Docking Methods Show Poor Transferability to Toxicity-Linked Targets

Anonymous Author(s)
Affiliation
Address
email

Abstract

Toxicity is a major cause of late-stage drug attrition, making accurate prediction of compound interactions with key safety targets essential. Molecular docking is widely used for this purpose, yet toxicity-related proteins are often underrepresented in standard benchmarks, raising concerns about generalizability. In this context, we benchmark seven docking methods, including classical baselines, AI-based approaches and hybrid methods. Focusing on scoring and ranking power we find that most methods perform suboptimally on toxicity-related targets. While several models show promise, their success is seemingly strongly influenced by binding pocket properties. These results demonstrate that performance reported on common benchmarks does not directly transfer to toxicity-focused tasks. Our study emphasizes the need for target-relevant evaluations of docking methods to improve computational toxicity prediction and support safer drug discovery.

1 Introduction

Toxicity is a major concern in drug development, as adverse effects can lead to late-stage attrition and regulatory withdrawal [1]. Early assessment of compounds against key safety targets helps guide safer lead selection and reduce clinical risk [2].

Accurate toxicity prediction requires reliable estimation of binding affinities for toxicity-related targets [3]. These affinities are usually predicted using molecular docking approaches [4]. However, toxicity-related targets are usually underrepresented in standard docking benchmarks (see Table **??** of the Appendix A for a detailed overview). Notably, CYP3A4 and CYP2D6 - key liver enzymes implicated in drug-induced liver injury [5, 6] - alongside androgen receptor (AR) and estrogen receptor (ER), which play central roles in hormone-mediated adverse effects [7, 5] are among primary targets in the toxicity screening of drug candidates. Focusing on these targets allows task-specific evaluation of docking methods, with direct implications for safety-driven lead selection and regulatory assessment.

Performance of docking methods is traditionally measured along four axes: docking, scoring, ranking, and screening power [8]. While AI-based methods such as Uni-Mol [9], SurfDock [10], and DiffDock [11] have shown strong pose prediction abilities on benchmarks like PoseX [12], often surpassing classical tools including Glide [13] and AutoDock Vina [14, 15], pose accuracy alone does not guarantee reliable affinity estimation [16, 17]. For toxicity-related applications, however, the central question is not whether a ligand can adopt a plausible binding pose, but whether relative binding preferences across a panel of targets can be captured. Scoring and ranking powers calculated as the correlation between predicted affinities and experimental activity data are therefore the most relevant for toxicity prediction. Moreover, prior work shows that methods with high docking power (e.g., GNINA 1.3 with $\sim$64% [12]) may still display weak screening power (nEF1% < 0.4 [18]), underscoring the need for multi-axis evaluation.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

In this study, we benchmark seven docking methods on the four toxicity-related targets, reporting scoring power and ranking power based on the correlations between predicted affinities and experimental ligand activities. Our contributions are as follows:

1. We provide a comparative analysis of classical baselines (QVina), state-of-the-art deep learning-based methods (PLAPT, Interformer, DynamicBind, Boltz-2), and hybrid approaches (Uni-Mol + AutoDock Vina, GNINA 1.3), focusing on their scoring and ranking power across toxicity-relevant targets.
2. We demonstrate that reported performance of the assessed methods on common docking benchmarks can not be fully transferred to toxicity-focused tasks prioritizing scoring and ranking abilities.
3. We identify two promising methods while showing that their performance strongly depends on the nature of a binding pocket.

The code and data used in this study are available at: `https://anonymous.4open.science/r/Docking-Methods-Show-Poor-Transferability-to-Toxicity-Linked-Targets-B541`

2 Methods

2.1 Targets

Four proteins were selected to benchmark docking methods. Two cytochrome P450 isoforms, CYP2D6 [19] (PDB ID: 4WNV) and CYP3A4 [20] (PDB ID: 1TQN), were included because of their central role in xenobiotic metabolism and frequent involvement in drug-induced liver injury. In addition, the Estrogen Receptor (ER, PDB ID: 1ERE) [21] and the Androgen Receptor (AR, PDB ID: 4K7A)[22] were selected as representative hormonal toxicity targets. Protein preparation is described in the Appendix B.1.

The inclusion of targets with clear binding pockets and a more complex case, such as CYP3A4, where ligands may adopt multiple, partially overlapping binding modes, strengthens our benchmark by introducing heterogeneity that mirrors real-world toxicological screening scenarios, in which ambiguous binding environments frequently arise [23].

2.2 Datasets

To estimate correlations of predicted binding affinities with experimental data, ligand activity datasets were obtained from the ChEMBL database [24] for each of the four studied proteins. For each target, two independent datasets were prepared based on different biological activity endpoints: the inhibition constant (K_i) and the half maximal inhibitory concentration (IC_{50}). This choice reflects the complementary nature of these measures: K_i provides a direct, assay-independent measure of binding affinity but is relatively scarce in public databases, whereas IC_{50} values are more abundant yet assay-dependent. Evaluating both allows us to leverage the broader coverage of IC_{50} data while also assessing performance on the more rigorous K_i subset. To ensure data quality and consistence, we removed duplicate entries, ligands with a molecular weight above 500 Da, ligands with activity reported in units other than nanomolar, and ligands with zero biological activity (see Appendix B.2).

2.3 Docking methods and metrics

Classical molecular docking was performed using QVina [25], an adaptation of AutoDock Vina [14] optimized for speed and efficiency. To assess modern deep learning (DL) approaches, we selected several recently published models that represent complementary strategies for protein–ligand docking. The first group includes methods that use different DL architectures to learn protein–ligand interaction patterns from data: DynamicBind [26], InterFormer [27], PLAPT [28], and Boltz-2 [29]. The second group includes hybrid approaches that combine DL-based and classical docking strategies. Specifically, we evaluated GNINA 1.3 [18], a hybrid method that augments a classical docking engine (Autodock Vina) with convolutional neural networks for pose ranking and scoring. In addition, we tested Uni-Mol [9] method, which originally allows only pose generation. In order to estimate binding affinities of predicted poses, we combined Uni-Mol with AutoDock Vina [15] run in `-local_only` mode, which refines pre-generated poses through local energy minimization without performing a

full redocking. Overall, this combination of classical, DL-based, and hybrid docking approaches provides a balanced framework to assess their relative strengths of methods in predicting binding affinities across toxicity-related targets.

We assess the performance of the above methods using scoring and ranking powers particularly significant for toxicity related docking. Scoring power was quantified as the Pearson correlation between predicted docking scores and experimental K_i/IC_{50} values, while ranking power was assessed as the Spearman correlation between docking scores and experimental activities. We also report confidence intervals for correlations values to ensure fair comparison across datasets of different sizes.

3 Results

The results for scoring power (r) are summarized in Table 1. For the considered targets, the scoring power of the evaluated methods was generally low, with most results not exceeding 0.2. Top results were provided by Boltz-2 on 1ERE protein ($r = 0.636$ for IC_{50} and $r = 0.643$ for K_i), DynamicBind on 4WNV ($r = 0.495$ for K_i), and Interformer on 1TQN ($r = 0.444$ for K_i). In contrast, Uni-Mol + AutoDock Vina approach demonstrated nearly zero correlation across all targets. Detailed information on the confidence intervals and ranking power results can be found in Figures 1–2 of the Appendix C.1.

Notably, no method showed consistent generalization ability: models excelling on Estrogen and Androgen receptors underperformed on cytochromes and vice versa. DL-based approaches provided better results compared to classical docking method in some cases, although these improvements were inconsistent.

Overall, scoring and ranking power across methods seem to remain suboptimal and highly target-dependent, with many correlations statistically indistinguishable from zero. These results highlight the difficulty of identifying a universally reliable docking strategy for toxicity-relevant assessments.

Table 1: Scoring power of the assessed docking methods. We separate classical, DL-based and hybrid approaches by horizontal lines. *For binding free energy–based methods, correlation signs were inverted so that higher positive values uniformly indicate better agreement with experimental data. **Binding affinities were computed for the best predicted pose (see Appendix C.2 for a comparison with the maximum-affinity selection strategy).

Method	1ERE		1TQN		4WNV		4K7A	
	IC50	K_i	IC50	K_i	IC50	K_i	IC50	K_i
QVina*	-0.196	0.079	-0.050	0.161	-0.012	0.326	0.059	0.121
Boltz-2*	**0.636**	**0.643**	0.051	-0.221	-0.065	0.296	0.100	0.311
DynamicBind (best pose)**	0.391	0.031	**0.099**	0.308	-0.012	**0.495**	**0.281**	**0.430**
PLAPT	0.451	0.152	0.045	0.045	-0.034	0.229	0.194	0.222
Interformer (best pose)**	0.058	0.217	0.007	**0.444**	**0.044**	0.079	0.268	0.367
GNINA 1.3 (best pose)**	-0.029	0.003	0.072	0.247	0.057	0.409	0.090	-0.081
Uni-Mol + AutoDock Vina*	-0.215	-0.087	0.052	0.129	-0.040	-0.059	0.090	-0.029

4 Discussion

4.1 Structural Determinants of Docking Accuracy

Boltz-2 and DynamicBind demonstrate high scoring power for compact Androgen (4K7A) and Estrogen (1ERE) receptor pockets, while correlations are lower for cytochromes. These patterns indicate that, although both models were initially trained on large numbers of proteins with different possible interactions (and, unlike DynamicBind [30], Boltz-2 was trained on PDBbind of 2023-06-01 with all four targets included [31]), their scoring functions seem to place increased importance on hydrophobic contacts. In order to further support this hypothesis, we analyzed the correlations between hydrophobic and hydrophilic ligands: Boltz-2 shows strong association with hydrophobic

contact recovery ($\rho \approx 0.545$, N = 1025, p < 10^{-10}) but negligible correlation with hydrophilic contacts ($\rho \approx 0.03$, N = 32), as shown in Table 2.

Structural characteristics of binding pockets further rationalize these trends. Nuclear receptors possess compact, moderately hydrophobic cavities (volume $\approx 680 \pm 45$ Å^3; hydrophobicity $\approx 0.62 \pm 0.08$), whereas CYPs are larger and more flexible (CYP2D6 volume $\approx 1650 \pm 120$ Å^3; hydrophobicity $\approx 0.41 \pm 0.07$), as summarized in Table5 of the Appendix D. These differences align with the observed performance: hydrophobic-biased scoring functions perform well in compact pockets, but their advantage diminishes in large, flexible cavities where flexibility-aware sampling and pose-confidence metrics become critical (Appendix C.2).

Overall, these findings highlight that docking outcomes are shaped by the interplay between algorithmic inductive biases and target-specific pocket properties. For future improvements, integrating hydrophobic-sensitive scoring with flexibility-aware sampling and pose-confidence evaluation may enhance generalizability across structurally diverse toxicity-relevant targets.

Table 2: Spearman correlation (ρ) of docking methods with experimental activities of hydrophobic (logP>2) and hydrophilic (logP<2) ligands, number of corresponding ligands (# ligands), and Mann–Whitney test results.

Method	Hydrophobic		Hydrophilic		Mann–Whitney $-log(p)$
	ρ	# ligands	ρ	# ligands	
QVina	-0.236	1965	-0.036	48	17
Boltz-2	0.545	1025	0.030	32	12
DynamicBind (best pose)	0.385	2031	0.066	49	17
PLAPT	0.457	2031	0.343	49	17
Interformer (best pose)	0.049	1044	-0.171	28	11
GNINA 1.3 (best pose)	0.047	2031	-0.222	49	1
Uni-Mol + AutoDock Vina	-0.229	1215	0.028	32	12

4.2 Limitations

The findings of this study should be interpreted in light of certain limitations. First, the current benchmark is restricted to only four toxicity-relevant targets. While these proteins are representative and widely studied in toxicological assessment, the obtained conclusions may not fully generalize to other toxicity-related targets. Secondly, not all ligands can be processed by all docking methods. For example, Interformer does not handle sulfur-containing ligands, GNINA 1.3 encounters difficulties when working with very large compounds, and Uni-Mol generates ligand compositions with internal spatial contradictions or unrealistic bond geometries. These problems, characteristic of individual implementations, lead to differences in ligand coverage of different methods and highlight the practical limitations of existing approaches. Third, as the test sets in this study were derived from ChEMBL, where structural binding site information for the targets is not available, we were not able to estimate the docking power of the considered methods.

5 Conclusion and Future Work

Our benchmark of classical, DL-based and hybrid docking methods shows that scoring and ranking power remain generally low and highly target-dependent regarding four considered toxicity-relevant targets (CYP2D6, CYP3A4, AR, ERα). Boltz-2 and DynamicBind demonstrate complementary strengths, excelling in compact hydrophobic pockets and flexible CYP cavities, respectively. High docking or pose accuracy does not guarantee reliable affinity prediction, highlighting the influence of pocket properties and utilized docking method.

Future work will focus on strengthening the generalizability of our findings. In particular, we plan to extend the analysis to a broader set of toxicity-related targets and to evaluate both screening and docking power in this context. Additionally, we aim to investigate the key factors underlying the performance differences we observed across docking methods, especially in comparison to results reported on existing datasets.

References

[1] Michael J Waring, Cheryl Arrowsmith, Andrew R Leach, Paul D Leeson, Samantha Mandrell, Ruth M Owen, Grace Pairaudeau, William D Pennie, Simon D Pickett, Xue Wang, Owen Wallace, Alistair Weir, Samuel Whitebread, Hongming Yang, and Man Hoi Lee. Risk factors for the clinical failure of experimental drugs: an analysis of 1,000 drug development projects. *Nature Reviews Drug Discovery*, 14(7):415–430, 2015.

[2] F Peter Guengerich. Mechanisms of drug toxicity and relevance to pharmaceutical development. *Drug metabolism and pharmacokinetics*, 26(1):3–14, 2011.

[3] YZ Chen and CY Ung. Prediction of potential toxicity and side effect protein targets of a small molecule by a ligand–protein inverse docking approach. *Journal of Molecular Graphics and Modelling*, 20(3):199–218, 2001.

[4] Isabella A Guedes, Camila S de Magalhães, and Laurent E Dardenne. Receptor–ligand molecular docking. *Biophysical reviews*, 6(1):75–87, 2014.

[5] G. Tarantino, M. N. D. Di Minno, and D. Capone. Drug-induced liver injury: Is it somehow foreseeable? *World Journal of Gastroenterology*, 15(23):2817–2833, 2009.

[6] Marina Villanueva-Paz, Laura Morán, Nuria López-Alcántara, Cristiana Freixo, Raúl J. Andrade, M. Isabel Lucena, and Francisco Javier Cubero. Oxidative stress in drug-induced liver injury (dili): From mechanisms to biomarkers for use in clinical practice. *Antioxidants*, 10(3):390, 2021.

[7] Athanasios Anestis et al. Are we there yet? understanding androgen receptor signaling in breast cancer. *npj Breast Cancer*, 8:4, 2022.

[8] Xuefeng Liu, Songhao Jiang, Xiaotian Duan, Archit Vasan, Chong Liu, Chih-chan Tien, Heng Ma, Thomas S. Brettin, Fangfang Xia, Ian T. Foster, and Rick L. Stevens. Binding affinity prediction: From conventional to machine learning-based approaches. arXiv preprint arXiv:2410.00709, 2024.

[9] Eric Alcaide, Zhifeng Gao, Guolin Ke, Yaqi Li, Linfeng Zhang, Hang Zheng, and Gengmo Zhou. Uni-mol docking v2: Towards realistic and accurate binding pose prediction. *arXiv preprint arXiv:2405.11769*, 2024.

[10] J. Zhang, X. Shen, et al. Surfdock: Surface-informed diffusion for reliable and accurate protein–ligand docking. *Nature Methods*, 2025. PMID: 39604569.

[11] Gabriele Corso, Hannes Stärk, Bowen Jing, Regina Barzilay, and Tommi Jaakkola. Diffdock: Diffusion steps, twists, and turns for molecular docking. *arXiv preprint arXiv:2210.01776*, 2022.

[12] Yize Jiang, Xinze Li, Yuanyuan Zhang, Jin Han, Youjun Xu, Ayush Pandit, Zaixi Zhang, Mengdi Wang, Mengyang Wang, Chong Liu, Guang Yang, Yejin Choi, Wu-Jun Li, Tianfan Fu, Fang Wu, and Junhong Liu. Posex: Ai defeats physics-based methods on protein–ligand cross-docking. *arXiv preprint arXiv:2505.01700*, 2025.

[13] R. A. Friesner, J. L. Banks, R. B. Murphy, T. A. Halgren, J. J. Klicic, D. T. Mainz, M. P. Repasky, E. H. Knoll, M. Shelley, J. K. Perry, D. E. Shaw, P. Francis, and P. S. Shenkin. Glide: a new approach for rapid, accurate docking and scoring. 1. method and assessment of docking accuracy. *Journal of Medicinal Chemistry*, 47(7):1739–1749, 2004.

[14] Oleg Trott and Arthur J. Olson. 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.

[15] Jacob Eberhardt, Diogo Santos-Martins, Andreas F. Tillack, and Stefano Forli. Autodock vina 1.2.0: new docking methods, expanded force field, and python bindings. *Journal of Chemical Information and Modeling*, 61(8):3891–3898, 2021.

[16] Jie Liu and et al. A comprehensive review of docking scoring functions: performance, limitations, and applications. *Journal of Chemical Information and Modeling*, 2023.

[17] M. Khamis, W. Gomaa, and B. Galal. Deep learning is competing random forest in computational docking. *arXiv preprint arXiv:1608.06665*, 2016.

[18] Andrew T McNutt, Yanjing Li, Rocco Meli, Rishal Aggarwal, and David R Koes. Gnina 1.3: the next increment in molecular docking with deep learning. *Journal of Cheminformatics*, 17(1):28, 2025.

[19] Anqi Wang, Nicole M. DeVore, et al. Human cytochrome p450 2d6 (cyp2d6) in complex with quinidine, 2006. RCSB PDB ID: 2F9Q.

[20] Irina F. Sevrioukova et al. Ritonavir-bound human cytochrome p450 3a4 (cyp3a4), 2017. RCSB PDB ID: 5VC0.

[21] A. K. Shiau et al. Estrogen receptor alpha ligand-binding domain in complex with 4-hydroxytamoxifen, 2006. RCSB PDB ID: 3ERT.

[22] Christoph E. Bohl et al. Androgen receptor ligand-binding domain in complex with bicalutamide, 2004. RCSB PDB ID: 1Z95.

[23] Anantha R. Nookala, Junhao Li, Anusha Ande, Lei Wang, Naveen K. Vaidya, Weihua Li, Santosh Kumar, and Anil Kumar. Effect of methamphetamine on spectral binding, ligand docking and metabolism of anti-hiv drugs with cyp3a4. *PLOS ONE*, 11(1):e0146529, 2016.

[24] David Mendez and et al. The chembl database in 2021. *Nucleic Acids Research*, 49:D1035–D1042, 2021.

[25] Amr Alhossary, Stephanus D. Handoko, Yuguang Mu, and Chee-Keong Kwoh. Fast, accurate, and reliable molecular docking with quickvina 2. *Bioinformatics*, 31(13):2214–2216, 2015.

[26] Wei Lu, Jixian Zhang, Weifeng Huang, et al. Dynamicbind: predicting ligand-specific protein–ligand complex structure with a deep equivariant generative model. *Nature Communications*, 15:1071, 2024.

[27] Unknown Author. Interformer: Protein–ligand interaction transformer, 2024. Preprint.

[28] X. Guo, Y. Sun, et al. Plapt: Protein–ligand alignment pre-training for docking and affinity prediction. *bioRxiv*, 2024. preprint; use arXiv/bioRxiv identifier 2411.02179.

[29] Jeremy Wohlwend and coauthors. Boltz-2 for conditional generation and scoring of protein–ligand poses. Whitepaper, 2025. Available online at author's website.

[30] Jie Li, Xingyi Guan, Oufan Zhang, Kunyang Sun, Yingze Wang, Dorian Bagni, and Teresa Head-Gordon. Leak proof pdbbind: A reorganized dataset of protein-ligand complexes for more generalizable binding affinity prediction. *arXiv*, 2023.

[31] Yingze Wang, Kunyang Sun, Jie Li, Xingyi Guan, Oufan Zhang, Dorian Bagni, and Teresa Head-Gordon. Pdbbind optimization to create a high-quality protein-ligand binding dataset for binding affinity prediction. *arXiv*, 2024.

[32] CASP15 Ligand Prediction Category Group. Assessment of protein–ligand complexes in casp15. *bioRxiv / Deep learning for protein–ligand docking: Are We There Yet?*, 2024. CASP15 PLI includes 23 protein-ligand complexes evaluated by lDDT-PLI and BiSyRMSD metrics.

[33] B. Rowan and colleagues. Boltz-1: Deep learning approach to protein-ligand interface prediction. *arXiv preprint*, 2023.

[34] Zhihai Liu et al. Pdb-wide collection of binding data: current status of the pdbbind database. *Bioinformatics*, 31(3):405–412, February 2015. Based on PDB contents as of October 2014; describes core and refined sets.

[35] Amr Alhossary, Sufian D. Handoko, Yuguang Mu, and Chee Keong Kwoh. Fast, accurate, and reliable molecular docking with quickvina 2. *Bioinformatics*, 31(13):2214–2216, 2015.

[36] R. Wang, X. Fang, Y. Lu, and S. Wang. The pdbbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *Journal of Medicinal Chemistry*, 47(12):2977–2980, June 2004.

[37] Unknown Author. Dynamicbind: Graph neural network model for docking, 2024. Preprint.

[38] Martin Buttenschoen, Garrett M. Morris, and Charlotte M. Deane. Posebusters: Ai-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chemical Science*, 15(9):3130–3139, 2023.

[39] J. Zhou, K. Wang, and colleagues. Uni-mol docking v2: Unified pretraining for molecular docking. *arXiv preprint*, 2024.

[40] Y. Wang, X. Zhang, and colleagues. Interformer: Transformer-based molecular docking with interaction-aware embeddings. *arXiv preprint*, 2025.

[41] Paul G. Francoeur, Tomohide Masuda, and David R. Koes. 3d convolutional neural networks and a crossdocked dataset for structure-based drug design. *Journal of Chemical Information and Modeling*, 2020. CrossDocked2020 provides 22.5 million docked ligand poses across non-cognate binding sites; data and splits available at GitHub.

[42] Andrew T. McNutt, Paul Francoeur, Rishal Aggarwal, Tomohide Masuda, Remi Meli, Matthew Ragoza, Jocelyn Sunseri, and David R. Koes. Gnina 1.0: molecular docking with deep learning. *Journal of Cheminformatics*, 13(1):43, 2021. Introduces GNINA 1.0 and evaluates docking performance on Redocked2020 and CrossDocked2020 benchmarks.

[43] M. M. Mysinger, M. Carchia, J. J. Irwin, and B. K. Shoichet. Directory of useful decoys, enhanced (dud-e): Better ligands and decoys for benchmarking. *Journal of Medicinal Chemistry*, 55(14):6582–6594, 2012.

[44] Andrew T. McNutt and collaborators. Gnina 1.3: Improved docking and screening with deep learning. *Journal of Cheminformatics*, 2025.

[45] David R. Koes. Meeko: A tool for preparing molecular mechanics-ready pdb files, 2022. GitHub repository.

[46] Unknown Author. Labox: Automated box calculation for docking, 2023. GitHub repository.

[47] Saida Perot, Olivier Sperandio, Mariya A Miteva, Anaïs C Camproux, and Bruno O Villoutreix. Computational analysis of protein pockets: definition, detection, and druggability. *Journal of Chemical Information and Modeling*, 50(9):1647–1662, 2010. Quote: "size, shape, and hydrophobicity as the global pocket descriptors are indeed important to automatically predict druggability.".

[48] Vincent Le Guilloux, Pascal Schmidtke, and Pierre Tuffery. Geometry-based methods for binding site identification and characterization. *Current Pharmaceutical Design*, 15(31):3573–3587, 2009. Discusses pocket depth, concavity, and their role in ligand binding.

[49] Vincent Le Guilloux, Pascal Schmidtke, and Pierre Tuffery. Fpocket: an open source platform for ligand pocket detection. *BMC Bioinformatics*, 10:168, 2009. Describes Fpocket algorithm for pocket detection and descriptor calculation.

[50] Wei Tian, Chu Chen, Xiang Lei, Jiayin Zhao, and Jie Liang. Castp 3.0: computed atlas of surface topography of proteins. *Nucleic Acids Research*, 46(W1):W363–W367, 2018. Describes CASTp tool for calculation of pocket volume, surface area, and mouth openings.

[51] Fabian Buchwald, Schuettelkopf CC, et al. The importance of ligand conformational energies in carbohydrate docking: Sorting the wheat from the chaff. *Journal of Chemical Information and Modeling*, 54(2):279–291, 2014.

[52] AL Hopkins, CR Groom, and A Alex. Drug efficiency indices for improvement of molecular docking scoring functions. *Journal of Medicinal Chemistry*, 52(5):1234–1247, 2009.

[53] Gregory L Warren, C William Andrews, Anna-Maria Capelli, et al. The consequences of scoring docked ligand conformations using free energy correlations. *Bioorganic & Medicinal Chemistry Letters*, 16(1):23–28, 2006.

[54] Tao Cheng, Xiaohong Li, Y. Li, and R. Wang. Variability in docking success rates due to dataset preparation. *Journal of Chemical Information and Modeling*, 52(5):1189–1198, 2012.

Appendix

A Benchmarks for docking methods

Table 3 provides a concise review of existing benchmarks commonly used to assess the performance of docking methods. It summarizes key properties of these benchmarks: the presence of key toxicity-related targets considered in this study (AR, ERα, CYP2D6, CYP3A4), as well as metrics of docking methods on these benchmarks reported in prior work. The analysis highlights multiple coverage gaps - particularly, the under-representation of toxicity-related targets - which motivated the creation of the proposed toxicity benchmark.

Table 3: Comparison of docking methods: benchmarks used, inclusion of selected safety-relevant targets (AR, ERα, CYP2D6, CYP3A4), and reported performance metrics.

Method	Benchmark	Toxicity-related targets (AR, ERα, CYP2D6, CYP3A4)	Performance and metrics
Boltz-1	CASP15 (PLI) [32]	None	LDDT-PLI $\approx$ 65%; DockQ $>$0.23 in 83% of complexes [33]
QVina 2	PDBbind core set (2014) [34]	None	RMSD $\leq$ 2Å: higher success than GOLD; strong energy correlation with AutoDock Vina ($r = 0.967$) [35]
DynamicBind	PDBbind[36]	None	Success rate (RMSD$<$2Å + low ClashScore) = 33% vs 19% for DiffDock [37]
	MDT (559 complexes) [37]	MDT includes nuclear receptors (likely AR/ERα), no CYPs	
Uni-Mol V2	PoseBusters[38]	None	Top-1 RMSD$<$2Å $\approx$ 77%; chemically valid poses retained [39]
Interformer	PoseBusters[38]	None	84% (RMSD$<$2Å)
	PDBbind split [40]	None	63.9% (RMSD$<$2Å) [40]
GNINA 1.3	CrossDocked2020[41]	None	40%
	Redocked2020[42]	None	67% (RMSD$<$2Å)
	DUD-E[43]	DUD-E includes AR, ERα, CYP3A4 (not CYP2D6)	AUC=0.78, EF1%=0.27; notably poor enrichment for AR/ERα [42, 44]

B Preparation of structures

B.1 Preparation of protein structures for docking

All protein structures were preprocessed using the `Meeko` framework [45] to ensure consistent protonation states. Binding site parameters were then derived with `LaBOX` [46]. For Estrogen receptor (1ERE), Androgen receptor (4K7A), and CYP2D6 (4WNV), the docking boxes were defined from the co-crystallized ligand, providing reliable reference pockets. In contrast, CYP3A4 (1TQN) was used in its apo form, where the ligand-binding cavity is large and flexible; here, cavity detection was applied to approximate the catalytic site without a priori bias toward a single binding mode.

B.2 Preparation of ligand structures for docking

To ensure data quality and comparability across targets, the ligand sets extracted from ChEMBL were subjected to a series of preprocessing steps. Specifically, we removed duplicate entries, ligands with molecular weight exceeding 500 Da, records with non-nanomolar activity units, and compounds with zero or undefined biological activity values. After filtering, the resulting curated datasets contained the following number of ligands:

- **K_i-based datasets:** 356 (4WNV), 422 (4K7A), 193 (1ERE), 64 (1TQN).

- **IC_{50}-based datasets:** 1971 (4WNV), 1512 (4K7A), 2080 (1ERE), 298 (1TQN).

These dataset sizes reflect the expected difference in coverage between K_i and IC_{50} data, with the latter being more abundant in public repositories. The curated ligand collections were subsequently used as the basis for docking and correlation analyses reported in the main text.

C Docking results

This subsection provides additional visualizations and statistical details: scatter plots comparing predicted scores to experimental affinities (Figures 1–2), full descriptions of the statistical measures used (Pearson and Spearman correlations, bootstrap confidence intervals), and guidance on interpreting effect sizes and statistical significance in the context of docking-based affinity prediction. All docking results were obtained using a server with an NVIDIA A6000 GPU.

C.1 Scoring and ranking power

Scoring and ranking power of the assessed docking methods with experimental K_i and IC_{50} values are shown in Figure 1 and Figure2, respectively.

C.2 The effect of a model's confidence in posture

Methods such as GNINA, Interformer, and DynamicBind provide both affinity predictions and internal estimates of pose confidence. For each ligand, multiple poses are generated with corresponding affinity and confidence scores, meaning that top results can be selected based on either maximum affinity or best pose confidence score. Our experiments reported in Table 4 demonstrate that correlations with experimental data are generally lower when using maximum affinity selection strategy for DynamicBind and GNINA 1.3. In contrast, correlations with binding affinities obtained for best poses are more consistent and reliable. Therefore, we choose to report only results for best pose strategy in the main text.

Table 4: Correlations of binding affinities with experimental data obtained using maximum affinity (max affinity) and best pose confidence score (best pose) selection strategies.

Method	1ERE		1TQN		4WNV		4K7A	
	IC50	K_i	IC50	K_i	IC50	K_i	IC50	K_i
DynamicBind (max affinity)	0.383	-0.060	0.090	0.278	**-0.005**	0.479	0.270	0.375
DynamicBind (best pose)	**0.391**	**0.031**	**0.099**	**0.308**	-0.012	**0.495**	**0.281**	**0.430**
Interformer (max affinity)	0.058	0.217	0.007	0.444	**0.059**	0.079	0.268	0.367
Interformer (best pose)	0.058	0.217	0.007	0.444	0.044	0.079	0.268	0.367
GNINA 1.3 (max affinity)	-0.040	-0.010	**0.083**	**0.266**	0.053	0.390	0.066	-0.143
GNINA 1.3 (best pose)	**-0.029**	**0.003**	0.072	0.247	**0.057**	**0.409**	**0.090**	**-0.081**

D Binding Site Characterization

To contextualize the docking performance across the four protein targets, we extracted structural descriptors of their ligand-binding sites directly from the respective PDB entries. Specifically, we retrieved pocket geometry—such as volume, surface area, and mouth opening dimensions—as well as pocket hydrophobicity, following established protocols in structure-based druggability analysis. The results are shown in Table5.

We referred to key works demonstrating the relevance of these descriptors: size, shape, and hydrophobicity are recognized as critical global features for automatic prediction of druggability [47]. Moreover, geometric properties such as pocket depth and concavity are known to enhance drug-like molecule interactions [48].

Computationally, we employed the `Fpocket` software to detect binding cavities and compute their descriptors in an automated, reproducible fashion [49]. Additionally, where more detailed geometric

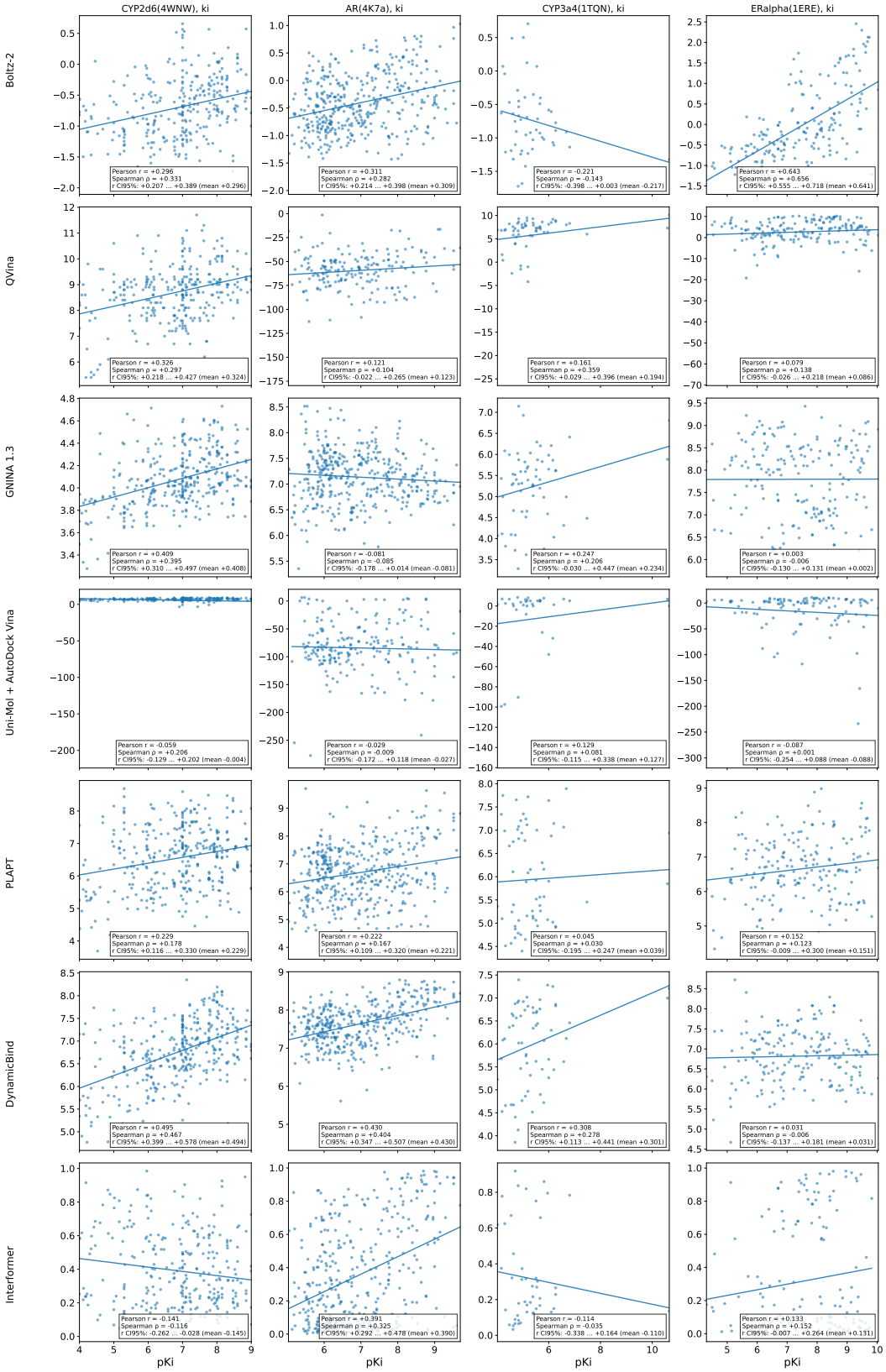

Figure 1: Scoring and ranking power of the assessed docking methods reported as Pearson and Spearman correlation of predicted binding affinities with experimental K_i values.

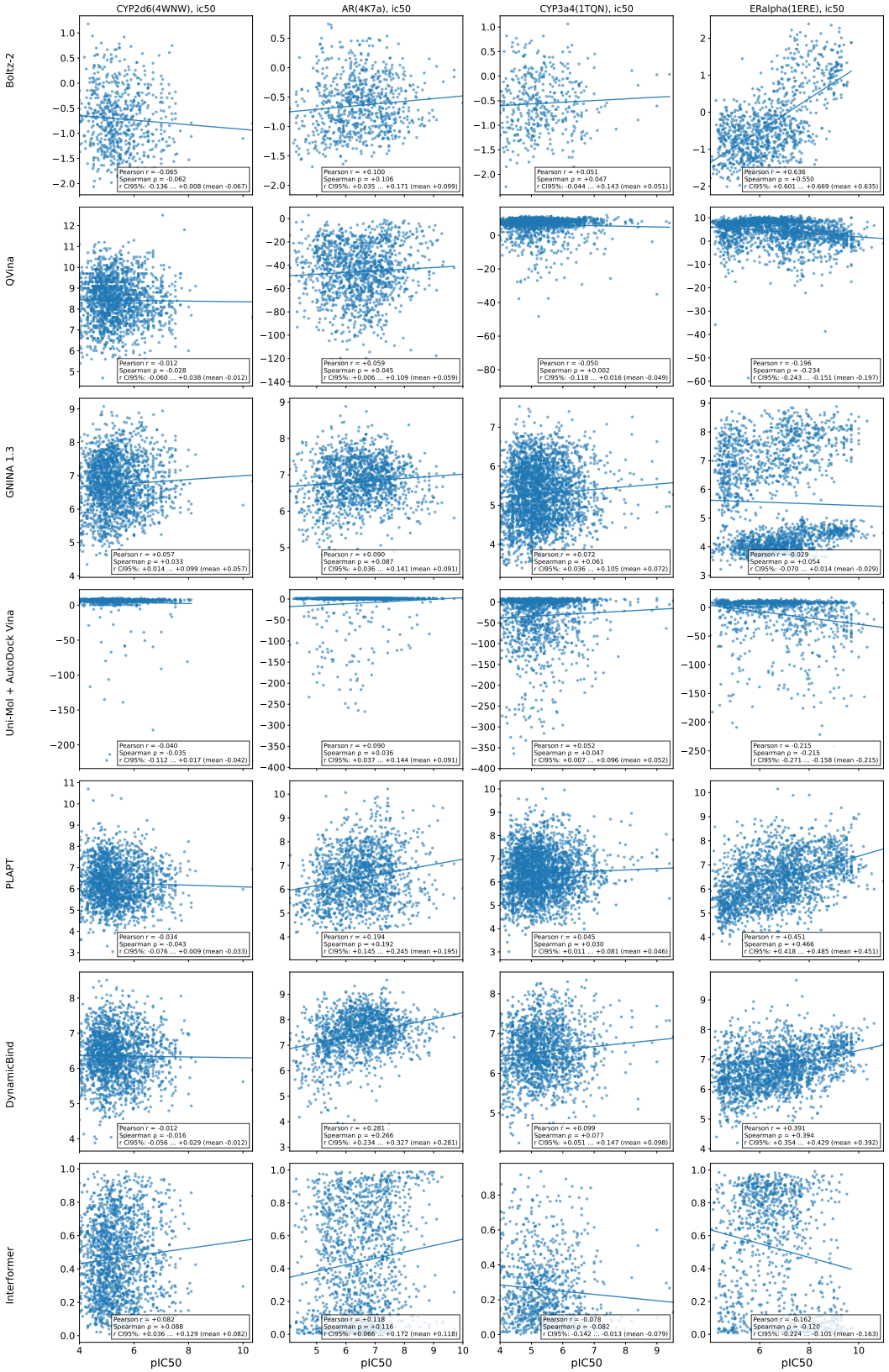

Figure 2: Scoring and ranking power of the assessed docking methods reported as Pearson and Spearman correlation of predicted binding affinities with experimental IC_{50} values.

information was needed—including pocket volume, surface area, and mouth opening characteristics—we used the CASTp web server [50]. These descriptors will be used in the Results section to interpret the observed differences in docking performance across methods and targets.

It has been shown that including internal strain or torsional energies can substantially improve the discrimination between near-native and incorrect poses [51]. For instance, torsional penalties increased the correlation with experimental RMSDs from <0.2 to >0.6. Moreover, ligand efficiency indices provide a size-normalized measure of binding that avoids overestimation of large, but inefficient ligands [52]. Several studies emphasize that relying solely on docking scores is insufficient, since near-native and incorrect poses often differ by as little as 1.7 kcal/mol [53, 54].

Table 5: Binding pocket properties in different proteins

PDB-ID	Druggability Score	Volume	Surface Area	Hydrophobicity	Polarity	Flexibility
1ERE (ERα)	0.741	4.810	102.394	55.7	5	0.043
4K7a (DHT)	0.825	4.85	103.349	50.9	5	0.0
4WNV (CYP2D6)	0.987	676.441	676.441	69.867	25	0.183
1TQN (CYP3a4)	0.612	1.64e+06	1.35e+05	0.117	0.282	22

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the claims made, including the contributions made in the paper and important assumptions and limitations.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed in section 4.2.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper provides complete disclosure of experimental settings, including dataset composition, targets, model architecture, and training parameters. Experimental details and hyperparameter settings are documented in Methods, Appendix B and GitHub https://anonymous.4open.science/r/Docking-Methods-Show-Poor-Transferability-to-Toxicity-Linked-Targets-B541

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The authors publish the preparation of structures for docking, as well as the code for analyzing the docking results. The code is available on GitHub.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: No models were trained in this study.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Uncertainties (bootstrap CI or std) are reported in Appendix C1 (Figers 1,2).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: This information is provided in the Appendix C.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The study complies with ethical standards.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: The study does not produce new methods which could have negative societal impact.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper clearly credits the sources of all external assets. References to the original publications are provided, and the use of these resources complies with their respective licenses and terms of use.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The newly introduced assets are described in detail within the paper and supplementary material. Clear documentation and usage instructions are provided alongside the released code repository, ensuring that other researchers can easily understand, reproduce, and extend the work.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.