DO DEEP LEARNING MODELS REALLY OUTPERFORM TRADITIONAL APPROACHES IN MOLECULAR DOCKING?

Yuejiang Yu, Shuqi Lu, Zhifeng Gao, Hang Zheng & Guolin Ke*
DP Technology
{yuyj, lusq, gaozf, zhengh, kegl}@dp.tech

ABSTRACT

Molecular docking is a technique widely used in drug design that predicts the binding mode of a protein-ligand complex, given a ligand molecule and a ligand binding site (called a "pocket") on a protein. Although many deep learning models have been developed for molecular docking, most of them perform docking on the whole protein, rather than on a given pocket as the traditional molecular docking approaches, which does not meet common needs. Many deep learning models have been developed for molecular docking, while most existing deep learning models perform docking on the whole protein, rather than on a given pocket as the traditional molecular docking approaches, which does not match common needs where pockets are mostly known. Moreover, these models claim to perform better than traditional molecular docking methods, but the comparison is unfair because traditional methods are not designed for docking on the whole protein without a given pocket. In this paper, we design a series of experiments to examine the actual performance of these deep learning models and traditional methods. For a fair comparison, we decompose the docking on the whole protein into two steps, pocket searching and docking on a given pocket, and build pipelines to evaluate traditional methods and deep learning methods respectively. Our findings suggest that deep learning models are good at pocket searching, but traditional methods are better than deep learning models at docking on given pockets. Overall, our work explicitly reveals some potential problems in current deep learning models for molecular docking and provides several suggestions for future improvements.

1 INTRODUCTION

Molecular docking is a widely used technique in drug design that predicts the binding mode of a protein-ligand complex, given a ligand molecule and a ligand binding site called "pocket." Over the last few decades, several molecular docking methods have been proposed, including Trott & Olson (2010); Alhossary et al. (2015); Hassan et al. (2017); Quiroga & Villarreal (2016).

Recently, deep learning has been increasingly used in drug design applications, such as molecular property prediction Zhou et al. (2022); Rong et al. (2020); Wang et al. (2022); Fang et al. (2022) and protein structure predition Jumper et al. (2021). Several recent works Stärk et al. (2022); Lu et al. (2022); Corso et al. (2022) have attempted to apply deep learning to molecular docking. However, most of these works did not follow traditional molecular docking settings. Instead of docking on a given pocket, these deep learning works directly perform docking on the whole protein (called "blind docking" [1]), which does not match common needs in drug design. Besides, although they claimed they are better than traditional molecular docking approaches, the experiment and the evaluation in their papers are questionable. Specifically, we have the following questions.

- *Are the comparisons with traditional approaches fair?* The traditional molecular docking approaches are not designed for blind docking. But the early works Stärk et al. (2022); Lu et al. (2022); Corso et al. (2022) directly apply traditional molecular docking approaches on the whole

[1] We use "molecular docking" to refer to the docking on a given pocket in this paper.

proteins, rather than the pockets, in the experiments. It is obvious that such a comparison is not fair for traditional molecular docking approaches.

- *What are these deep learning models really good at? Pocket searching or molecular docking?* In traditional methods, blind docking is generally divided into two steps: pocket searching and docking. Although these works combine pocket searching and molecular docking together, they simply evaluate the performance of the final complex conformation. We do not know whether the gain is brought by better pocket searching or better molecular docking. Thus, we actually do not know whether these models can outperform the traditional approaches in molecular docking or not, given the same protein pockets.

To answer the above questions, we design a series of experiments in Sec. 3, and have the following findings.

- *Traditional molecular docking approaches are still better than deep learning models, when given the same pockets.* Besides, the two-stage traditional approaches (traditional pocket searching + molecular docking) can outperform most of the deep learning models.

- *Deep learning models are actually good at pocket searching.* Although their performance on molecular docking is worse, deep blind docking models actually perform well in pocking searching. Besides, we find there is still a large room to reach the performance upper bound of pocket searching.

- *Deep Learning models are full of potential.* DiffDock Corso et al. (2022) is currently the best method in pocket searching, and its performance in molecular docking is very close to traditional approaches. It is no doubt that deep learning models will outperform traditional approaches in the future.

2 Related Work

Pocket Searching Approaches To find potential new biological targets on proteins, several pocket searching approaches have been proposed. Fpocket Le Guilloux et al. (2009) uses Voronoi tessellation and alpha spheres to locate, rank and describe pockets. P2Rank Krivák & Hoksza (2018) is a convolutional neural network based tool. PointSite Yan et al. (2022) further leverages the 3D spatial information, based on 3D U-Net model Çiçek et al. (2016).

Traditional Molecular Docking Approaches AutoDock and its variants Trott & Olson (2010); Ravindranath et al. (2015); Eberhardt et al. (2021); Santos-Martins et al. (2021) are popular tools in molecular docking to efficiently predict binding poses and affinities of ligands according to the scoring function and fast search methods. Vinardo Quiroga & Villarreal (2016), QVina2 Alhossary et al. (2015), QVina-W Hassan et al. (2017), and Smina Koes et al. (2013) are developed from AutoDock to further improve docking's searching and scoring power. To handle ultra-large ligand datasets, GPU-accelerated docking engines such as Uni-Dock Yu et al. (2022) and AutoDock-GPU Santos-Martins et al. (2021) have broadened docking's throughput dramatically.

Deep Learning Models for Docking Recent advances in deep learning-based molecular docking focus on blind docking, where the protein's pocket is unknown, and the molecular positions and conformations are directly predicted. EquiBind Stärk et al. (2022) proposed a SE(3)-equivariant geometric deep learning model to predict the molecular positions and conformations by directly predicting 3D atom coordinates. Several later works Zhang et al. (2022); Lu et al. (2022); Corso et al. (2022) also focus on blind docking. TANKBind Lu et al. (2022) proposes a two-stage deep docking framework that segments the whole protein into functional blocks and predicts their interactions with the ligand using a trigonometry-aware architecture. Then, the binding structure is prioritized based on the predicted interactions. DiffDock Corso et al. (2022) randomly samples a molecular conformation and predicts the molecular atom coordinates with a denoising diffusion probability model starting from the random molecular conformation.

3 Experiments

We design a series of experiments, to ensure a fair comparison between traditional molecular docking approaches and deep learning based models. To address the limitations of traditional approaches in searching pockets, we have incorporated additional pocket searching tools or utilized the pockets identified by deep learning models. The details of the experimental designs are in the following subsection.

3.1 Experimental Settings

Data We used the same benchmark dataset as Equibind Stärk et al. (2022) and DiffDock Corso et al. (2022) with the test set collected from PDBBind Liu et al. (2017). The conformations of molecular ligands are initialized by RDKit, and the protein structures are sourced from https://anonymous.4open.science/r/DiffDock.

Deep learning models The below models are used as baselines.

- EquiBind Stärk et al. (2022), a docking tool that predicts the conformation without binding site knowledge. We directly use the number reported in DiffDock Corso et al. (2022) paper.
- TANKBind Lu et al. (2022), we used the number reported in DiffDock Corso et al. (2022) paper.
- TANKBind*, we reproduced TANKBind's results, based on its officially released model weights and source codes [2]. We ran the model three times with different random seeds, and reported the mean and std like in DiffDock Corso et al. (2022).
- DiffDock Corso et al. (2022), a generative model of binding pose prediction under the setting from the corresponding paper Corso et al. (2022), which generates 40 samples with 20 diffusion steps. We directly used the reported number in their paper.
- DiffDock*, we reproduced DiffDock's results, based on its officially released model weights and source codes [3]. We ran the results three times with different random seeds, and reported the mean and std the same way as TANKBind*.

Traditional approaches To have a fair comparison with deep learning models in the blind docking setting, we applied the two-stage solution for the traditional approaches: pocket searching, followed by molecular docking. For the molecular docking, we use Uni-dock Yu et al. (2022), a highly efficient GPU-accelerated docking tool. We set up five different configurations using different pocket searching methods, as follows:

- 1. Fpocket + Uni-dock, we used pockets found by Fpocket Le Guilloux et al. (2009). Specifically, we used the pocket with the best "fpocket score". We created an axis-parallel cube with 30 Å edge size for molecular docking, using the geometric center of the predicted pocket atoms as the center.
- 2. P2Rank + Uni-dock, we used pockets found by P2Rank Krivák & Hoksza (2018). Specifically, we used the rank-1 pocket predicted by P2Rank. We created an axis-parallel cube with 30 Å edge size for molecular docking, using the geometric center of the predicted pocket atoms as the center.
- 3. PointSite + Uni-dock, we used PointSite Yan et al. (2022), a 3D U-Net model to find the pocket atoms. PointSite only predicts one pocket for one protein, which we used directly. We created an axis-parallel cube with 30 Å edge size for molecular docking, using the geometric center of the predicted pocket atoms as the center.
- 4. DiffDock* + Uni-dock, we used pockets found by DiffDock. Specifically, we selected the top-1 conformation for each protein based on DiffDock*'s predicted confidence scores. Then, we created a minimal axis-parallel rectangular cuboid that could cover all ligand molecular atoms. Finally, we enlarged the cuboid in three axes by 5 Å, and used it for molecular docking. Note that here we only used the pockets found from DiffDock in this setting, rather than its predicted molecular conformations.

[2] https://github.com/luwei0917/TankBind

[3] https://github.com/gcorso/DiffDock

- 5. GT pocket + Uni-dock, we used the ground-truth (GT) pockets directly. Similar to DiffDock's pockets, we created a minimal axis-parallel rectangular cuboid that could cover all ground-truth molecular atoms. Then, we enlarged the cuboid in three axes by 5 Å, and used the enlarged cuboid for molecular docking. We set up this experiment to demonstrate the performance upper bound of traditional docking approaches when given the correct pockets.

Each of the above settings was run three times with different random seeds, and we reported the means and standard deviations for evaluation metrics. We also provided a repository [4] for reproducing our results. Notably, in the repository, we used the open-source Autodock Vina as the traditional docking engine, as Uni-dock is not open-sourced yet. However, the repository should be able to reproduce our results, considering that Uni-dock can be recognized as the GPU-accelerated version of Autodock Vina.

Evaluation metrics The evaluation also follows DiffDock Corso et al. (2022). In particular, we first select the top-k ($k \in \{1, 5\}$) conformations based on docking scores (or confidence scores predicted by deep learning models), for each protein. Then, we compute the heavy atoms' RMSD for each of these conformations, using ground-truth conformations. Next, we select the best conformation with the minimal RMSD for each protein. Finally, we report the percentage of best conformation with RMSD $< m$Å ($m \in \{1, 2\}$) across all proteins, as well as the median RMSD of the best conformation on all proteins. Notably, the RMSD between two conformations considers the symmetry permutations, which is consistent with DiffDock's evaluation.

3.2 Results

Blind docking performance From the end-to-end blind docking performance shown in Table 1, we have the following findings. 1) Firstly, traditional approaches (with P2Rank and PointSite) outperform deep learning approaches for accurate docking conformations (percentage of RMSD < 1Å). 2) Even for less accurate metrics (percentage of RMSD < 2Å), traditional approaches still outperform EquiBind and TankBind. 3) although DiffDock performs well on RMSD < 2Å, Uni-Dock with DiffDock's pocket performs even better, indicating that its performance is attributed to the better pocket searching ability.

Molecular docking performance We can compare DiffDock* with "DiffDock* +Uni-dock", to examine the performance of molecular docking, when given the same pockets. From the results, it is clear that "DiffDock* +Uni-dock" consistently outperforms DiffDock*. This suggests that traditional molecular docking approaches are still better than deep learning models when using the same pocket.

Pocket searching performance We can compare Uni-dock with different pocket searching methods, to examine the pocket searching performance. 1) It is easy to find that PointSite is the best, and fpocket is the worst. 2) The pockets found by DiffDock* are quite good, outperforming all existing pocket searching tools. We suppose the gain is from the additional molecule inputs used in DiffDock, since fpocket, P2Rank, and PointSite only take the protein as input. 3) When using the ground-truth pockets, the performance of Uni-Dock largely outperforms all other methods. This indicates there is still a large room to reach the performance upper bound of pocket searching.

3.3 Discussions

Based on the experimental results presDented above, it can be concluded that current deep learning models excel at pocket searching, but not molecular docking. We have the following suggestions for future deep learning models for molecular docking:

- Focus on molecular docking with given pockets, rather than blind docking. In real-world applications, pockets are usually known and fixed in a drug design project.

[4] https://github.com/pkuyyj/Blind_docking

[5] Here, we use the confidence model provided in https://github.com/gcorso/DiffDock to identify molecular conformations with RMSD < 1Å. Although the confidence model here is originally trained based on molecular conformations with RMSD < 2Å, it still provides a reasonable indicator of model performance.

Table 1: Performance of blind docking.

	Method	Top-1 RMSD(Å)			Top-5 RMSD(Å)		
		% < 1Å (↑)	% < 2Å (↑)	Med. (↓)	% < 1Å (↑)	% < 2Å (↑)	Med. (↓)
Deep Learning	EquiBind	-	5.5±1.2	6.2±0.3	-	-	-
	TANKBind		20.4±2.1	4.0±0.2		24.5±2.1	3.4±0.1
	TANKBind*	2.66±0.26	18.18±0.6	4.2±0.05	4.13±0.0	20.39±0.45	3.5±0.04
	DiffDock		38.2±2.5	3.30±0.3		44.7±2.6	2.40±0.2
	DiffDock*	15.41±0.49[5]	36.62±0.35	3.31±0.03	21.58±0.38[5]	44.19±0.49	2.37±0.06
Traditional	Fpocket + Uni-dock	13.33±0.4	18.7±0.13	13.2±0.26	19.16±0.39	27.32±0.69	8.3±0.25
	P2Rank + Uni-dock	19.31±1.07	28.6±1.17	6.4±0.22	27.76±1.03	39.18±1.03	3.76±0.06
	PointSite + Uni-dock	21.36±1.65	32.12±0.93	5.54±0.46	31.38±0.86	46.06±0.69	2.52±0.18
Better Pocket + Traditional	DiffDock* + Uni-dock	25.49±0.6	38.93±0.23	4.14±0.07	36.97±1.05	51.07±1.06	1.93±0.12
	GT pocket + Uni-dock	32.77±0.38	51.11±0.6	1.89±0.04	47.5±0.23	67.59±0.94	1.11±0.02

- Pocket searching itself is an important problem, and still has plenty of room for improvement. Moreover, if the additional ligand molecules can be used as model inputs (like in DiffDock), the performance may be further improved.
- For end-to-end blind docking models, it's important to ensure a fair comparison in experiments. Specifically, you should first use pocket searching approaches (or directly use the same pockets as your models), then apply the traditional molecular docking methods in the pockets, rather than in the whole proteins.
- The deep learning models, particularly DiffDock, show tremendous potential. From the results, we can find that DiffDock is the best model in pocket searching, and achieve almost comparable performance with traditional approaches. We believe the deep learning models can be further improved in the near future.

4 Conclusion

Several deep learning models have been proposed for molecular docking. However, these models focus on blind docking, which differs from the docking in a given pocket in traditional approaches. Besides, in their experiments, traditional approaches are often used on the whole protein rather than in a given pocket, making the comparison with deep learning models unfair. To examine the actual performance of deep learning models, we design a series of experiments to compare them with traditional molecular docking approaches. Our experimental results indicate that traditional molecular docking approaches still outperform deep learning models when using the same pockets. Based on our findings, we suggest the community and future works on molecular docking can correctly evaluate the traditional approaches. Besides, since blind docking actually does not align with common real-world applications, it may be more effective to address pocket searching and molecular docking (on given pockets) separately.

Acknowledgments

We thank Gengmo Zhou, Junhan Chang and many colleagues in DP Technology for their great help in this project. Most experiments are carried out on the cloud platform Bohrium™(https://bohrium.dp.tech).

References

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

Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*, pp. 424–432. Springer, 2016.

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.

Jerome 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.

Xiaomin Fang, Lihang Liu, Jieqiong Lei, Donglong He, Shanzhuo Zhang, Jingbo Zhou, Fan Wang, Hua Wu, and Haifeng Wang. Geometry-enhanced molecular representation learning for property prediction. *Nature Machine Intelligence*, 4(2):127–134, 2022.

Nafisa M Hassan, Amr A Alhossary, Yuguang Mu, and Chee-Keong Kwoh. Protein-ligand blind docking using quickvina-w with inter-process spatio-temporal integration. *Scientific reports*, 7 (1):1–13, 2017.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

David Ryan Koes, Matthew P Baumgartner, and Carlos J Camacho. Lessons learned in empirical scoring with smina from the csar 2011 benchmarking exercise. *Journal of chemical information and modeling*, 53(8):1893–1904, 2013.

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

Vincent Le Guilloux, Peter Schmidtke, and Pierre Tuffery. Fpocket: an open source platform for ligand pocket detection. *BMC bioinformatics*, 10(1):1–11, 2009.

Zhihai Liu, Minyi Su, Li Han, Jie Liu, Qifan Yang, Yan Li, and Renxiao Wang. Forging the basis for developing protein–ligand interaction scoring functions. *Accounts of chemical research*, 50 (2):302–309, 2017.

Wei Lu, Qifeng Wu, Jixian Zhang, Jiahua Rao, Chengtao Li, and Shuangjia Zheng. Tankbind: Trigonometry-aware neural networks for drug-protein binding structure prediction. *bioRxiv*, pp. 2022–06, 2022.

Rodrigo Quiroga and Marcos A Villarreal. Vinardo: A scoring function based on autodock vina improves scoring, docking, and virtual screening. *PloS one*, 11(5):e0155183, 2016.

Pradeep Anand Ravindranath, Stefano Forli, David S Goodsell, Arthur J Olson, and Michel F Sanner. Autodockfr: advances in protein-ligand docking with explicitly specified binding site flexibility. *PLoS computational biology*, 11(12):e1004586, 2015.

Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying Wei, Wenbing Huang, and Junzhou Huang. Self-supervised graph transformer on large-scale molecular data. *Advances in Neural Information Processing Systems*, 33:12559–12571, 2020.

Diogo Santos-Martins, Leonardo Solis-Vasquez, Andreas F Tillack, Michel F Sanner, Andreas Koch, and Stefano Forli. Accelerating autodock4 with gpus and gradient-based local search. *Journal of chemical theory and computation*, 17(2):1060–1073, 2021.

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

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.

Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4(3):279–287, 2022.

Xu Yan, Yingfeng Lu, Zhen Li, Qing Wei, Xin Gao, Sheng Wang, Song Wu, and Shuguang Cui. Pointsite: A point cloud segmentation tool for identification of protein ligand binding atoms. *Journal of Chemical Information and Modeling*, 62(11):2835–2845, 2022. doi: 10.1021/acs.jcim.1c01512. URL `https://doi.org/10.1021/acs.jcim.1c01512`. PMID: 35621730.

Yuejiang Yu, Chun Cai, Zhengdan Zhu, and Hang Zheng. Uni-dock: A gpu-accelerated docking program enables ultra-large virtual screening. 2022.

Yangtian Zhang, Huiyu Cai, Chence Shi, Bozitao Zhong, and Jian Tang. E3bind: An end-to-end equivariant network for protein-ligand docking. *arXiv preprint arXiv:2210.06069*, 2022.

Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-mol: A universal 3d molecular representation learning framework. 2022.