

How accurately can one predict drug binding modes using AlphaFold models?

Reviewed Preprint

Published from the original preprint after peer review and assessment by eLife.

[About eLife's process](#)

Reviewed preprint posted

August 8, 2023 (this version)

Sent for peer review

May 30, 2023

Posted to bioRxiv

May 23, 2023

Masha Karelina, Joseph J. Noh, Ron O. Dror 

Biophysics Program, Stanford University, Stanford, CA 94305, USA • Department of Computer Science, Stanford University, Stanford, CA, USA • Department of Molecular and Cellular Physiology, Stanford University School of Medicine, Stanford, CA, USA • Department of Structural Biology, Stanford University School of Medicine, Stanford, CA, USA • Institute for Computational and Mathematical Engineering, Stanford University, Stanford, CA, USA

 https://en.wikipedia.org/wiki/Open_access

 <https://creativecommons.org/licenses/by/4.0/>

Abstract

Computational prediction of protein structure has been pursued intensely for decades, motivated largely by the goal of using structural models for drug discovery. Recently developed machine learning methods such as AlphaFold 2 (AF2) have dramatically improved protein structure prediction, with reported accuracy approaching that of experimentally determined structures. To what extent do these advances translate to an ability to predict more accurately how drugs and drug candidates bind to their target proteins? Here we carefully examine the utility of AF2 protein structure models for predicting binding poses of drug-like molecules at the largest class of drug targets, the G protein-coupled receptors. We find that AF2 models capture binding pocket structures much more accurately than traditional homology models, with errors nearly as small as differences between structures of the same protein determined experimentally with different ligands bound. Strikingly, however, the accuracy of ligand binding poses predicted by computational docking to AF2 models is not significantly higher than when docking to traditional homology models and is much lower than when docking to structures determined experimentally without these ligands bound. These results have important implications for all those who might use predicted protein structures for drug discovery.

eLife assessment

This **important** study presents findings with broad implications for the use of AlphaFold2 models in ligand binding pose modeling, a common task in protein structure modeling. The computational experiments and analyses provide **compelling** results for the GPCR protein family data, but the conclusions are likely to apply also to other proteins and they will therefore be of interest to biophysicists, physical chemists, structural biologists, and anyone interested or involved in structure-based ligand discovery.

Recent breakthroughs in machine learning have substantially improved the accuracy of protein structure prediction, to the point that some have declared the problem solved (Baek et al., 2021 [DOI](#); Jumper et al., 2021 [DOI](#); Ourmazd et al., 2022 [DOI](#)). Indeed, in the most recent round of the Community Assessment of Structure Prediction (CASP)—a blind protein structure prediction competition—AlphaFold 2 (AF2) demonstrated an unprecedented ability to predict protein structures with atomic accuracy, sometimes rivaling the accuracy of certain experimental methods for structure determination (Kryshtafovych et al., 2021 [DOI](#); Tejero et al., 2022 [DOI](#)).

These advances have generated tremendous excitement, focused in particular on their potential impact on drug discovery (Lowe, 2022 [DOI](#); Thornton et al., 2021 [DOI](#); Toews, 2021 [DOI](#)). The vast majority of drug targets are proteins, and structures play a critical role in rational drug design (Anderson, 2003 [DOI](#); Kolb et al., 2009 [DOI](#)). In particular, rational drug design strategies frequently require determining structures of the target protein bound to many ligands, as information on how these ligands bind is critical both to discovering early-stage drug candidates and to optimizing their properties (Maveyraud & Mourey, 2020 [DOI](#); Warren et al., 2012 [DOI](#)). Determining these structures experimentally is typically slow and expensive, and sometimes proves impossible (Slabinski et al., 2007 [DOI](#)). Tremendous effort has thus gone into development of computational docking methods for predicting ligand binding modes, which have dramatically accelerated this process in cases where a high-resolution experimentally determined structure of the target protein is available (Ferreira et al., 2015 [DOI](#); Pinzi & Rastelli, 2019 [DOI](#)).

However, the metrics generally used to evaluate protein structure prediction methods—including all of those used in CASP—focus on the overall accuracy of predicted structures, rather than on their utility in predicting ligand binding. This has left the large community of academic and industrial researchers who wish to use protein structures to design drugs and other ligands unsure of whether, when, and how to use the new structural models. Improvements in protein structure prediction are widely assumed to lead to better models of ligand binding pockets and thus more accurate predictions of ligand binding (Beuming & Sherman, 2012 [DOI](#); Bordogna et al., 2011 [DOI](#); Erickson et al., 2004 [DOI](#)), but to what extent do these assumptions hold?

Here, we address these questions by evaluating (1) the structural accuracy of ligand binding pockets in modeled proteins and (2) the accuracy of ligand binding poses predicted using these models, where a ligand's binding pose is specified by the 3D coordinates of its atoms when bound to the target protein. We evaluate models generated both by AF2 and by a traditional template-based modeling strategy, systematically comparing both to experimentally determined structures.

Several recent studies have experimented with the use of models generated by AF2—and related protein structure prediction approaches such as RoseTTAFold (Baek et al., 2021 [DOI](#))—for tasks related to drug design, with mixed results (Díaz-Rovira et al., 2023 [DOI](#); He et al., 2022 [DOI](#); Heo & Feig, 2022 [DOI](#); C. Lee et al., 2022 [DOI](#); S. Lee et al., 2022 [DOI](#); Liang et al., 2022 [DOI](#); Qiao et al., 2022 [DOI](#); Scardino et al., 2022 [DOI](#); Wong et al., 2022 [DOI](#)). Our study stands apart in two regards. First, we use structural models generated without any use of known structures of the target protein. For machine learning methods, this requires ensuring that no structure of the target protein was used to train the method. Second, we perform a systematic comparison that takes into account the variation between experimentally determined structures of the same protein when bound to different ligands.

Our results reveal both strengths and weaknesses of AF2 models. The structural accuracy of ligand binding pockets in AF2 models is usually substantially higher than that of traditional homology models. Indeed, the typical difference between corresponding binding pockets in an AF2 model and in an experimentally determined structure is only slightly larger than the typical difference between two experimentally determined structures of the same protein with different ligands

bound. Surprisingly, however, ligand binding poses predicted given AF2 models were not significantly more accurate than those predicted given traditional models and were much less accurate than those predicted by docking computationally to experimentally determined protein structures. Our results provide guidelines as to how AF2 models should—and should not—be used for effective ligand binding prediction. Our findings also suggest opportunities for improving structure prediction methods to maximize their impact on drug discovery.

Results

Selection of proteins and models to ensure a fair comparison

Evaluating the accuracy of predicted ligand binding poses requires that we examine protein–ligand complexes whose structures have been determined experimentally. We wish to ensure, however, that the predicted structures of a given protein are not informed by experimentally determined structures of that same protein.

AF2 and other recent neural-network-based protein structure prediction methods use experimentally determined protein structures in two ways: first, to train the neural network and second, as templates for prediction of individual protein structures (Baek et al., 2021 [↗](#); Jumper et al., 2021 [↗](#)). The AF2 neural network was trained on structures in the Protein Data Bank (PDB) (Berman et al., 2000 [↗](#)) as of April 30, 2018 so we evaluated structural models of proteins for which no experimentally determined structure was available as of that date. In addition, when generating models using AF2, we used only structures that were available on that date as templates (see Methods). For a fair comparison, we also ensured that the traditional template-based models we used were not informed by any structures that became available after that date. In particular, we used traditional models that were published before April 30, 2018.

We selected all proteins, ligands, and structural models used in this study in a systematic manner (see Methods). We focused on one class of drug targets—G protein–coupled receptors (GPCRs)—for three reasons. First, GPCRs represent the largest class of drug targets by a substantial margin, with 35% of FDA approved drugs targeting a GPCR (Sriram & Insel, 2018 [↗](#)). Second, GPCR binding pockets are very diverse, reflecting the fact that GPCRs have evolved to bind very different ligands with high specificity (Venkatakrisnan et al., 2013 [↗](#)). Third, the GPCRdb project maintains a historical database of template-based models of GPCRs (Pándy-Szekeres et al., 2018 [↗](#)). By selecting the template-based models listed in GPCRdb in April 2018, we ensured that these models were not informed by any structures that became available after April 30, 2018.

Binding pocket structures are much more accurate in AF2 models than traditional models

Because proteins are flexible, each protein can adopt multiple structures. In particular, the structure of a protein—especially the structure of the binding pocket—depends on which ligand is bound. To quantify this natural variation, we compared all pairs of experimentally determined structures of the same protein to one another, using root mean squared deviation (RMSD) as a metric (see Methods).

AF2 does not allow one to specify a bound ligand when generating a protein model, and the template-based models listed in GPCRdb were also generated without specification of a bound ligand. We thus evaluated the accuracy of each model by computing its RMSD to that of all available experimental structures of the same protein. We compared the resulting distribution of RMSDs to the distribution of RMSDs between structures of the same protein with different ligands bound.

As expected based on previously reported results for other proteins (Jumper et al., 2021 [DOI](#)), we find that the global structural accuracy of the AF2 models is substantially better than that of the traditional template-based models. In particular, when computing RMSDs based on all non-hydrogen atoms in each protein, we find a median RMSD of 2.9 Å between AF2 models and corresponding experimentally determined structures, compared to a median RMSD of 4.3 Å for traditional models (Supplementary Figure S1).

Next, we turn our attention to the orthosteric binding pocket of each protein—that is, the region where native ligands and most other ligands, including those in our dataset, bind. AF2 was trained to optimize global structural accuracy, with an emphasis on correctly predicting the structure of the protein backbone. The accuracy of binding pocket prediction, on the other hand depends heavily on side-chain conformations in this local region. Nevertheless, we find that AF2 predictions of binding pocket structures are typically very accurate.

In fact, the binding pocket RMSD between an AF2 model and an experimentally determined structure is typically nearly as low as the RMSD between two experimentally determined structures of the same protein with different ligands bound (**Figure 1** [DOI](#)). By contrast, binding pockets in the traditional template-based models are much less accurate, with a median RMSD of 3.3 Å for traditional models vs. 1.3 Å for AF2 models. We note that the binding pockets of a few AF2 models are highly inaccurate (approximately 5 Å RMSD), as is the case for several of the traditional models.

Binding pose prediction using AlphaFold 2 models or traditional models yields similar accuracy

Next, we assess the accuracy of ligand binding poses predicted by computational docking to AF2 models and traditional template-based models. For each ligand we consider, the binding pose to a target protein is known based on an experimentally determined structure of the ligand bound to the protein, which we call the *reference structure*. Using industry-standard software, we docked each ligand to both AF2 and GPCRdb models of the protein. A predicted pose is considered correct if its RMSD from the experimentally determined pose is ≤ 2.0 Å (**Fig. 2** [DOI](#)), a widely used criterion for correct pose prediction (Erickson et al., 2004 [DOI](#); Gohlke et al., 2000 [DOI](#); Paggi et al., 2021 [DOI](#)).

For comparison, we also dock each ligand to structures of the same protein determined experimentally with other ligands bound. Note that when calculating pose prediction accuracy (**Fig. 2** [DOI](#)), we exclude cases where the ligand is docked back to the reference structure; such “self-docking” yields much higher accuracy (Supplementary Figure S2) but is of little practical utility, because in practice one wishes to predict binding poses that have not already been determined experimentally.

Strikingly, binding pose prediction accuracy was similar when using AF2 models or traditional template-based models, despite the fact that binding pocket structures were substantially more accurate in AF2 models. Although the fraction of ligands docked correctly was slightly higher when using AF2 models (16%) than when using traditional models (11%), this difference was not statistically significant ($P > 0.5$). On the other hand, binding pose prediction accuracy was far higher when docking to experimentally determined protein structures (48%), even though these structures were determined in complex with ligands different from those being docked.

The docked poses analyzed above were determined with the widely used commercial docking software package Glide, run in SP mode (for “standard precision”) (Friesner et al., 2004 [DOI](#)). To verify the robustness of our results, we repeated all docking experiments using a completely different docking software package—Rosetta docking (Park et al., 2021 [DOI](#))—as well as with Glide run in XP mode (for “extra precision”) (Friesner et al., 2006 [DOI](#)). In both cases, we again found that,

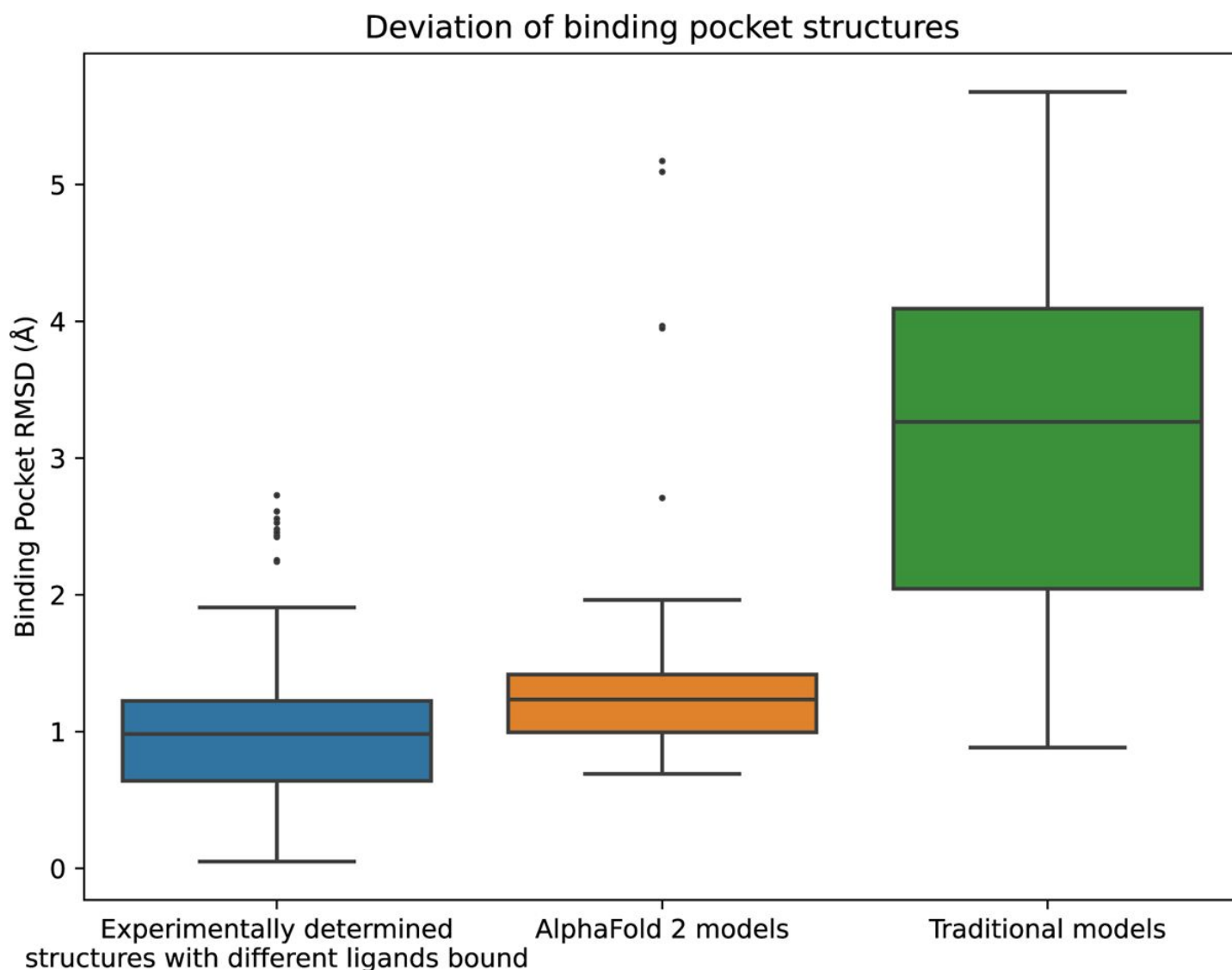


Figure 1.

Structural accuracy of modeled binding pockets.

Binding pockets are defined to include all amino acid residues with any atom within 5 Å of the ligand in an experimentally determined structure. We compute all-atom binding pocket RMSDs between each modeled structure and all experimentally determined structures of the same protein. For comparison, we also compute binding pocket RMSDs between all pairs of experimentally determined structures of the same protein with different ligands bound. The middle line of each box in the plot is the median RMSD, with the box extending from the 1st to the 3rd quartile and defining the “interquartile range.” Whiskers extend to last data points that are within 150% of the interquartile range, and outlier data points beyond those are shown individually.

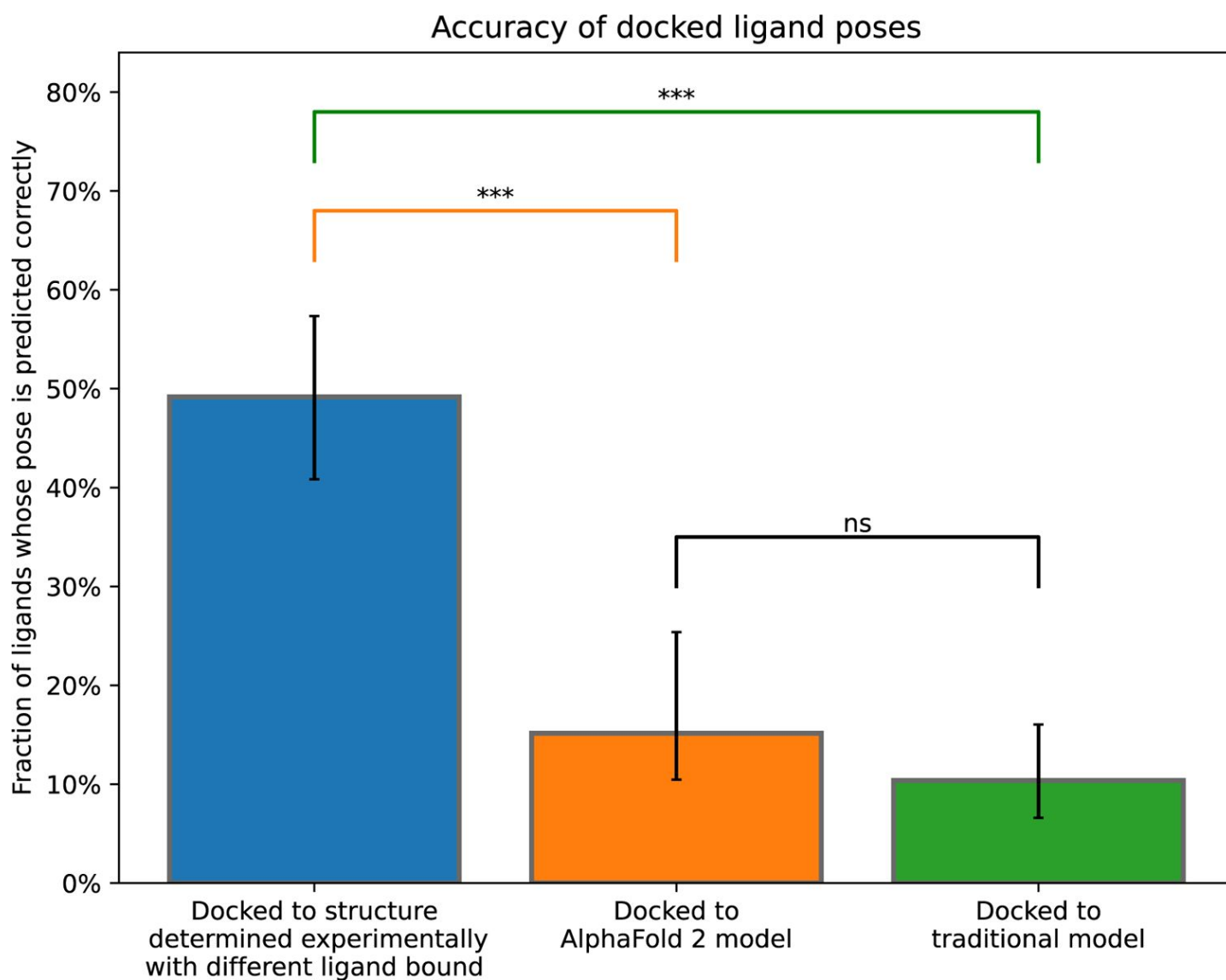


Figure 2.

Accuracy of ligand binding poses predicted by computational docking to AlphaFold 2 models, traditional template-based models, or protein structures determined experimentally in complex with a ligand different from the one being docked.

We plot the fraction of docked ligands whose pose is predicted correctly (see Methods). Error bars are 90% confidence intervals calculated via bootstrapping. *** for P values < 0.001. ns for P values > 0.05.

when used for ligand binding pose prediction, AF2 models substantially underperformed experimentally determined structures and did not significantly outperform traditional models (Supplementary Figure S3). In fact, the difference in binding pose prediction accuracy between AF2 models and traditional models was even smaller when using Rosetta docking or Glide XP than when using Glide SP.

Closer look at discrepancy between structural accuracy and pose prediction accuracy

The results of the previous two sections contrast sharply with one another. AF2 models nearly match experimentally determined structures—and are much better than traditional template-based models—in binding pocket accuracy, as measured by a standard structural metric. Yet AF2 models fare much worse than experimentally determined structures—do not significantly outperform traditional models—when used to predict ligand binding poses. **Figures 3** and **4** provide illustrative examples.

To further quantify these results, we compare pose prediction accuracy for structures and models with similar binding pocket accuracy. In particular, we calculate pose prediction accuracy as a function of binding pocket accuracy (RMSD) for both experimentally determined structures and models. As expected, we find that average pose prediction accuracy increases as binding pocket accuracy increases (i.e., RMSD to reference structure decreases). This trend holds for experimentally determined structures, AF2 models, and traditional template-based models (**Figure 5** and Supplementary Figure S4). Yet for the same binding pocket accuracy (RMSD from the reference structure), AF2 models yield lower average pose prediction accuracy than experimentally determined structures (**Fig. 3**). By contrast, in cases where traditional models and experimentally determined structures have similar binding pocket accuracies, they yield similar pose prediction accuracies (Supplementary Figure S4).

We also considered the possibility that the improved performance of experimentally determined structures relative to AF2 models might be because, in some cases, the structure used for docking was determined with a bound ligand similar (though not identical) to the ligand being docked. Eliminating such cases from consideration does lead to decreases in both aggregate binding pocket accuracy and pose prediction accuracy of experimentally determined structures, but experimentally determined structures continue to yield significantly more accurate pose predictions than AF2 models ($P < 0.01$) (Supplementary Figures S2, S3, S5).

Discussion

Several recent papers have concluded that AF2 models fall short of experimentally determined structures when used for ligand docking (Díaz-Rovira et al., 2023; He et al., 2022; Heo & Feig, 2022; Scardino et al., 2022). Our results agree with this conclusion but go beyond it in several ways. We compare AF2 models and traditional template-based models, rigorously ensuring that none of these models was informed by experimentally determined structures of the protein being modeled. We find that, in terms of structural accuracy of the binding pocket, AF2 models are much better than traditional models; indeed, the RMSD between an AF2 model and an experimentally determined structure is typically comparable to the RMSD between two structures determined experimentally with different ligands bound. Yet, when used for docking, AF2 models yield pose prediction accuracy similar to that of traditional models and much worse than that of experimentally determined structures.

A few caveats are in order. First, building a traditional template-based model is not always possible. We only considered template-based models in cases where a structure was available for another protein with at least 40% sequence identity. However, 80% of human drug targets have at

Docking to an AF2 model vs. experimentally determined structure

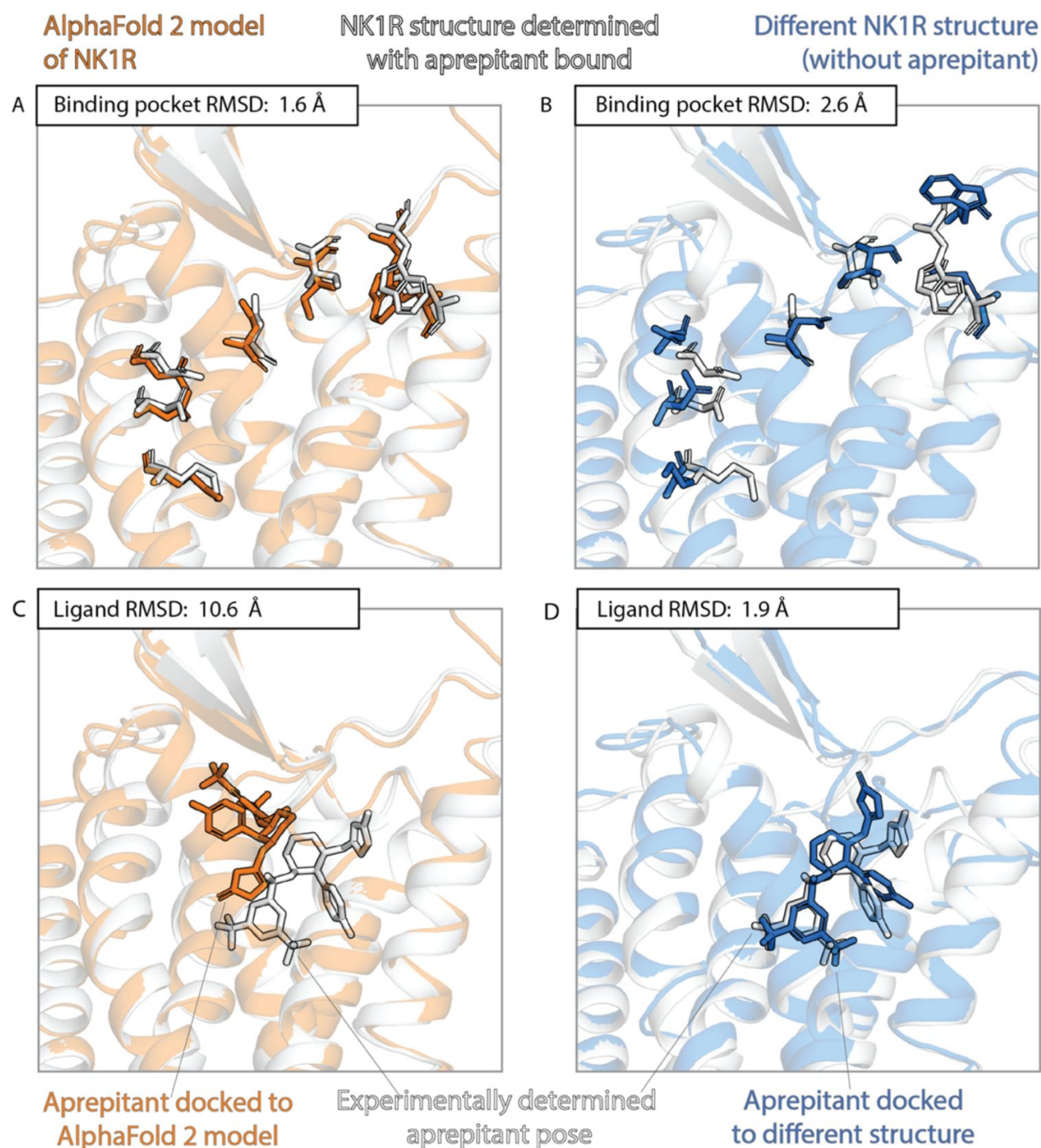


Figure 3.

An example in which docking to an AF2 model yields poor results even though the model's binding pocket has high structural accuracy. We predict the binding pose of the drug aprepitant to its target, the neurokinin-1 receptor (NK1R) given either the AF2 model (orange) of NK1R or the experimentally determined structure (blue) of NK1R bound to a different ligand, L760735. (A, B) The binding pocket of the AF2 model is more similar (lower RMSD) than the binding pocket of the L760735-bound structure to the binding pocket of the aprepitant-bound structure (the "reference structure," white). Amino acid residues whose positions differ most from the reference structure are shown in sticks (see Methods). (C, D) The aprepitant binding pose predicted by docking is much less accurate (higher RMSD) when using the AF2 model than when using the L760735-bound structure. Ligand L760735 shares a scaffold with aprepitant; for completeness, we include another example with highly dissimilar ligands in Supplementary Figure S6.

Docking to an AF2 model vs. traditional model

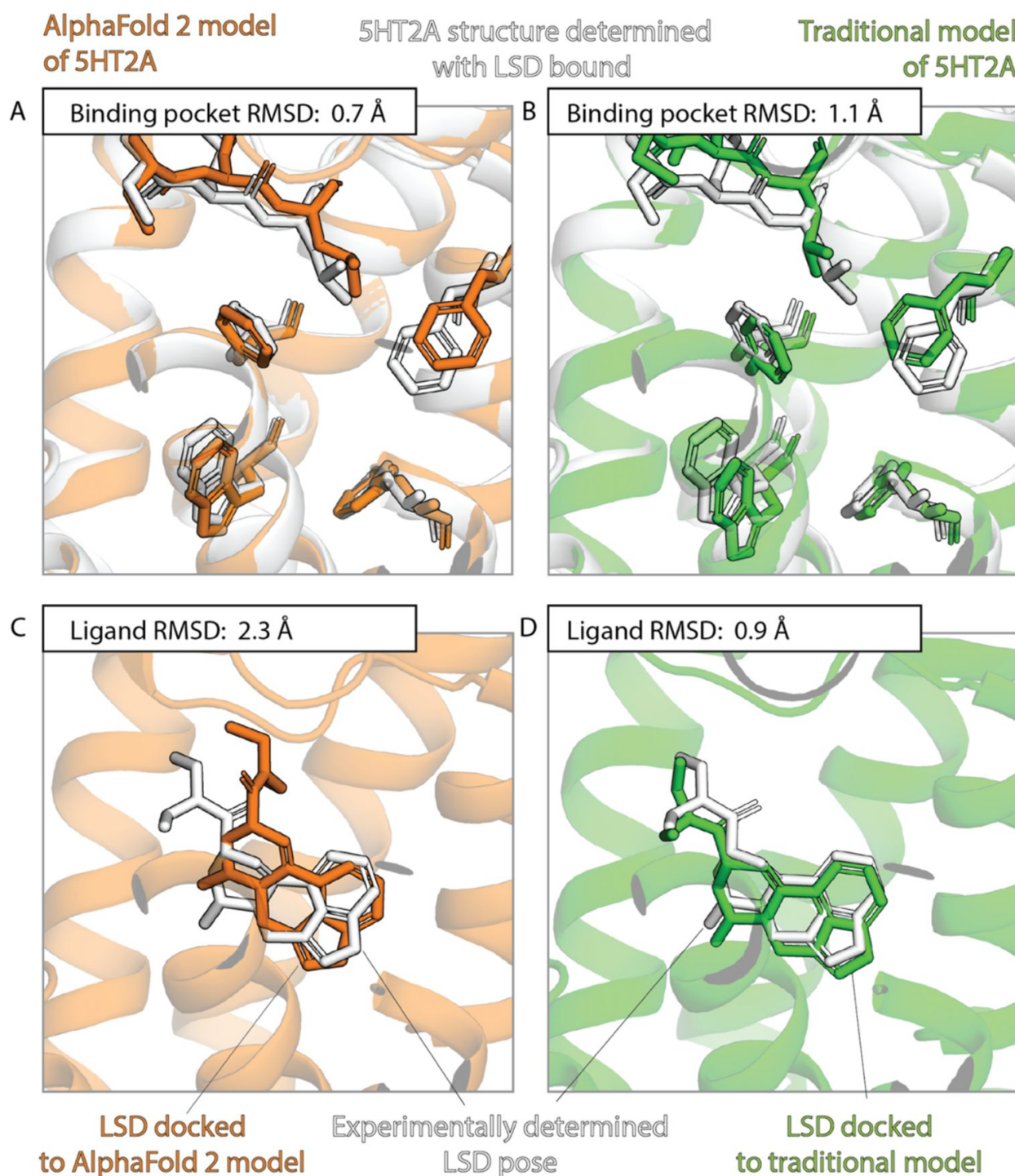


Figure 4.

An example in which docking to a traditional template-based model yields better results than docking to an AF2 model, even though the AF2 model's binding pocket has higher structural accuracy. We predict the binding pose of the psychedelic LSD to its primary target, the serotonin 2A receptor (5HT2A) given either the AF2 model (orange) or a traditional model (green) of 5HT2A. (A, B) The binding pocket of the AF2 model is more similar (lower RMSD) than that of the traditional model to the binding pocket of the experimentally determined LSD-bound structure (the "reference structure," white). Amino acid residues that differ most from the reference structure are shown in sticks. (C, D) The LSD binding pose predicted by docking is less accurate (higher RMSD) when using the AF2 model than when using the traditional model.

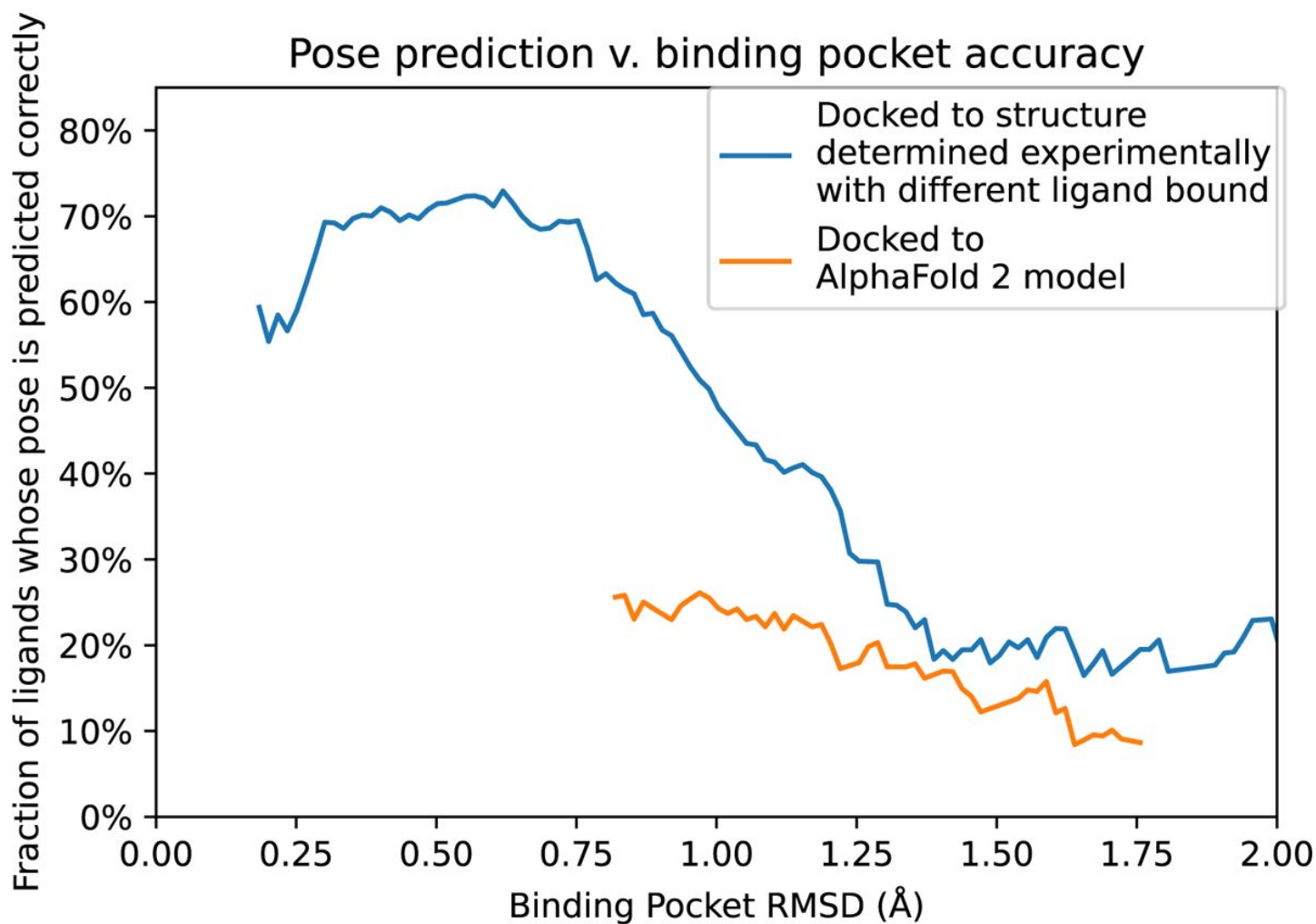


Figure 5.

Pose prediction accuracy as a function of binding pocket structural accuracy when docking to AF2 models or experimentally determined structures.

Docking to an experimentally determined structure generally leads to more accurate pose prediction than docking to an AF2 model with the same binding pocket RMSD. The difference between the two curves is statistically significant for all binding pocket RMSD values below 1.1 Å (see Methods).

least 50% sequence identity to a protein whose structure was known in 2017 (Somody et al., 2017), and that number is even higher today. In addition, we note that the traditional models we used were generated relatively quickly by an academic group along with models for hundreds of other proteins. An expert working on a drug discovery project could frequently build a better template-based model—for example, by examining the predictive utility of several models (Carlsson et al., 2011 [↗](#); Haddad et al., 2020 [↗](#)).

Second, our study examined only GPCRs. The results might in principle be different for other classes of drug targets. However, GPCRs have very diverse binding pockets, which have evolved to binding an extremely broad range of ligands, ranging from very small molecules to peptides and proteins, with high specificity (Venkatakrishnan et al., 2013 [↗](#)). Our conclusions thus likely apply to many non-GPCR drug targets—and GPCR targets alone account for roughly half of drug discovery projects. We also note that, for GPCRs specifically, one can often improve docking performance by using a model or an experimentally determined structure in the activation state to which a given ligand preferentially binds (Heo & Feig, 2022 [↗](#); S. Lee et al., 2022 [↗](#)). In this study, we did not limit the ligands docked to each structure or model to those that preferentially bind one specific activation state. We made this choice for several reasons: many other classes of drug targets do not have well-defined activation states; in practical applications, the preferred activation state for a given ligand may not be known when one docks it; and AF2 does not produce models in specific activation states by default.

Third, we have not examined performance of the full range of available computational docking methods (Pagadala et al., 2017 [↗](#)). Our results are, however, consistent across the three docking methods we examined: Glide SP, Glide XP, and Rosetta docking. In each case, AF2 models yield docking accuracy similar to that of traditional models and substantially worse than that of experimentally determined structures.

Our results have important implications for practitioners of structure-based drug discovery. Our findings suggest that, despite dramatic recent improvements in protein structure prediction, structural models (in particular, AF2 models) yield substantially poorer accuracy than experimentally determined structures in prediction of ligand binding modes. Moreover, when a reasonable structural template is available—as is typically the case—AF2 models should not be expected to yield significantly better pose prediction accuracy than traditional models. Indeed, with sufficient effort, an expert may well be able to build a traditional template-based model that yields better performance than an AF2 model.

Despite these findings, we believe that AF2 and other recently developed protein structure prediction methods will prove valuable in certain structure-based drug discovery efforts. Up-to-date AF2 models for any human protein are readily available for download (Tunyasuvunakool et al., 2021 [↗](#); Varadi et al., 2022 [↗](#)). By contrast, expertise is required to build a decent template-based model, and up-to-date, downloadable repositories of such models are not available for all drug targets. Also, in the relatively rare cases when no structural template is available for a drug target, approaches like AF2 may be the only realistic option short of solving a structure experimentally.

Our results also point to opportunities for developing improved computational modeling methods. Existing deep-learning methods for protein structure prediction were developed with the aim of maximizing structural accuracy for models. We find that the RMSD of a model—a classic structural accuracy metric—is not a good predictor of the accuracy that model yields for pose prediction. Future protein structure prediction methods might instead be designed to maximize utility for ligand binding prediction, while still exploiting the deep learning advances of AF2 and related recent methods. We also note that the vast majority of test cases used in the design of currently

available docking software involved experimentally determined protein structures. One might be able to redesign docking methods to yield better performance given computationally predicted models.

Methods

Protein and ligand selection

We selected structures from the aggregated list of all experimentally determined human GPCR structures from GPCRdb (Isberg et al., 2014 [DOI](#)). We removed all proteins for which a structure was published in the PDB before April 30, 2018, including two Class C GPCRs for which structures of the extracellular domain (to which ligands bind) had been published before that date. We also removed proteins for which structures were available in complex with fewer than two unique orthosteric ligands. This left us with 18 GPCRs (17 class A, one class B). The full list of proteins is in Supplementary Table 1.

Obtaining structural models

AlphaFold 2

For each protein, we took the full protein sequence from Uniprot (The UniProt Consortium, 2023 [DOI](#)) to generate 5 AF2 models, as was done for CASP (Jumper et al., 2021 [DOI](#)), and picked the top-scoring prediction as our model. We set the template cutoff date to 2018-04-30.

GPCRdb

We downloaded template-based models from the GPCRdb archive dated 2018-04-06. We used all the inactive-state and active-state models available, and averaged results as described in the statistical tools section below. We excluded models for which the minimum sequence identity of the templates was below 40%, as such template-based models are typically not used for drug binding predictions.

This led to exclusion of models for four GPCRs (AGRG3, TA2R, PE2R2, PE2R4). GPCRdb did not include models for two of the GPCRs (GPBAR and PE2R3).

Ligand preparation

Ligands were extracted from PDB structures using the Schrödinger Python API and then manually inspected to make sure we had chosen the ligand at the orthosteric site. With the same API, ligands were converted to SMILES strings. Ligands were then prepared with Schrödinger's LigPrep with default command line parameters (Schrödinger Release 2021-1, 2021 [DOI](#)).

Ligand similarity was defined as a ratio of the size of the maximum common substructure to the size of the smaller molecule. Ligand pairs for which this ratio was less than 0.5 were deemed very different ligands.

Protein preparation

Experimentally determined structures were downloaded from the PDB. Only a single chain containing the ligand was kept, and all waters were removed. All structures and models were prepared with Schrödinger Protein Preparation Wizard (Schrödinger Release 2021-1, 2021 [DOI](#)) following the same protocols as Paggi et al. (2021 [DOI](#)). With this process, all structures underwent a minimization this restrains all non-hydrogen atoms to 0.3 Å from the initial structure with a

harmonic potential. All the experimentally determined structures were determined with a ligand bound, and we retained this ligand during minimization. Computationally predicted models did not include a ligand.

Glide Docking

For Glide XP and Glide SP docking, we follow the docking protocol described in [Paggi et al. \(2021\)](#). The grid is centered at the geometric centroid of the ligand and is defined with an inner box with 15 Å sides and outer box with 30 Å sides. For predicted structural models the center of the box is determined by aligning the model to an experimentally determined structure of that protein and using the centroid of its ligand. For this alignment, we pick the structure with the alphabetically first PDB code.

Rosetta Docking

For Rosetta docking, we used the GALigandDock protocol ([Park et al., 2021](#)). To prepare the proteins, PyMol ([Schrödinger, LLC, 2015](#)) was used to remove heteroatoms and all alternative locations for each atom. CONECT information was included. UCSF Chimera v1.16 ([Pettersen et al., 2004](#)) was used to initially remove all hydrogens, followed by their Dock Prep software to prepare the structure for docking.

For ligands, Open Babel v2.4.0 ([O'Boyle et al., 2011](#)) was used to generate 3-D structures from ligand SMILES and add hydrogens at a pH of 7 ([O'Boyle et al., 2011](#)). Ambertools v22.0 ([Case, 2022](#)) was used to add AM1-BCC partial charges. The ligand was translated such that the center of mass of the ligand was positioned at center of the binding pocket of the prepared structure. The center of the binding pocket was defined as the center of mass of the ligand in the experimentally determined structure. Models were aligned to the experimentally determined structure of the protein which had the alphabetically first PDB code. This structure's ligand center of mass was used to define the model's binding pocket.

Structural Comparisons

To calculate ligand pose RMSD, structures were first aligned on amino acid residues within 15 Å of bound orthosteric ligands, using Schrödinger's structalign tool ([Schrödinger Release 2021-1, 2021](#)). This alignment was used to calculate the RMSD of each docked ligand pose from the reference ligand pose.

Binding pocket RMSDs were calculated with a PyMOL script, considering all residues that are within 5 Å of the ligand in the reference structure (for both alignment and RMSD calculation). We included all non-hydrogen atoms in this calculation, except that for the backbone-only binding-pocket RMSDs shown in Supplementary Figure S5, we only backbone atoms.

We computed full-structure RMSDs with a similar PyMOL script, but taking into account all non-hydrogen atoms in the protein.

In [Figure 4](#), [Figure 5](#) and Supplementary Figure S5, we highlight residues in the binding pocket whose positions differ most from the reference structure. To identify these residues, we aligned to the reference structure the structure and model (or the two models) being docked to. For each structure/model, we computed the RMSD for each residue within 5 Å of the ligand, and then for each residue determined the maximum RMSD across the two structures/models. We selected the residues with the largest maximum RMSDs.

Statistical tools

P values were computed with two sided paired t-tests, using SciPy (Virtanen et al., 2020 [DOI](#)). Bootstrapping 90% confidence intervals were computed with the default 9999 resamples, also using SciPy.

In **Figure 2** [DOI](#), the docking accuracy is calculated as a weighted average across all docking results. The weighting for the average is such that each ligand is equally represented in the average. That is, there were cases where proteins with multiple experimental structures or traditional template models available for a ligand to dock into. The average is calculated as a weighted average in which each pose prediction is weighted as the inverse of the number of template-based models or experimentally determined structures the given ligand was docked into.

In **Figure 5** [DOI](#) and Supplementary Figure S4, the smoothing function was created by running kernel smoothing over docking outcomes. The smoothing was done over data such that each data point was the binding pocket RMSD, and the binary value indicating whether or not the predicted poses was correct. An Epanechnikov kernel (Hastie et al., 2009 [DOI](#)) was used, with a width of 0.25 Å. To avoid inaccurate boundary condition behaviors each end the kernel was only evaluated when the center of the kernel overlapped with the first or last data point on the X-axis. To calculate statistical significance and P values, a bootstrapping significance test was conducted for each interval on the X-axis. Additionally, in Supplementary Figure S4, we show 90% confidence intervals that are also computed with bootstrapping.

Author Contributions

Conceptualization: M.K. and R.O.D. Data curation and preparation: M.K. Investigation: M.K. and J.N. Analysis: M.K. Writing: M.K. and R.O.D. Supervision: R.O.D.

Acknowledgements

We thank Joseph Paggi, John Wang, Akshat Nigam, and all members of the Dror lab for assistance and insightful comments. This work was supported by Novo Nordisk and an NSF Graduate Research Fellowship to M.K..

References

- Anderson A. C. (2003) **The Process of Structure-Based Drug Design** *Chemistry & Biology* **10**:787–797 <https://doi.org/10.1016/j.chembiol.2003.09.002>
- Baek M., DiMaio F., Anishchenko I., Dauparas J., Ovchinnikov S., Lee G. R., Wang J., Cong Q., Kinch L. N., Schaeffer R. D., Millán C., Park H., Adams C., Glassman C. R., DeGiovanni A., Pereira J. H., Rodrigues A. V., van Dijk A. A., Ebrecht A. C., Baker D. (2021) **Accurate prediction of protein structures and interactions using a three-track neural network** *Science* **373**:871–876 <https://doi.org/10.1126/science.abj8754>
- Berman H. M., Westbrook J., Feng Z., Gilliland G., Bhat T. N., Weissig H., Shindyalov I. N., Bourne P. E. (2000) **The Protein Data Bank** *Nucleic Acids Research* **28**:235–242 <https://doi.org/10.1093/nar/28.1.235>
- Beuming T., Sherman W. (2012) **Current assessment of docking into GPCR crystal structures and homology models: Successes, challenges, and guidelines** *Journal of Chemical Information and Modeling* **52**:3263–3277 <https://doi.org/10.1021/ci300411b>
- Bordogna A., Pandini A., Bonati L. (2011) **Predicting the accuracy of protein-ligand docking on homology models** *Journal of Computational Chemistry* **32**:81–98 <https://doi.org/10.1002/jcc.21601>
- Carlsson J., Coleman R. G., Setola V., Irwin J. J., Fan H., Schlessinger A., Sali A., Roth B. L., Shoichet B. K. (2011) **Ligand discovery from a dopamine D3 receptor homology model and crystal structure** *Nature Chemical Biology* **7** <https://doi.org/10.1038/nchembio.662>
- Díaz-Rovira A. M., Martín H., Beuming T., Díaz L., Guallar V., Ray S. S. (2023) **Are Deep Learning Structural Models Sufficiently Accurate for Virtual Screening? Application of Docking Algorithms to AlphaFold2 Predicted Structures** *Journal of Chemical Information and Modeling* **63**:1668–1674 <https://doi.org/10.1021/acs.jcim.2c01270>
- Erickson J. A., Jalaie M., Robertson D. H., Lewis R. A., Vieth M. (2004) **Lessons in Molecular Recognition: The Effects of Ligand and Protein Flexibility on Molecular Docking Accuracy** *Journal of Medicinal Chemistry* **47**:45–55 <https://doi.org/10.1021/jm030209y>
- Ferreira L. G., dos Santos R. N., Oliva G., Andricopulo A. D. (2015) **Molecular Docking and Structure-Based Drug Design Strategies** *Molecules* **20**:13384–13421 <https://doi.org/10.3390/molecules200713384>
- Friesner R. A., Banks J. L., Murphy R. B., Halgren T. A., Klicic J. J., Mainz D. T., Repasky M. P., Knoll E. H., Shelley M., Perry J. K., Shaw D. E., Francis P., Shenkin P. S. (2004) **Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy** *Journal of Medicinal Chemistry* **47**:1739–1749 <https://doi.org/10.1021/jm0306430>
- Friesner R. A., Murphy R. B., Repasky M. P., Frye L. L., Greenwood J. R., Halgren T. A., Sanschagrin P. C., Mainz D. T. (2006) **Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein–Ligand Complexes** *Journal of Medicinal Chemistry* **49**:6177–6196 <https://doi.org/10.1021/jm051256o>

- Gohlke H., Hendlich M., Klebe G. (2000) **Knowledge-based scoring function to predict protein-ligand interactions** Edited by R. Huber *Journal of Molecular Biology* **295**:337–356 <https://doi.org/10.1006/jmbi.1999.3371>
- Haddad Y., Adam V., Heger Z. (2020) **Ten quick tips for homology modeling of high-resolution protein 3D structures** *PLoS Computational Biology* **16** <https://doi.org/10.1371/journal.pcbi.1007449>
- Hastie T., Tibshirani R., Friedman J. H., Friedman J. H. (2009) **The elements of statistical learning: Data mining, inference, and prediction**
- He X., You C., Jiang H., Jiang Y., Xu E., Cheng X. (2022) **AlphaFold2 versus experimental structures: Evaluation on G protein-coupled receptors** :1–7 <https://doi.org/10.1038/s41401-022-00938-y>
- Heo L., Feig M. (2022) **Multi-state modeling of G-protein coupled receptors at experimental accuracy** *Proteins: Structure, Function, and Bioinformatics* **90**:1873–1885 <https://doi.org/10.1002/prot.26382>
- Hillisch A., Pineda L. F., Hilgenfeld R. (2004) **Utility of homology models in the drug discovery process** *Drug Discovery Today* **9**:659–669 [https://doi.org/10.1016/S1359-6446\(04\)03196-4](https://doi.org/10.1016/S1359-6446(04)03196-4)
- Isberg V., Vroeling B., van der Kant R., Li K., Vriend G., Gloriam D. (2014) **GPCRDB: An information system for G protein-coupled receptors** *Nucleic Acids Research* **42**:D422–D425 <https://doi.org/10.1093/nar/gkt1255>
- Jumper J., Evans R., Pritzel A., Green T., Figurnov M., Ronneberger O., Tunyasuvunakool K., Bates R., Žídek A., Potapenko A., Bridgland A., Meyer C., Kohl S. A. A., Ballard A. J., Cowie A., Romera-Paredes B., Nikolov S., Jain R., Adler J., Hassabis D. (2021) **Highly accurate protein structure prediction with AlphaFold** *Nature* **596**:583–589 <https://doi.org/10.1038/s41586-021-03819-2>
- Kolb P., Ferreira R. S., Irwin J. J., Shoichet B. K. (2009) **Docking and chemoinformatic screens for new ligands and targets** *Current Opinion in Biotechnology* **20**:429–436 <https://doi.org/10.1016/j.copbio.2009.08.003>
- Kryshtafovych A., Schwede T., Topf M., Fidelis K., Moult J. (2021) **Critical assessment of methods of protein structure prediction (CASP)-Round XIV** *Proteins* **89**:1607–1617 <https://doi.org/10.1002/prot.26237>
- Lee C., Su B.-H., Tseng Y. J. (2022) **Comparative studies of AlphaFold, RoseTTAFold and Modeller: A case study involving the use of G-protein-coupled receptors** *Briefings in Bioinformatics* **23** <https://doi.org/10.1093/bib/bbac308>
- Lee S., Kim S., Lee G. R., Kwon S., Woo H., Seok C., Park H. (2022) **Evaluating GPCR modeling and docking strategies in the era of deep learning-based protein structure prediction** *Computational and Structural Biotechnology Journal* **21**:158–167 <https://doi.org/10.1016/j.csbj.2022.11.057>
- Liang S., Liang Z., Wu Z., Huang F., Wang X., Tan Z., He R., Lu Z., Cai Y., Huang B., Wang X. (2022) **Liang, S., Liang, Z., Wu, Z., Huang, F., Wang, X., Tan, Z., He, R., Lu, Z., Cai, Y., Huang, B., & Wang, X. (2022). A benchmark study of protein folding algorithms on nanobodies (p. 2022.08.07.503071). bioRxiv. 10.1101/2022.08.07.503071** <https://doi.org/10.1101/2022.08.07.503071>

Lowe D. (2022) **AlphaFold Excitement**

Maveyraud L., Mourey L. (2020) **Protein X-ray Crystallography and Drug Discovery** *Molecules* **25** <https://doi.org/10.3390/molecules25051030>

O'Boyle N. M., Banck M., James C. A., Morley C., Vandermeersch T., Hutchison G. R. (2011) **Open Babel: An open chemical toolbox** *Journal of Cheminformatics* **3** <https://doi.org/10.1186/1758-2946-3-33>

Ourmazd A., Moffat K., Lattman E. E. (2022) **Structural biology is solved—Now what?** *Nature Methods* **19** <https://doi.org/10.1038/s41592-021-01357-3>

Pagadala N. S., Syed K., Tuszyński J. (2017) **Software for molecular docking: A review** *Biophysical Reviews* **9**:91–102 <https://doi.org/10.1007/s12551-016-0247-1>

Paggi J. M., Belk J. A., Hollingsworth S. A., Villanueva N., Powers A. S., Clark M. J., Chemparathy A. G., Tynan J. E., Lau T. K., Sunahara R. K., Dror R. O. (2021) **Leveraging nonstructural data to predict structures and affinities of protein–ligand complexes** *Proceedings of the National Academy of Sciences* **118** <https://doi.org/10.1073/pnas.2112621118>

Pándy-Szekeres G., Munk C., Tsonkov T. M., Mordalski S., Harpsøe K., Hauser A. S., Bojarski A. J., Gloriam D. E. (2018) **GPCRdb in 2018: Adding GPCR structure models and ligands** *Nucleic Acids Research* **46**:D440–D446 <https://doi.org/10.1093/nar/gkx1109>

Park H., Zhou G., Baek M., Baker D., DiMaio F. (2021) **Force Field Optimization Guided by Small Molecule Crystal Lattice Data Enables Consistent Sub-Angstrom Protein–Ligand Docking** *Journal of Chemical Theory and Computation* **17**:2000–2010 <https://doi.org/10.1021/acs.jctc.0c01184>

Pettersen E. F., Goddard T. D., Huang C. C., Couch G. S., Greenblatt D. M., Meng E. C., Ferrin T. E. (2004) **UCSF Chimera—A visualization system for exploratory research and analysis** *Journal of Computational Chemistry* **25**:1605–1612 <https://doi.org/10.1002/jcc.20084>

Pinzi L., Rastelli G. (2019) **Molecular Docking: Shifting Paradigms in Drug Discovery** *International Journal of Molecular Sciences* **20** <https://doi.org/10.3390/ijms20184331>

Qiao Z., Nie W., Vahdat A., Miller T. F., Anandkumar A. (2022) **Qiao, Z., Nie, W., Vahdat, A., Miller, T. F., & Anandkumar, A. (2022). Dynamic-Backbone Protein-Ligand Structure Prediction with Multiscale Generative Diffusion Models. 10.48550/ARXIV.2209.15171** <https://doi.org/10.48550/ARXIV.2209.15171>

Scardino V., Di Filippo J. I., Cavasotto C. (2022) **Scardino, V., Di Filippo, J. I., & Cavasotto, C. (2022). How good are AlphaFold models for docking-based virtual screening? [Preprint]. Chemistry. 10.26434/chemrxiv-2022-sgj8c** <https://doi.org/10.26434/chemrxiv-2022-sgj8c>

Schrödinger LLC (2015) **The PyMOL Molecular Graphics System, Version 1.8**

Schrödinger Release 2021-1 (2021) **Schrödinger, LLC**

Slabinski L., Jaroszewski L., Rodrigues A. P. C., Rychlewski L., Wilson I. A., Lesley S. A., Godzik A. (2007) **The challenge of protein structure determination—Lessons from structural genomics** *Protein Science : A Publication of the Protein Society* **16**:2472–2482 <https://doi.org/10.1110/ps.073037907>

- Sriram K., Insel P. A. (2018) **G Protein-Coupled Receptors as Targets for Approved Drugs: How Many Targets and How Many Drugs?** *Molecular Pharmacology* **93**:251–258 <https://doi.org/10.1124/mol.117.111062>
- Tejero R., Huang Y. J., Ramelot T. A., Montelione G. T. (2022) **AlphaFold Models of Small Proteins Rival the Accuracy of Solution NMR Structures** *Frontiers in Molecular Biosciences* **9** <https://doi.org/10.3389/fmolb.2022.877000>
- The UniProt Consortium (2023) **UniProt: The Universal Protein Knowledgebase in 2023** *Nucleic Acids Research* **51**:D523–D531 <https://doi.org/10.1093/nar/gkac1052>
- Thornton J. M., Laskowski R. A., Borkakoti N. (2021) **AlphaFold heralds a data-driven revolution in biology and medicine** *Nature Medicine* **27** <https://doi.org/10.1038/s41591-021-01533-0>
- Toews R. (2021) **AlphaFold Is The Most Important Achievement In AI—Ever**
- Tunyasuvunakool K., Adler J., Wu Z., Green T., Zielinski M., Žídek A., Bridgland A., Cowie A., Meyer C., Laydon A., Velankar S., Kleywegt G. J., Bateman A., Evans R., Pritzel A., Figurnov M., Ronneberger O., Bates R., Kohl S. A. A., Hassabis D. (2021) **Highly accurate protein structure prediction for the human proteome** *Nature* **596** <https://doi.org/10.1038/s41586-021-03828-1>
- Varadi M., Anyango S., Deshpande M., Nair S., Natassia C., Yordanova G., Yuan D., Stroe O., Wood G., Laydon A., Žídek A., Green T., Tunyasuvunakool K., Petersen S., Jumper J., Clancy E., Green R., Vora A., Lutfi M., Velankar S. (2022) **AlphaFold Protein Structure Database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models** *Nucleic Acids Research* **50**:D439–D444 <https://doi.org/10.1093/nar/gkab1061>
- Venkatakrishnan A. J., Deupi X., Lebon G., Tate C. G., Schertler G. F., Babu M. M. (2013) **Molecular signatures of G-protein-coupled receptors** *Nature* **494** <https://doi.org/10.1038/nature11896>
- Virtanen P., Gommers R., Oliphant T. E., Haberland M., Reddy T., Cournapeau D., Burovski E., Peterson P., Weckesser W., Bright J., van der Walt S. J., Brett M., Wilson J., Millman K. J., Mayorov N., Nelson A. R. J., Jones E., Kern R., Larson E., SciPy 1.0 Contributors (2020) **SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python** *Nature Methods* **17**:261–272 <https://doi.org/10.1038/s41592-019-0686-2>
- Vyas V. K., Ukawala R. D., Ghatge M., Chintla C. (2012) **Homology Modeling a Fast Tool for Drug Discovery: Current Perspectives** *Indian Journal of Pharmaceutical Sciences* **74**:1–17 <https://doi.org/10.4103/0250-474X.102537>
- Warren G. L., Do T. D., Kelley B. P., Nicholls A., Warren S. D. (2012) **Essential considerations for using protein–ligand structures in drug discovery** *Drug Discovery Today* **17**:1270–1281 <https://doi.org/10.1016/j.drudis.2012.06.011>
- Wong F., Krishnan A., Zheng E. J., Stärk H., Manson A. L., Earl A. M., Jaakkola T., Collins J. J. (2022) **Benchmarking AlphaFold-enabled molecular docking predictions for antibiotic discovery** *Molecular Systems Biology* **18** <https://doi.org/10.15252/msb.202211081>

Author information

Masha Karelina

Biophysics Program, Stanford University, Stanford, CA 94305, USA, Department of Computer Science, Stanford University, Stanford, CA, USA, Department of Molecular and Cellular Physiology, Stanford University School of Medicine, Stanford, CA, USA, Department of Structural Biology, Stanford University School of Medicine, Stanford, CA, USA, Institute for Computational and Mathematical Engineering, Stanford University, Stanford, CA, USA
ORCID iD: [0000-0003-1880-4536](https://orcid.org/0000-0003-1880-4536)

Joseph J. Noh

Department of Computer Science, Stanford University, Stanford, CA, USA, Department of Molecular and Cellular Physiology, Stanford University School of Medicine, Stanford, CA, USA, Department of Structural Biology, Stanford University School of Medicine, Stanford, CA, USA, Institute for Computational and Mathematical Engineering, Stanford University, Stanford, CA, USA

Ron O. Dror

Biophysics Program, Stanford University, Stanford, CA 94305, USA, Department of Computer Science, Stanford University, Stanford, CA, USA, Department of Molecular and Cellular Physiology, Stanford University School of Medicine, Stanford, CA, USA, Department of Structural Biology, Stanford University School of Medicine, Stanford, CA, USA, Institute for Computational and Mathematical Engineering, Stanford University, Stanford, CA, USA

For correspondence: ron.dror@stanford.edu

ORCID iD: [0000-0002-6418-2793](https://orcid.org/0000-0002-6418-2793)

Editors

Reviewing Editor

Aaron Frank

Arrakis Therapeutics, United States of America

Senior Editor

Qiang Cui

Boston University, United States of America

Reviewer #1 (Public Review):

The authors assess the accuracy of AlphaFold2 (AF2) structures for small molecule ligand pose prediction versus the accuracy with traditional template-based homology models and experimental crystal structures (with a different ligand). They take a careful, rigorous approach leveraging the wealth of structural data around the GPCR protein family and using state-of-the-art docking methods. They find that binding sites are significantly more accurately modeled by AF2 compared to traditional template-based approaches, but this does not translate to greater accuracy in small-molecule docking pose prediction. The important findings around this conclusion have broad implications for the use of AF2 models in ligand binding pose prediction for proteins and drug design.

Strengths:

The strength of the work is the rigor and careful, thoughtful comparison that cleverly leverages the cut-off date of April 30th, 2018 used in the training of AF2. While the authors list their limited number of docking methods as a caveat, the fact that they use three state-of-the-

art ligand docking methods is a strength of the work; many studies use just one. The rigorous analysis of the binding site RMSD and docked ligand pose RMSD is novel to my knowledge and is particularly insightful.

Weaknesses:

The authors are rigorous in their approach by using state-of-the-art workflows that are high-throughput in nature. However, human expert-refined models and expert selection from multiple models could improve the results of ligand pose prediction when using protein models. The authors alluded to this for traditional models but this can also be true when starting from AF2 models. This is difficult to test systematically and rigorously, however. One possible experiment is to explore whether using multiple AF2 models (there are five by default) would have an effect on pose accuracy, perhaps for selected examples such as NK1R, 5HT2A, and DRD1 to help build out the discussion further. Another possible weakness is that the authors focus only on GPCRs for reasons they state but make a good argument as to why the conclusions are likely to extend to other protein classes.

Context:

One of the most common and impactful uses of protein structures is in small molecule therapeutic chemical tool design. When no experimental structure is available, models are frequently used and such models include traditional template-based homology models and, more recently, AF2 models. AF2 is widely recognized as an inflection point in protein structure prediction due to the unprecedented accuracy of the protein structure models produced automatically. However, understanding whether this unprecedented accuracy translates to better small molecule ligand pose prediction has been an open question, and this study directly addresses the question in a systematic, rigorous approach.

Reviewer #2 (Public Review):

While the question of 'are AlphaFold-predicted structures useful for drug design' has largely seen comparisons of AF versus experimental protein structures, this paper takes a less explored (but perhaps more practically important) angle of 'are AlphaFold-predicted structures any better than the previous generation of homology modeling tools' to the protein-ligand (rigid) docking problem. The conclusions of this work will be of largest interest to the audience less familiar with the precision required for successful rigid docking, while the expert crowd might find them obvious, yet a good summary of results previously shown in the literature. Further work, understanding the structural objectives/metrics that should be placed on future AlphaFold-like models for better pose prediction performance, would greatly expand the practicality of the observations made here.

The main conclusion of the paper, that structural accuracy (expressed as RMSD) of the protein model is not a good predictor of the accuracy the model will show in rigid docking protein-ligand pose prediction, is a good reminder of the well-appreciated need for high-quality side chain placements in docking. The expected phenomenon of AlphaFold predicting 'more apo-like structures' is often discussed in the field, and readers should be cautious about drawing conclusions from the rigid (rather than flexible, as in some previous works) docking done here.

The authors have very clearly communicated that the use of AlphaFold-generated structures in traditional docking might not be a good idea, and motivated that the time of a molecular designer might be better spent preparing a high-quality homology model. The visual presentation of the conclusions is very clear but might leave the reader wanting a more in-depth discussion of which structural elements of the AF models lead to bad docking outcomes. For example, Fig. 3 presents an example of a very accurate AlphaFold prediction leading to the ligand being docked completely outside of the binding pocket. Close inspection of the Figure suggests a clash of the ligand with the slightly displaced tryptophan residue in the AF model that might be to blame, as can be confirmed by comparison of the model and

PDB structure by the reader themselves but has not been discussed by the authors. Only a few examples of the systems used are shown even visually, leaving the reader unable to study more interesting cases in depth without re-doing the work themselves.

The authors acknowledged that several recent studies exist in this space. They point out two advancements made in their work, worthy of further review. Similarly, it's important to evaluate the novelty of this work's claims vs previously available results, and the diversity of information made available to the reader.

"First, we use structural models generated without any use of known structures of the target protein. For machine learning methods, this requires ensuring that no structure of the target protein was used to train the method." This is done by limiting the scope of the work to GPCRs whose structures became available only after the training date of AlphaFold (April 30, 2018), as well as not using templates available after that date during prediction. The use of a time limit seems less preferable than the approach taken in Ref. 1 of discarding templates above a sequence identity cutoff. On the other hand, the 'ablation test' performed in Ref. 2 showed no loss in accuracy when no templates were used at all. Authors should discuss in more detail whether these modeling choices could change anything in their conclusions and why they made their choices compared to those in previous work.

"Second, we perform a systematic comparison that takes into account the variation between experimentally determined structures of the same protein when bound to different ligands." Cross-docking is indeed a more appropriate comparison than self-docking (as done in previous works), and the observation that the accuracy of AF models is similar to that between different holo structures of the same protein is interesting. Previous literature on cross-docking should however be discussed, and the well-known conclusions from it that small variations in side-chain positions, in otherwise highly similar structures, can lead to large changes in docked poses. It is important to realize that AlphaFold models are 'just another structure' - if previous literature is sufficient to show the sensitivity of rigid docking, doing it again on AF structures does not add to our understanding. Further, Ref. 3 might have already addressed the question of correlation between binding site RMSD and docking pose prediction accuracy - see e.g. Supplementary Figure 3 there (also Figure S15 in Ref. 2).

Further, the authors should discuss the commonly brought up problem of AlphaFold generating 'more apo-like structures' - are the models used here actually 'holo-like' because of the low RMSDs? (what RMSD differences are to be expected between apo and holo structures of these systems?) How are the volumes of the pockets affected? The position on this problem taken by previous works is worth mentioning - "much higher rmsd values are found when using the AF2 models (...), which reflect the difficulties in performing docking into apo-like structures" in Ref. 1 and "computational model structures were predicted without consideration of binding ligands and resulted in apo structures" in Ref. 2.

Because of this 'apo problem', Ref. 2 assumed that rigid docking (as done here) would not succeed and used flexible docking where "two sidechains at the binding site were set to be flexible". In fact, the reader of this new paper will be left to wonder if it is not simply presenting a subset of the results already seen in Ref. 2, where "the success ratios dropped significantly for them because misoriented sidechains prevented a ligand from docking (Figure S14)". While this conclusion is not made as clear in Ref. 2 as it is here, a comparison of Figures 4 and S14 there will lead the reader to the same conclusion, and more – that flexible docking meaningfully improves the performance of AF models, and more so than homology models.

Finally, certain data analyses present in previous works but not here should be necessary to make this work more informative to the readers:

a) Consideration of multiple top poses, e.g., in Ref. 2, Figures 4 and S14 mentioned before, comparison of success rates in top 1 and top 3 docked poses add much context.

b) Notes on the structural features preventing successful docking, see e.g., in Ref. 1, Table 2 or in Ref. 4, Tables 2 and 4.

This work has the potential to become an important piece of the puzzle, if deeper insights into the reasons for AF model failures are drawn by the authors. These could include a discussion of the problematic structural elements (clashes of side chain with ligands, missing interactions/waters, etc.), potential solutions with some preliminary data (flexible docking, softening interactions, etc.), or proposals for metrics better than RMSD to score the soundness of pockets generated by AF for docking.

References:

1. Díaz-Rovira, A. M., Martín, H., Beuming, T., Díaz, L., Guallar, V., & Ray, S. S. (2023). Are Deep Learning Structural Models Sufficiently Accurate for Virtual Screening? Application of Docking Algorithms to AlphaFold2 Predicted Structures. *Journal of Chemical Information and Modeling*, 63(6), 1668-1674. <https://doi.org/10.1021/acs.jcim.2c01270>
2. Heo, L., & Feig, M. (2022). Multi-state modeling of G-protein coupled receptors at experimental accuracy. *Proteins: Structure, Function, and Bioinformatics*, 90(11), 1873-1885. <https://doi.org/10.1002/prot.26382>
3. Beuming, T., & Sherman, W. (2012). Current assessment of docking into GPCR crystal structures and homology models: Successes, challenges, and guidelines. *Journal of Chemical Information and Modeling*, 52(12), 3263-3277. <https://doi.org/10.1021/ci300411b>
4. Scardino, V., Di Filippo, J. I., & Cavasotto, C. (2022). How good are AlphaFold models for docking-based virtual screening? [Preprint]. *Chemistry*. <https://doi.org/10.26434/chemrxiv-2022-sgj8c>

Author Response

We thank the reviewers and editor for their careful evaluation of our manuscript, and we appreciate their favorable assessment of our work. Below, we clarify a few points concerning the relationship between our study and previous studies evaluating ligand docking to protein models.

As reviewer 2 correctly notes, several previous assessments of AF2 models have simply excluded templates above a sequence identity cutoff when using AF2 to predict structures. Such AF2 predictions are still informed by all structures in the PDB before April 30, 2018, because these structures were used to train AF2—that is, to determine the tens of millions of parameters (“weights”) in the AF2 neural network. Machine learning methods nearly always perform better when evaluated on the data used to train them than when evaluated on other data. For this reason, we consider AF2 models only for proteins whose structures were not used to train AF2—that is, for proteins whose structures were not available in the PDB before April 30, 2018.

Previous papers (including Beuming and Sherman, 2012, <https://doi.org/10.1021/ci300411b>) have shown a clear correlation between the binding pocket RMSD of a protein model and pose prediction accuracy based on that model. Our main findings are unexpected in light of these previous reports: we find that AF2 models yield pose prediction accuracy similar to that of traditional homology models despite having much better binding pocket RMSDs, and that AF2 models yield substantially worse pose prediction accuracy than experimentally determined structures with different ligands bound despite having similar binding pocket RMSDs.

Reviewer 2 also correctly notes that previous papers have described AF2 models as “apo models,” because these models do not include coordinates for bound ligands. As noted by the AF2 developers (e.g., <https://alphafold.ebi.ac.uk/faq>), however, AF2 is designed to predict coordinates of protein atoms as they might appear in the PDB, and AF2 models are thus

frequently consistent with structures in the presence of ligands even though those ligands are not included in the models. GPCR structures in the PDB, including those used to train AF2, nearly always contain a ligand in the orthosteric binding pocket. An AF2 model of a GPCR should thus not be viewed as an attempt to predict the GPCR's structure in the unliganded (apo) state.