UMD-fit: Generating Realistic Ligand Conformations for Distance-Based Deep Docking Models

Anonymous Author(s)
Affiliation
Address
email

Abstract

Recent advances in deep learning have enabled fast and accurate prediction of protein-ligand binding poses through methods such as Uni-Mol Docking. These techniques utilize deep neural networks to predict interatomic distances between proteins and ligands. Subsequently, ligand conformations are generated to satisfy the predicted distance constraints. However, directly optimizing atomic coordinates often results in distorted, and thus invalid, ligand geometries; which are disastrous in actual drug development. We introduce UMD-fit as a practical solution to this problem applicable to all distance-based methods. We demonstrate it as an improvement to Uni-Mol Docking, which retains the overall distance prediction pipeline while optimizing ligand positions, orientations, and torsion angles instead. Experimental evidence shows that UMD-fit resolves the vast majority of invalid conformation issues while maintaining accuracy.

1 Introduction

Molecular docking refers to the precise prediction of protein-ligand binding configurations. Successful docking methods enable vast applications in drug design, from fast virtual screening of small molecules to improved insights of structure-activity relationships (SAR), which can help medicinal chemists understand the binding mechanism of molecules with target proteins . Traditional docking software, such as AutoDock Vina and Schrödinger GLIDE, [8, 3], has relied on algorithms that optimize the conformation and orientation of the ligand within the protein binding site. However, they are unable to describe certain interactions due to the simplified scoring functions owing to maintain a reasonable cost and speed [10]. Recent advances in deep learning shed light on new possibilities for predicting ligand binding poses. When applied to molecular docking, they could model large, highly flexible ligands such as peptides and long-range interactions without incurring in prohibitive costs, enabling higher accuracy than traditional docking methods.

Recently, many studies have proposed deep learning-based protein-ligand complex structure prediction, achieving significant improvements in quantitative metrics such as RMSD [6, 2, 12]. However, as pointed out by [1] they have considerable defects in the rationality of small molecule conformations such as abnormal bond lengths, changes in chirality or wrong geometries in aromatic rings; which is unacceptable in drug development applications and hinders their applicability in SAR studies. Therefore, an optimization method to prevent these issues that respects the flexibility around rotatable bonds and preserves the stereochemistry, thus producing plausible ligand conformations by default, would increase reliability and adoption of deep learning in molecular docking.

Herein, we identify many of the issues related to unreasonable ligand conformations pointed out by [1] to be a consequence of direct optimization of coordinates over a model-parametrized loss function; and propose UMD-fit these problems in protein-ligand binding pose prediction. UMD-fit optimizes ligand translation, orientation, and inner torsion angles instead of directly optimizing

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

atomic coordinates. Stereochemical configurations are also enforced during the optimization process. This allows the resulting conformation to intrinsically meet rigid geometry requirements. To address issues arising from equivalence between atoms, we used the symmetric RMSD as the final metric. We combined Uni-Mol Docking[12] and UMD-fit (Uni-Mol Docking with fit conformations) for a practical application, modifying the final optimization from the predicted distance matrix, while maintaining the rest of the inference pipeline intact. Experiments with different sets of protein-ligand complexes confirmed the improved plausibility of predictions while showing little degradation in quantitative metrics such as RMSD.

2 Methods

Uni-Mol Docking. We adapted the trained model in Uni-Mol [12] and modified the inputs and optimization process in the inference setup. To briefly summarize, the original model presents three main blocks: a protein pocket module, a ligand module, and a joint protein-ligand block which predicts a final inter-atomic distance matrix. Specifically, the distance matrix has shape $d_{ij} \in R^{N \times N}$, where $N = N_l + N_p$ the sum of protein and ligand atoms. Ligand conformations in terms of atomic coordinates ($\boldsymbol{C}_l = \{\boldsymbol{c}_1, \cdots, \boldsymbol{c}_{N_l}\}^\top \in R^{N_l \times 3}$) are then obtained by means of gradient-based optimization with a weighted loss function and the LBFGS optimizer. The optimization problem is formulated in Uni-Mol Docking as

$$\min_{\boldsymbol{C}} \sum_{i,j=1}^{N_l} \| \|\boldsymbol{c}_i - \boldsymbol{c}_j\| - d_{ij} \|_2^2 \cdot w_{ij}, w_{ij} = \begin{cases} 1 & \text{if } d_{ij} < 8.0\text{Å} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

without further introducing geometric constraints. The major difference between UMD-fit and Uni-Mol Docking is that we use a different parameterization for ligand conformations.

6+T Parametrization. Instead of using *free gas parameterization* of conformations like Uni-Mol Docking[12] or [6], i.e. directly optimizing atomic coordinates, UMD-fit introduces the 6+T parameterization which retains the intrinsic degrees of freedom (*d.o.f.*) of ligand conformations. As is well-known in rigid docking (rigid protein, flexible molecule), the protein-ligand structure can be accurately described by the molecular conformation of the ligand with T torsional *d.o.f.* $(t_1, \cdots, t_T)$, and the relative pose of the ligand to the protein, parameterized as a roto-translation $(\boldsymbol{R}, \boldsymbol{x}) \in SO(3) \times \mathbb{R}^3$ with 6 *d.o.f.*. This new parameterization introduces a total of $6 + T$ *d.o.f.*, much lower than the $3 \times N_l$ *d.o.f.* in the free gas parameterization. The optimization problem can then be formulated as

$$\min_{\boldsymbol{R}, \boldsymbol{x}, \{t_i\}_{r=1}^T} \sum_{i,j=1}^{N_l} \| \|\tilde{\boldsymbol{c}}_i - \tilde{\boldsymbol{c}}_j\| - d_{ij} \|_2^2 \cdot w_{ij}, w_{ij} = \begin{cases} 1 & \text{if } d_{ij} < 8.0\text{Å} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where atomic coordinates $\{\tilde{\boldsymbol{c}}_i\}_{i=1}^{N_l}$ are obtained from $(\boldsymbol{R}, \boldsymbol{x}, \{t_i\}_{r=1}^T)$ and known rigid parameters of bond lengths and angles in a differentiable manner. As such, optimizing Eq. (2) is equivalent to optimizing Eq. (1) under the geometric constraints of rigid substructures, thus yielding more realistic ligand conformations. Notably, similar approaches have been explored in other recent deep docking methods [4, 2]. Crucially, we further introduce a kabsch alignment after torsion updates, thus making the degree of freedoms in translation, rotation and torsional orthogonal in the tangent space. This step is not needed for practical convergence, although it accelerates it. The process is exemplified in pseudo-code in the appendix. Unlike previous works which used gradient-free techniques such as differential evolution[4, 7], we keep differentiability and use the LBFGS algorithm as in [12] for fast convergence.

Stereochemistry Preservation. Previous work [13] described an inexpensive protocol to produce diverse molecular conformations with open source library RDKit[5]. A similar process was used for the generation of diverse conformers as an input to Uni-Mol [12]. However, we identified that some molecules exhibit changes in stereochemistry when their torsions are randomized following the protocol in [13]. Let a torsion be composed by atoms i, j, k, l the RDKit torsion update utility only rotates l and its linked atoms. This is problematic when k has multiple atoms bonded as it can change the stereochemistry. Therefore the torsion update was modified such that under a torsion update all atoms closer to k than j would move together, thus ensuring all the resulting diverse conformers presented the same stereochemistry.

3 Results

Evaluation. The datasets used for evaluation were CASF-2016 [9] test set and the PoseBusters [1] set. In both datasets, the same protocol settings described in [12] except for the optimization routine was followed. Mean RMSD and percentage of compounds with RMSD lower than different cutoffs (0.5, 1.0, 1.5, 2.0, 3.0, 5.0 Å) are used as the primary performance quantitative metric to control potential degradation relative to the baseline model. For qualitative and quantitative assessment of outputs plausibility, as well as error identification, we use adapted scripts from [1]. Notably, we introduce an improvement over the original Uni-Mol paper[12] in the RMSD calculation, as we introduce symmetric RMSD ("symRMSD") to take into account symmetric molecular structures and not incur in excessive penalties.

Representative unrealistic conformations caused by the original method proposed in [12] and their plausible counterparts with 6+T+S strategy are depicted Figure 1, (left) as well as a docking result showcasing poor chemical accuracy in specific functional groups of Uni-Mol Docking baseline despite the correct overall placement, and the correction under the 6+T+S strategy (right).

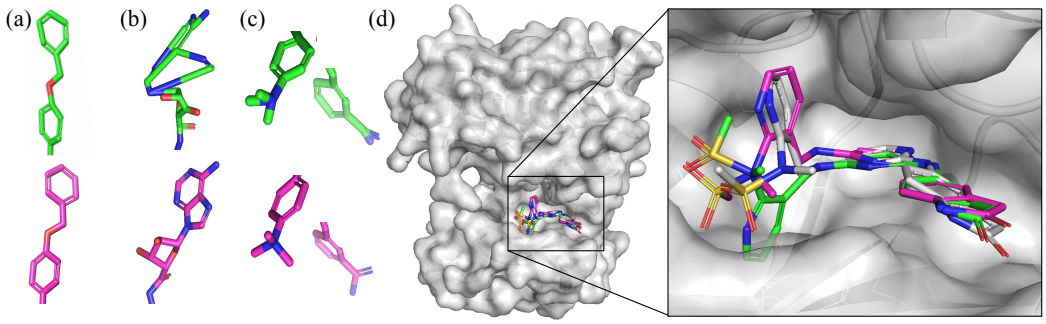

Figure 1: Uni-Mol Docking (green) and UMD-fit (fuchsia) outputs. (a) abnormal ring geometries in the phenyl and benzene core (PDB: 3MSS). (b) invalid bond lengths, ring geometries (purine) and chirality changes (hydrofuran) (PDB: 3AG9). (c) invalid bond lengths and internal steric clashes in terminal groups (trimethyl, amidine) (PDB: 1LPG). (d) docking result of Uni-Mol Docking and UMD-fit against human focal adhesion kinase (FAK) (PDB: 6YT6) (grey); where the baseline presents unrealistic geometries in the sulfonamide group, and the oxidanylidene is not in the same plane as the indole, contrary to what is expected for an aromatic ring.

Quantitative results for the PoseBusters set are shown in Table 1, and details on the failure modes following the [1] report style are given in Figure 2 for both PoseBusters and CASF-2016 test set. A plausibility comparison between UMD-fit and relevant deep learning methods evaluated in [1] is provided in Table 2. CASF-2016 test set quantitative results are detailed in the Appendix.

As shown in Figure 2, UMD-fit addresses the majority of failing plausibility checks resulting in unphysical conformations, except the steric clashes with the protein. UMD-fit can effectively increase the number of total protein-ligand complexes with a correct pose (RMSD $\leq$ 2.0 Å) passing automated plausibility tests by more than 2-fold, in the CASF-2016 test set, and a similar relative improvement is observed for the PoseBusters test set. Furthermore, as evidenced in Table 1, UMD-fit which builds on top of the 6+T protocol does not negatively affect the overall RMSD of the docking method in a significant way, especially when symmetric RMSD is considered, indicating that the 6+T parametrization is not an obstacle for accurate conformer optimization when combined with the LBFGS optimizer, even under complex and non-smooth loss functions.

Table 1: Performance for baseline Uni-Mol Docking and 6+T+S strategy in the PoseBusters set

			% $\leq$ symRMSD					
Strategy	RMSD (Å)	symRMSD (Å)	0.5Å	1.0Å	1.5Å	2.0Å	3.0Å	5.0Å
Baseline	3.59	3.51	0.93	12.61	28.97	39.01	59.35	75.70
UMD-fit	3.62	3.53	1.40	10.98	28.73	40.42	57.24	74.53

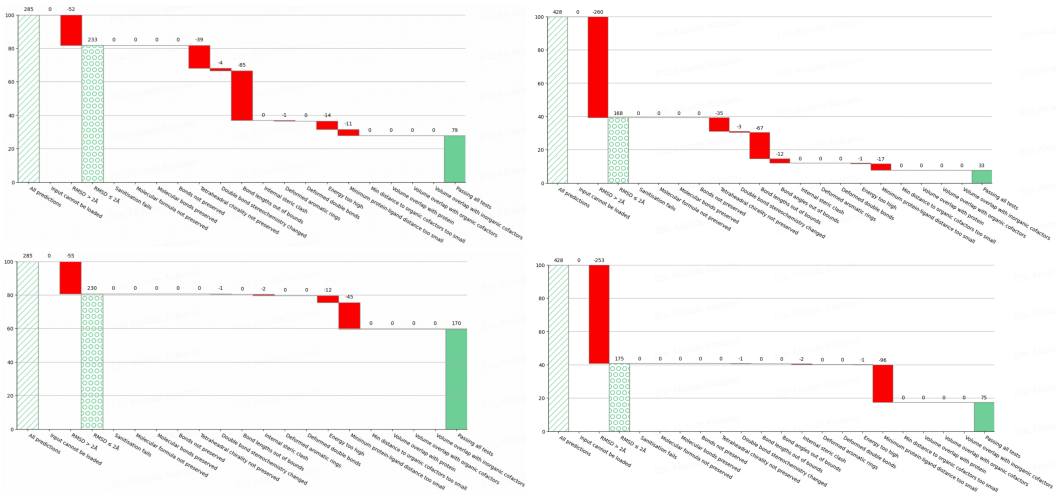

Figure 2: Failure mode analysis of Uni-Mol inference in the CASF-2016 [9] test set (left) and PoseBusters [1] set (right) with baseline (top) and UMD-fit (bottom) strategies.

Table 2: Plausibility of predictions in the PoseBusters set (428 complexes); comparison data from [1]

Method	UMD-fit	DiffDock[2]	DeepDock[7]	TankBind[6]	Uni-Mol[12]
# RMSD $\leq$ 2Å	**175**	162	72	64	98
# Passing all tests	**75**	58	20	11	8

4 Discussion

For the challenging PoseBusters set, there is a difference in the performance reported in this work and that of [1] for the Uni-Mol Docking baseline, which might be explained by differences in the inference protocol. We followed the protocol described in the original Uni-Mol paper [12] for both the CASF-2016 [9], where the results are in line with the original paper, and the PoseBusters [1] set.

The described improvements make UMD-fit, when coupled with Uni-Mol Docking, the best-performing machine learning method of those reported in [1] for the PoseBusters test set as shown in Table 2, followed by DiffDock [2], with the caveat that UMD-fit has been combined with a pocket-only model whereas DiffDock is a whole protein docking model. UMD-fitis also model agnostic, being applicable to 3 ([12, 7, 6]) out of 5 deep learning models tested in [1] Further work might explore the combination of an automated pocket finding tool or a whole protein docking model and UMD-fit for extension to blind docking.

Among predictions below 2Å RMSD in UMD-fit, the majority of remaining failure cases can be rooted to steric clashes between the ligand and protein atoms. Such cases could be resolved by combining a learning method like UMD-fit and a more physics-inspired traditional docking method such as UniDock with biased docking. Recent works have explored this direction with great success[11].

Nonetheless, a substantial drop in performance can be observed when comparing the CASF-2016[9] test set and the PoseBusters set, in line with results from [1] showing reduced performance of machine learning models compared to classical docking algorithms. The performance gap can be explained by the low difficulty of CASF-2016[9] test set, already pointed out by [14], and by the lack of generalization of ML docking models. Such an issue might benefit from more work characterizing the data efficiency and extrapolation ability of different architectures and training regimes.

References

[1] 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, 2023.

[2] Gabriele Corso, Hannes Stärk, Bowen Jing, Regina Barzilay, and Tommi Jaakkola. Diffdock: Diffusion steps, twists, and turns for molecular docking. *International Conference on Learning Representations (ICLR)*, 2023.

[3] Leonardo Ferreira, Ricardo dos Santos, Glaucius Oliva, and Adriano Andricopulo. Molecular docking and structure-based drug design strategies. *Molecules*, 20(7):13384–13421, 2015.

[4] Bowen Jing, Gabriele Corso, Jeffrey Chang, Regina Barzilay, and Tommi Jaakkola. Torsional diffusion for molecular conformer generation. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 24240–24253. Curran Associates, Inc., 2022.

[5] Gregory A. Landrum. Rdkit: Open-source cheminformatics. release 2014.03.1. 2014.

[6] 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*, 2022.

[7] Oscar Méndez-Lucio, Mazen Ahmad, Ehecatl Antonio del Rio-Chanona, and Jörg Kurt Wegner. A geometric deep learning approach to predict binding conformations of bioactive molecules. *Nature Machine Intelligence*, 3(12):1033–1039, 2021.

[8] Nataraj S. Pagadala, Khajamohiddin Syed, and Jack Tuszynski. Software for molecular docking: A review. *Biophysical Reviews*, 9(2):91–102, 2017.

[9] Minyi Su, Qifan Yang, Yu Du, Guoqin Feng, Zhihai Liu, Yan Li, and Renxiao Wang. Comparative assessment of scoring functions: The casf-2016 update. *Journal of Chemical Information and Modeling*, 59(2):895–913, 2019.

[10] Zhe Wang, Huiyong Sun, Xiaojun Yao, Dan Li, Lei Xu, Youyong Li, Sheng Tian, and Tingjun Hou. Comprehensive evaluation of ten docking programs on a diverse set of protein–ligand complexes: The prediction accuracy of sampling power and scoring power. *Physical Chemistry Chemical Physics*, 18(18):12964–12975, 2016.

[11] He Yang, Hongrui Lin, Yannan Yuan, Yaqi Li, Rongfeng Zou, Gengmo Zhou, Linfeng Zhang, and Hang Zheng. *Synergistic application of molecular docking and machine learning for improved protein-ligand binding pose prediction*, 2023.

[12] Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-mol: A universal 3d molecular representation learning framework. In *The Eleventh International Conference on Learning Representations*, 2023.

[13] Gengmo Zhou, Zhifeng Gao, Zhewei Wei, Hang Zheng, and Guolin Ke. Do deep learning methods really perform better in molecular conformation generation?, 2023.

[14] Hui Zhu, Jincai Yang, and Niu Huang. Assessment of the generalization abilities of machine-learning scoring functions for structure-based virtual screening. *Journal of Chemical Information and Modeling*, 62(22):5485–5502, 2022.