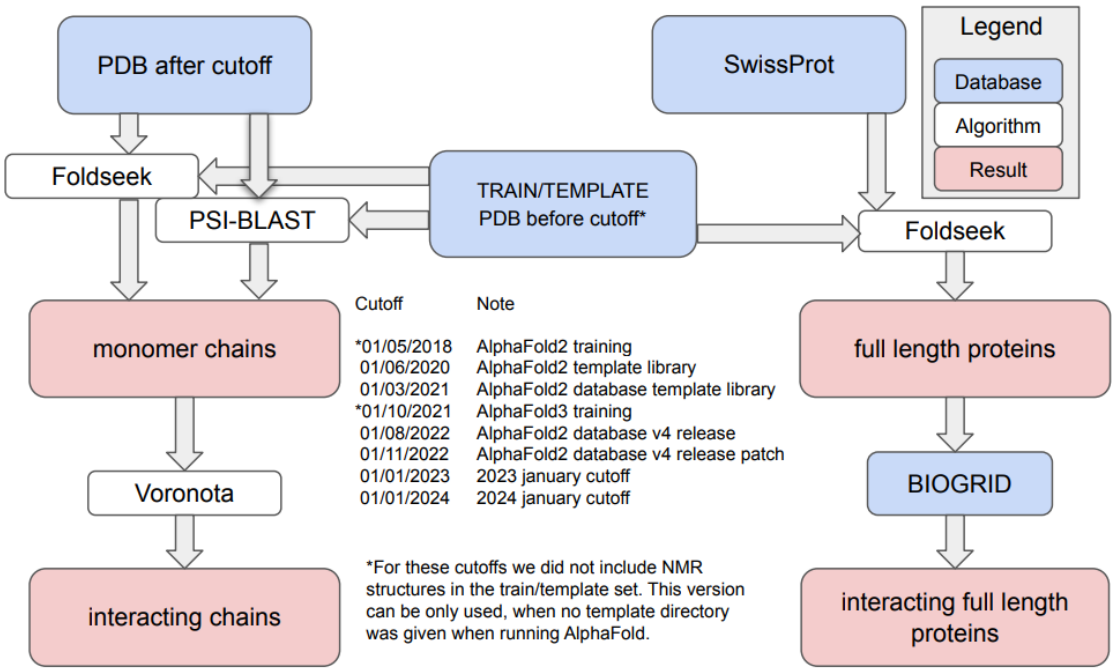


Figure 1: Pipeline used to define the Benchmarking Evaluation Test for AlphaFold (BETA).

A case study: disordered prediction method

To demonstrate the importance of our methodology, we assessed how prediction methods of intrinsically disordered proteins/regions (IDPs/IDRs) would be affected by utilizing BETA (Figure 2/A). Previously, the confidence measure Local-Distance Difference Test (pLDDT) was shown to be a good predictor of protein disorder [10]. We downloaded all PDB structures, where the most probable oligomeric state was a monomer according to PDBe [6]. Sequences were filtered to 40% identity using CD-HIT [17], and disorder of residues was inferred, if coordinates of their side-chains were missing from the PDB structures. The remaining residues (with side-chain coordinates) were labeled as ordered. We used SIFTS [18] to cross-reference UniProt identifiers and positions with PDB chains and residues, then downloaded AF2 structures from the AlphaFold protein structure database. Structures were categorized into 'Homologous structures' and 'BETA structures', considering the AF database training/template cutoff, which in this case is 01/03/2021. Supplementary Table 1 contains all residues incorporated in the analysis (with cross-referenced PDB and AF positions), with ordered and disordered residue classifications (inferred from PDB structures), AF2 pLDDT values, as well as BETA and homologous structure classifications.

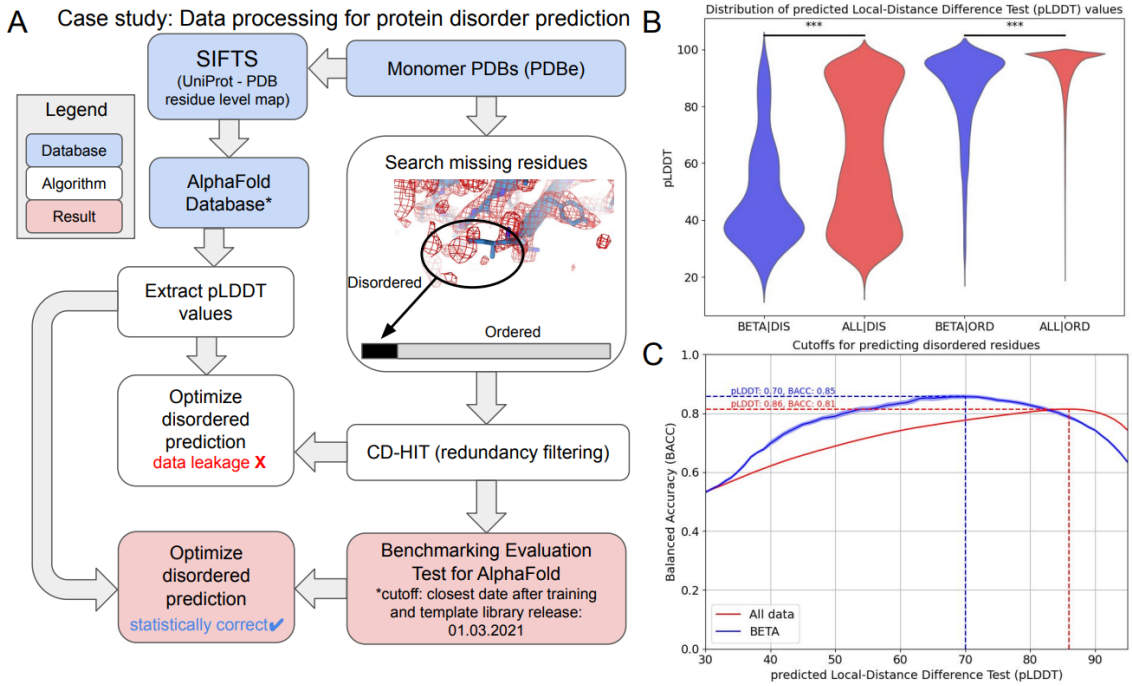


Figure 2. A) Pipeline to define disordered residues from PDB and incorporating BETA into the pipeline. B) Distribution of pLDDT values of disordered and ordered residues in all monomeric structures (red) and BETA structures (blue). There is highly significant differences between pLDDT values in the case ofBETA against all structures, both for ordered and disordered residues. C) Correspondence between pLDDT and Balanced accuracy considering all structures and BETA structures. The highest Balanced accuracy (with suggested pLDDT cutoff is shown). The small area around curves show the standard error calculated from bootstrapping 50% of the data 5 times, while the curves show the mean value at a given point.

Results

Benchmarking Evaluation Test for AlphaFold (BETA)

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 AlphaFold (BETA), a dataset of structures and sequences that were not used in AF training (https://beta.pbrg.hu, https://github.com/brgenzim/Beta). 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 (Figure 1).

BETA can be utilized in many different ways to help construct a benchmark set 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 AF via the AlpaFold database [3], ColabFold [19] or the AlphaFold server [11], 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 AF (AlphaFold database, ColabFold, AlphaFold server or running on one's own computer), and then select the first date after this cutoff to avoid using structures used during AF2/AF3 training or prediction. To demonstrate the general applicability of BETA, we outline a highly informative case study of predicting structural disorder, out of five relevant Use Cases (more details are available at Supplemantary Material S1 and https://beta.pbrg.hu/): (1) Building a protein disorder prediction method (further detailed as a case study); (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.

Case study: the effect of using BETA for disordered prediction

To set the precise pLDDT threshold to predict IDRs, we can define a set of proteins where we can assign labels to each residue, whether they belong to IDRs or not. Using a benchmark set of PDB structures (see Methods), we looked for the distribution of pLDDT values in ordered and disordered residues (Figure 2/B, red distributions). As the two distributions are different, one can assume that pLLDT value is an indeed good predictor for protein disorder. However, limiting the examined residues to those that are available in the BETA database, there is also a significance difference between ordered residues in BETA and all ordered residues (Kolmogorov-Smirnov (KS) Test, $p<0.0001$), as well as between residues in IDRs (BETA against all structures, K-S Test, $p<0.0001$) (Figure 2/B). To further highlight the gravity of differences caused, we also looked for the optimal pLDDT cutoff to achieve the highest balanced accuracy (Figure 2/C). Indeed, the cutoff for the highest accuracy is different for all structures (pLDDT=0.70) from the cutoff defined using only BETA structures (pLDDT=0.86). Notably, the number of residues in IDRs is relatively low in the BETA set defined by monomeric PDBs (total 1001 residues), which is likely caused by the fact that most novel structures deposited to PDB correspond to big complexes. To avoid statistical bias, we randomly sampled 50% of the residues five times to define the pLDDT cutoffs.

We have to add that we do not aim to build a perfect IDR predictor here, and these steps can be further optimized. Our goal was to show that defining a stringent benchmark set may heavily influence an application built on AF.

Conclusion

While AF2 is already available through multiple resources, AF3 is currently available as a webserver (or the standalone program), with a broad range of applications for predicting various features on the horizon.. Our comparison on predicting protein disorder demonstrates that our conclusion and suggestion are useful for future applications of AF, 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. Since AF was developed, multiple other methods were published [20-22] – although BETA cannot offer an exact cutoff date for the training date of each method, we suggest researchers rely on BETA (with the nearest training cutoff) when building applications on structure prediction tools.

Data availability

Supplementary Table 1 is available at https://zenodo.org/records/14338331

Funding

NKFIH [PD-146564 and K-146314 to L.D. and G.E.T., respectively].

References

1. Kryshtafovych A, Schwede T, Topf M, et al. Critical assessment of methods of protein structure prediction (CASP)-Round XIV. Proteins 2021; 89:1607–1617
2. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. Nature 2021; 596:583–589
3. Varadi M, Anyango S, Deshpande M, et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res. 2022; 50:D439–D444
4. Pak MA, Markhieva KA, Novikova MS, et al. Using AlphaFold to predict the impact of single mutations on protein stability and function. PLoS One 2023; 18:e0282689
5. Stein RA, Mchaourab HS. SPEACH_AF: Sampling protein ensembles and conformational heterogeneity with Alphafold2. PLoS Comput. Biol. 2022; 18:e1010483
6. PDBe-KB consortium. PDBe-KB: a community-driven resource for structural and functional annotations. Nucleic Acids Res. 2020; 48:D344–D353
7. Bryant P, Pozzati G, Zhu W, et al. Predicting the structure of large protein complexes using AlphaFold and Monte Carlo tree search. Nat. Commun. 2022; 13:6028
849. Lee CY, Hubrich D, Varga JK, et al. Systematic discovery of protein interaction

interfaces using AlphaFold and experimental validation. Mol. Syst. Biol. 2024; 20:75–97
9. Bordin N, Sillitoe I, Nallapareddy V, et al. AlphaFold2 reveals commonalities and novelties in protein structure space for 21 model organisms. Commun Biol 2023; 6:160
10. Conte AD, Mehdiabadi M, Bouhraoua A, et al. Critical assessment of protein intrinsic disorder prediction (CAID) - Results of round 2. Proteins 2023; 91:1925–1934
11. Abramson J, Adler J, Dunger J, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature 2024; 630:493–500
12. UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res 2023; 51:D523–D531
13. Oughtred R, Stark C, Breitkreutz B-J, et al. The BioGRID interaction database: 2019 update. Nucleic Acids Res 2019; 47:D529–D541
14. Altschul SF, Madden TL, Schäffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997; 25:3389–3402
15. van Kempen M, Kim SS, Tumescheit C, et al. Fast and accurate protein structure search with Foldseek. Nat Biotechnol 2024; 42:243–246
16. Olechnovič K, Venclovas C. Voronota: A fast and reliable tool for computing the vertices of the Voronoi diagram of atomic balls. J Comput Chem 2014; 35:672–681
17. Fu L, Niu B, Zhu Z, et al. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 2012; 28:3150–3152
18. Dana JM, Gutmanas A, Tyagi N, et al. SIFTS: updated Structure Integration with Function, Taxonomy and Sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins. Nucleic Acids Res 2019; 47:D482–D489
19. Mirdita M, Schütze K, Moriwaki Y, et al. ColabFold: making protein folding accessible to all. Nat. Methods 2022; 19:679–682
20. Wohlwend J, Corso G, Passaro S, et al. Boltz-1 Democratizing Biomolecular Interaction Modeling. bioRxiv 2024;
21. Lin Z, Akin H, Rao R, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 2023; 379:1123–1130
22. Wu R, Ding F, Wang R, et al. High-resolution de novo structure prediction from primary sequence. bioRxiv 2022; 2022.07.21.500999