Deep Learning for Protein-Ligand Docking: Are We There Yet?

Alex Morehead [1] **Nabin Giri** [1] **Jian Liu** [1] **Jianlin Cheng** [1]

Abstract

The effects of ligand binding on protein structures and their *in vivo* functions carry numerous implications for modern biomedical research and biotechnology development efforts such as drug discovery. Although several deep learning (DL) methods and benchmarks designed for protein-ligand docking have recently been introduced, to date no prior works have systematically studied the behavior of docking methods within the *practical* context of (1) using predicted (apo) protein structures for docking (e.g., for broad applicability); (2) docking multiple ligands concurrently to a given target protein (e.g., for enzyme design); and (3) having no prior knowledge of binding pockets (e.g., for pocket generalization). To enable a deeper understanding of docking methods' real-world utility, we introduce POSEBENCH, the first comprehensive benchmark for *practical* protein-ligand docking. POSEBENCH enables researchers to rigorously and systematically evaluate DL docking methods for apo-to-holo protein-ligand docking and protein-ligand structure generation using *both* single and multi-ligand benchmark datasets, the latter of which we introduce for the first time to the DL community. Empirically, using POSEBENCH, we find that all recent DL docking methods but one fail to generalize to multi-ligand protein targets and also that template-based docking algorithms perform equally well or better for multi-ligand docking as recent single-ligand DL docking methods, suggesting areas of improvement for future work. Code, data, tutorials, and benchmark results are available at https://github.com/BioinfoMachineLearning/PoseBench.

[1]Department of Electrical Engineering & Computer Science, NextGen Precision Health, University of Missouri-Columbia, Columbia, Missouri, USA. Correspondence to: Alex Morehead <acmwhb@umsystem.edu>.

The 2024 ICML AI for Science Workshop. Vienna, Austria, 2024.
Copyright 2024 by the author(s).

1. Introduction

The field of drug discovery has long been challenged with a critical task: determining the structure of ligand molecules in complex with proteins and other key macromolecules (Warren et al., 2012). As accurately identifying such complex structures (in particular multi-ligand structures) can yield advanced insights into the binding dynamics and functional characteristics (and thereby, the medicinal potential) of numerous protein complexes *in vivo*, in recent years, significant resources have been spent developing new experimental and computational techniques for protein-ligand structure determination (Du et al., 2016). Over the last decade, machine learning (ML) methods for structure prediction have become indispensable components of modern structure determination at scale, with AlphaFold 2 for protein structure prediction being a recent hallmark example (Jumper et al., 2021).

As the field has gradually begun to investigate whether proteins in complex with other types of molecules can faithfully be modeled with ML (and particularly deep learning (DL)) techniques (Dhakal et al., 2022; Harris et al., 2023; Krishna et al., 2024), several new works in this direction have suggested the promising potential of such approaches to protein-ligand structure determination (Corso et al., 2022; Lu et al., 2024; Qiao et al., 2024; Abramson et al., 2024). Nonetheless, to date, it remains to be shown whether such DL methods can adequately generalize in the context of *apo* (i.e., unbound) protein structures and multiple interacting ligand molecules (e.g., which can alter the chemical functions of various enzymes) as well as whether such methods are more accurate than traditional techniques for protein-ligand structure determination (for brevity hereafter referred to interchangeably as structure generation or docking) such as template-based (Pang et al., 2023) or molecular docking software tools (Xu et al., 2023).

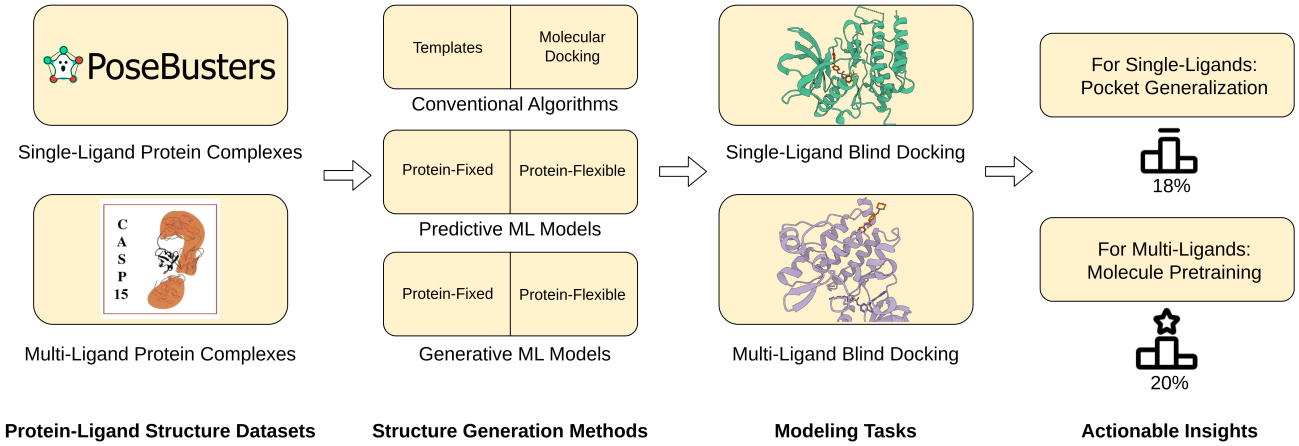


Figure 1. Overview of POSEBENCH, our comprehensive benchmark for *practical* ML modeling of single and multi-ligand protein complex structures in the context of apo (predicted) protein structures without known binding pockets (i.e., blind docking).

To bridge this knowledge gap, our contributions in this work are as follows:

- We introduce the first unified benchmark for protein-ligand structure generation that evaluates the performance of both recent DL-based methods as well as conventional methods for single and *multi*-ligand docking.
- In contrast to several recent works on protein-ligand docking (Buttenschoen et al., 2024; Corso et al., 2024a), the benchmark results we present in this work are all within the context of *apo* (i.e., predicted) protein structures without known binding pockets, which notably enhances the practicality and real-world utility of this study's findings.
- Our newly proposed benchmark, POSEBENCH, enables specific insights into necessary areas of future work for accurate and generalizable protein-ligand structure generation, including that molecule pretraining seems to be *key* to generalizing to multi-ligand docking targets.
- Our benchmark's results also show that template-based algorithms for protein-ligand structure generation surpass the multi-ligand docking performance of several recent DL methods for protein-ligand docking, which suggests the importance of *directly* training and evaluating future DL methods on multi-ligand targets.

2. Related work

Structure prediction of protein-ligand complexes. The field of DL-driven protein-ligand structure determination was largely sparked with the development of geometric deep learning methods such as EquiBind (Stärk et al., 2022) and TANKBind (Lu et al., 2022) for direct (i.e., regression-based) prediction of bound ligand structures in protein complexes. Notably, these predictive methods could estimate localized ligand structures in complex with multiple protein chains as well as the associated complexes' binding affinities. However, in addition to their limited predictive accuracy, they have more recently been found to frequently produce steric clashes between protein and ligand atoms, notably hindering their widespread adoption in modern drug discovery pipelines.

Protein-ligand structure generation and docking. Shortly following the first wave of predictive methods for protein-ligand structure determination, DL methods such as DiffDock (Corso et al., 2022) demonstrated the utility of a new approach to this problem by reframing protein-ligand docking as a generative modeling task, whereby multiple ligand conformations can be generated for a particular protein target and rank-ordered using a predicted confidence score. This approach has inspired many follow-up works offering alternative formulations of this generative approach to the problem (Lu et al., 2024; Plainer et al., 2023; Zhu et al., 2024), with some of such follow-up works also being capable of accurately modeling protein flexibility upon ligand binding or predicting binding affinities to a high degree of accuracy.

Benchmarking efforts for protein-ligand complexes. In response to the large number of new methods that have been developed for protein-ligand structure generation, recent works have introduced several new datasets and metrics with which to evaluate newly developed methods, with some of such benchmarking efforts focusing on modeling single-ligand protein interactions (Buttenschoen et al., 2024) and with others specializing in the assessment of multi-ligand protein interactions (Robin et al., 2023). One of the primary

Table 1. POSEBENCH evaluation datasets for protein-(multi-)ligand structure generation.

Name	Type	Source	Size (Total # Ligands)
Astex Diverse	Single-Ligand	(Hartshorn et al., 2007)	80
PoseBusters Benchmark	Single-Ligand	(Buttenschoen et al., 2024)	280
DockGen	Single-Ligand	(Corso et al., 2024a)	91
CASP15	Multi-Ligand		102 (across 19 complexes)
			→ 6 (13) single (multi)-ligand complexes

aims of this work is to bridge this gap by systematically assessing a selection of the latest (pocket-blind) structure generation methods within both interaction regimes in the context of unbound protein structures and *ab initio* complex structure prediction, efforts we describe in greater detail in the following section.

3. POSEBENCH

The overall goal of POSEBENCH, our newly proposed benchmark for protein-ligand structure generation, is to provide the ML research community with a centralized resource with which one can systematically measure, in a variety of macromolecular contexts, the methodological advancements of new DL methods proposed for this problem. In the remaining sections, we describe POSEBENCH's design and composition (as illustrated in Figure 1), how we have used POSEBENCH to evaluate several recent DL methods (as well as conventional algorithms) for protein-ligand structure modeling, and what actionable insights we can derive from POSEBENCH's benchmark results with these latest DL methods.

3.1. Preprocessed datasets

POSEBENCH provides users with four datasets with which to evaluate existing or new protein-ligand structure generation methods, the Astex Diverse and PoseBusters Benchmark (DockGen) datasets previously curated by Buttenschoen et al. (2024) ((Corso et al., 2024a)) as well as the CASP15 protein-ligand interaction (PLI) dataset that we have manually curated in this work.

Astex Diverse dataset. The Astex Diverse dataset (Hartshorn et al., 2007) is a collection of 85 protein-ligand complexes composed of various drug-like molecules known to be of pharmaceutical or agrochemical interest, where a single representative ligand is present in each complex. This dataset can be considered an easy benchmarking set for many DL-based docking methods in that several of its proteins are known to overlap with the commonly used PDBBind (time-split) training dataset. Nonetheless, including this dataset for benchmarking allows one to determine the performance "upper bound" of each method's docking capabilities for single-ligand protein complexes.

To perform *apo* docking with this dataset, we used ESMFold (Lin et al., 2023) to predict the complex structure of each of its proteins, where 5 of these 85 complexes were excluded from the effective benchmarking set due to being too large for structure prediction on an 80GB NVIDIA A100 GPU. For the remaining 80 complexes, we then optimally aligned their predicted protein structures to the corresponding ground-truth (holo) protein-ligand structures using the PLI-weighted root mean square deviation (RMSD) alignment algorithm originally proposed by Corso et al. (2022).

PoseBusters Benchmark dataset. The PoseBusters Benchmark dataset (Buttenschoen et al., 2024) contains 308 recent protein-ligand complexes released from 2021 onwards. Like the Astex Diverse set, each complex in this dataset contains a single ligand for prediction. In contrast to Astex Diverse, this dataset can be considered a harder benchmark set since its proteins do not directly overlap with the commonly used PDBBind (time-split) training dataset composed of protein-ligand complexes with release dates up to 2019.

Likewise to Astex Diverse, for the PoseBusters Benchmark set, we used ESMFold to predict the *apo* complex structures of each of its proteins. After filtering out 28 complexes for which the corresponding protein structure could not be predicted on an 80GB A100 GPU, we RMSD-aligned the remaining 280 predicted protein structures while optimally weighting each complex's protein-ligand interface in the alignment. For the **DockGen** dataset, we refer readers to Appendix G.1.

CASP15 dataset. To assess the multi-ligand modeling capabilities of recent methods for protein-ligand structure generation, in this work, we introduce a curated version of the CASP15 PLI dataset introduced as a first-of-its-kind prediction category in the 15th Critical Assessment of Structure Prediction (CASP) competition (Robin et al., 2023) held in 2022. The CASP15 PLI set is originally comprised of 23 protein-ligand complexes, where we subsequently filter out 4 complexes based on (1) whether the CASP organizers ultimately assessed predictions for the complexes; (2) whether they are RNA-ligand complexes with no interacting protein chains; or (3) whether we could obtain a reasonably accurate prediction of the complex's multimeric protein chains using

either ESMFold or AlphaFold-based structure prediction on an 80GB A100 GPU (selecting for each complex the prediction which yielded the lowest-RMSD protein complex structure). Following this initial filtering step, we optimally align each remaining complex's predicted protein structures to the corresponding ground-truth protein-(multi-)ligand structures, weighting *each* of the complex's protein-ligand binding sites in the structural alignment.

The 19 remaining protein-ligand complexes, which contain a total of 102 (fragment) ligands, consist of a variety of ligand types including single-atom (metal) ions and large drug-sized molecules with up to 92 atoms in each (fragment) ligand. As such, this dataset is appropriate for assessing how well structure generation methods can model interactions between different (fragment) ligands in the same complex, which can yield insights into the (protein-ligand and ligand-ligand) steric clash rates of each method.

Sequence identity overlap. Note that for all four of the test datasets described above and listed in Table 1, we do not perform an analysis of the sequence identity overlap between these test datasets' proteins and those of e.g., the PDBBind (time-split) training dataset, as (1) not all DL-based docking methods use PDBBind as their respective training datasets and (2) leaving the test complexes unfiltered according to sequence identity should, in principle, reflect many real-world use cases of these methods in which several (new) protein targets they are presented with may or may not be similar to what the methods have "seen" during training. Nevertheless, for an investigation of the sequence identity overlap between e.g., the PoseBusters Benchmark set and PDBBind, we refer interested readers to Buttenschoen et al. (2024). Furthermore, in Appendix F, we analyze the different types and frequencies of protein-ligand interactions natively found within the Astex Diverse, PoseBusters Benchmark, DockGen, and CASP15 datasets, respectively, to quantify the diversity of the (predicted) interactions each dataset can be used to evaluate.

3.2. Formulated tasks

In this work, we have developed POSEBENCH to focus our analysis on the behavior of different DL methods for protein-ligand docking in a variety of macromolecular contexts (e.g., with or without inorganic cofactors present). With this goal in mind, below we formulate the structure generation tasks currently available in POSEBENCH.

Single-ligand blind docking. For single-ligand blind docking, each benchmark method is provided with a (multi-chain) protein sequence and an optional *apo* (predicted) protein structure as input along with a corresponding ligand SMILES string for each complex. In particular, no knowledge of the complex's protein-ligand binding pocket is provided to evaluate how well each method can (1) identify the correct binding pockets and (2) propose the correct ligand conformation within each predicted pocket.

Multi-ligand blind docking. For multi-ligand blind docking, each benchmark method is provided with a (multi-chain) protein sequence and an optional *apo* (predicted) protein structure as input along with the corresponding (fragment) ligand SMILES strings. As in single-ligand blind docking, no knowledge of the protein-ligand binding pocket is provided, which offers the opportunity to not only evaluate binding pocket and conformation prediction precision but also multimeric steric clash rates.

4. Methods and experimental setup

Overview. Our benchmark is designed to explore answers to specific modeling questions for protein-ligand docking such as (1) which types of methods are best able to identify the correct binding pocket(s) in target proteins and (2) which types of methods most accurately produce multi-ligand structures without steric clashes? In the following sections, we describe in detail which types of methods we evaluate in our benchmark, what the input and output formats look like for each method, and how we evaluate each method's predictions for particular protein complex targets.

Method categories. As illustrated in Figure 1, we divide the benchmark methods included in POSEBENCH into one of three categories: (1) conventional algorithms, (2) predictive (i.e., regression-based) ML algorithms, and (3) generative (i.e., distributional) ML algorithms.

As representative algorithms for conventional protein-ligand docking, we include AutoDock Vina (v1.2.5) (Trott & Olson, 2010) as well as a template-based modeling method for accurate ligand-protein complex structure prediction (TULIP) that we introduce in this work. To represent predictive ML docking algorithms, we include FABind (Pei et al., 2024) as well as the recently released version of RoseTTAFold 2 for all-atom structural modeling (i.e., RoseTTAFold-All-Atom) (Krishna et al., 2024). Lastly, for generative ML docking algorithms, we include DynamicBind (Lu et al., 2024), NeuralPLexer (Qiao et al., 2024), and the latest version of DiffDock referred to as DiffDock-L (Corso et al., 2024a) which is designed with binding site generalization as a key aim. Notably, AlphaFold 3 (Abramson et al., 2024) does not support *generic* SMILES string inputs, so we cannot benchmark it.

Additionally, we provide a method ensembling baseline (Ensemble) that uses (multi-)ligand structural consensus ranking (Con) (Roy et al., 2023) to rank its ligand structure predictions selected from the (intrinsically method-ranked) top-40 ligand conformations produced by each conventional, predictive, and generative ML method. This ensembling baseline is included to answer the question, "Which method

produces the most consistent conformations in interaction with a protein complex?".

Input and output formats.

1. Formats for conventional methods are as follows:
   (a) Template-based methods such as **TULIP** are provided with an *apo* (predicted) protein structure and (fragment) ligand SMILES strings and are tasked with retrieving (PDB template (Bank, 1971)) ligand conformations residing in the same coordinate system as the given (predicted) protein structure following optimal molecular and structural alignment (Hu et al., 2018) with corresponding RDKit conformers of the input (query) ligand SMILES strings, where structural similarity with the query ligands is used to rank-order the selected (PDB template) conformations.
   (b) Molecular docking tools such as **AutoDock Vina**, which require specification of protein binding sites, are provided with not only a predicted protein structure but also the centroid coordinates of each (DiffDock-L-)predicted protein-ligand binding site residue. Such binding site residues are classified using a 4 Å protein-ligand heavy atom interaction threshold and using a 25 Å ligand-ligand heavy atom interaction threshold to define a "group" of ligands belonging to the same binding site and therefore residing in the same 25 $Å^3$-sized binding site input voxel for AutoDock Vina. For interested readers, in Appendix G.1, we additionally report results using P2Rank (Krivák & Hoksza, 2018) to predict AutoDock Vina's binding site centroid inputs.
2. Formats for predictive methods are as follows:
   (a) **FABind** is provided with a predicted protein structure as well as a ligand SMILES string, and it is then tasked with producing a (single) ligand conformation in complex with the given protein.
   (b) **RoseTTAFold-All-Atom** is provided with a (multi-chain) protein sequence as well as (fragment) ligand SMILES strings, and it is subsequently tasked with producing not only a (single) bound ligand conformation but also the bound protein conformation (as a representative *ab initio* structure generation method).
3. Formats for generative methods are as follows:
   (a) **DiffDock-L** is provided with a predicted protein structure and (fragment) ligand SMILES strings and is then tasked with producing (multiple rank-ordered) ligand conformations (for each fragment) for the given protein. Note that DiffDock-L does not natively support multi-ligand SMILES string inputs, so in this work, we propose a modified inference procedure for DiffDock-L which *autoregressively* presents each (fragment) ligand SMILES string to the model while providing the same predicted protein structure to the model in each inference iteration (reporting for each complex the average confidence score over all iterations). Notably, as an inference-time modification, this sampling formulation permits multi-ligand sampling yet cannot model multi-ligand interactions directly and therefore often produces ligand-ligand steric clashes.
   (b) As a single-ligand generative docking method, **DynamicBind** adopts the same input and output formats as DiffDock-L with the following exceptions: (1) the predicted input protein structure can be modified in response to (fragment) ligand docking; (2) the autoregressive inference procedure we adapted from that of DiffDock-L now provides DynamicBind with its own most recently generated protein structure in each (fragment) ligand inference iteration, thereby providing the model with partial multi-ligand interaction context; and (3) iteration-averaged confidence scores *and* predicted affinities are reported for each complex. Nonetheless, for both DiffDock-L and DynamicBind, such modified inference procedures highlight the importance in future work of retraining such generative methods directly on multi-ligand complexes to address such inference-time compromises.
   (c) Lastly, as a natively multi-ligand structure generation model pretrained using various 3D molecular and protein data sources, **NeuralPLexer** receives as its inputs a (multi-chain) protein sequence, a predicted protein (template) structure, as well as (fragment) ligand SMILES strings. The method is then tasked with producing (multiple rank-ordered) protein-ligand structure conformations for each input complex, using the method's average predicted per-ligand heavy atom local Distance Difference Test (lDDT) score (Mariani et al., 2013) for rank-ordering.

Prediction and evaluation procedures. Using the prediction formats above, the protein-ligand complex structures each method produces are subsequently evaluated using various structural accuracy and molecule validity metrics depending on whether the targets are single or multi-ligand complexes.

Single-ligand evaluation. For single-ligand targets, we

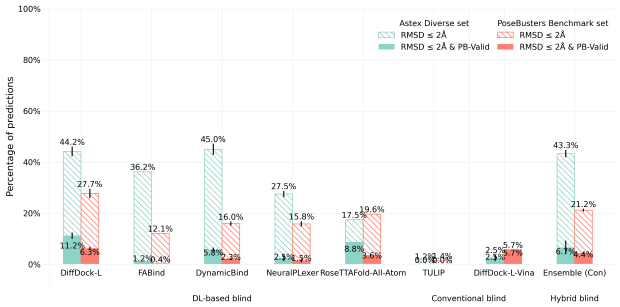


Figure 2. Astex & PoseBusters dataset results for successful single-ligand docking.

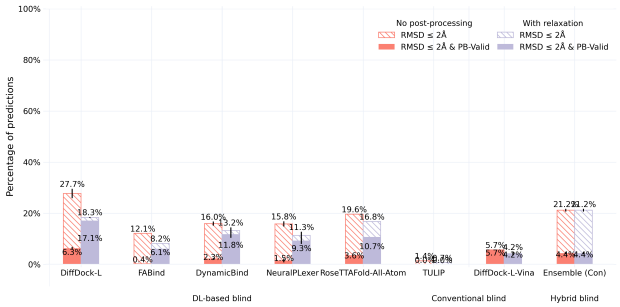


Figure 3. PoseBusters dataset results for successful single-ligand docking with relaxation.

report each method's percentage of (top-1) ligand conformations within 2 Å of the corresponding ground-truth ligand structure (RMSD $\leq$ 2 Å) as well as the percentage of such "correct" ligand conformations that are also considered to be chemically and structurally valid according to the PoseBusters software suite (Buttenschoen et al., 2024) (RMSD $\leq$ 2 Å & PB-Valid).

Multi-ligand evaluation. Following CASP15's official scoring procedure for protein-ligand complexes (Robin et al., 2023), for multi-ligand targets, we report each method's percentage of "correct" (binding site-superimposed) ligand conformations (RMSD $\leq$ 2 Å) as well as violin plots of the RMSD and PLI-specific lDDT scores of its protein-ligand conformations across all (fragment) ligands within the benchmark's multi-ligand complexes (see Appendix G for these plots). Notably, this final metric, referred to lDDT-PLI, allows one to evaluate specifically how well each method can model protein-ligand structural interfaces. We refer readings to Appendix D for formal definitions of these metrics. In the remainder of this work, we will discuss our benchmark's results and their implications for the development of future complex structure generation methods.

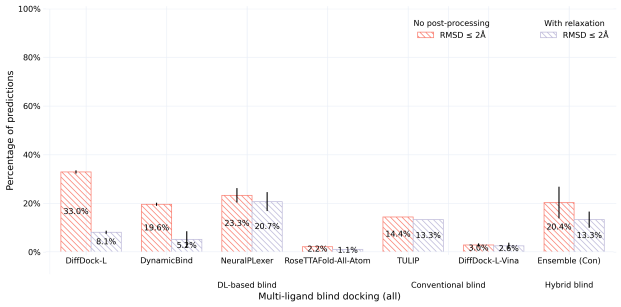


Figure 4. CASP15 dataset results for successful multi-ligand docking with relaxation.

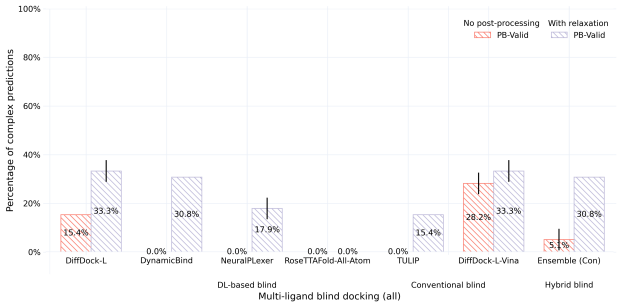


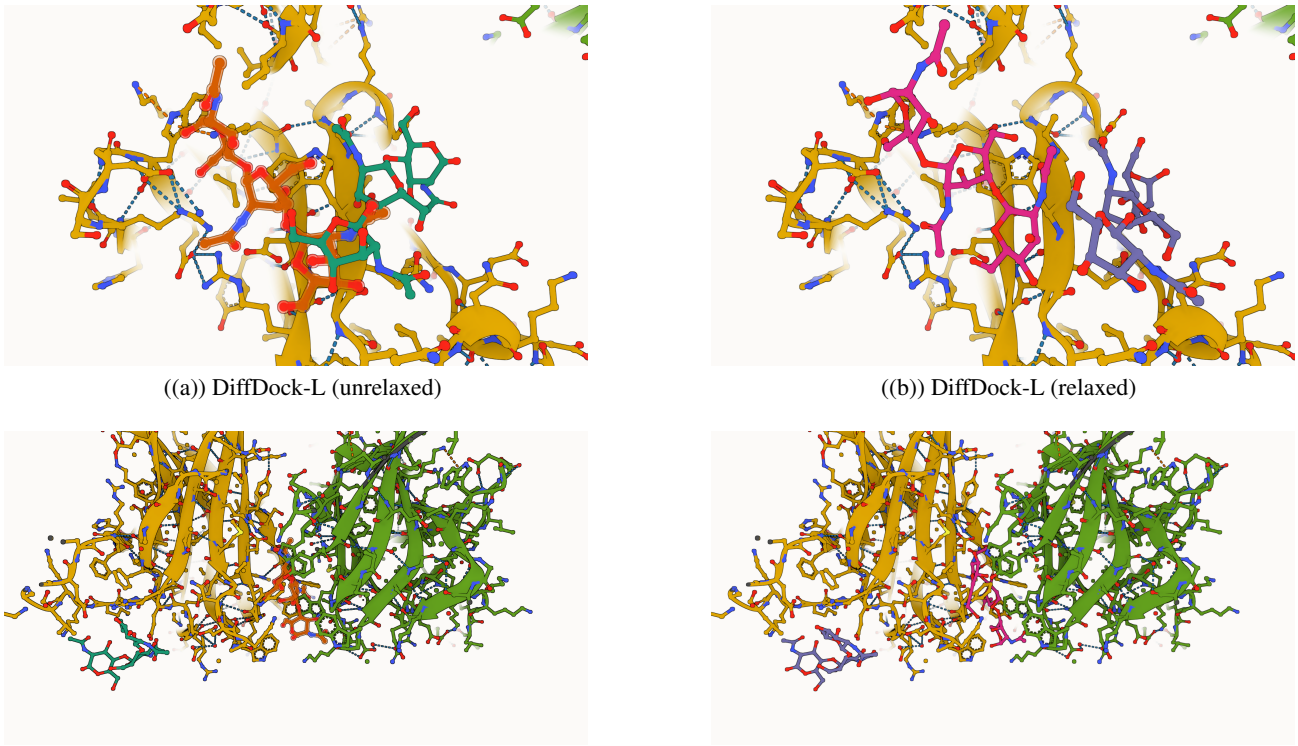
Figure 5. CASP15 dataset results for multi-ligand PoseBusters validity rates with relaxation.

5. Results and discussions

In this section, we present POSEBENCH's results for single and multi-ligand protein-ligand structure generation and discuss their implications for future work. Note that across all the experiments, for generative methods (or methods that use generative inputs to make their predictions), we report their performance metrics in terms of the mean and standard deviation across *three* independent runs of the method to gain insights into its inter-run stability and consistency. For interested readers, in Appendix C, we report the average runtime and memory usage of each baseline method to determine which methods are the most practical for real-world docking applications.

5.1. Generalization to new binding pockets implies single-ligand docking performance

We begin our investigations by evaluating the performance of each baseline method for single-ligand docking using the Astex Diverse and PoseBusters Benchmark datasets. Notably, for results on the PoseBusters Benchmark dataset, we perform an additional analysis where we apply post-prediction (fixed-protein) relaxation to each method's generated ligand conformations using molecular dynamics simulations (Eastman & Pande, 2010), as originally proposed by Buttenschoen et al. (2024). Additionally, for inter-



((a)) DiffDock-L (unrelaxed)

((b)) DiffDock-L (relaxed)

((c)) NeuralPLexer (unrelaxed)

((d)) NeuralPLexer (relaxed)

Figure 6. DiffDock-L and NeuralPLexer multi-ligand predictions for CASP15 target T1187.

ested readers, in Appendix G.1 we include benchmark results for flexible-protein relaxation as implemented by Lu et al. (2024). As shown in Figures 2 and 3, DiffDock-L achieves the best overall performance across the two datasets both with and without applying relaxation to its generated structures. Closely behind in performance for the PoseBusters Benchmark dataset are DynamicBind and RoseTTAFold-All-Atom following structural relaxation. Interestingly, without relaxation, AutoDock Vina combined with DiffDock-L's predicted binding pockets achieves the second-best performance on the PoseBusters Benchmark dataset, which suggests that DiffDock-L is currently the *only* single-ligand deep learning method that presents a better intrinsic understanding of biomolecular physics for docking than conventional modeling tools. For interested readers, in Appendix G.2, we report e.g., pocket-only PoseBusters Benchmark experiments and RMSD violin plots for both the Astex Diverse and PoseBusters Benchmark datasets.

5.2. Molecule pretraining implies multi-ligand docking performance

We now turn to investigating the performance of various deep learning and conventional methods for *multi*-ligand docking. In Figure 4, we see that although DiffDock-L initially appears to achieve the best performance in this context, after applying structural relaxation its performance quickly diminishes. This trend holds for similar deep learning methods such as DynamicBind that were specifically trained on single-ligand protein complexes, as achieving a low RMSD for each (fragment) ligand does not rule out the existence of protein-ligand and particularly ligand-ligand steric clashes. In contrast to this trend, however, NeuralPLexer does not lose a significant fraction of accurate (fragment) ligand predictions following structural relaxation, which suggests that its multi-ligand conformations are already largely free of steric clashes. Figure 6 illustrates these steric clash trends using top-1 predictions from DiffDock-L and NeuralPLexer for CASP15 target T1187 as a case study.

Another interesting observation is that TULIP (following structural relaxation) outperforms the docking success rates of single-ligand deep learning docking methods such as DiffDock-L and DynamicBind in the context of multi-ligand docking, which suggests room for future improvement in the multi-ligand modeling capabilities of these recent deep learning baselines. To further inspect each method's understanding of biomolecular physics for docking, in Figure 5 we report each method's percentage of predicted complexes (whether correct or not) for which all ligand conformations in the complex are jointly considered valid according to the PoseBusters software suite (i.e., PB-Valid). In short, in

the context of multi-ligands, we find that AutoDock Vina followed by DiffDock-L are tied in terms of their PoseBusters validity rates following structural relaxation, with our ensembling consensus baseline (i.e., Ensemble (Con)) as well as DynamicBind shortly behind. As an addendum, we note that NeuralPLexer's (DynamicBind's) predictions seem to be frequently selected by Ensemble (Con) for single (multi)-ligand targets, which suggests that NeuralPLexer (DynamicBind) produces the most consistent (i.e., similar) ligand poses for a given single (multi)-ligand protein complex. For interested readers, in Appendix G.3, we report additional results e.g., in terms of lDDT-PLI and RMSD violin plots for both the total available CASP15 targets as well as the publicly available ones.

Conclusions

contrast In this work, we introduced POSEBENCH, the first deep learning benchmark for *practical* protein-ligand docking. Experimental results with POSEBENCH suggest the importance of developing new multi-ligand structure generation methods for enhanced generalization in future work. Moreover, based on these benchmark results, we posit that advances in protein-ligand docking will also likely be driven by advances in modeling macromolecular structures (Abramson et al., 2024) (e.g., by training models on full protein-nucleic acid-ligand complexes), in contrast to current methods that are trained primarily on one type of biomolecular complex (e.g., protein-ligand complexes). Key limitations of this study include its reliance on the accuracy of its predicted protein structures, its limited number of multi-ligand prediction targets available for benchmarking, and its inclusion of only a subset of all available protein-ligand docking baselines to focus on the most recent deep learning algorithms designed specifically for docking. In future work, we aim to expand not only the number of baseline methods but also the number of available multi-ligand targets while maintaining a diverse composition of heterogeneous (ionic) complexes. As a publicly available resource, POSEBENCH is flexible to accommodate new datasets and methods for protein-ligand structure generation.

Availability. The POSEBENCH codebase, documentation, and tutorial notebooks are available on GitHub under a permissive MIT license, with further licensing discussed in Appendix A.

Acknowledgments. This work was supported by a U.S. NSF grant (DBI2308699) and two U.S. NIH grants (R01GM093123 and R01GM146340). Additionally, this work was performed using high performance computing infrastructure provided by Research Support Services at the University of Missouri-Columbia (DOI: 10.32469/10355/97710). We would like to thank Martin Buttenschoen and Andriy Kryshtafovych for graciously allowing us to use in Figure 1 the PoseBusters and CASP logos, respectively. We would also like to specifically thank Martin Buttenschoen for their assistance in running the PoseBusters software suite for various protein-ligand complexes. Lastly, we would like to thank Arian Jamasb for providing insightful feedback on an early version of this manuscript.

References

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, pp. 1–3, 2024.

Akhtar, M., Benjelloun, O., Conforti, C., Giner-Miguelez, J., Jain, N., Kuchnik, M., Lhoest, Q., Marcenac, P., Maskey, M., Mattson, P., Oala, L., Ruyssen, P., Shinde, R., Simperl, E., Thomas, G., Tykhonov, S., Vanschoren, J., Vogler, S., and Wu, C.-J. Croissant: A metadata format for ml-ready datasets, 2024.

Bank, P. D. Protein data bank. *Nature New Biol*, 233(223): 10–1038, 1971.

Buttenschoen, M., Morris, G. M., and Deane, C. M. PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences, August 2023. URL https://doi.org/10.48550/arXiv.2308.05777.

Buttenschoen, M., Morris, G. M., and Deane, C. M. Posebusters: Ai-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chemical Science*, 2024.

Corso, G., Stärk, H., Jing, B., Barzilay, R., and Jaakkola, T. Diffdock: Diffusion steps, twists, and turns for molecular docking. *arXiv preprint arXiv:2210.01776*, 2022.

Corso, G., Deng, A., Fry, B., Polizzi, N., Barzilay, R., and Jaakkola, T. Deep confident steps to new pockets: Strategies for docking generalization. *arXiv preprint arXiv:2402.18396*, 2024a.

Corso, G., Deng, A., Fry, B., Polizzi, N., Barzilay, R., and Jaakkola, T. The Discovery of Binding Modes Requires Rethinking Docking Generalization, February 2024b. URL https://doi.org/10.5281/zenodo.10656052.

Dhakal, A., McKay, C., Tanner, J. J., and Cheng, J. Artificial intelligence in the prediction of protein–ligand interactions: recent advances and future directions. *Briefings in Bioinformatics*, 23(1):bbab476, 2022.

Du, X., Li, Y., Xia, Y.-L., Ai, S.-M., Liang, J., Sang, P., Ji, X.-L., and Liu, S.-Q. Insights into protein–ligand interactions: mechanisms, models, and methods. *International journal of molecular sciences*, 17(2):144, 2016.

Eastman, P. and Pande, V. Openmm: A hardware-independent framework for molecular simulations. *Computing in science & engineering*, 12(4):34–39, 2010.

Harris, C., Didi, K., Jamasb, A. R., Joshi, C. K., Mathis, S. V., Lio, P., and Blundell, T. Benchmarking generated poses: How rational is structure-based drug design with generative models? *arXiv preprint arXiv:2308.07413*, 2023.

Hartshorn, M. J., Verdonk, M. L., Chessari, G., Brewerton, S. C., Mooij, W. T., Mortenson, P. N., and Murray, C. W. Diverse, high-quality test set for the validation of protein-ligand docking performance. *Journal of medicinal chemistry*, 50(4):726–741, 2007.

Hu, J., Liu, Z., Yu, D.-J., and Zhang, Y. Ls-align: an atom-level, flexible ligand structural alignment algorithm for high-throughput virtual screening. *Bioinformatics*, 34 (13):2209–2218, 2018.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Krishna, R., Wang, J., Ahern, W., Sturmfels, P., Venkatesh, P., Kalvet, I., Lee, G. R., Morey-Burrows, F. S., Anishchenko, I., Humphreys, I. R., et al. Generalized biomolecular modeling and design with rosettafold all-atom. *Science*, pp. eadl2528, 2024.

Krivák, R. and Hoksza, D. P2rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. *Journal of cheminformatics*, 10:1–12, 2018.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637): 1123–1130, 2023.

Lu, W., Wu, Q., Zhang, J., Rao, J., Li, C., and Zheng, S. Tankbind: Trigonometry-aware neural networks for drug-protein binding structure prediction. *Advances in neural information processing systems*, 35:7236–7249, 2022.

Lu, W., Zhang, J., Huang, W., Zhang, Z., Jia, X., Wang, Z., Shi, L., Li, C., Wolynes, P. G., and Zheng, S. Dynamicbind: predicting ligand-specific protein-ligand complex structure with a deep equivariant generative model. *Nature Communications*, 15(1):1071, 2024.

Mariani, V., Biasini, M., Barbato, A., and Schwede, T. lddt: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, 29(21):2722–2728, 2013.

Morehead, A., Giri, N., Liu, J., and Cheng, J. Deep Learning for Protein-Ligand Docking: Are We There Yet?, June 2024. URL https://doi.org/10.5281/zenodo.11477766.

Pang, M., He, W., Lu, X., She, Y., Xie, L., Kong, R., and Chang, S. Codock-ligand: Combined template-based docking and cnn-based scoring in ligand binding prediction. *BMC bioinformatics*, 24(1):444, 2023.

Pei, Q., Gao, K., Wu, L., Zhu, J., Xia, Y., Xie, S., Qin, T., He, K., Liu, T.-Y., and Yan, R. Fabind: Fast and accurate protein-ligand binding. *Advances in Neural Information Processing Systems*, 36, 2024.

Plainer, M., Toth, M., Dobers, S., Stark, H., Corso, G., Marquet, C., and Barzilay, R. Diffdock-pocket: Diffusion for pocket-level docking with sidechain flexibility. *NeurIPS 2023 Machine Learning in Structural Biology Workshop*, 2023.

Qiao, Z., Nie, W., Vahdat, A., Miller III, T. F., and Anandkumar, A. State-specific protein–ligand complex structure prediction with a multiscale deep generative model. *Nature Machine Intelligence*, pp. 1–14, 2024.

Robin, X., Studer, G., Durairaj, J., Eberhardt, J., Schwede, T., and Walters, W. P. Assessment of protein–ligand complexes in casp15. *Proteins: Structure, Function, and Bioinformatics*, 91(12):1811–1821, 2023.

Roy, R. S., Liu, J., Giri, N., Guo, Z., and Cheng, J. Combining pairwise structural similarity and deep learning interface contact prediction to estimate protein complex model accuracy in casp15. *Proteins: Structure, Function, and Bioinformatics*, 91(12):1889–1902, 2023.

Stärk, H., Ganea, O., Pattanaik, L., Barzilay, R., and Jaakkola, T. Equibind: Geometric deep learning for drug binding structure prediction. In *International conference on machine learning*, pp. 20503–20521. PMLR, 2022.

Stärk, H., Jing, B., Barzilay, R., and Jaakkola, T. Harmonic self-conditioned flow matching for multi-ligand docking and binding site design. *arXiv preprint arXiv:2310.05764*, 2023.

Trott, O. and Olson, A. J. Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*, 31(2):455–461, 2010.

Wankowicz, S. A., de Oliveira, S. H., Hogan, D. W., van den Bedem, H., and Fraser, J. S. Ligand binding remodels protein side-chain conformational heterogeneity. *eLife*, 11:e74114, mar 2022. ISSN 2050-084X. doi: 10.7554/eLife.74114. URL https://doi.org/10.7554/eLife.74114.

Warren, G. L., Do, T. D., Kelley, B. P., Nicholls, A., and Warren, S. D. Essential considerations for using protein–ligand structures in drug discovery. *Drug Discovery Today*, 17(23-24):1270–1281, 2012.

Xu, X., Duan, R., and Zou, X. Template-guided method for protein–ligand complex structure prediction: Application to casp15 protein–ligand studies. *Proteins: Structure, Function, and Bioinformatics*, 91(12):1829–1836, 2023.

Zhu, J., Gu, Z., Pei, J., and Lai, L. Diffbindfr: An se (3) equivariant network for flexible protein-ligand docking. *Chemical Science*, 2024.

A. Availability

The POSEBENCH codebase and tutorial notebooks are available under an MIT license at https://github.com/BioinfoMachineLearning/PoseBench. Preprocessed datasets and benchmark method predictions are available on Zenodo (Morehead et al., 2024) under a CC-BY 4.0 license, of which the Astex Diverse and PoseBusters Benchmark datasets (Buttenschoen et al., 2024) are associated with a CC-BY 4.0 license; of which the DockGen dataset (Corso et al., 2024a) is available under an MIT license; and of which the CASP15 dataset (Robin et al., 2023), as a mixture of publicly and privately available resources, is partially licensed. In particular, 15 (4 single-ligand and 11 multi-ligand targets) of the 19 CASP15 protein-ligand complexes evaluated with POSEBENCH are publicly available, whereas the remaining 4 (2 single-ligand and 2 multi-ligand targets) are confidential and, for the purposes of future benchmarking and reproducibility, must be requested directly from the CASP organizers. Lastly, our use of the PoseBusters software suite for molecule validity checking is permitted under a BSD-3-Clause license.

B. Broader impacts

Our benchmark unifies protein-ligand structure generation datasets, methods, and tasks to enable enhanced insights into the real-world utility of such methods for accelerated drug discovery and energy research. We acknowledge the risk that, in the hands of "bad actors", such technologies may be used with harmful ends in mind. However, it is our hope that efforts in elucidating the performance of recent protein-ligand structure generation methods in various macromolecular contexts will disproportionately influence the positive societal outcomes of such research such as improved medicines and subsequent clinical outcomes as opposed to possible negative consequences such as the development of new bioweapons.

C. Compute resources

To produce the results presented in this work, we concurrently utilized 4 80GB NVIDIA A100 GPUs for 4 weeks in total to run inference with each baseline method three times (where applicable), where each baseline deep learning method required approximately 3-7 days of GPU compute to complete its inference runs (except for FABind which completed its inference runs in the span of several hours). Notably, due to RoseTTAFold-All-Atom's significant storage requirements for running inference with its multiple sequence alignment databases, we utilized approximately 3 TB of solid-state storage space in total to benchmark all baseline methods. Lastly, in terms of CPU requirements, our experiments utilized approximately 192 concurrent CPU processes for AutoDock Vina inference (as an upper bound) and 128 GB of CPU RAM. Note that an additional 1-2 weeks of compute were spent performing initial (failed) versions of each experiment during POSEBENCH's initial phase of development.

As a more formal investigation of the computational resources required to run each baseline method in this work, in Table 2 we list the average runtime (in seconds) and peak CPU (GPU) memory usage (in GB) consumed by each method when running them on a 25% subset of the Astex Diverse dataset.

D. Metrics

In this work, we reference two key metrics in the field of structural bioinformatics: RMSD and lDDT. The RMSD between a predicted ligand 3D conformation (with atomic positions $\hat{x}_i$ for each of the ligand's n heavy atoms) and the ground-truth conformation (x_i) is defined as:

$$\text{RMSD} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\|\hat{x}_i - x_i\|^2}. \tag{1}$$

The lDDT score, which is commonly used to compare predicted and ground-truth protein 3D structures, is defined as:

$$\text{lDDT} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{4}\sum_{k=1}^{4}\left(\frac{1}{|\mathcal{N}_i|}\sum_{j\in\mathcal{N}_i}\Theta(|\hat{d}_{ij} - d_{ij}| < \Delta_k)\right), \tag{2}$$

where N is the total number of heavy atoms in the ground-truth structure; $\mathcal{N}_i$ is the set of neighboring atoms of atom i

Table 2. The average runtime (in seconds) and peak memory usage (in GB) of each baseline method on a 25% subset of the Astex Diverse dataset (using an NVIDIA 80GB A100 GPU for benchmarking). The symbol - denotes a result that could not be estimated. Where applicable, an integer enclosed in parentheses indicates the number of samples drawn from a particular (generative) baseline method.

Method	Runtime (s)	CPU Memory Usage (GB)	GPU Memory Usage (GB)
DiffDock-L (40)	130.53	9.67	63.07
FABind	4.01	5.00	8.44
DynamicBind (40)	187.00	5.36	79.11
NeuralPLexer (40)	223.65	11.31	42.61
RoseTTAFold-All-Atom	862.60	49.78	78.97
TULIP	-	-	-
DiffDock-L-Vina	13.05	0.80	0.00
P2Rank-Vina	17.83	2.13	0.00
Ensemble (Con)	-	-	-

within the inclusion radius $R_o = 15$ $\AA$ in the ground-truth structure, excluding atoms from the same residue; $\hat{d}_{ij}$ (d_{ij}) is the distance between atoms i and j in the predicted (ground-truth) structure; Δ_k are the distance tolerance thresholds (i.e., 0.5 Å, 1 Å, 2 Å, and 4 Å); $\Theta(x)$ is a step function that equals 1 if x is true, and 0 otherwise; and $|\mathcal{N}_i|$ is the number of neighboring atoms for atom i.

As originally proposed by Robin et al. (2023), in this study, we adopt the PLI-specific variant of lDDT, which calculates lDDT scores to compare predicted and ground-truth protein-ligand complex structures following optimal structural alignment of the predicted and ground-truth protein-ligand binding pockets.

E. Documentation for datasets

Below, we provide detailed documentation for each dataset included in our benchmark, summarised in Table 1. Each dataset is freely available for download from the benchmark's accompanying Zenodo data record (Morehead et al., 2024) under a CC-BY 4.0 license. In lieu of being able to create associated metadata for each of our macromolecular datasets using an ML-focused library such as Croissant (Akhtar et al., 2024) (due to file type compatibility issues), instead, we report structured metadata for our preprocessed datasets using Zenodo's web user interface (Morehead et al., 2024). Note that, for all datasets, we authors bear all responsibility in case of any violation of rights regarding the usage of such datasets.

E.1. Astex Diverse Set - Single-Ligand Docking (Difficulty: *Easy*)

A common drug discovery task is to screen several novel drug-like molecules against a target protein in rapid succession. The Astex Diverse dataset was originally developed with this application in mind, as it features many therapeutically relevant 3D molecules for computational modeling.

- **Motivation** Several downstream drug discovery efforts rely on having access to high-quality molecular data for docking.
- **Collection** For this dataset, which was originally compiled by Hartshorn et al. (2007), we adopt the version further prepared by Buttenschoen et al. (2024).
- **Composition** The dataset consists of 85 (80) single-ligand protein complexes (for which we could obtain high-accuracy predicted protein structures using AlphaFold/ESMFold).
- **Hosting** Our preprocessed version of the dataset (`https://doi.org/10.5281/zenodo.11477766`) can be downloaded from the benchmark's Zenodo data record at `https://zenodo.org/records/11477766/files/astex_diverse_set.tar.gz`.
- **Licensing** We have released our preprocessed version of the dataset under a CC-BY 4.0 license. The original dataset is available under a CC-BY 4.0 license on Zenodo (Buttenschoen et al., 2023).

- **Maintenance** We will announce any errata discovered in or changes made to the dataset using the benchmark's GitHub repository at https://github.com/BioinfoMachineLearning/PoseBench.
- **Uses** This dataset of holo (and predicted-apo) protein PDB and holo ligand SDF files can be used for single-ligand docking or protein-ligand structure generation.
- **Metric** Ligand RMSD ≤ 2 Å & PoseBusters-Valid (PB-Valid).

E.2. PoseBusters Benchmark Set - Single-Ligand Docking (Difficulty: *Intermediate*)

Like the Astex Diverse dataset, the PoseBusters Benchmark dataset was originally developed for docking individual ligands to target proteins. However, this dataset features a larger and more challenging collection of protein-ligand complexes for computational modeling.

- **Motivation** Data sources of challenging single-ligand protein complexes for molecular docking are critical for the development of future docking methods.
- **Collection** For this dataset, we adopt the version introduced by Buttenschoen et al. (2024).
- **Composition** The dataset consists of 308 (280) single-ligand protein complexes (for which we could obtain high-accuracy predicted protein structures using AlphaFold/ESMFold).
- **Hosting** Our preprocessed version of the dataset (https://doi.org/10.5281/zenodo.11477766) can be downloaded from the benchmark's Zenodo data record at https://zenodo.org/records/11477766/files/posebusters_benchmark_set.tar.gz.
- **Licensing** We have released our preprocessed version of the dataset under a CC-BY 4.0 license. The original dataset is available under a CC-BY 4.0 license on Zenodo (Buttenschoen et al., 2023).
- **Maintenance** We will announce any errata discovered in or changes made to the dataset using the benchmark's GitHub repository at https://github.com/BioinfoMachineLearning/PoseBench.
- **Uses** This dataset of holo (and predicted-apo) protein PDB and holo ligand SDF files can be used for single-ligand docking or protein-ligand structure generation.
- **Metric** Ligand RMSD ≤ 2 Å & PoseBusters-Valid (PB-Valid).

E.3. DockGen Set - Single-Ligand Docking (Difficulty: *Challenging*)

The DockGen dataset was originally developed for docking individual ligands to target proteins in the context of novel protein binding pockets. As such, this dataset is useful for evaluating how well each baseline method can generalize to distinctly different binding pockets compared to those on which it may have been trained.

- **Motivation** Data sources of protein-ligand complexes representing novel single-ligand binding pockets are critical for the development of generalizable docking methods.
- **Collection** For this dataset, we adopt the version introduced by Corso et al. (2024a).
- **Composition** The dataset originally consists of 189 single-ligand protein complexes, after which we perform additional filtering down to 91 complexes based on ESMFold structure prediction accuracy (< 5 Å Cα atom RMSD for the primary protein interaction chain).
- **Hosting** Our preprocessed version of the dataset (https://doi.org/10.5281/zenodo.11477766) can be downloaded from the benchmark's Zenodo data record at https://zenodo.org/records/11477766/files/dockgen_set.tar.gz.
- **Licensing** We have released our preprocessed version of the dataset under a CC-BY 4.0 license. The original dataset is available under an MIT license on Zenodo (Corso et al., 2024b).

- **Maintenance** We will announce any errata discovered in or changes made to the dataset using the benchmark's GitHub repository at https://github.com/BioinfoMachineLearning/PoseBench.
- **Uses** This dataset of holo (and predicted-apo) protein PDB and holo ligand PDB files can be used for single-ligand docking or protein-ligand structure generation.
- **Metric** Ligand RMSD $\leq$ 2 Å & PoseBusters-Valid (PB-Valid).

E.4. CASP15 Set - Multi-Ligand Docking (Difficulty: *Challenging*)

As the most complex of our benchmark's four test datasets, the CASP15 protein-ligand interaction dataset was created to represent the new protein-ligand modeling category in the 15th Critical Assessment of Structure Prediction (CASP) competition. Whereas the Astex Diverse and PoseBusters Benchmark datasets feature solely single-ligand protein complexes, the CASP15 dataset provides users with a variety of challenging organic (e.g., drug molecules) and inorganic (e..g., ion) cofactors for *multi*-ligand biomolecular modeling.

- **Motivation** Multi-ligand evaluation datasets for molecular docking provide the rare opportunity to assess how well baseline methods can model intricate protein-ligand interactions while avoiding troublesome protein-ligand and ligand-ligand steric clashes. Additionally, more accurate modeling of multi-ligand complexes in future works may lead to improved techniques for computational enzyme design and regulation (Stärk et al., 2023).
- **Collection** For this dataset, we manually collect each publicly and privately available CASP15 protein-bound ligand complex structure compatible with protein-ligand (e.g., non-RNA) benchmarking.
- **Composition** The dataset consists of 102 (86) fragment ligands contained within 19 (15) separate (publicly available) protein complexes, of which 6 (2) and 13 (2) of these complexes are single and multi-ligand complexes, respectively.
- **Hosting** Our preprocessed version of (the publicly available version of) this dataset (https://doi.org/10.5281/zenodo.11477766) can be downloaded from the benchmark's Zenodo data record at https://zenodo.org/records/11477766/files/casp15_set.tar.gz.
- **Licensing** We have released our preprocessed version of the (public) dataset under a CC-BY 4.0 license. The original (public) dataset is free for download via the RCSB PDB (Bank, 1971).
- **Maintenance** We will announce any errata discovered in or changes made to the dataset using the benchmark's GitHub repository at https://github.com/BioinfoMachineLearning/PoseBench.
- **Uses** This dataset of holo (and predicted-apo) protein PDB and holo ligand PDB files can be used for multi-ligand docking or protein-ligand structure generation.
- **Metric** (Fragment) Ligand RMSD $\leq$ 2 Å & (Complex) PoseBusters-Valid (PB-Valid).

F. Analysis of protein-ligand interactions

Inspired by a similar analysis presented in the PoseCheck benchmark (Harris et al., 2023), in this section, we study the frequency of different types of protein-ligand interactions such as Van der Waals contacts and hydrophobic interactions occurring natively within the Astex Diverse, PoseBusters Benchmark, DockGen, and CASP15 datasets, respectively. In particular, these measures allow us to better understand the diversity of interactions each baseline method within the POSEBENCH benchmark is tasked to model, within the context of each test dataset. Figure 7 displays the results of this analysis.

Overall, we find that the Astex Diverse, PoseBusters Benchmark, and DockGen datasets contain similar types and frequencies of interactions, with the PoseBusters Benchmark and DockGen datasets containing slightly more hydrogen bond acceptors ($\sim$3 vs 1) and Van der Waals contacts ($\sim$13 vs 8) on average compared to the Astex Diverse dataset. However, we note a significant difference in interaction types and frequencies between the CASP15 dataset and the three other datasets. Specifically, we find it contains a significantly higher proportion of hydrogen bond acceptors and donors ($\sim$40), Van der Waals contact ($\sim$200), and hydrophobic interactions ($\sim$15) on average. Particularly interesting to note is the CASP15

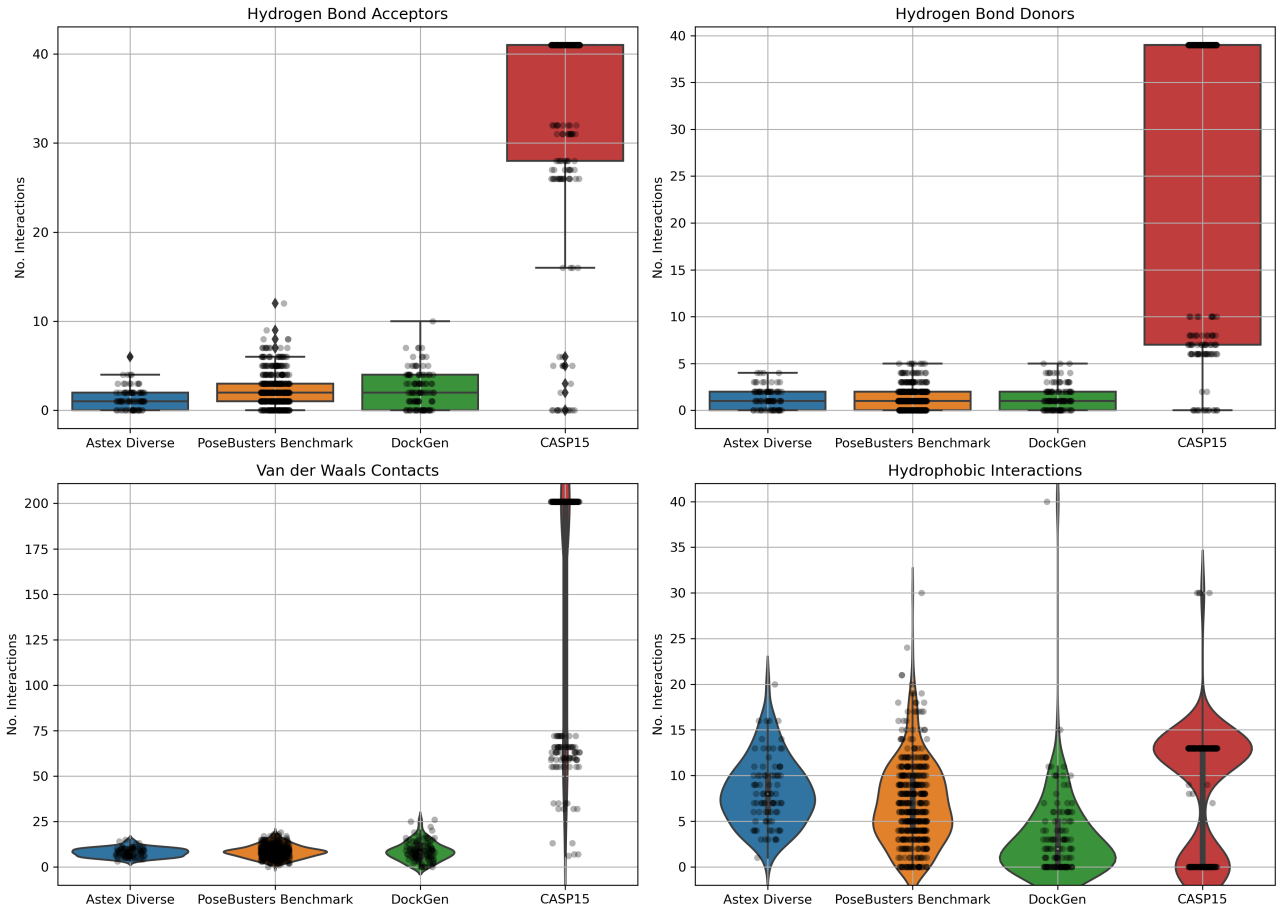


Figure 7. Astex, PoseBusters, DockGen, and CASP15 dataset analysis of protein-ligand interactions.

dataset's bimodal distribution of hydrophobic interactions, suggesting that the dataset contains two primary classes of interacting ligands giving rise to hydrophobic interactions. One possible explanation for this phenomenon is that the CASP targets, in contrast to the Astex Diverse, PoseBusters Benchmark, and DockGen targets, consist of a variety of both organic (e.g., drug-like molecules) and inorganic (e.g., metal) cofactors.

G. Additional results

In this section, we provide additional results for each baseline method using the Astex Diverse, PoseBusters Benchmark, and DockGen datasets as well as the CASP15 ligand targets. Note that for all violin plots listed in this section, we curate them using combined results across each method's three independent runs (where applicable), in contrast to this section's bar charts where we instead report mean and standard deviation values across each method's three independent runs.

G.1. DockGen results

DockGen dataset. The DockGen dataset (Corso et al., 2024a) contains 189 diverse single-ligand protein complexes, each representing a novel type of protein-ligand binding pocket. This dataset can be considered the most difficult single-ligand benchmark set since its protein binding sites are distinctly different from those found in the training datasets of most deep learning-based docking methods to date.

For this dataset, we once again used ESMFold to predict the *apo* complex structures of each of its proteins. We performed additional filtering down to 91 of the dataset's complexes, as using ESMFold not all 189 of its protein complex structures could be accurately predicted (i.e., achieving < 5 Å Cα atom RMSD for the primary protein interaction chains). After predicting each structure, we RMSD-aligned these *apo* structures while optimally weighting each complex's protein-ligand interface in the alignment.

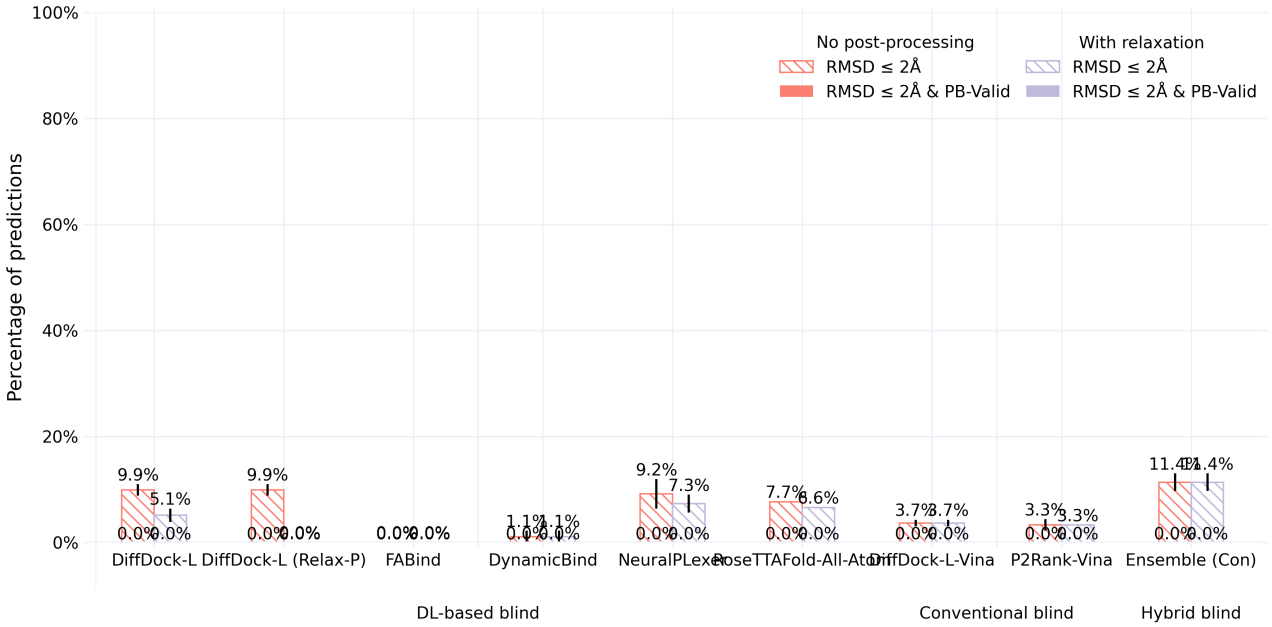


Figure 8. DockGen dataset results for successful single-ligand docking with relaxation.

Benchmark results. Figures 8 and 9 show that DiffDock-L, NeuralPLexer, and Ensemble (Con) provide the best pocket generalization capabilities compared to all other baseline methods. In particular, Ensemble (Con) seems to provide the most stable performance in this setting. Moreover, the results for DiffDock-L-Vina and P2Rank-Vina suggest that DiffDock-L generalizes to new binding pockets slightly better than P2Rank for conventional docking with AutoDock-Vina. Additionally, DiffDock-L's results with protein-flexible relaxation applied (i.e., DiffDock-L (Relax-P)) demonstrate that fixed-protein relaxation (albeit unideal from a theoretical e.g., protein side chain perspective (Wankowicz et al., 2022)) yields less accuracy degradation to DiffDock-L's original ligand conformations compared to side chain-flexible relaxation. However, we note that none of the baseline methods generate *any* PB-valid ligand conformations, suggesting that such "correct" poses are approximately accurate yet physically implausible in certain measurable ways.

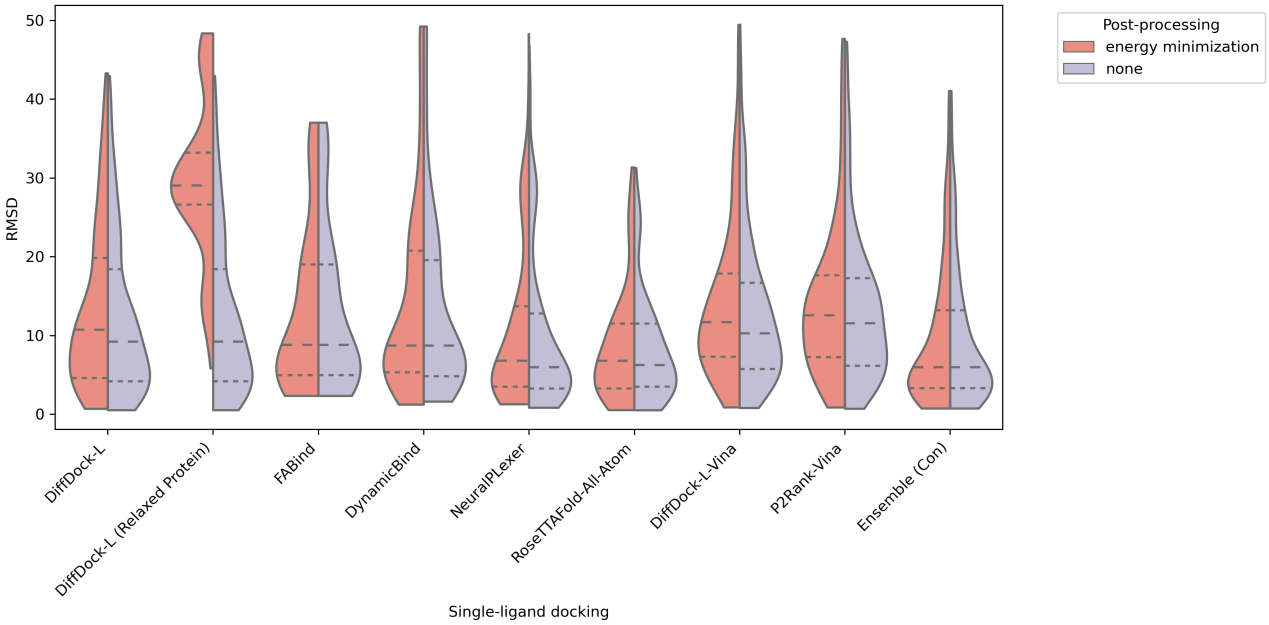


Figure 9. DockGen dataset results for single-ligand docking RMSD.

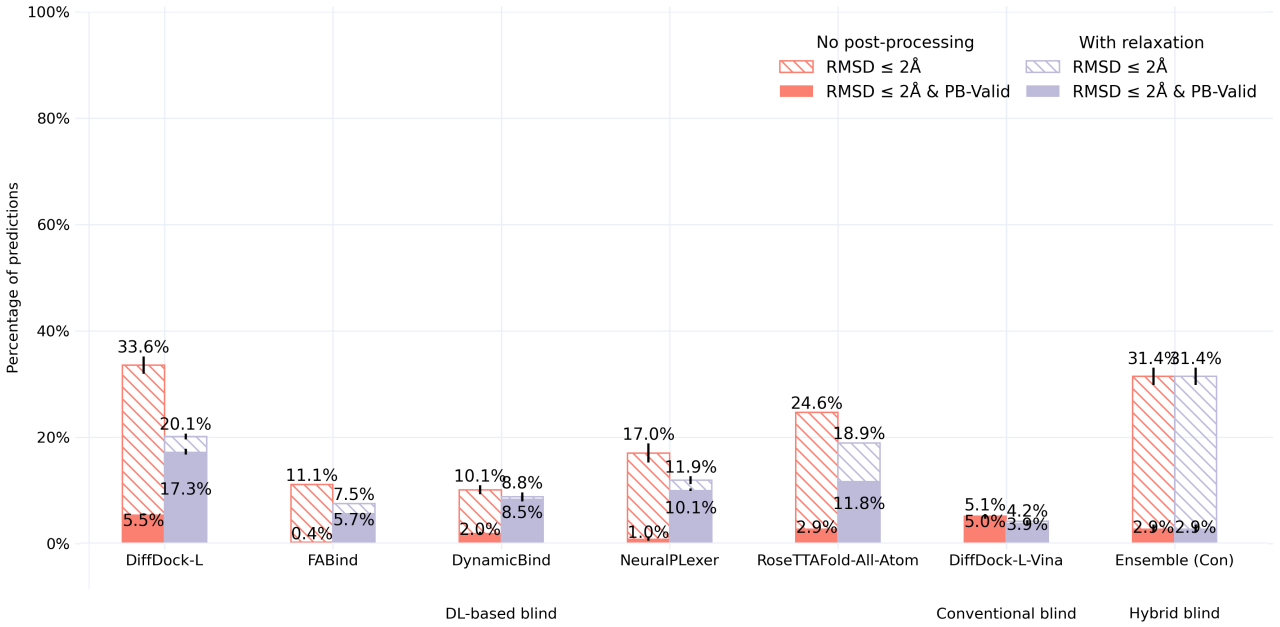


Figure 10. Pocket-only PoseBusters dataset results for successful single-ligand docking with relaxation.

G.2. Expanded Astex & PoseBusters results

G.2.1. Pocket-only PoseBusters results

Figures 10 and 11 illustrate the impact of reducing the binding pocket search space of each baseline docking method by providing each method with alternative versions of the predicted PoseBusters Benchmark protein structures that have been cropped to contain only ligand-interacting (< 4 Å heavy atom distance) protein residues and their (7) sequence-adjacent neighbors. Overall, we find that performing such *pocket-level* docking increases the docking success rates and favorably narrows the ligand RMSD distributions of DiffDock-L, NeuralPLexer, RoseTTAFold-All-Atom, and Ensemble (Con), whereas for all other baselines, performance is either maintained or degraded marginally. This finding highlights that methods such as DiffDock-L, NeuralPLexer, and RoseTTAFold-All-Atom are better at identifying the correct local structural conformations of each ligand compared to other baseline methods such as FABind and DynamicBind.

G.2.2. Astex & PoseBusters RMSD results

In Figures 12 and 13, we report the ligand RMSD values of each baseline method across the Astex Diverse and PoseBusters Benchmark datasets, with relaxation being applied in the context of the PoseBusters Benchmark dataset. In short, we see that most methods are relatively similar in terms of their ligand RMSD distributions, with RoseTTAFold-All-Atom and our ensembling consensus baseline (i.e., Ensemble (Con)), however, offering more condensed distributions overall. Interestingly, for Astex Diverse, TULIP also appears to produce a uniquely confined ligand RMSD distribution.

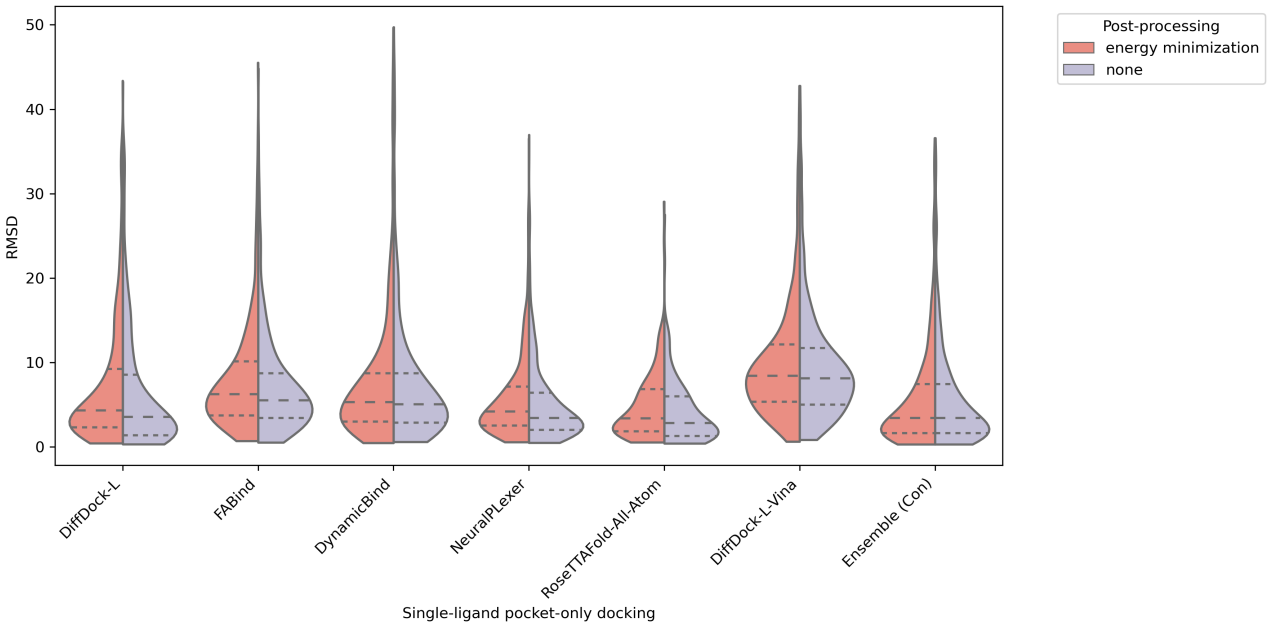


Figure 11. Pocket-only PoseBusters dataset results for single-ligand docking RMSD.

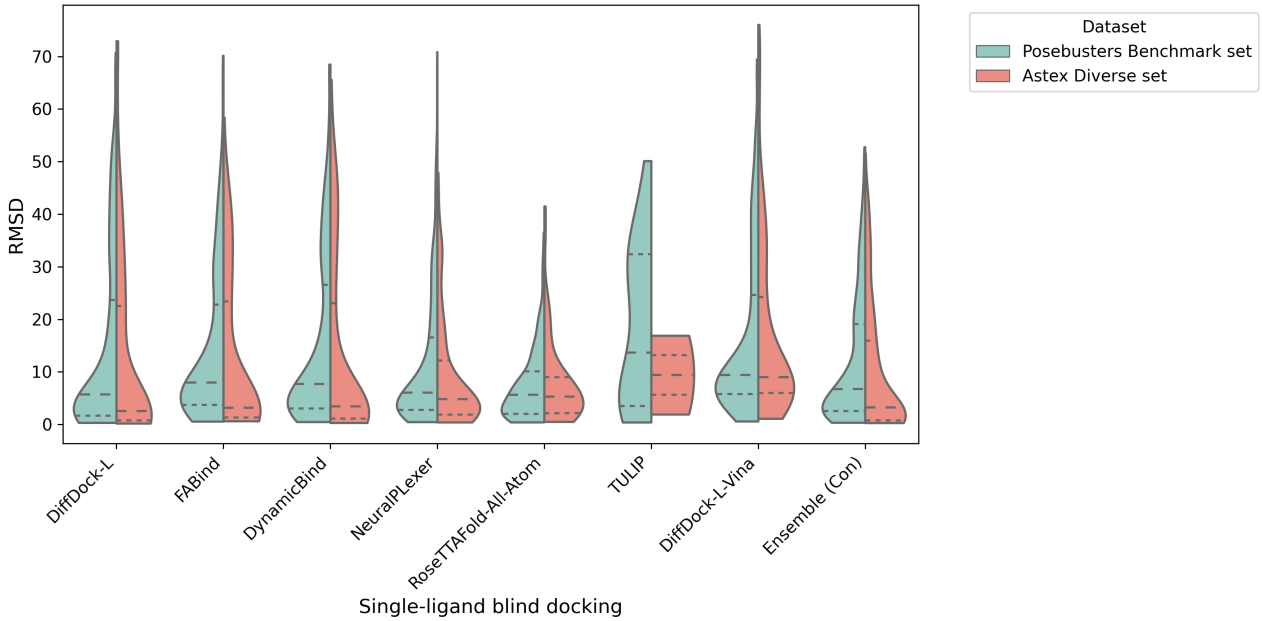


Figure 12. Astex & PoseBusters dataset results for single-ligand docking RMSD.

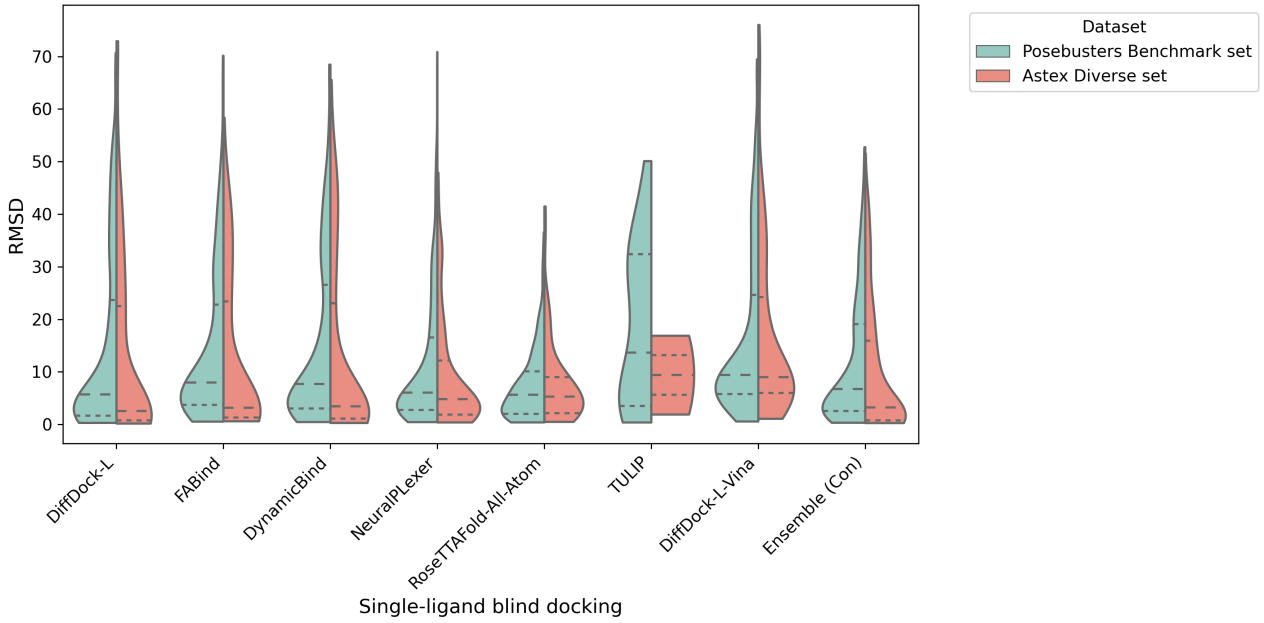


Figure 13. PoseBusters dataset results for single-ligand docking RMSD with relaxation.

G.3. Expanded CASP15 results

G.3.1. Overview of expanded results

In this section, we begin by reporting additional CASP15 benchmarking results in terms of each baseline method's multi-ligand RMSD and lDDT-PLI distributions as violin plots. Subsequently, we report successful ligand docking success rates as well as RMSD and lDDT-PLI results specifically for the single-ligand CASP15 targets. Lastly, we report all the above single and multi-ligand results specifically using only the CASP15 targets for which the ground-truth (experimental) structures are publicly available, to enable reproducible future benchmarking and follow-up works.

G.3.2. Multi-ligand RMSD and lDDT-PLI

To start, Figures 14 and 15 report each method's multi-ligand RMSD and lDDT-PLI distributions with and without relaxation. We see that NeuralPLexer and Ensemble (Con) produce the most tightly bound and accurate RMSD and lDDT-PLI distributions overall.

G.3.3. All single-ligand results

Next, Figures 16, 17, 18, and 19 display each method's single-ligand CASP15 docking success rates, PoseBusters validity rates, docking RMSD, and docking lDDT-PLI distributions, respectively. In summary, we can make a few respective observations from these plots. (1) DiffDock-L and Ensemble (Con) are the only methods capable of successfully docking *any* single-ligand CASP15 complexes. (2) AutoDock Vina produces the most PB-valid single-ligand complexes overall, with DiffDock-L shortly behind. (3) DiffDock-L, AutoDock Vina, and Ensemble (Con) appear to achieve the most tightly bound and accurate RMSD distributions. (4) In contrast to (3), only Ensemble (Con) appears to achieve top results in terms of lDDT-PLI compared to the other baseline methods.

G.3.4. Single and multi-ligand results for *public* targets

Lastly, for completeness and reproducibility, Figures 20, 21, 22, and 23 present corresponding multi-ligand results for the public CASP15 targets, whereas Figures 24, 25, 26, and 27 report corresponding single-ligand results for the public CASP15 targets. Overall, we observe marginal differences between the full and public CASP15 target results for multi-ligand complexes. However, we notice more striking performance drops between the full and public *single*-ligand CASP15 target results, suggesting that some of the private single-ligand complexes are easier prediction targets than most of the publicly available single-ligand complexes.

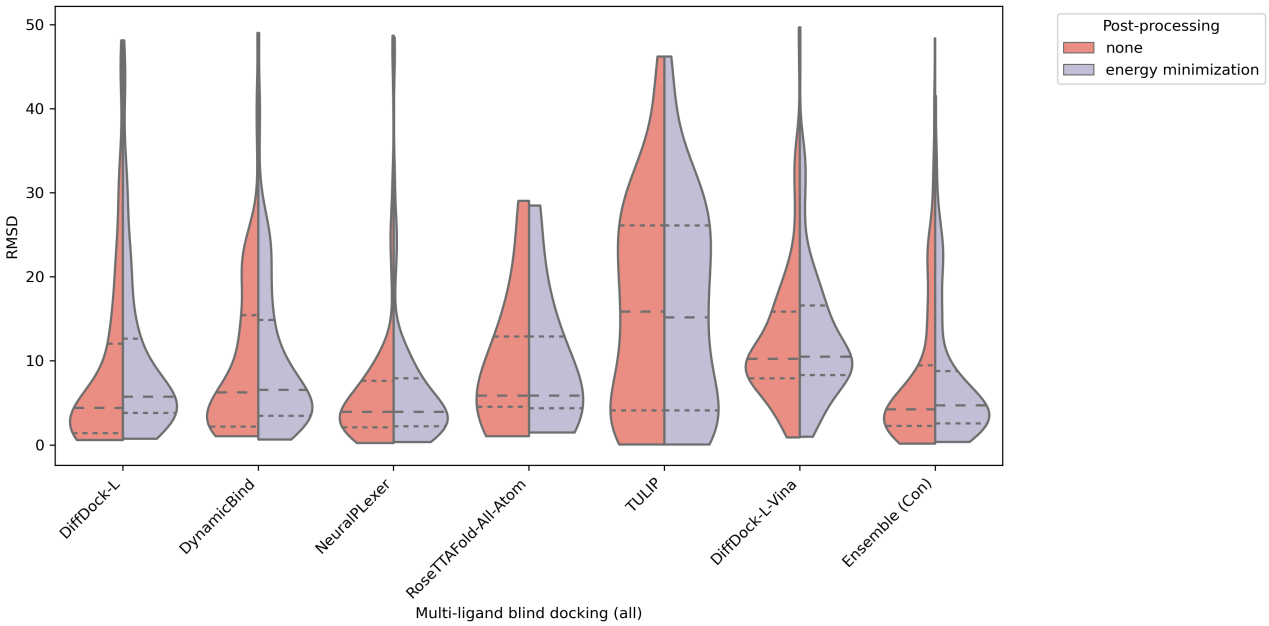


Figure 14. CASP15 dataset results for multi-ligand docking RMSD with relaxation.

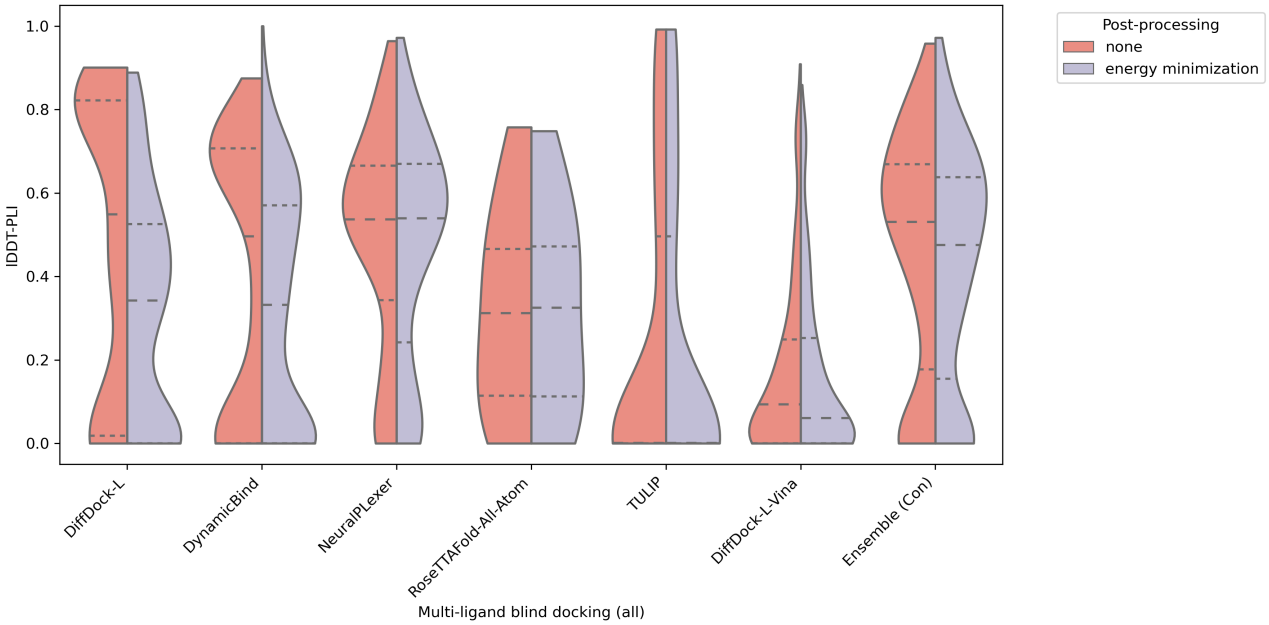


Figure 15. CASP15 dataset results for multi-ligand docking lDDT-PLI with relaxation.

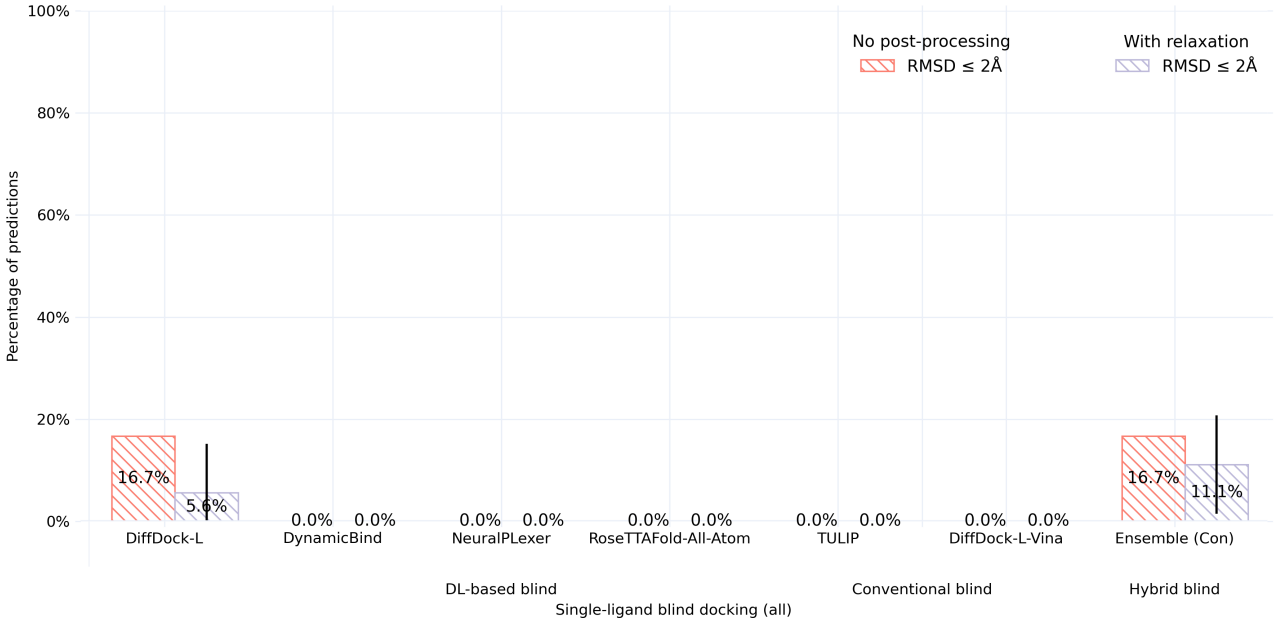


Figure 16. CASP15 dataset results for successful single-ligand docking with relaxation.

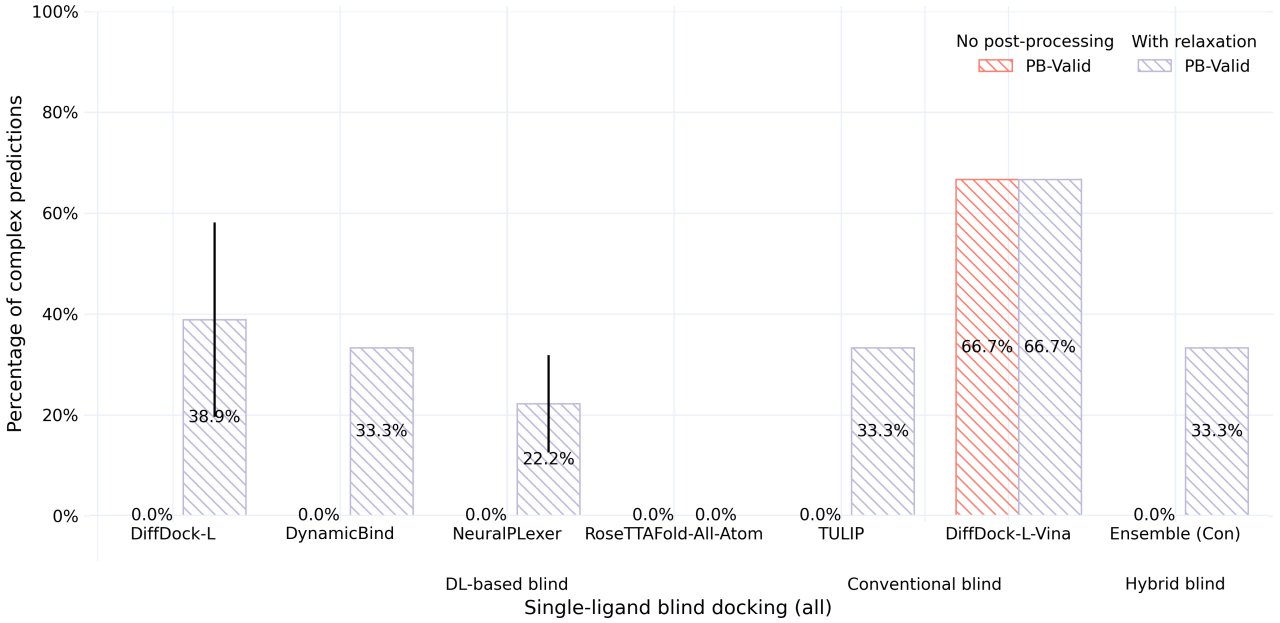


Figure 17. CASP15 dataset results for single-ligand PoseBusters validity rates with relaxation.

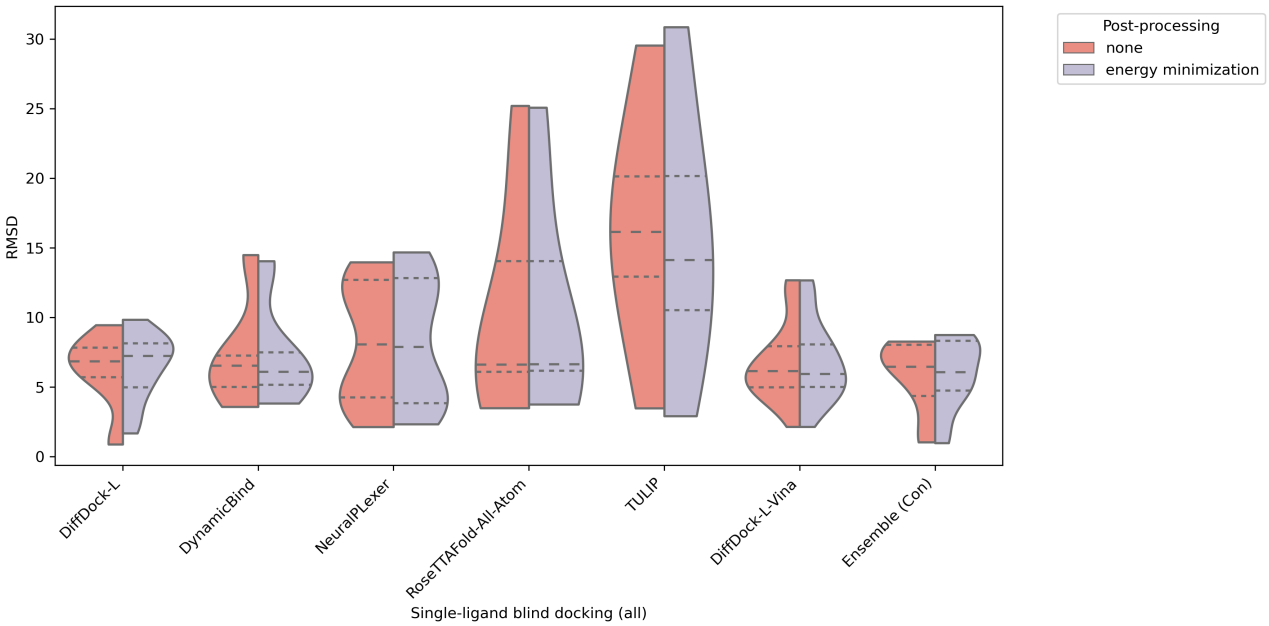


Figure 18. CASP15 dataset results for single-ligand docking RMSD with relaxation.

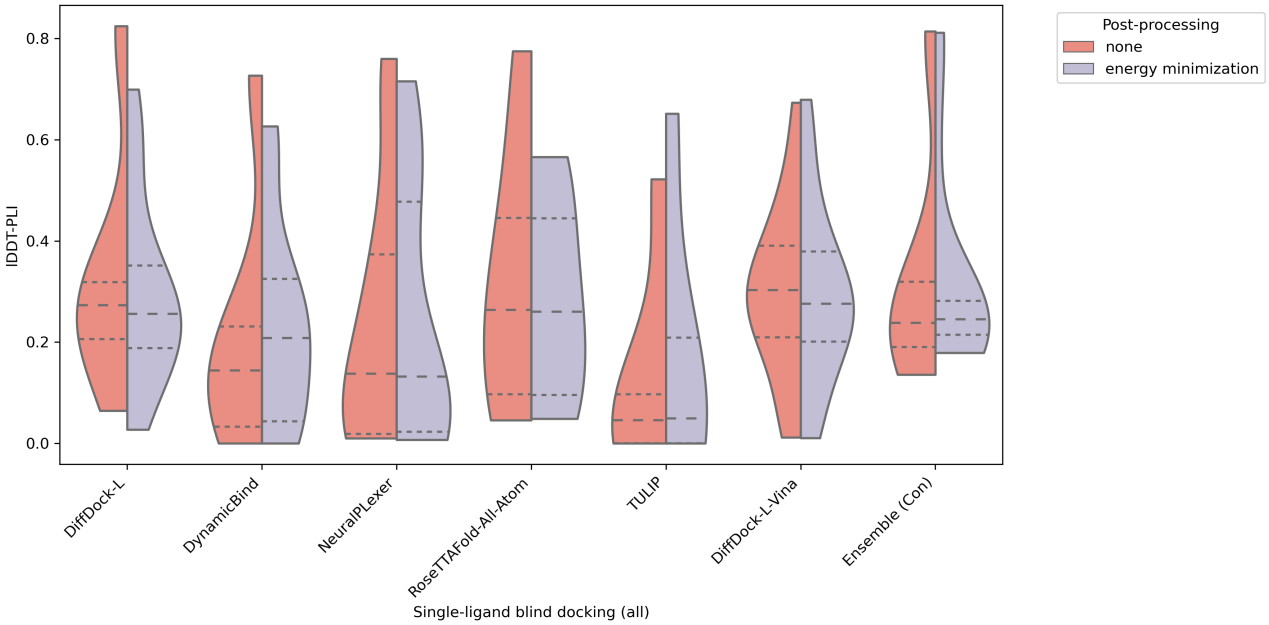


Figure 19. CASP15 dataset results for single-ligand docking lDDT-PLI with relaxation.

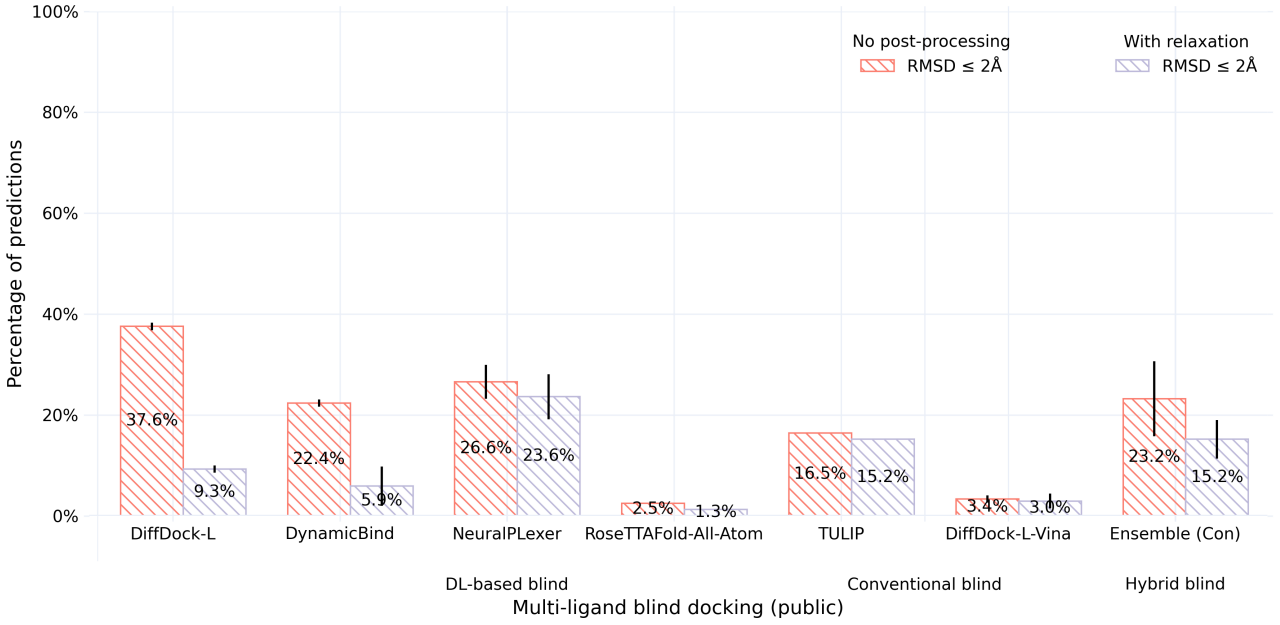


Figure 20. CASP15 public dataset results for successful multi-ligand docking with relaxation.

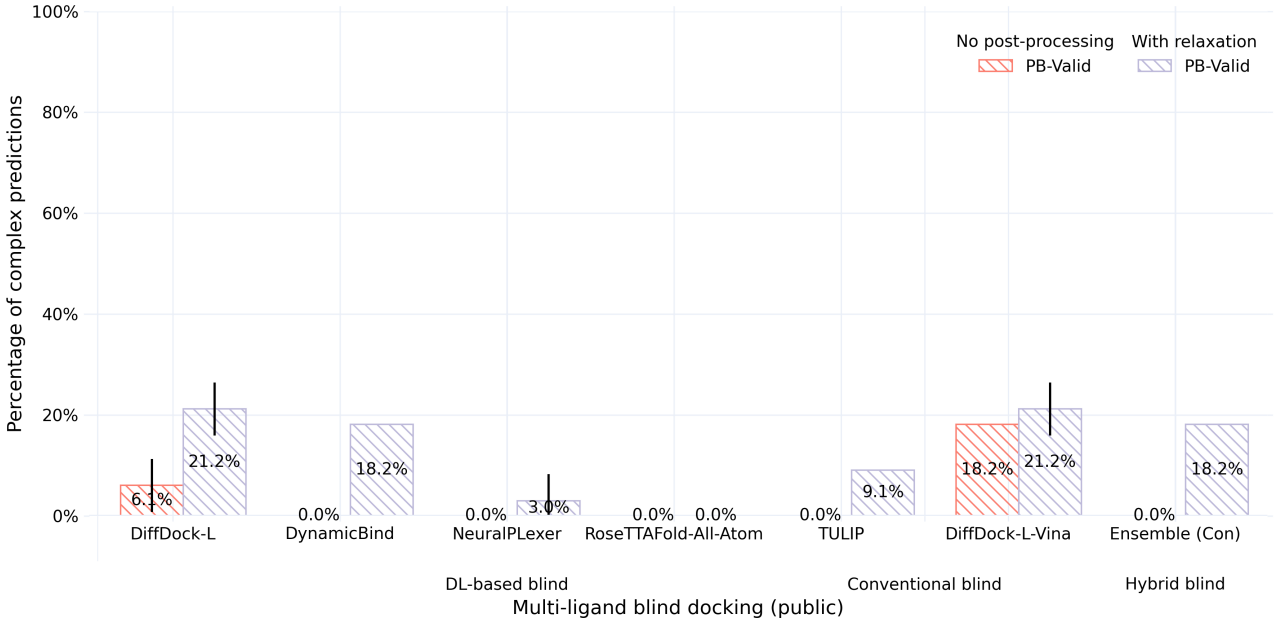


Figure 21. CASP15 public dataset results for multi-ligand PoseBusters validity rates with relaxation.

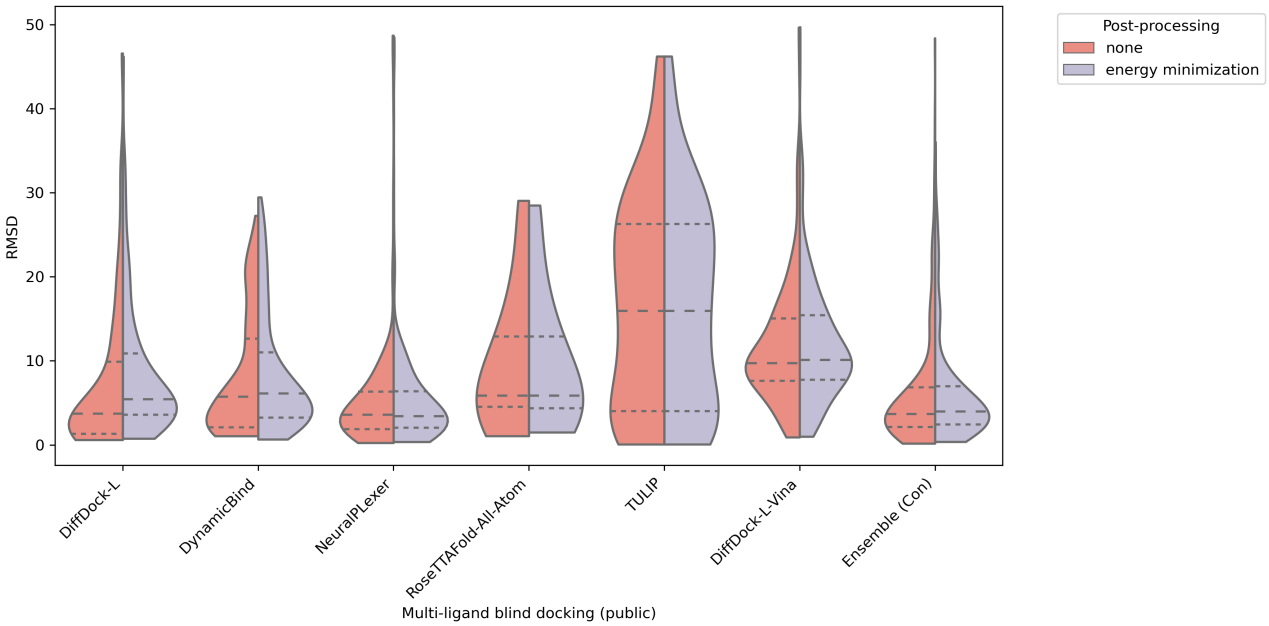


Figure 22. CASP15 public dataset results for multi-ligand docking RMSD with relaxation.

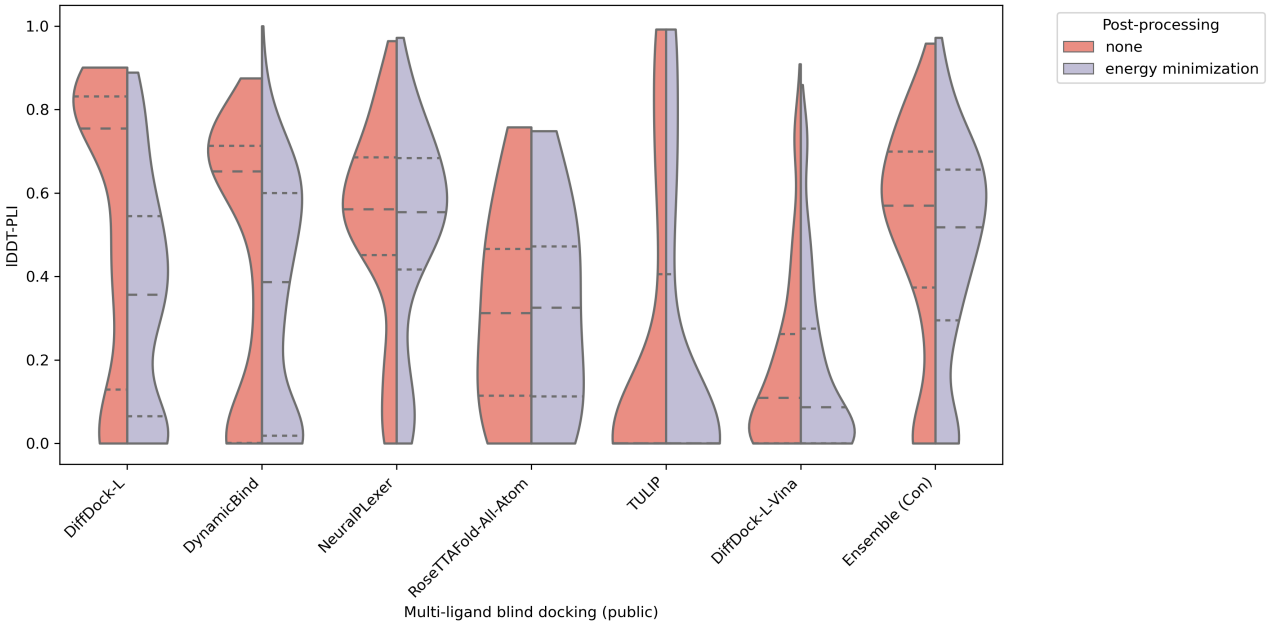


Figure 23. CASP15 public dataset results for multi-ligand docking lDDT-PLI with relaxation.

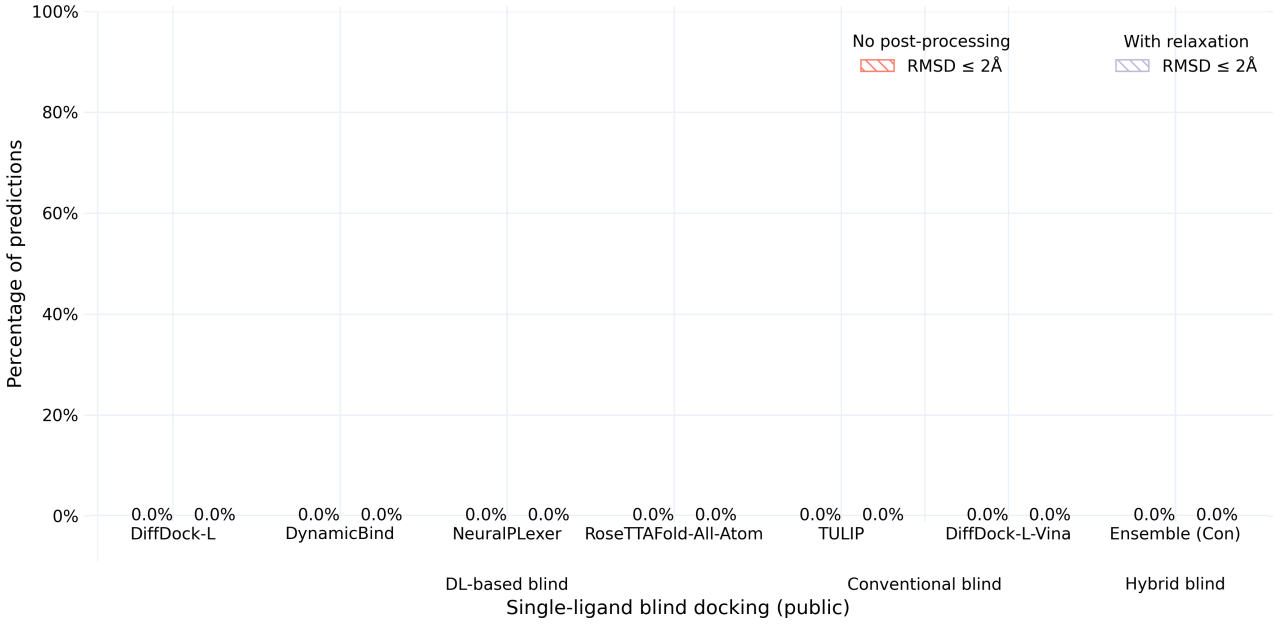


Figure 24. CASP15 public dataset results for successful single-ligand docking with relaxation.

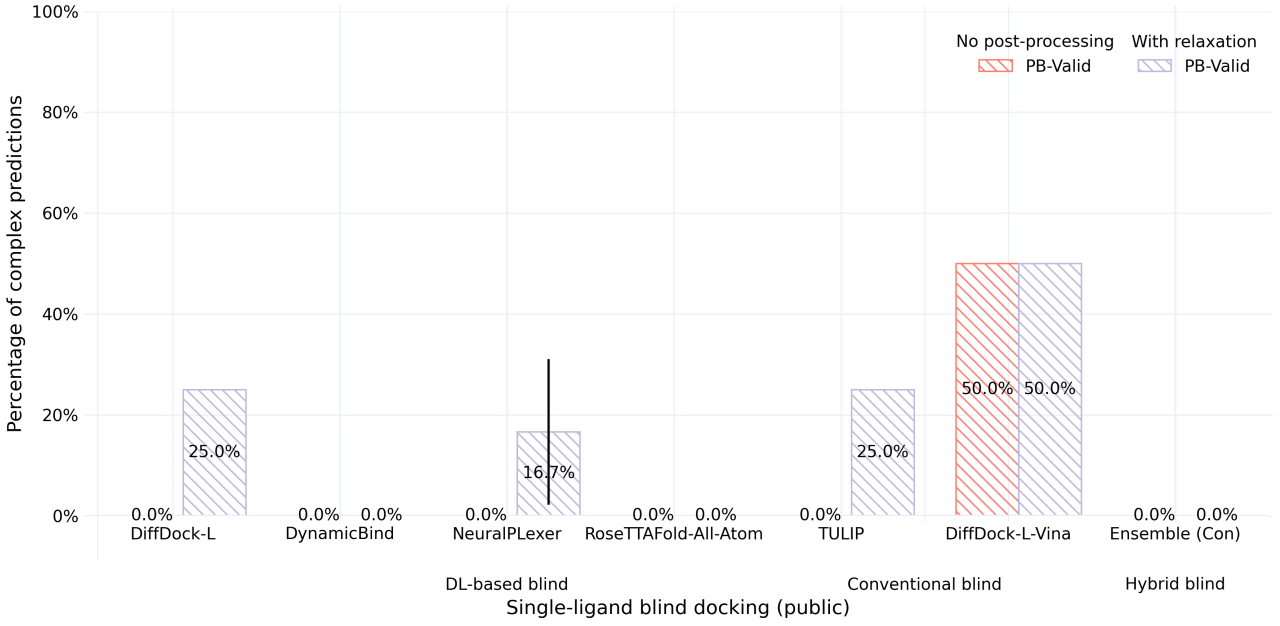


Figure 25. CASP15 public dataset results for single-ligand PoseBusters validity rates with relaxation.

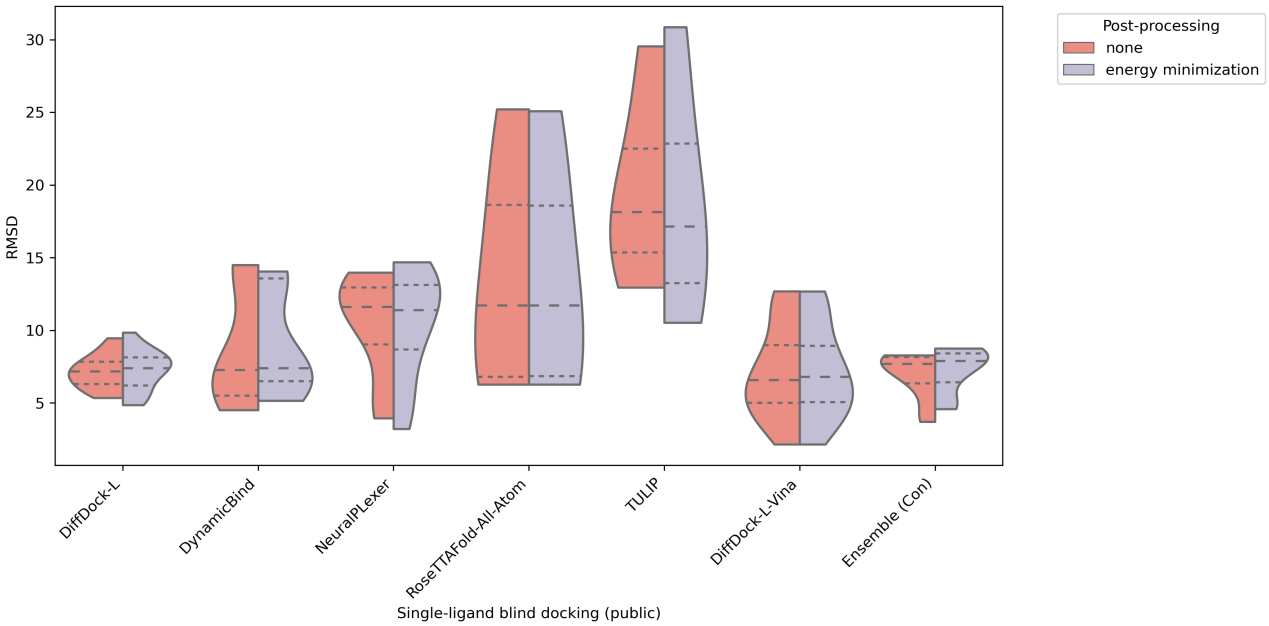


Figure 26. CASP15 public dataset results for single-ligand docking RMSD with relaxation.

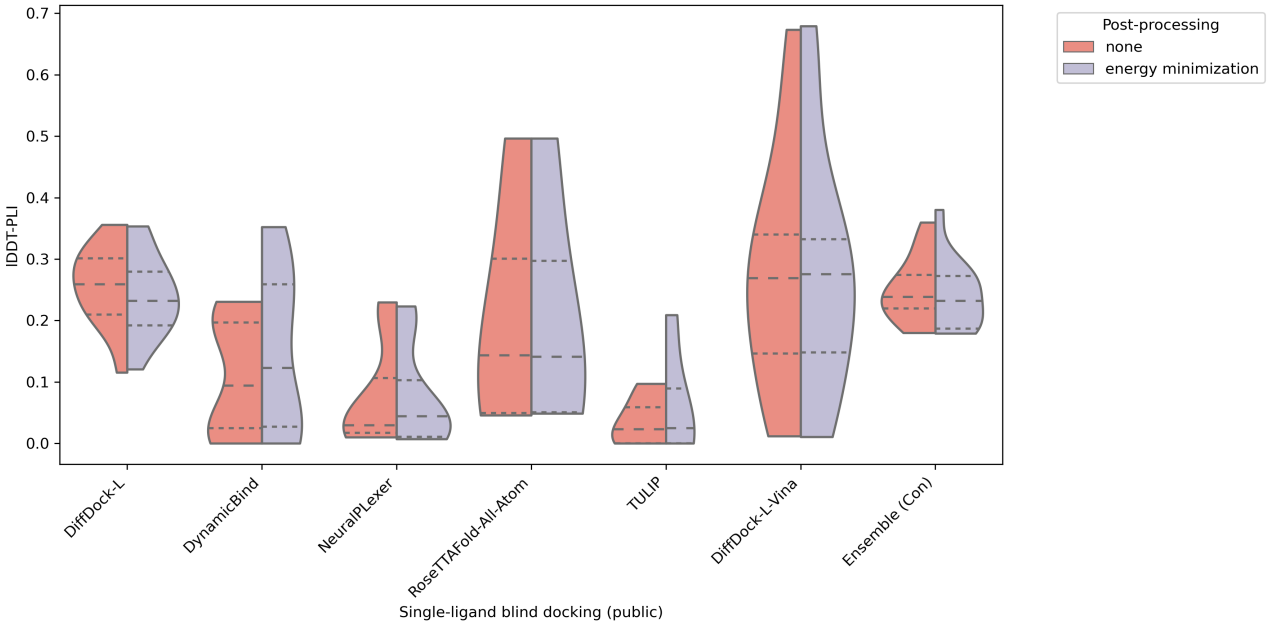


Figure 27. CASP15 public dataset results for single-ligand docking lDDT-PLI with relaxation.