Assessing interaction recovery of predicted protein-ligand poses

David Errington **Constantin Schneider** **Cédric Bouysset** **Frédéric A. Dreyer***
Exscientia, Oxford Science Park, Oxford, OX4 4GE, UK
{derrington,cschneider,cbouysset}@exscientia.co.uk
dreyer.frederic@gene.com

Abstract

The field of protein-ligand pose prediction has seen significant advances in recent years, with machine learning-based methods now being commonly used in lieu of classical docking methods or even to predict all-atom protein-ligand complex structures. Most contemporary studies focus on the accuracy and physical plausibility of ligand placement to determine pose quality, often neglecting a direct assessment of the interactions observed with the protein. In this work, we demonstrate that ignoring protein-ligand interaction fingerprints can lead to overestimation of model performance, most notably in recent protein-ligand cofolding models which often fail to recapitulate key interactions.

1 Introduction

Recent advances in AI-based docking hold the potential to generate accurate protein-ligand poses at often a fraction of the computational cost of classical docking algorithms. Additionally, cofolding models that can directly predict the full protein-ligand complex structure have emerged as a promising alternative, circumventing the need for docking while providing the capability to model conformational changes to the protein.

As these machine learning (ML) methods are typically trained on the PDBBind General dataset [Liu et al., 2017] released in 2020, it has become commonplace to benchmark them using the PoseBusters test suite [Buttenschoen et al., 2024] which consists of 308 protein-ligand complexes released after 2021 and that are, therefore, outside their training data.

It has previously been noted [Cole et al., 2005, Buttenschoen et al., 2024, Harris et al., 2023, Baillif et al., 2024, Morehead et al., 2024] that ML methods lack the necessary inductive bias to generate realistic poses, even though they can often obtain low root-mean-squared deviation (RMSD) values from the crystal structure ground truth. Performing further quality checks on the ligand chemistry and the physical plausibility of the pose, notably through the PoseBusters benchmark, is therefore an important test for ML-based docking tools.

However, from the perspective of computational chemists, a physically plausible pose with low RMSD is a necessary but not sufficient condition for that ligand to be of interest. In particular, these conditions ensure that the ligand is close to where it should be and adopts a sensible pose within the pocket, but for that pose to be of biological relevance, it must also create key interactions between the protein and the ligand. These interactions are in fact often used to constrain classical docking tools, an option that is not currently available in ML docking methods. Such interactions are typically classified using protein-ligand interaction fingerprints (PLIFs), which identify the protein residue, the interaction type and, optionally, the ligand atom involved in the interaction. Several tools exist to

*Currently at Prescient Design, Genentech, New York City, USA.

Foundation Models for Science Workshop,38th Conference on Neural Information Processing Systems (NeurIPS 2024).

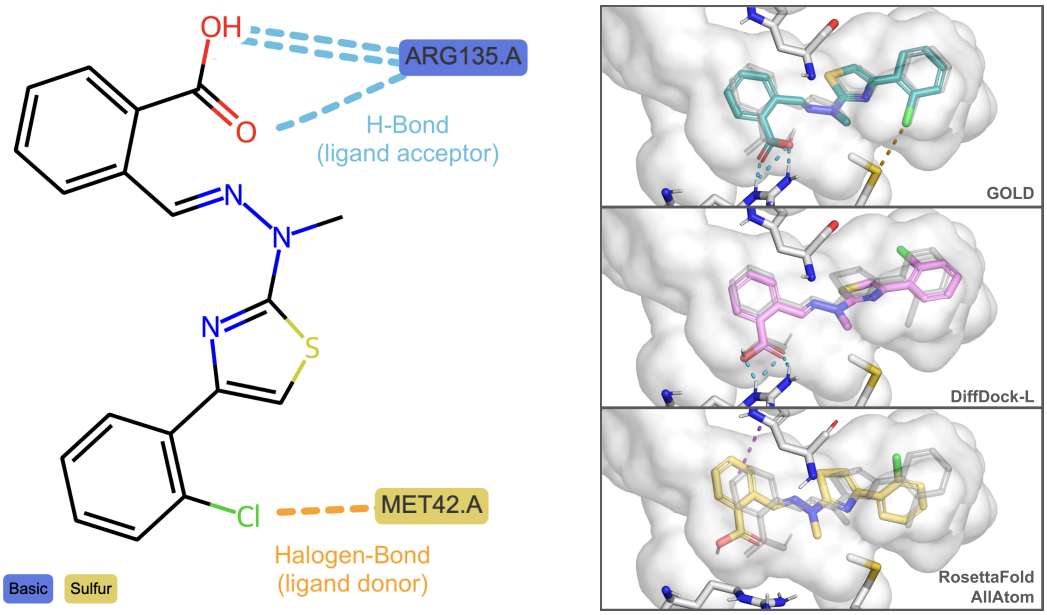

Figure 1: Left: Two-dimensional representation of the ligand EZO and its four interactions with the crystal structure 6M2B. Basic residues are shown in blue and residues containing a sulfur atom are shown in yellow. Right: Docked poses generated with GOLD, DiffDock-L and RosettaFold-AllAtom showing the calculated interactions for each model, with the ground truth ligand in grey.

detect PLIFs [Salentin et al., 2015, Jubb et al., 2017, Wójcikowski et al., 2015, Da Silva et al., 2018] and in this work we use the ProLIF package [Bouysset and Fiorucci, 2021].

In Figure 1, we show on the left a visualisation of the PLIFs detected in the crystal structure of the protein target 6M2B with ligand EZO, and on the right the 3D poses generated by GOLD (classical docking), DiffDock-L (ML docking) and RoseTTAFold-AllAtom (ML cofolding). The ground truth complex has hydrogen bonds and a halogen bond. In this example, both GOLD and DiffDock-L are able to identify PoseBuster-valid (PB-valid) poses with RMSD$\leq$2Å, but whilst DiffDock-L recovers 75% of the PLIFs from the crystal pose, missing the halogen bond interaction with the Chlorine atom, GOLD is able to recover all of them. DiffDock-L also changes the conformation of the ligand so that the hydrogen bonding involves a different set of atoms, while GOLD recovers the exact ground truth pose. RoseTTAFold-AllAtom meanwhile, which has the more challenging task of also reconstructing the protein, finds a pose with a RMSD of 2.19Å and steric clashes, which also fails to recover any of the ground truth crystal interactions.

ML methods do learn indirectly about protein-ligand interactions but without an explicit term to this effect in the loss function, the training signal is weak, and ML docked ligands can often end up with key functional groups pointing in the wrong direction. In contrast, classic docking algorithms are, through the design of their scoring functions, inherently interaction-seeking; their top scoring poses are those that achieve certain key interactions. In this paper, we aim to motivate PLIF recovery as a useful metric for assessing model quality and use them to benchmark a number of modern pose prediction tools.

2 Method

2.1 Protein-ligand interaction fingerprint

Interaction fingerprints summarize the three-dimensional interactions present in a molecular complex. In the context of small molecule drug discovery, we are primarily interested in interactions that a ligand achieves with the protein pocket of interest, for which PLIFs provide a vectorized representation. This representation typically consists of a mapping between protein residues and a ligand along with a bitvector that can encode different types of interactions, such as hydrophobic, π-stacking, π-cation,

ionic, and hydrogen bonds. PLIFs were calculated with the ProLIF package [Bouysset and Fiorucci, 2021] considering only hydrogen and halogen bonds (donor and acceptor), π-stacking, cation-π and π-cation, and ionic interactions (anionic and cationic), excluding the less specific hydrophobic interactions and Van der Waals contacts. These are more rarely considered by computational chemists as key interactions that must be recapitulated as they are non-directional and therefore highly correlated with RMSD. This is because these latter interactions are much more promiscuous than the others, which would result in a weaker signal from polar interactions despite their critical importance in ligand-protein binding. Custom distance thresholds were used for hydrogen bonds (3.7Å), cation-π (5.5Å) and ionic (5Å) interactions while all other parameters are the defaults in ProLIF v2.0.3.

We note that while not all the other PLIF-calculation tools previously mentioned require explicit hydrogens to be present in the input files [Salentin et al., 2015, Jubb et al., 2017], they end up adding them if not present, although the optimisation of the hydrogen bond network is either not enabled by default, or not available.

The interactions detected by PLIF-libraries are very sensitive to the protonation state of both the protein and the ligand as it can decide whether an interaction gets labelled as ionic or hydrogen bond. Whilst the classical docking methods can model hydrogens explicitly, their scoring functions often infer the potential for interactions such as hydrogen bonds from the geometry of the heavy atoms alone. Meanwhile, ML methods typically model only heavy atoms. In order to treat all methods equally, we place explicit hydrogens on the protein structure using PDB2PQR [Jurrus et al., 2018], as well as on the ligand pose if not already present, using RDKit [Landrum et al., 2024]. We then performed a short minimisation of the ligand inside the pocket, defined as protein residues within 6Å of the ligand, whilst keeping the heavy atoms fixed, using RDKit's implementation of the Merck Molecular Force Field (MMFF) [Tosco et al., 2014, Halgren, 1999]. This is a consistent way to optimise the hydrogen bond network of the docked/cofolded pose and gives each method the best possible chance to make interactions from the proposed heavy atom positions.

2.2 Classical docking algorithms

Classical molecular docking aims to predict plausible ligand poses when binding to a protein target, leveraging computational algorithms to accurately simulate molecular interactions, as pioneered by the development of the DOCK [Kuntz et al., 1982] and AutoDock [Goodsell and Olson, 1990] algorithms. In this analysis, we use the FRED, HYBRID and GOLD algorithms which are more modern approaches to classical docking.

FRED and HYBRID are docking programs from the OEDocking suite [OpenEye, a] and are rarely included in ML docking benchmarks. Both algorithms work by first generating an ensemble of conformations which then undergo rigid docking into a specified pocket. FRED is an unbiased docking program that uses only the structure of the target protein to position and score molecules, whilst HYBRID is a biased docking program that also uses the structure of the reference ligand to find the optimal docked pose [McGann, 2012]. HYBRID is typically used in a lead optimisation campaign to dock novel compounds that differ minimally from a reference ligand. For the self-docking task we consider in this work, HYBRID has an unfair advantage over the other methods and we include it here mainly to validate this advantage over FRED.

Finally we include CCDC GOLD [Jones et al., 1997]. Unlike the OEDocking tools, GOLD generates ligand conformations on the fly as it places the ligand in the pocket.

FRED and HYBRID both use the ChemGauss4 scoring function [OpenEye, a] whilst GOLD uses the PLP scoring function [Verkhivker et al., 2000] to identify the optimal pose. In both cases, these scoring functions pay close attention to the shape and hydrogen bond complementarity of poses within the active site. In contrast to ML methods, classical docking methods explicitly seek interactions and we hypothesise this will lead to improved PLIF recovery and ultimately more favourable poses.

For all three classical methods we return 10 poses and then select the pose with the top docking score for our subsequent analysis.

Additionally, we note that existing benchmarks in the literature often perform classical docking with minimal processing of the PDB files, overlooking refinement steps to address issues like missing loops, alternate conformations, flipped functional groups, and adding explicit hydrogen atoms to the ligand and protein structures consistently with their titration states. A suitable preparation of input

files ensures that the active site residues and the ligand are ready for docking, making the simulations more accurate and predictive of ground-truth interactions. Since we are using OpenEye docking tools in this work, we performed structure preparation using the Spruce CLI from OpenEye [OpenEye, b]. We note however, that other structure preparation tools do exist such as Reduce [Word et al., 1999], the CSD API from CCDC [Groom et al., 2016] and the Protein Preparation Wizard from Schrödinger [Schrödinger, 2024].

2.3 ML docking algorithms

The application of ML to accelerate molecular docking and find more accurate binding poses has received a lot of interest in recent years [Stärk et al., 2022, Lu et al., 2022, Zhou et al., 2023].

In this work, we consider DiffDock-L [Corso et al., 2024], the latest version of DiffDock which uses confidence bootstrapping to improve significantly on previous versions. DiffDock-L is a state-of-the-art ML docking model that uses a diffusion model over the non-Euclidean manifold parameterizing the ligand degrees of freedom in order to generate plausible orientations and conformations. DiffDock-L uses its confidence model to assign a score to each sampled pose and so, as with the classical methods, we sample 10 poses for each ligand and use the highest-confidence pose for our subsequent analysis.

It is worth highlighting that the confidence model underpinning DiffDock-L is a GNN classifier trained to identify poses with RMSD$\leq$2Å and, whilst this will indirectly capture some information about interactions, it does not explicitly rank poses based on PLIFs in the same way as classical scorers.

2.4 Protein-ligand cofolding

Several structure prediction models have recently incorporated the description of more general biomolecular assemblies beyond simple protein polypeptide chains, including the capability of cofolding a protein and small molecule simultaneously and predict all corresponding atomic coordinates [Krishna et al., 2024, Bryant et al., 2024, Abramson et al., 2024, Qiao et al., 2024]. We test two such models, Umol [Bryant et al., 2024] and RoseTTAFold All-Atom (RFAA) [Krishna et al., 2024]. Unlike the docking methods described in previous sections, Umol and RFAA will return a protein different to that in the crystal structure and the protein structure is also an output of the model.

Cofolding is a complex problem and, whilst there has been much progress recently, it is still a relatively nascent field. As a result, it is not uncommon for the output structures to have issues such as steric clashes, overabundance of cis-peptide bonds or gaps in the protein, or to fail to preserve the chemistry of the input ligand (*e.g.*, flipped stereochemistry). By default, Umol performs postprocessing in a attempt to fix this whereby it generates conformers of the input ligand and then returns the conformer with the best Kabsch alignment against the predicted atom positions. This approach guarantees the chemistry of the output ligand matches the chemistry of the input ligand. Finally, Umol then places hydrogens and uses OpenMM [Eastman et al., 2017] to optimise the protein-ligand system and it is this optimised complex that we use in our subsequent analysis.

In contrast, RFAA does no such postprocessing out of the box. Consequently, we often find that the output ligand either has invalid stereochemistry or it has valid stereochemistry (making it PB-valid) but this stereochemistry is different to that of the input ligand. To ensure we are assessing the correct ligand, and for consistency with Umol, we add a similar postprocessing pipeline to RFAA, but with minimization in the YASARA2 forcefield [Krieger and Vriend, 2015], which we found to be more tolerant than OpenMM to unphysical structures.

2.5 Data and Metrics

The original PoseBusters test suite identified 308 high-quality protein-ligand complexes released after 2021 and therefore outside the training data of most ML methods [Buttenschoen et al., 2024]. We excluded 37 data points due to limitations in compute time for cofolding involving large targets, and 8 due to either structure preparation or forcefield failures. A further 7 targets were found to have no relevant interactions in the crystal pose, which arises when the ligand and pocket residues exclusively have hydrophobic interactions which we do not calculate, or the interactions in the complex are slightly outside of the distance and angles thresholds used to generate PLIFs. Altogether, this leaves 256 PoseBuster complexes for our analysis.

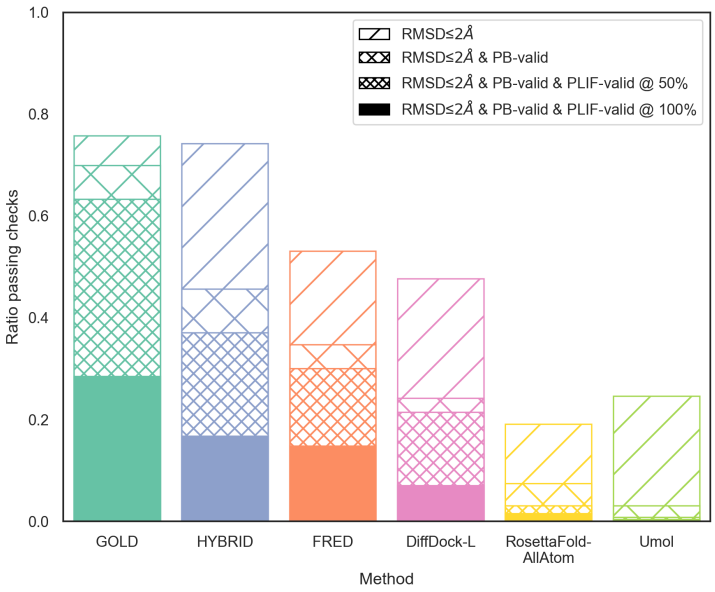

Figure 2: The ratio of predicted protein-ligand complex structures for each model passing checks on ligand positioning (RMSD$\leq$2Å), physicality (PoseBuster-valid) and interaction recovery (PLIF-valid).

We run each of the methods on the PoseBusters dataset and record the following properties

- RMSD to crystal pose
- PoseBuster validity
- PLIFs of the predicted pose

The central contribution of this paper is the introduction of a PLIF recovery rate metric. This metric measures the percentage of interactions in the crystal pose that are successfully replicated in the docked or cofolded pose, as measured by PLIFs generated by ProLIF (see Section 2.1), and captures how well each method can account for protein-ligand interactions.

3 Results

We now turn to the evaluation of interaction recovery in predicted ligand poses with classical docking, ML docking and protein-ligand cofolding on the PoseBusters dataset.

3.1 PoseBusters benchmark

Figure 2 shows the overall results of the six methods on the PoseBusters benchmark set. As in the original PoseBusters paper [Buttenschoen et al., 2024], we show performance according to different metrics. The striped region shows the percentage of poses with RMSD$\leq$2Å, whilst the coarse crosschecked region shows the percentage of poses that are also physically plausible and successfully pass the PoseBuster validity checks. Our new additions are the fine crosschecked and solid regions which show the percentage of poses that are additionally "PLIF-valid" and succeed in also recovering at least 50% and 100% of the interactions present in the crystal pose respectively.

One observation in the original PoseBusters study is that by RMSD alone, the original DiffDock model is shown to outperform GOLD but, when physical plausability metrics are added, GOLD is substantially better. However, Figure 2 shows that in our study, GOLD does substantially better than the latest DiffDock-L model, even on the RMSD criteria. This is because we perform structure preparation on the protein before docking as described in Section 2.2. This is more typical of how

traditional docking tools are used in a drug discovery campaign, while classical methods are often used somewhat naively when benchmarking ML algorithms [Rich et al., 2024].

Turning our attention to the full set of results, it is clear that the three traditional docking algorithms outperform the three ML algorithms across every metric, with GOLD achieving the best results. Indeed, GOLD finds more poses successfully recovering at least 50% of the crystal interactions than any of the ML methods are able to produce falling within 2Å RMSD. As expected, HYBRID outperforms FRED due to its ability to use prior knowledge from the crystal ligand pose. Interestingly though, despite being the only method to have this prior advantage, HYBRID is still outperformed by GOLD.

We find that cofolding methods achieve substantially worse interaction recovery than DiffDock-L. Umol achieves a higher fraction of ligands placed within 2Å RMSD than RFAA though it should be noted that, unlike RFAA, Umol receives pocket residues as input. However, Figure 2 shows that for both cofolding tools, but especially for Umol, the vast majority of these poses are physically implausible and missing key interactions.

3.2 Interaction recovery rates

Whilst the previous section focused on the number of poses that successfully recovered either 50% or 100% of the PLIFs in the crystal pose, here we look at the distribution of PLIF recovery rates across all PoseBuster data points.

In Figure 3 we show a histogram of PLIF recovery rates for every method. We use normalized histograms to highlight the impact on this distribution of the RMSD and PoseBuster validity criteria.

We see a noticeable difference in skew between the histograms for the classical methods and the histograms for the ML methods, confirming that classical methods are much more successful at recovering the crystal interactions.

Under the premise that protein-ligand interactions are what we are actually interested in, we can ask the question whether either the RMSD$\leq$2Å filter or the RMSD$\leq$2Å and PB-valid filter are sufficient to leave only poses that make key interactions. If so, we would see a large change in the skew of the histogram as we apply these filters as poses with low PLIF recovery would get filtered out. We observe a noticeable change to all distributions when applying the RMSD filter, which removes ligands placed too far from the ground truth pose for any interactions to be recovered. In the case of GOLD, the PLIF recovery rate is relatively unaffected by the PB-valid filter. The change in skew is more noticeable in HYBRID, FRED and ML-based methods, though the latter have a sample size after filtering too small to be conclusive. It is however clear that many poses with few recovered PLIFs remain after these filters, confirming that interaction recovery can provide a useful orthogonal metric to PoseBuster validity. Further analysis of the correlation between RMSD and PLIF recovery is shown in Appendix A.

3.3 Recovery of different interaction types

Up until this point in our analysis we have not distinguished between different types of protein-ligand interactions. In Figure 4 we show a breakdown by model of the predictions for different types of interactions. The solid region shows the recall for each type of interaction whilst the striped region shows the ratio of detected PLIFs in the proposed pose relative to the PLIFs in the crystal pose.

Looking at the solid regions, we see that the classical methods produce poses that are better at recovering every type of interaction being considered with the exception of cationic interactions where DiffDock outperforms FRED. We hypothesise that this is because classical methods have scoring functions that explicitly seek interactions.

Hydrogen bonds are the most important kind of interactions to consider [Bissantz et al., 2010] and, as shown in Table 1, they are the most prevalent in our dataset, so it is worth emphasising the difference in recall observed across models in this case. It was previously noted that ligands produced by ML *generative* methods do not make as many hydrogen bonds as found in reference datasets [Harris et al., 2023]. Our results here confirm that this is also true for the simpler task of ML *docking* where the reference ligand is given and the model is simply tasked with finding the optimal pose. Again, the reason that ML methods consistently recover fewer hydrogen bonds than classical methods is likely

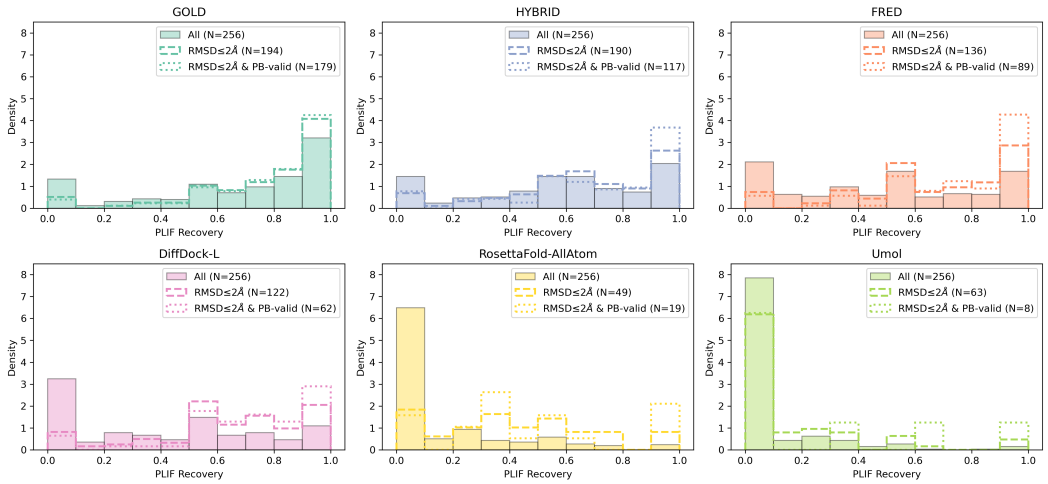

Figure 3: Recovery of protein-ligand interaction fingerprint for each model. The distribution of PLIF recovery among poses that pass the RMSD and PoseBuster test are shown in dashed and dotted lines.

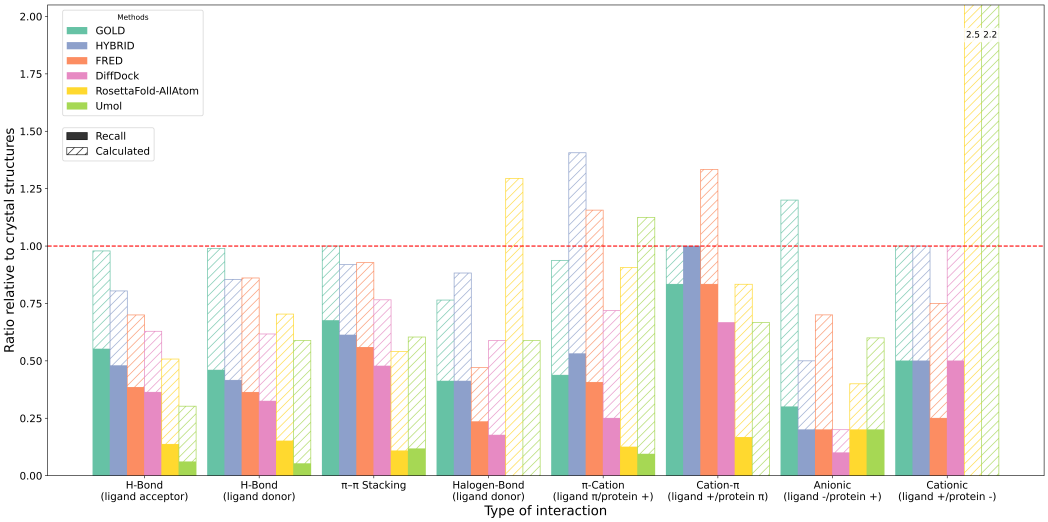

Figure 4: Ratio to the ground truth of calculated and correctly recovered (recall) interactions shown separately for each interaction types.

because the scoring functions driving classical methods are carefully optimised to prioritise hydrogen bonds.

Turning to the striped bars, we can also observe that ML methods generally produce much fewer hydrogen bonds and π-stacking interactions, which are the most frequent interactions in ligand-protein docking as shown in Table 1. The outliers in calculated cationic interactions for RosettaFold-AllAtom and Umol are due to a completely different orientation of the docking pose with respect to the crystal ligand, often replacing cation-π interactions found in the crystal structure with cationic interactions.

4 Discussion

In this paper, we have considered interaction fingerprints in protein bound small molecules. It has become commonplace to consider both ligand RMSD and PoseBuster validity as a proxy for model accuracy. These metrics however do not fully capture the recapitulation of key interactions. We studied how accurately different protein-ligand pose prediction tools, notably classical docking, ML

Table 1: Number of interaction types across PoseBuster crystal structures

Interaction	Frequency
H-Bond (ligand acceptor)	843
H-Bond (ligand donor)	496
π-π Stacking	111
Halogen-Bond (ligand donor)	17
π-Cation (ligand π / protein $+$)	32
Cation-π (ligand $+$ / protein π)	6
Anionic (ligand $-$ / protein $+$)	10
Cationic (ligand $+$ / protein $-$)	4

docking and protein-ligand structure prediction models, can recover ground truth interactions. PLIF recovery provides a useful metric, orthogonal to those used in existing benchmarks, which can further assess validity of predicted poses and is particularly valuable in drug discovery applications.

We showed that classical docking algorithms tend to substantially outperform ML-based methods in generating physically plausible poses, and recover relevant interactions with much higher success rate. This result highlights the fact that classical docking benchmarks are rarely run competitively in the literature. In contrast, cofolding models, where the coordinates of all atoms of the protein and ligand are jointly predicted, while often placing the ligand in the right location, rarely generate physically plausible poses that recover meaningful interactions with the target protein [Masters et al., 2024]. Protein-ligand structure prediction is a harder task than docking, and also claims a much wider set of use cases, such as being able to adapt the conformation of the protein to accommodate different ligands or accurately model cryptic pockets, where the druggable pocket is absent in the apo structure and becomes exposed through interaction with the ligand [Oleinikovas et al., 2016, Meller et al., 2023a,b]. However our results here suggest that in order for this emerging technique to be successful, considerably more attention is needed to ensure the predicted poses form key interactions. This could be achieved by incorporating an explicit PLIF or pharmacophore-sensitive loss to the training of ML models. We note that it is possible to infer all interactions, including hydrogen bonds, from the geometry of the heavy atoms only and so we see potential to introduce geometric terms to the loss functions of ML methods to encourage this. Another simpler option would be to use a weighted RMSD that assigns a higher contribution to atoms matching specific pharmacophoric features (*e.g.* hydrogen bond donors and acceptors, charged atoms, and π-rings).

The code used in this study is made available online at `https://github.com/Exscientia/plif_validity`, along with all prepared protein structures at `https://doi.org/10.5281/zenodo.13843798`.

Acknowledgements

We are grateful to Henry Kenlay, Daniel Cutting, Gail Bartlett, Daniel Nissley, Lukáš Pravda, Ben Butt, Richard Bradshaw, Francis Atkinson, Douglas Pires and Hagen Triendl for useful discussions.

References

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J. Ballard, Joshua Bambrick, Sebastian W. Bodenstein, David A. Evans, Chia-Chun Hung, Michael O'Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvilė Žemgulytė, Eirini Arvaniti, Charles Beattie, Ottavia Bertolli, Alex Bridgland, Alexey Cherepanov, Miles Congreve, Alexander I. Cowen-Rivers, Andrew Cowie, Michael Figurnov, Fabian B. Fuchs, Hannah Gladman, Rishub Jain, Yousuf A. Khan, Caroline M. R. Low, Kuba Perlin, Anna Potapenko, Pascal Savy, Sukhdeep Singh, Adrian Stecula, Ashok Thillaisundaram, Catherine Tong, Sergei Yakneen, Ellen D. Zhong, Michal Zielinski, Augustin Žídek, Victor Bapst, Pushmeet Kohli, Max Jaderberg, Demis Hassabis, and John M. Jumper. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 2024. doi: 10.1038/s41586-024-07487-w. URL https://doi.org/10.1038/s41586-024-07487-w.

Benoit Baillif, Jason Cole, Patrick McCabe, and Andreas Bender. Benchmarking structure-based three-dimensional molecular generative models using genbench3d: ligand conformation quality matters, 2024. URL https://arxiv.org/abs/2407.04424.

Caterina Bissantz, Bernd Kuhn, and Martin Stahl. A medicinal chemist's guide to molecular interactions. *Journal of Medicinal Chemistry*, 53(14):5061–5084, March 2010. ISSN 1520-4804. doi: 10.1021/jm100112j. URL http://dx.doi.org/10.1021/jm100112j.

Cédric Bouysset and Sebastien Fiorucci. Prolif: a library to encode molecular interactions as fingerprints. *Journal of Cheminformatics*, 13, 09 2021. doi: 10.1186/s13321-021-00548-6.

Patrick Bryant, Atharva Kelkar, Andrea Guljas, Cecilia Clementi, and Frank Noé. Structure prediction of protein-ligand complexes from sequence information with umol. *Nature Communications*, 15 (1), May 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-48837-6. URL http://dx.doi.org/10.1038/s41467-024-48837-6.

Martin Buttenschoen, Garrett M. Morris, and Charlotte M. Deane. Posebusters: Ai-based docking methods fail to generate physically valid poses or generalise to novel sequences, 2024. URL http://dx.doi.org/10.1039/D3SC04185A.

Jason C. Cole, Christopher W. Murray, J. Willem M. Nissink, Richard D. Taylor, and Robin Taylor. Comparing protein–ligand docking programs is difficult. *Proteins: Structure, Function, and Bioinformatics*, 60(3):325–332, 2005. doi: https://doi.org/10.1002/prot.20497. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.20497.

Gabriele Corso, Arthur Deng, Benjamin Fry, Nicholas Polizzi, Regina Barzilay, and Tommi Jaakkola. Deep confident steps to new pockets: Strategies for docking generalization, 2024.

Franck Da Silva, Jeremy Desaphy, and Didier Rognan. Ichem: A versatile toolkit for detecting, comparing, and predicting protein–ligand interactions. *ChemMedChem*, 13(6):507–510, 2018. doi: 10.1002/cmdc.201700505.

Peter Eastman, Jason Swails, John D. Chodera, Robert T. McGibbon, Yutong Zhao, Kyle A. Beauchamp, Lee-Ping Wang, Andrew C. Simmonett, Matthew P. Harrigan, Chaya D. Stern, Rafal P. Wiewiora, Bernard R. Brooks, and Vijay S. Pande. Openmm 7: Rapid development of high performance algorithms for molecular dynamics. *PLOS Computational Biology*, 13(7):e1005659, July 2017. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1005659. URL http://dx.doi.org/10.1371/journal.pcbi.1005659.

David S. Goodsell and Arthur J. Olson. Automated docking of substrates to proteins by simulated annealing. *Proteins: Structure, Function, and Bioinformatics*, 8(3):195–202, 1990. doi: 10.1002/prot.340080302.

Colin R. Groom, Ian J. Bruno, Matthew P. Lightfoot, and Suzanna C. Ward. The cambridge structural database. *Acta Crystallographica Section B Structural Science, Crystal Engineering and Materials*, 72(2):171–179, April 2016. ISSN 2052-5206. doi: 10.1107/s2052520616003954. URL http://dx.doi.org/10.1107/S2052520616003954.

Thomas A. Halgren. Mmff vi. mmff94s option for energy minimization studies. *Journal of Computational Chemistry*, 20(7):720–729, 1999. doi: 10.1002/(SICI)1096-987X(199905)20:7<720::AID-JCC7>3.0.CO;2-X.

Charles Harris, Kieran Didi, Arian R. Jamasb, Chaitanya K. Joshi, Simon V. Mathis, Pietro Lio, and Tom Blundell. Benchmarking generated poses: How rational is structure-based drug design with generative models?, 2023. URL https://arxiv.org/abs/2308.07413.

Gareth Jones, Peter Willett, Robert C Glen, Andrew R Leach, and Robin Taylor. Development and validation of a genetic algorithm for flexible docking11edited by f. e. cohen. *Journal of Molecular Biology*, 267(3):727–748, 1997. ISSN 0022-2836. doi: https://doi.org/10.1006/jmbi.1996.0897. URL https://www.sciencedirect.com/science/article/pii/S0022283696908979.

Harry C Jubb, Alicia P Higueruelo, Bernardo Ochoa-Montaño, Will R Pitt, David B Ascher, and Tom L Blundell. Arpeggio: A web server for calculating and visualising interatomic interactions in protein structures. *Journal of Molecular Biology*, 429(3):365–371, 2017. ISSN 0022-2836. doi: https://doi.org/10.1016/j.jmb.2016.12.004. Computation Resources for Molecular Biology.

Elizabeth Jurrus, Dave Engel, Keith Star, Kyle Monson, Juan Brandi, Lisa E. Felberg, David H. Brookes, Leighton Wilson, Jiahui Chen, Karina Liles, Minju Chun, Peter Li, David W. Gohara, Todd Dolinsky, Robert Konecny, David R. Koes, Jens Erik Nielsen, Teresa Head-Gordon, Weihua Geng, Robert Krasny, Guo-Wei Wei, Michael J. Holst, J. Andrew McCammon, and Nathan A. Baker. Improvements to the apbs biomolecular solvation software suite. *Protein Science*, 27(1):112–128, 2018. doi: 10.1002/pro.3280.

Elmar Krieger and Gert Vriend. New ways to boost molecular dynamics simulations. *Journal of Computational Chemistry*, 36(13):996–1007, 2015. doi: https://doi.org/10.1002/jcc.23899. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.23899.

Rohith Krishna, Jue Wang, Woody Ahern, Pascal Sturmfels, Preetham Venkatesh, Indrek Kalvet, Gyu Rie Lee, Felix S. Morey-Burrows, Ivan Anishchenko, Ian R. Humphreys, Ryan McHugh, Dionne Vafeados, Xinting Li, George A. Sutherland, Andrew Hitchcock, C. Neil Hunter, Alex Kang, Evans Brackenbrough, Asim K. Bera, Minkyung Baek, Frank DiMaio, and David Baker. Generalized biomolecular modeling and design with rosettafold all-atom. *Science*, 384(6693), April 2024. ISSN 1095-9203. doi: 10.1126/science.adl2528. URL http://dx.doi.org/10.1126/science.adl2528.

I D Kuntz, J M Blaney, S J Oatley, R Langridge, and T E Ferrin. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.*, 161(2):269–288, October 1982.

Greg Landrum, Paolo Tosco, Brian Kelley, Ric, David Cosgrove, sriniker, Riccardo Vianello, gedeck, NadineSchneider, Gareth Jones, Eisuke Kawashima, Dan Nealschneider, Andrew Dalke, Brian Cole, Matt Swain, Samo Turk, Aleksandr Savelev, Alain Vaucher, Maciej Wójcikowski, Ichiru Take, Vincent F. Scalfani, Daniel Probst, Kazuya Ujihara, Rachel Walker, guillaume godin, Axel Pahl, Juuso Lehtivarjo, Francois Berenger, strets123, and jasondbiggs. rdkit/rdkit: 2023_09_6 (q3 2023) release, March 2024.

Z. Liu, M. Su, H. Liang, J. Liu, Q. Yang, Y. Li, and R. Wang. Forging the basis for developing protein–ligand interaction scoring functions. *Accounts of Chemical Research*, 50:302–309, 2017. doi: 10.1021/acs.accounts.6b00491.

Wei Lu, Qifeng Wu, Jixian Zhang, Jiahua Rao, Chengtao Li, and Shuangjia Zheng. Tankbind: Trigonometry-aware neural networks for drug-protein binding structure prediction. *Advances in Neural Information Processing Systems*, 2022.

Matthew R. Masters, Amr H. Mahmoud, and Markus A. Lill. Do deep learning models for co-folding learn the physics of protein-ligand interactions? *bioRxiv*, 2024. doi: 10.1101/2024.06.03.597219. URL https://www.biorxiv.org/content/early/2024/06/04/2024.06.03.597219.

M. McGann. Fred and hybrid docking performance on standardized datasets. *Journal of Computer-Aided Molecular Design*, 26:897–906, 2012. doi: 10.1007/s10822-012-9584-8.

Artur Meller, Soumendranath Bhakat, Shahlo Solieva, and Gregory R. Bowman. Accelerating cryptic pocket discovery using alphafold. *Journal of Chemical Theory and Computation*, 19 (14):4355–4363, Jul 2023a. ISSN 1549-9618. doi: 10.1021/acs.jctc.2c01189. URL https://doi.org/10.1021/acs.jctc.2c01189.

Artur Meller, Michael Ward, Jonathan Borowsky, Meghana Kshirsagar, Jeffrey M. Lotthammer, Felipe Oviedo, Juan Lavista Ferres, and Gregory R. Bowman. Predicting locations of cryptic pockets from single protein structures using the pocketminer graph neural network. *Nature Communications*, 14(1):1177, Mar 2023b. ISSN 2041-1723. doi: 10.1038/s41467-023-36699-3. URL https://doi.org/10.1038/s41467-023-36699-3.

Alex Morehead, Nabin Giri, Jian Liu, and Jianlin Cheng. Deep learning for protein-ligand docking: Are we there yet?, 2024. URL https://arxiv.org/abs/2405.14108.

Vladimiras Oleinikovas, Giorgio Saladino, Benjamin P. Cossins, and Francesco L. Gervasio. Understanding cryptic pocket formation in protein targets by enhanced sampling simulations. *Journal of the American Chemical Society*, 138(43):14257–14263, Nov 2016. ISSN 0002-7863. doi: 10.1021/jacs.6b05425. URL https://doi.org/10.1021/jacs.6b05425.

OpenEye. *OEDOCKING 4.3.0.3. Cadence Molecular Sciences, Inc., Santa Fe, NM. http://www.eyesopen.com.* a.

OpenEye. *Spruce 1.6.0.0. OpenEye, Cadence Molecular Sciences, Santa Fe, NM. http://www.eyesopen.com.* b.

Zhuoran Qiao, Weili Nie, Arash Vahdat, Thomas F. Miller, and Animashree Anandkumar. State-specific protein–ligand complex structure prediction with a multiscale deep generative model. *Nature Machine Intelligence*, 6(2):195–208, Feb 2024. ISSN 2522-5839. doi: 10.1038/s42256-024-00792-z. URL https://doi.org/10.1038/s42256-024-00792-z.

A. Rich, B. Birnbaum, and J. Haimson. Approaching AlphaFold 3 docking accuracy in 100 lines of code. https://www.inductive.bio/blog/strong-baseline-for-alphafold-3-docking, 2024. [Accessed 07-08-2024].

Sebastian Salentin, Sven Schreiber, V. Joachim Haupt, Melissa F. Adasme, and Michael Schroeder. PLIP: fully automated protein–ligand interaction profiler. *Nucleic Acids Research*, 43(W1): W443–W447, 04 2015. ISSN 0305-1048. doi: 10.1093/nar/gkv315.

LLC Schrödinger. Schrödinger release 2024-3: Protein preparation wizard; epik, schrödinger, llc, new york, ny, 2024; impact, schrödinger, llc, new york, ny; prime, schrödinger, llc, new york, ny, 2024, 2024. Schrödinger, LLC, New York, NY.

Hannes Stärk, Octavian Ganea, Lagnajit Pattanaik, Regina Barzilay, and Tommi Jaakkola. Equibind: Geometric deep learning for drug binding structure prediction. In *International Conference on Machine Learning*, pages 20503–20521. PMLR, 2022.

Paolo Tosco, Nikolaus Stiefl, and Gregory Landrum. Bringing the mmff force field to the rdkit: implementation and validation. *Journal of Cheminformatics*, 6(1):37, Jul 2014. ISSN 1758-2946. doi: 10.1186/s13321-014-0037-3.

Gennady M. Verkhivker, Djamal Bouzida, Daniel K. Gehlhaar, Paul A. Rejto, Sandra Arthurs, Anthony B. Colson, Stephan T. Freer, Veda Larson, Brock A. Luty, Tami Marrone, and Peter W. Rose. *Journal of Computer-Aided Molecular Design*, 14(8):731–751, 2000. ISSN 0920-654X. doi: 10.1023/a:1008158231558. URL http://dx.doi.org/10.1023/a:1008158231558.

J.Michael Word, Simon C. Lovell, Jane S. Richardson, and David C. Richardson. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation 1 1edited by j. thornton. *Journal of Molecular Biology*, 285(4):1735–1747, January 1999. ISSN 0022-2836. doi: 10.1006/jmbi.1998.2401. URL http://dx.doi.org/10.1006/jmbi.1998.2401.

Maciej Wójcikowski, Piotr Zielenkiewicz, and Pawel Siedlecki. Open Drug Discovery Toolkit (ODDT): a new open-source player in the drug discovery field. *Journal of Cheminformatics*, 7(1): 26, 06 2015. ISSN 1758-2946. doi: 10.1186/s13321-015-0078-2.

Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-mol: A universal 3d molecular representation learning framework. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=6K2RM6wVqKu.

A Correlation betweeen PLIF recovery and RMSD

In Figure 5, we show a scatter plot of the PLIF recovery rate against the RMSD.

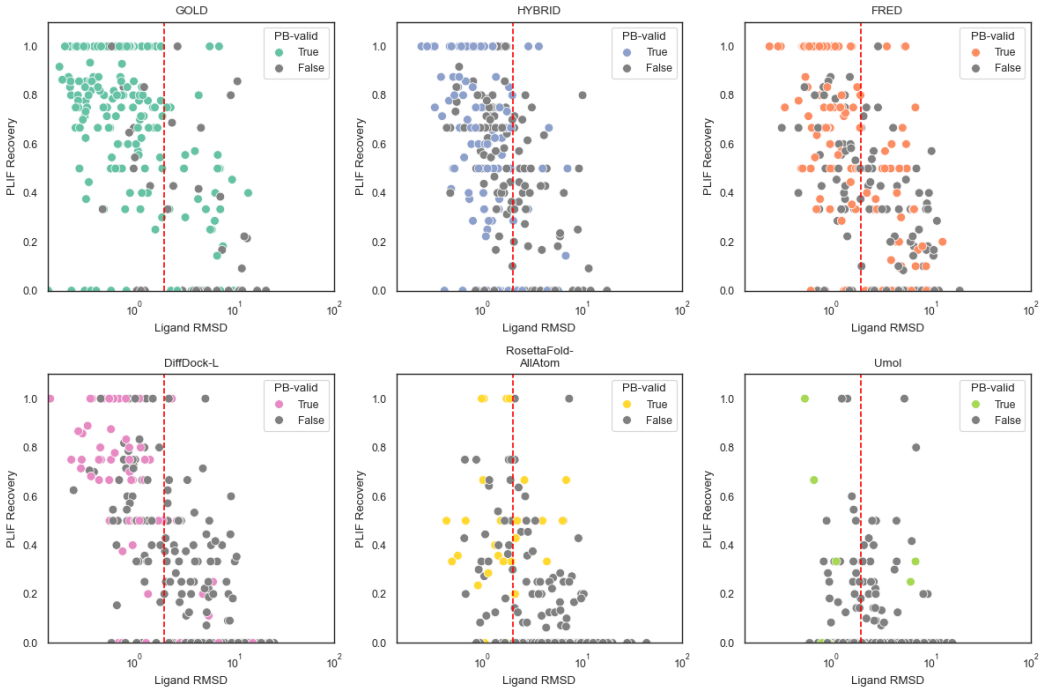

Figure 5: PLIF recovery rate and RMSD, highlighting data points which are PoseBuster-valid. Note that we use a modified definition of PB-validity that excludes ligand RMSD. The red line indicates a ligand RMSD of 2Å.