Journal of Cheminformatics

RESEARCH Open Access

Assessing interaction recovery of predicted protein-ligand poses

David Errington[1], Constantin Schneider[2], Cédric Bouysset[1] and Frédéric A. Dreyer[2,3*]

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.

Scientific Contribution The interaction analysis used in this study is provided as a python package at https://github.com/Exscientia/plif_validity.

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 Protein Data Bank (PDB) [1] with a cutoff date of September 30, 2021, or on the PDBBind General dataset [2] released in 2020, it has become commonplace to benchmark them using the PoseBusters test suite [3] which consists of 308 protein-ligand complexes released after 2021 and that are, therefore, outside their training data.

*Correspondence:
Frédéric A. Dreyer
dreyer.frederic@gene.com
[1] Recursion, Oxford, UK
[2] Exscientia, Oxford Science Park, Oxford OX4 4GE, UK
[3] Present Address: Prescient Design, Genentech, New York, USA

It has previously been noted [3–7] 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. They also tend to perform poorly on structures that do not have high similarity to their training set [8]. 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 [9, 10]. 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 detect PLIFs [11–14] and in this work we use the ProLIF package [15].

In Fig. 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≤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.

Method

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 [15] 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, and including them would result in a weaker signal from polar interactions (as demonstrated in Appendix A Figs. 5 and 6, which show trends similar to another analysis of interactions on small molecule-protein complexes from the PDB [16]) despite their critical importance in ligand-protein binding. Custom distance thresholds were used for hydrogen bonds (3.7Å),

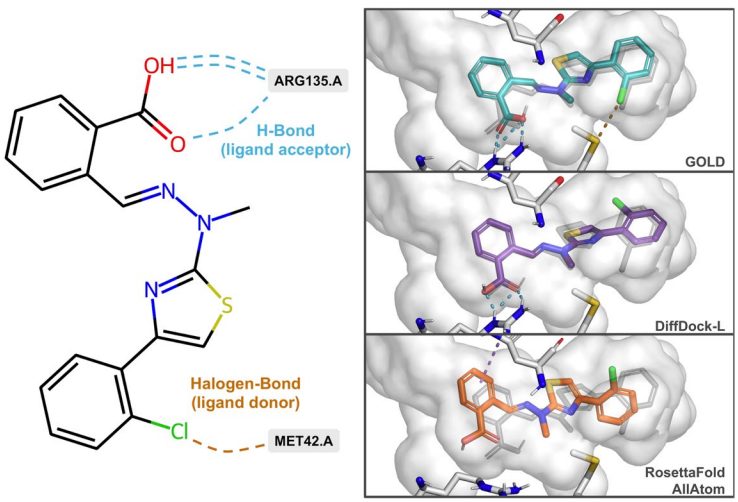

Fig. 1 Left: Two-dimensional representation of the ligand EZO and its four interactions with the crystal structure 6M2B. Basic residues are shown in blue and residues containing a sulfur atom are shown in yellow. Right: Docked poses generated with GOLD, DiffDock-L and RosettaFold-AllAtom showing the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

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 [12, 13], 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 [17], as well as on the ligand pose if not already present, using RDKit [18]. 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) [19, 20]. 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.

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 [21] and AutoDock [22] algorithms. In this analysis, we use the FRED, HYBRID2 and GOLD algorithms which are more modern approaches to classical docking.

FRED and HYBRID2 are docking programs from the OEDocking suite [23] 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 HYBRID2 is a biased docking program that also uses the structure of the reference ligand to find the optimal docked pose [24]. HYBRID2 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, HYBRID2 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 [25]. Unlike the OEDocking tools, GOLD generates ligand conformations on the fly as it places the ligand in the pocket.

FRED and HYBRID2 both use the ChemGauss4 scoring function [23] whilst GOLD uses the PLP scoring function [26] 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 [27]. We note however, that other structure preparation tools do exist such as Reduce [28], the CSD API from CCDC [29] and the Protein Preparation Wizard from Schrödinger [30].

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 [31–33].

In this work, we consider DiffDock-L [34], 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.

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 [35–40]. We test three such models, Umol [36], RoseTTAFold All-Atom (RFAA) [38] and Chai-1 [37]. Unlike the docking methods described in previous sections, Umol, RFAA and Chai-1 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 [41] 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 [42], which we found to be more tolerant than OpenMM to unphysical structures.

For our comparisons with Chai-1, we use as is the dataset of predicted PoseBuster structures provided by the authors in Section 6.2 of [37].

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 [3]. We excluded 37 data points due to limitations in compute time for cofolding involving large targets, and 10 due to either structure preparation or forcefield failures. A further 7 targets were found to have no relevant interactions in the crystal pose (corresponding to the 2.7% of crystal structures with 0 interactions seen in Appendix A Fig. 8), 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 254 PoseBuster complexes for our analysis. Detailed information on compute resources used for this benchmark are given in Appendix B.

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 fraction of interactions in the crystal pose that are successfully replicated in the docked or cofolded pose, as measured by PLIFs generated by ProLIF (see Sect. Protein-ligand interaction fingerprint), and captures how well each method can account for protein-ligand interactions. Concretely,

$$\text{PLIF Recovery} = \frac{\sum_{i,r} \min(C_{i,r}, P_{i,r})}{\sum_{i,r} C_{i,r}}$$

where $C_{i,r}$ and $P_{i,r}$ correspond to the counts for interaction type i and residue r in the crystal structure and docked/cofolded pose respectively. To be concrete, if we consider $i =$ H-bond interactions with residue $r =$ TYR123 and find that the crystal structure has 3 such interactions whilst the docked/cofolded pose has just 1 such interaction then we contribute $\min(3, 1) = 1$ to the sum in the numerator. Since the numbering and chain used for residues may differ between the crystal structure and docked/cofolded pose, we perform a systematic sequence alignment of the relevant chains to correct these differences.

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.

PoseBusters benchmark

Figure 2 shows the overall results of the six methods on the PoseBusters benchmark set. As in the original PoseBusters paper [3], we show performance according to different metrics. The striped region shows the percentage of poses with RMSD≤2Å, whilst the coarse crosschecked region shows the percentage of poses that are also

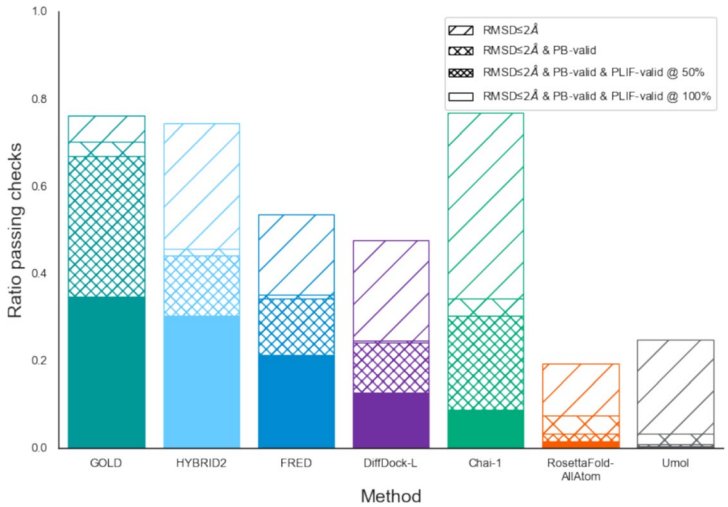

Fig. 2 The ratio of predicted protein-ligand complex structures for each model passing checks on ligand positioning (RMSD≤2Å), physicality (PoseBuster-valid) and interaction recovery (PLIF-valid)

physically plausible and successfully pass the PoseBuster validity checks. Our new additions are the fine cross-checked 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.

Figure 2 shows that GOLD performs best across all metrics and does substantially better than the ML methods, even on the RMSD criteria alone. This is because we perform structure preparation on the protein before docking as described in Sect. Classical docking algorithms. 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 [43].

Turning our attention to the full set of results, it is clear that the three traditional docking algorithms outperform all ML algorithms across every metric, except on RMSD only, where Chai-1 achieves comparable recovery to classical docking. GOLD achieves the best overall results and finds more poses successfully recovering at least 50% of the crystal interactions than the remaining three ML methods are able to produce falling within 2Å RMSD. As expected, HYBRID2 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, HYBRID2 is still outperformed by GOLD.

With the exception of Chai-1, we find that other 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, Fig. 2 shows that the vast majority of poses predicted by RFAA and Umol are physically implausible and missing key interactions.

Whilst the analysis in the main text of this paper considers only the top-scoring docked pose, we explore the effect of including more poses from the docking methods in Appendix C.

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 Fig. 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. Further comparisons of PLIF recovery against RMSD are shown in Appendix D.

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≤2Å filter or the RMSD≤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

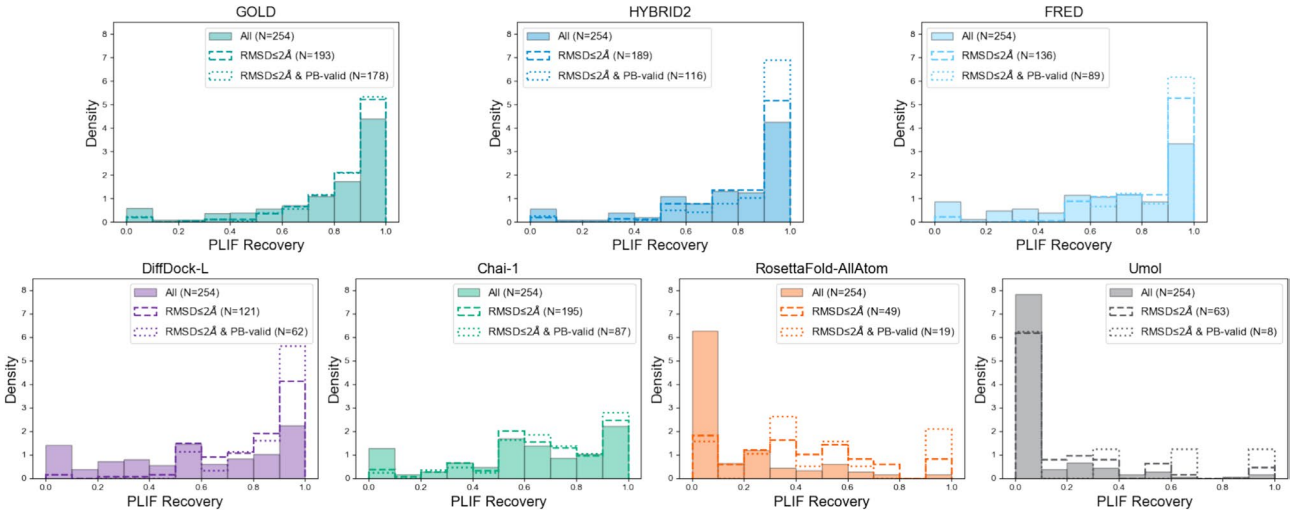

Fig. 3 Recovery of protein-ligand interaction fingerprint for each model. The distribution of PLIF recovery among poses that pass the RMSD and PoseBuster test are shown in dashed and dotted lines

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 HYBRID2, 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 D.

Recovery of different interaction types

Up until this point in our analysis we have not distinguished between different types of protein-ligand interactions. In Fig. 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

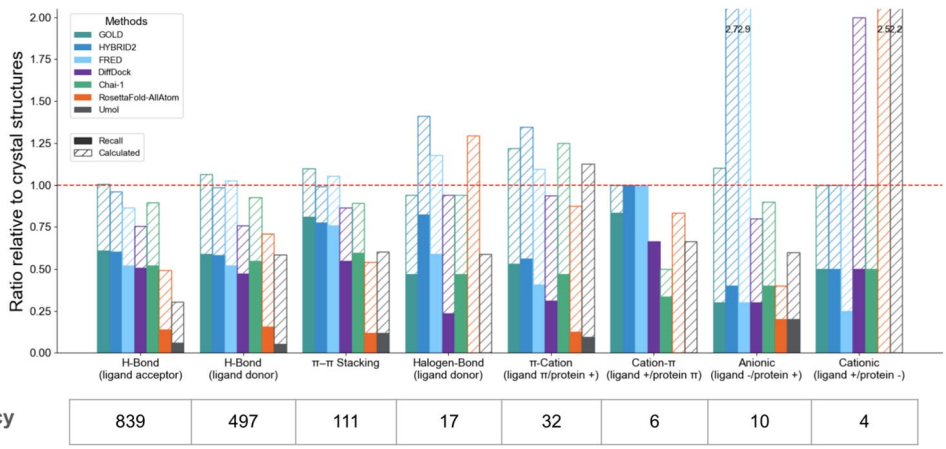

Fig. 4 Ratio to the ground truth of calculated and correctly recovered (recall) interactions shown separately for each interaction types

classical methods have scoring functions that explicitly seek interactions.

Hydrogen bonds are the most important kind of interactions to consider [44] and, as shown in Fig. 4, 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 [6]. 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 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 at the bottom of Fig. 4. 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 as seen in the Appendix E Fig. 12.

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 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. Further assessments, e.g. based on recovery of electrostatic complementarity [45] or weighted combinations of different interaction types, including hydrophobic ones, could provide interesting avenues for future studies.

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 [46]. 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 [47–49]. 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 or lDDT-PLI [50] 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.

Interactions analysis

Figure 5 shows a boxplot of the count for different interaction types across crystal structures from the PoseBusters dataset. Note that hydrophobic and van der Waals interactions are presented on a different axis as they were on average 10 times more prevalent than the other more specific interactions that were ultimately chosen for the PLIF recovery metric. Given that some structures are known to have a low interaction count when excluding hydrophobic and van der Waals contacts, an alternative recovery score including the latter interactions with a lower weighting could broaden the applicability of the metric to more complexes.

In Fig. 6 we show an alternative representation to the interaction count seen in Fig. 5 using the fraction of ligand atoms involved in each interaction rather than the interaction count. Hydrophobic and van der Waals contacts are still predominant compared to the other interactions used in the proposed recovery metric.

In Fig. 7 we show how the hydrophobic interactions and van der Waals contacts vary between methods.

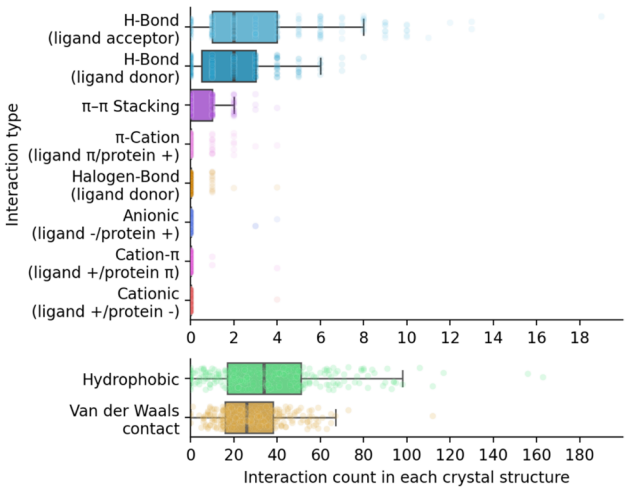

Fig. 5 Boxplot of the counts for each interaction type for crystal structures. Individual dots on the plot represent individual crystal structures

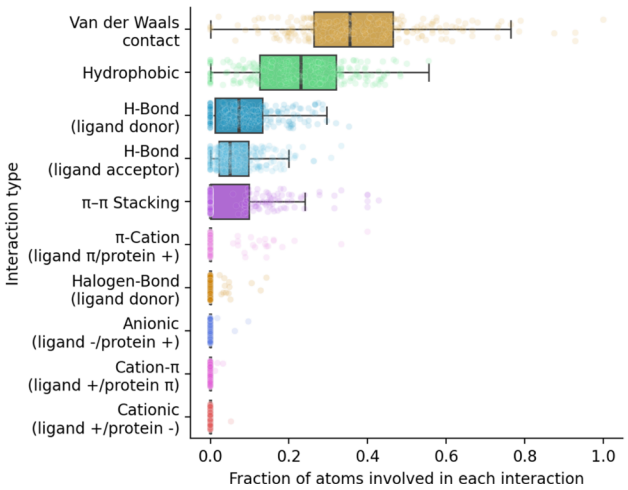

Fig. 6 Fraction of atoms involved in each interaction. $\pi - \pi$ stacking and π − cation involve 6 atoms on the ligand, and H-Bond donor and Halogen-Bond involve 2 so their fractions are higher (given a similar occurrence) compared to all other interactions that only involve a single atom

Umol tends to make less of these interactions as some of the residues are pointing away from the ligand, as seen in Fig. 12. This effect gets amplified when ring-systems are involved as they can provide multiple atoms to participate in these interactions. For RosettaFold-AllAtom, the extreme values are due to the ligand clashing with residues, vastly amplifying the number of van der Waals contacts and, to a lesser extent, the number of hydrophobic interactions.

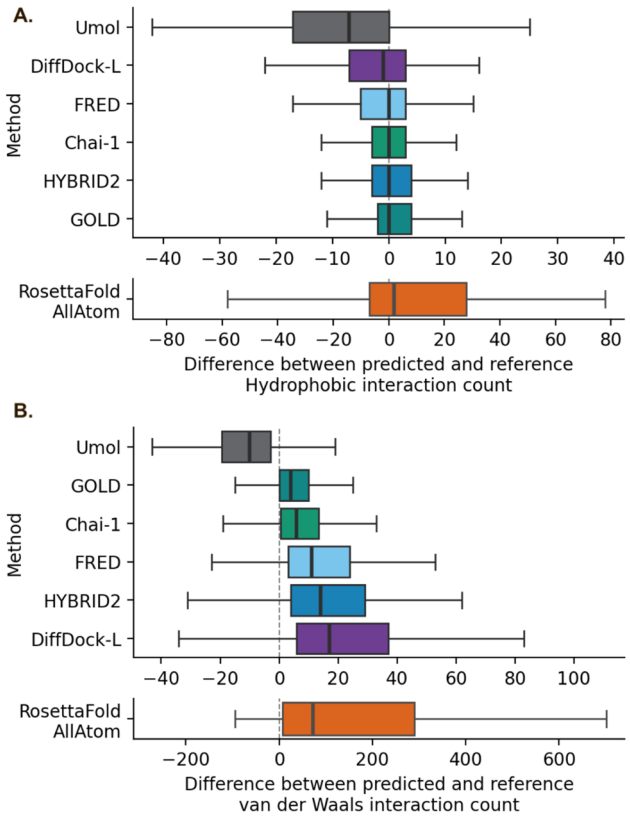

Fig. 7 Difference in the counts for Hydrophobic (**A**) and van der Waals (**B**) interactions between the predicted complexes and the crystal structure

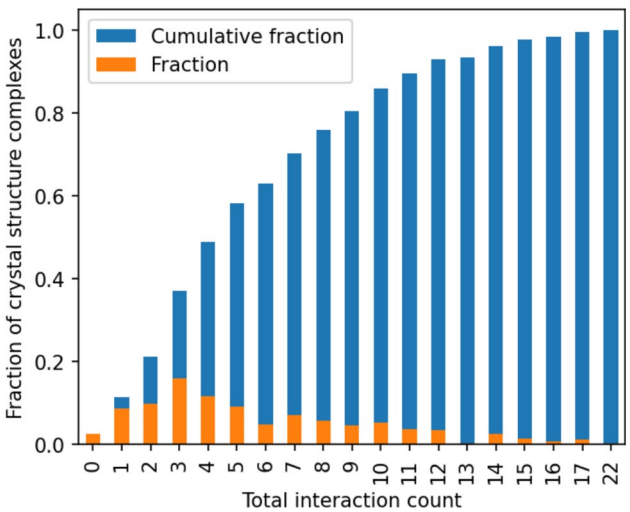

Fig. 8 Fraction of crystal structures having a specific total interactions count. Hydrophobic and van der Waals contacts are excluded from the total

In Fig. 8 we show the fraction of crystal structures in the PoseBusters dataset as a function of the total count of interactions. We see that around 20 percent of crystal structures have two or less interactions, meaning that our proposed PLIF recovery metric may not be a good fit for all complexes, especially ones governed by hydrophobic interactions as we do not account for those in the metric presented here.

Compute requirements

Each individual GOLD, HYBRID2 and FRED docking was run on a single CPU. Details of the compute used for the ML tools is given in Table 1. We do not provide details of running times, as these were not logged and differ substantially across methods based on e.g. protein sequence length.

The PLIF calculations can be performed locally. On an M1 MacBook Pro (MK183B/A) laptop, it takes 8-9 s to optimize the hydrogen bond network and generate the fingerprint for a single complex, then approximately 140ms to calculate the recovery rate given fingerprints for a crystal and docked complex.

Table 1 Instances used for inference on ML methods

Method	Instance
DiffDock-L	g5.8xlarge
RosettaFold-AllAtom	m6i.8xlarge (MSA) & g5.xlarge (structure prediction)
Umol	m6i.8xlarge (MSA) & g5.xlarge (structure prediction)

Effect of number of docking poses on PLIF recovery

In Fig. 9 we examine how including more of the top-ranked docking poses affects our PLIF recovery metric. We see that GOLD has the highest PLIF recovery regardless of the number of poses kept, but the gap narrows significantly as we include more poses from HYBRID2. FRED and DiffDock-L also demonstrate an improved PLIF recovery but the gap to GOLD remains relatively constant.

In Fig. 10 we explore on what percentage of targets the best pose by docking score is also the best pose by PLIF recovery and again we find that GOLD performs best.

We cannot show Umol or RFAA in this analysis since these methods only generate a single pose.

Correlation betweeen PLIF recovery and RMSD

In Fig. 11, we show a scatter plot of the PLIF recovery rate against the RMSD.

Examples of limitations

Figure 12 shows some limitations encountered during the analysis, either from the docking/cofolding methods (panels A, B and C) or due to the requirement for explicit hydrogen atoms to calculate PLIFs (panel D). Panels A and B illustrate, respectively, some of the

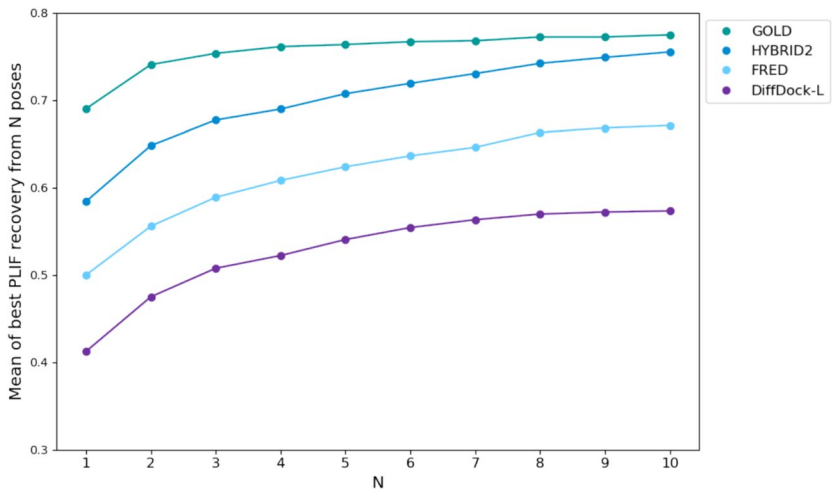

Fig. 9 Mean across targets of the best PLIF recovery from keeping *N* poses for each target

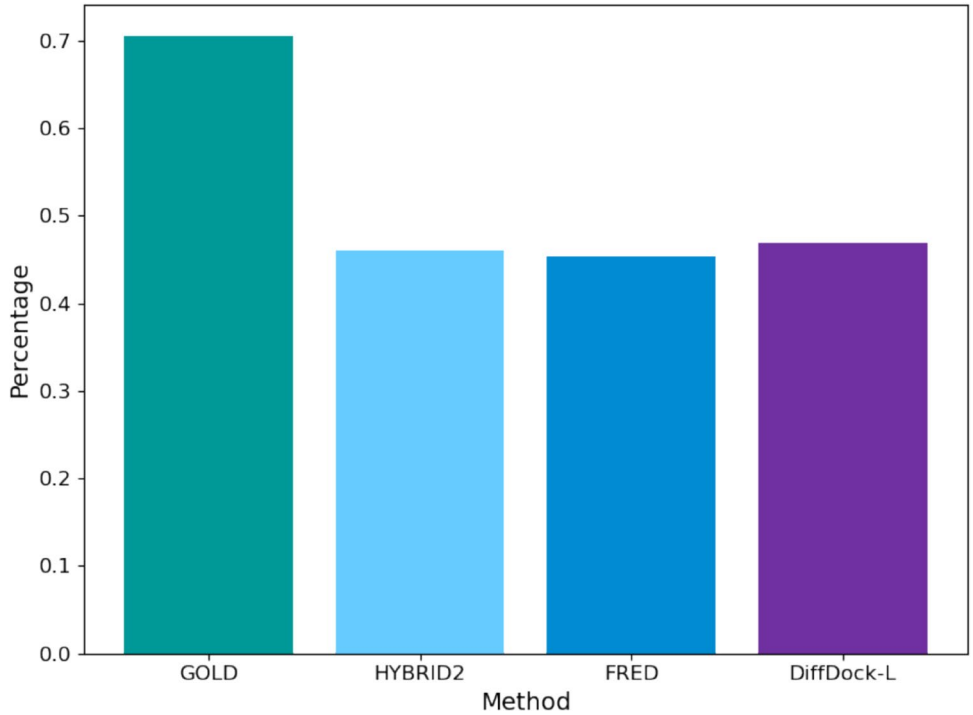

Fig. 10 Percentage of best docking poses that achieve the best PLIF recovery

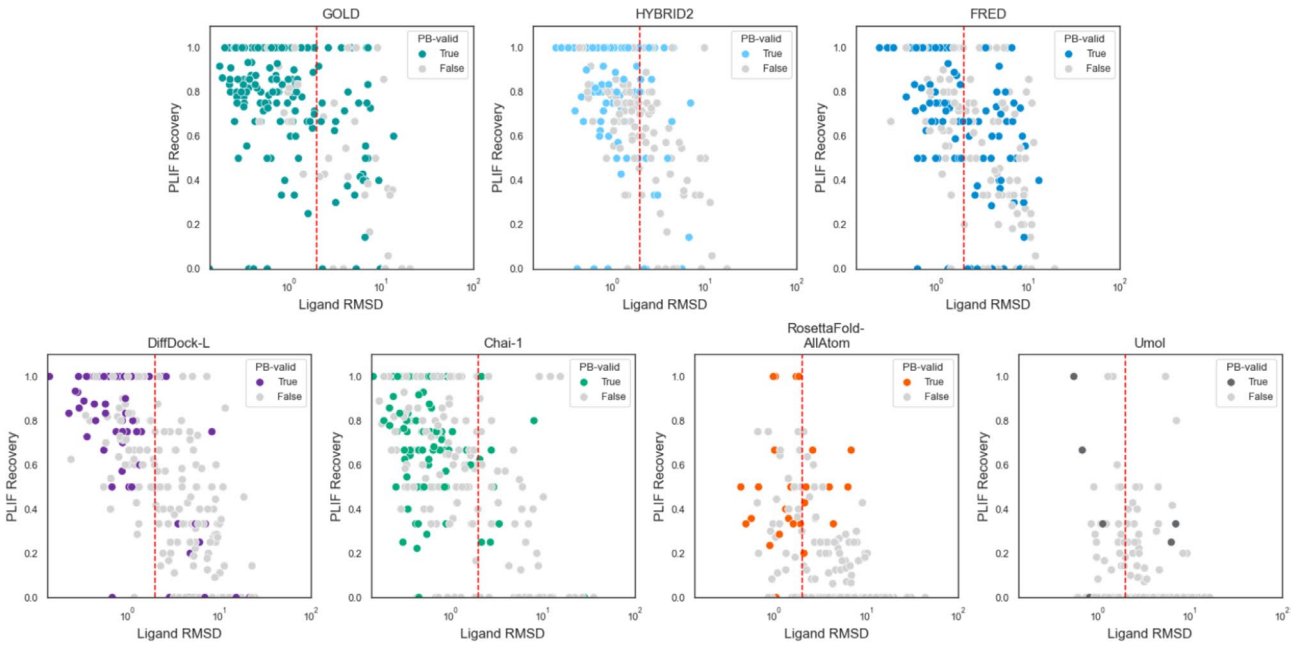

Fig. 11 PLIF recovery rate and RMSD, highlighting data points which are PoseBuster-valid. Note that we use a modified definition of PB-validity that excludes ligand RMSD. The red line indicates a ligand RMSD of 2Å

anionic and cationic outliers seen in Fig. 4 for Gold and RosettaFold-AllAtom that can result from an incorrect posing of the ligand and/or folding of residues in the binding site. Panel C shows a reasonable pose for the ligand generated by Umol, but key residues are facing the opposite direction resulting in no interactions

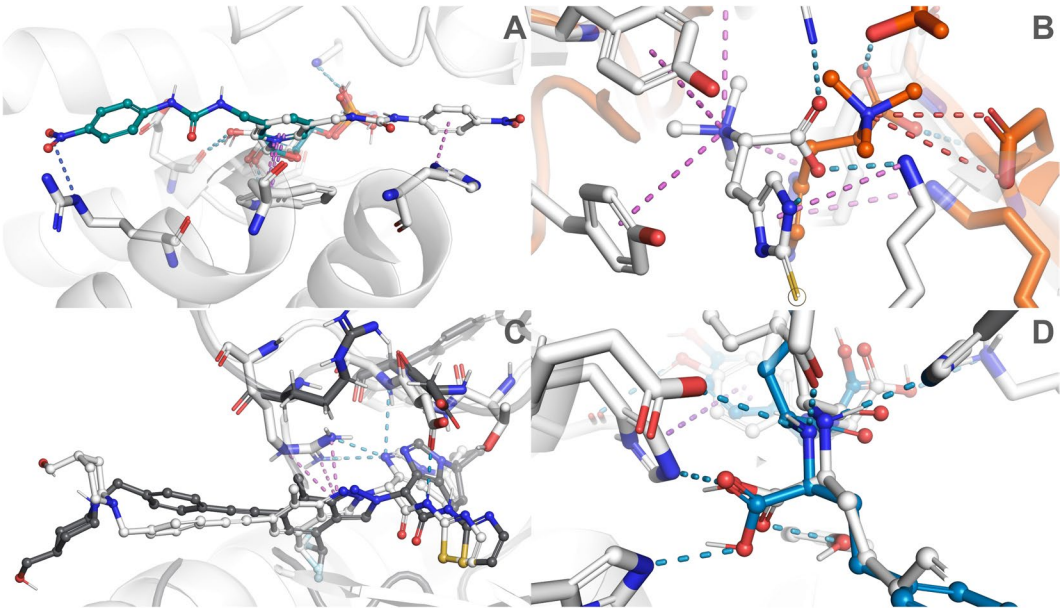

Fig. 12 Limitations of the proposed methodology as well as docking and cofolding methods. Ligands are shown in ball and stick representation, proteins in cartoon, and key residues as stick, with the crystal structure in white. **A**: In dark green, pose produced by Gold, preferring an anionic interaction (dark blue dashes) over a π-cation one (pink dashes, PDB 7M6K with ligand YRJ). **B**: In orange, cofolding results from RosettaFold-AllAtom where the conformation of the ligand and residues result in cationic interactions (red dashes) being detected, instead of cation-π interactions (pink dashes, PDB 7TXK with ligand LW8). **C**: In grey, cofolding results from Umol where despite a good overlap of the ligand with the crystal structure, the residues are facing the opposite direction resulting in the absence of interactions (PDB 8A2D with ligand KXY). **D**: In blue, HYBRID2 pose where the position of hydrogen atoms resulting from the protonation and hydrogen-bond optimisation differed and resulted in a poor interaction recovery score (PDB 7Q25 with ligand 8J9)

detected. This illustrates the additional difficulty that cofolding methods have to overcome in this study as a correct positioning of the ligand is necessary but not sufficient for recovering ground truth interactions. Panel D focuses on limitations in our current methodology, as PLIFs require explicit hydrogens to be present to evaluate hydrogen bonds, yet in some cases such as carboxylic acids, which oxygen atom gets protonated may result in different interactions being detected as seen in the bottom left. In the center of this same panel, we can see a secondary amine with a hydrogen oriented in opposite directions, facing towards a histidine for the crystal structure, or towards a glutamate for the HYBRID2 pose, resulting in different interactions detected despite the proximity of the nitrogen atom between both conformations. This showcases that our protonation and hydrogen-bond optimisation workflow can still be improved, or highlights the need for a PLIF methodology that only accounts for heavy atom positioning for the detection of interactions.

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

Author contributions
D.E. wrote the main text and performed the analysis. C.S. prepared figures and curated the dataset. C.B. wrote software and performed part of the analysis. F.D. wrote part of the main text and supervised the project. All authors read and approved the manuscript.

Data availability
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.

Declarations

Competing interests
The authors declare no competing interests.

Received: 13 October 2024 Accepted: 11 April 2025
Published online: 19 May 2025

References

1. Berman Helen M, John Westbrook, Zukang Feng, Gary Gilliland, Bhat TN, Helge Weissig, Shindyalov Ilya N, Bourne Philip E (2000) The protein data bank. Nucleic Acids Res 28(1):235–242. https://doi.org/10.1093/nar/28.1.235.
2. Liu Z, Su M, Liang H, Liu J, Yang Q, Li Y, Wang R (2017) Forging the basis for developing protein-ligand interaction scoring functions. Acc Chem Res 50:302–309. https://doi.org/10.1021/acs.accounts.6b00491

3. Buttenschoen M, Morris GM, Deane CM (2024) Posebusters: Ai-based docking methods fail to generate physically valid poses or generalise to novel sequences. Chem Sci. https://doi.org/10.1039/D3SC04185A
4. Baillif B, Cole J, McCabe P, Bender A. Benchmarking structure-based three-dimensional molecular generative models using genbench3d: ligand conformation quality matters, 2024. URL https://arxiv.org/abs/2407.04424
5. Cole Jason C, Murray Christopher W, Willem Nissink J, M, Taylor Richard D, Taylor Robin, (2005) Comparing protein-ligand docking programs is difficult. Proteins Struct, Funct Bioinform 60(3):325–332. https://doi.org/10.1002/prot.20497
6. Harris C, Didi K, Jamasb AR, Joshi CK, Mathis SV, Lio P, Blundell T. Benchmarking generated poses: How rational is structure-based drug design with generative models?, 2023. URL https://arxiv.org/abs/2308.07413
7. Morehead A, Giri N, Liu J, Cheng J. Deep learning for protein-ligand docking: are we there yet?, 2024. URL https://arxiv.org/abs/2405.14108
8. Škrinjar P, Eberhardt J, Durairaj J, Schwede T. Have protein-ligand co-folding methods moved beyond memorisation? 2025. *bioRxiv*, https://doi.org/10.1101/2025.02.03.636309. URL https://www.biorxiv.org/content/early/2025/02/10/2025.02.03.636309
9. Zhan Deng, Claudio Chuaqui, Juswinder Singh (2004) Structural interaction fingerprint (sift): a novel method for analyzing three-dimensional protein-ligand binding interactions. J Med Chem 47(2):337–344. https://doi.org/10.1021/jm030331x.
10. Marcou G, Rognan D (2007) Optimizing fragment and scaffold docking by use of molecular interaction fingerprints. J Chem Inf Model 47(1):195–207. https://doi.org/10.1021/ci600342e
11. Da Silva F, Desaphy J, Rognan D (2018) Ichem: a versatile toolkit for detecting, comparing, and predicting protein-ligand interactions. ChemMedChem 13(6):507–510. https://doi.org/10.1002/cmdc.201700505
12. Jubb Harry C, Higueruelo Alicia P, Bernardo Ochoa-Montaño, Pitt Will R, Ascher David B, Blundell Tom L (2017) Arpeggio: a web server for calculating and visualising interatomic interactions in protein structures. J Mol Biol 429(3):365–371. https://doi.org/10.1016/j.jmb.2016.12.004
13. Salentin S, Schreiber S, Haupt VJ, Adasme MF, Schroeder M (2015) PLIP: fully automated protein-ligand interaction profiler. Nucl Acids Res 43(W1):W443–W447. https://doi.org/10.1093/nar/gkv315
14. Wójcikowski M, Zielenkiewicz P, Siedlecki P (2015) Open Drug Discovery Toolkit (ODDT): a new open-source player in the drug discovery field. J Cheminform 7(1):26. https://doi.org/10.1186/s13321-015-0078-2
15. Bouysset Cédric, Fiorucci Sebastien (2021) Prolif: a library to encode molecular interactions as fingerprints. J Cheminform 13:09. https://doi.org/10.1186/s13321-021-00548-6
16. de Freitas RF, Schapira M (2017) A systematic analysis of atomic protein-ligand interactions in the pdb. RSC Med Chem 8:1970–1981. https://doi.org/10.1039/C7MD00381A
17. Jurrus E, Engel D, Star K, Monson K, Brandi J, Felberg LE, Brookes DH, Wilson L, Chen J, Liles K, Chun M, Peter Li, Gohara David W, Dolinsky T, Konecny R, Koes DR, Nielsen JE, Head-Gordon T, Geng W, Krasny R, Wei GW, Holst MJ, McCammon JA, Baker NA (2018) Improvements to the APBS biomolecular solvation software suite. Protein Sci 27(1):112–128. https://doi.org/10.1002/pro.3280
18. Landrum G, Tosco P, Kelley B, Ric, Cosgrove D, Sriniker, Vianello R, Gedeck, Nadine S, Jones G, Kawashima E, Nealschneider D, Dalke A, Cole B, Swain M, Turk S, Savelev A, Vaucher A, Wójcikowski M, Take I, Scalfani VF, Probst D, Ujihara K, Walker R, Godin G, Pahl A, Lehtivarjo J, Berenger F, strets123, jasondbiggs. rdkit/rdkit: 2023_09_6 (q3 2023) release. 2024
19. Halgren TA (1999) Mmff vi. mmff94s option for energy minimization studies. J Computat Chem 20(7):720–729
20. Tosco P, Stiefl N, Landrum G (2014) Bringing the MMFF force field to the RDKit: implementation and validation. J Cheminform 6(1):37. https://doi.org/10.1186/s13321-014-0037-3
21. Kuntz ID, Blaney JM, Oatley SJ, Langridge R, Ferrin TE (1982) A geometric approach to macromolecule-ligand interactions. J Mol Biol 161(2):269–288
22. Goodsell DS, Olson AJ (1990) Automated docking of substrates to proteins by simulated annealing. Proteins: Struct, Funct, Bioinf 8(3):195–202. https://doi.org/10.1002/prot.340080302
23. OpenEye. *OEDOCKING 4.3.0.3. Cadence Molecular Sciences, Inc., Santa Fe, NM. http://www.eyesopen.com*. a
24. McGann M (2012) Fred and hybrid docking performance on standardized datasets. J Comput Aided Mol Des 26:897–906. https://doi.org/10.1007/s10822-012-9584-8
25. Jones G, Willett P, Glen RC, Leach AR, Taylor R (1997) Development and validation of a genetic algorithm for flexible docking11edited by f. e. cohen. J Mol Biol 267(3):727–748
26. Verkhivker GM, Bouzida D, Gehlhaar DK, Rejto PA, Arthurs S, Colson AB, Freer ST, Larson V, Luty BA, Tami Marrone PW, Rose. (2000) Deciphering common failures in molecular docking of ligand-protein complexes. J Comput Aided Mol Des 14(8):731–751. https://doi.org/10.1023/a:1008158231558
27. OpenEye. *Spruce 1.6.0.0. OpenEye, Cadence Molecular Sciences, Santa Fe, NM. http://www.eyesopen.com*. b
28. Word JM, Lovell SC, Richardson JS, Richardson DC (1999) Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation 1 1edited by j. thornton. J Mol Biol 285(4):1735–1747. https://doi.org/10.1006/jmbi.1998.2401
29. Groom CR, Bruno IJ, Lightfoot MP, Ward SC (2016) The Cambridge structural database. Acta Crystallogr Sect B Struct Sci, Cryst Eng Mater 72(2):171–179. https://doi.org/10.1107/S2052520616003954
30. Schrödinger LLC. 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. Schrödinger, LLC, New York, NY. 2024
31. Lu W, Wu Q, Zhang J, Rao J, Li C, Zheng S (2022) Tankbind: trigonometry-aware neural networks for drug-protein binding structure prediction. Adv Neural Inf Process Syst 35:7236
32. Stärk H, Ganea O, Pattanaik L, Barzilay R, Jaakkola T. Equibind: Geometric deep learning for drug binding structure prediction. In *International Conference on Machine Learning*, PMLR. 2022. p 20503–20521
33. Zhou G, Gao Z, Ding Q, Zheng H, Xu H, Wei Z, Zhang L, Ke G. 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
34. Corso G, Deng A, Fry B, Polizzi N, Barzilay R, Jaakkola T. Strategies for docking generalization, Deep confident steps to new pockets. 2024
35. Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, Ronneberger O, Willmore L, Ballard AJ, Bambrick J, Bodenstein SW, Evans DA, Hung CC, O'Neill M, Reiman D, Tunyasuvunakool K, Wu Z, Zemgulyte A, Arvaniti E, Beattie C, Bertolli O, Bridgland A, Cherepanov A, Congreve M, Cowen-Rivers AI, Cowie A, Figurnov M, Fuchs FB, Gladman H, Jain R, Khan YA, Low CMR, Perlin K, Potapenko A, Savy P, Singh S, Stecula A, Thillaisundaram A, Tong C, Yakneen S, Zhong ED, Zielinski M, Zidek A, Bapst V, Kohli P, Jaderberg M, Hassabis D, Jumper JM (2024) Accurate structure prediction of biomolecular interactions with alphafold 3. Nature. https://doi.org/10.1038/s41586-024-07487-w
36. Bryant P, Kelkar A, Guljas A, Clementi C, Noé F (2024) Structure prediction of protein-ligand complexes from sequence information with umol. Nat Commun. https://doi.org/10.1038/s41467-024-48837-6
37. Chai D, Boitreaud J, Dent J, McPartlon M, Meier J, Reis V, Rogozhonikov A, Wu K. Chai-1: Decoding the molecular interactions of life. 2024. *bioRxiv*, https://doi.org/10.1101/2024.10.10.615955. URL https://www.biorxiv.org/content/early/2024/10/11/2024.10.10.615955
38. Krishna R, Wang J, Ahern W, Sturmfels P, Venkatesh P, Kalvet I, Lee GR, Morey-Burrows FS, Anishchenko I, Humphreys IR, McHugh R, Vafeados D, Li X, Sutherland GA, Hitchcock A, Hunter CN, Kang A, Brackenbrough E, Bera AK, Baek M, DiMaio F, Baker D (2024) Generalized biomolecular modeling and design with rosettafold all-atom. Science. https://doi.org/10.1126/science.adl2528
39. Qiao Z, Nie W, Vahdat A, Miller TF, Anandkumar A (2024) State-specific protein-ligand complex structure prediction with a multiscale deep generative model. Nat Mach Intell 6(2):195–208. https://doi.org/10.1038/s42256-024-00792-z
40. Wohlwend J, Corso G, Passaro S, Reveiz M, Leidal K, Swiderski W, Portnoi T, Chinn I, Silterra J, Jaakkola T, Barzilay R. Boltz-1 democratizing biomolecular interaction modeling. *bioRxiv*, https://doi.org/10.1101/2024.11.19.624167. 2024. URL https://www.biorxiv.org/content/early/2024/11/20/2024.11.19.624167
41. Eastman P, Swails J, Chodera JD, McGibbon RT, Zhao Y, Beauchamp KA, Wang LP, Simmonett AC, Harrigan MP, Stern CD, Wiewiora RP, Brooks BR, Pande VS (2017) Openmm 7: rapid development of high performance

algorithms for molecular dynamics. PLoS Comput Biol 13(7):e1005659. https://doi.org/10.1371/journal.pcbi.1005659
42. Krieger E, Vriend G (2015) New ways to boost molecular dynamics simulations. J Comput Chem 36(13):996–1007. https://doi.org/10.1002/jcc.23899
43. Rich A, Birnbaum B, Haimson J. Approaching AlphaFold 3 docking accuracy in 100 lines of code. 2024. https://www.inductive.bio/blog/strong-baseline-for-alphafold-3-docking, Accessed 07 Aug 2024
44. Caterina Bissantz, Bernd Kuhn, Martin Stahl (2010) A medicinal chemist's guide to molecular interactions. J Med Chem 53(14):5061–5084. https://doi.org/10.1021/jm100112j.
45. Bauer Matthias R, Mackey Mark D (2019) Electrostatic complementarity as a fast and effective tool to optimize binding and selectivity of protein-ligand complexes. J Med Chem 62(6):3036–3050. https://doi.org/10.1021/acs.jmedchem.8b01925.
46. Masters MR, Mahmoud AH, Lill MA. Do deep learning models for co-folding learn the physics of protein-ligand interactions? 2024. *bioRxiv*, https://doi.org/10.1101/2024.06.03.597219. URL https://www.biorxiv.org/content/early/2024/06/04/2024.06.03.597219
47. Meller A, Bhakat S, Solieva S, Bowman GR (2023) Accelerating cryptic pocket discovery using alphafold. J Chem Theory Comput 19(14):4355–4363. https://doi.org/10.1021/acs.jctc.2c01189
48. Meller A, Ward M, Borowsky J, Kshirsagar M, Lotthammer JM, Oviedo F, Ferres JL, Bowman GR (2023) Predicting locations of cryptic pockets from single protein structures using the pocketminer graph neural network. Nat Commun 14(1):1177. https://doi.org/10.1038/s41467-023-36699-3
49. Oleinikovas V, Saladino G, Cossins BP, Gervasio FL (2016) Understanding cryptic pocket formation in protein targets by enhanced sampling simulations. J Am Chem Soc 138(43):14257–14263. https://doi.org/10.1021/jacs.6b05425.
50. Robin X, Studer G, Durairaj J, Eberhardt J, Schwede T, Patrick W (2023) Assessment of protein-ligand complexes in casp15. Proteins Struct, Funct Bioinform 91(12):1811–1821. https://doi.org/10.1002/prot.26601

Publisher's Note