ArtiDock: fast and accurate machine learning approach to protein-ligand docking based on multimodal data augmentation

Taras Voitsitskyi[1,4], *Semen Yesylevskyy*[1,3,4,6], *Volodymyr Bdzhola*[2], *Roman Stratiichuk*[1,5], *Ihor Koleiev*[1,4], *Zakhar Ostrovsky*[1], *Volodymyr Vozniak*[1], *Ivan Khropachov*[1], *Pavlo Henitsoi*[1], *Leonid Popryho*[1], *Roman Zhytar*[1], *Alan Nafiiev*[1], *Serhii Starosyla*[1]

[1] - Receptor.AI Inc., 20-22 Wenlock Road, London N1 7GU, United Kingdom.
[2] - Institute of Molecular Biology and Genetics of The National Academy of Sciences of Ukraine, 150 Zabolotnogo Str., 03143, Kyiv, Ukraine.
[3] - Institute of Organic Chemistry and Biochemistry, Czech Academy of Sciences, CZ-166 10 Prague 6, Czech Republic.
[4] - Department of Physics of Biological Systems, Institute of Physics of The National Academy of Sciences of Ukraine, 46 Nauky Ave., 03038, Kyiv, Ukraine.
[5] - Department of Biophysics and Medical Informatics, Educational and Scientific Centre “Institute of Biology and Medicine", Taras Shevchenko Kyiv National University, 64 Volodymyrska Str., 01601, Kyiv, Ukraine.
[6] - Department of Physical Chemistry, Faculty of Science, Palacký University Olomouc, 17. listopadu 12, 771 46 Olomouc, Czech Republic.

We present ArtiDock - the deep learning technique for predicting ligand poses in the protein binding pockets (aka “AI docking”), which is based on augmenting inherently limited training data with algorithmically generated artificial binding pockets and the ensembles of representative conformations of the ligand-protein complexes obtained from MD simulations. Performance of ArtiDock is compared systematically with other AI docking techniques and conventional docking programs on the PoseBusters dataset, which is dedicated for benchmarking the AI pose prediction algorithms. ArtiDock outperforms the best AI docking techniques and the major conventional docking programs, being at least an order of magnitude faster while providing superior accuracy in terms of RMSD and additional ligand pose correctness metrics. The influence of data augmentation on the model performance is evaluated and the perspectives of further development are discussed.

1. Introduction

Classical docking is one of the foundational tools in computational drug discovery for more than two decades. During this extended period of time the docking technology has witnessed an impressive scaling up and a universal adoption in academia and industry, while remaining

remarkably unchanged in terms of used algorithms. Indeed, all the major docking scoring functions were established decades ago and did not evolve much since then. Their fundamental limitations, such as poor handling of the metal atoms, coordination bonds, polarization, charge transfer and entropic contribution from water also persist with no clear trend for improvement. There is currently an implicit consensus in the community that classical docking has reached the practical plateau of accuracy and no significant progress could be made without a paradigm change [1,2].

Such a change has emerged in recent years with the appearance of the Machine Learning (ML) techniques for ligand pose prediction, often colloquially called an “AI docking”. In contrast to the classical docking, which is based on minimization of some physics-based scoring function, these techniques leverage a completely data-centric approach by learning from experimentally determined protein-ligand complexes.

The first generation of the ML docking techniques used simple and lightweight model architectures and demonstrated results that were subpar to conventional docking, while being much faster [3,4]. The issue with their accuracy was attributed to insufficiently sophisticated internals, which missed some important hidden correlations between the optimal ligand pose and the structures of the ligand and the target protein. This resulted in the appearance of the second generation of ML docking models based on the Deep Learning (DL) approach. This shift was largely inspired by the phenomenal success of the AlphaFold [5,6] and the overall rise of generative AI based on transformers architecture. The DL docking models like DiffDock [7], UniMol [8] and AlphaFold-latest [9] have shown an impressive boost of accuracy that comes, however, at the expense of very complex architectures, large model sizes and slow training and inference.

Currently the most “heavyweight” ML models (such as AlphaFold-latest) are more accurate than classical docking while being several orders of magnitude slower. The “midweight” models (such as DiffDock) are on par in terms of speed but have some issues with accuracy. The “lightweight” models (such as TankBind) are very fast but very inaccurate. As a result, ML docking is unable to replace the classical one in the practical high-throughput screening tasks because of an unfavorable accuracy-to-speed ratio. In other words, we still did not reach the much anticipated docking paradigm shift.

In this paper we introduce ArtiDock - the novel ML docking technique, which overcomes these limitations. At the time of writing ArtiDock outperformed all tested classical and ML-based docking techniques in terms of speed to accuracy ratio on the PoseBusters dataset thus pretending to be a method of choice in the real-world high-throughput virtual screening applications.

The main idea of ArtiDock is to utilize multiple data augmentation modalities to overcome the inherently limited number of available experimentally resolved protein-ligand complexes, while keeping the model architecture simple and lightweight by means of careful feature selection.

ArtiDock used two sources of augmented data:

- Algorithmic generation of artificial “binding pockets” for a diverse set of small molecule ligands (including those, not present in any experimental structures), which

closely follow statistical distributions of the protein-ligand non-bond interactions deduced from experimentally resolved complexes [10].

- Ensembles of representative conformations obtained from a massive Molecular Dynamics (MD) Simulations of about 17,000 protein-ligand complexes from a PDB data bank.

Our approach could be considered as opposite to those used by AlphaFold-latest and DiffDock. Instead of complicating the model architecture in an attempt to deal with limited data, we augment the data to provide a much larger and better balanced training set for a simple and fast model.

In this paper we describe the ArtiDock architecture and data augmentation techniques, provide comprehensive performance comparison with a number of classical docking and ML-docking techniques and discuss the future directions.

2. Methods

2.1. Preprocessing of proteins and ligands

The Python API of the RDKit v.2022.9.1 and Open Babel [11] v.3.1.0 were utilized for loading, processing, and feature generation of small molecules. Before feature extraction, all the explicit hydrogen atoms were removed.

A custom protein processing module was developed to extract protein data from the PDB files and generate the necessary features for model training. This module utilizes the PDB atom names to obtain atom-level graph features, rather than relying on third-party software to infer them. This approach decreases the exclusion rate for processed proteins due to inevitable inconsistencies in the PDB files. For the model training and inference, we extracted protein binding pockets from the protein-ligand complexes defined as all the residues within 6 Å of any heavy ligand atom. Subsequently, only the binding pocket was used for feature extraction.

2.2. Training data

2.2.1. Inclusion criteria for protein-ligand complexes

We filtered out experimentally determined protein-ligand complexes with at least one of the following conditions:

- More than one ligand molecule in the binding pocket;
- Ligand containing less than 5 or more than 100 heavy atoms;
- Ligand containing more than 50 rotatable bonds;
- Protein binding pocket containing less than 5 residues;
- Protein-ligand steric clashes;
- Complex with a significant fraction of ligand atoms too far away from any protein atom.

The filtering conditions were selected to avoid low-quality complexes or non-drug-like ligands while keeping as much experimental data as possible.

2.2.2. PDBbind database

The PDBbind database [12,13] (v.2020) provides a collection of biomolecular complexes from the PDB with experimentally measured binding affinity data. The database contains 19,443 protein-ligand complexes. We used a preprocessed version of the complexes published previously with EquiBind [4] and additionally filtered as described above.

2.2.3. Binding MOAD database

The Binding MOAD database [14–16] (v.2020) provides a subset of the PDB, containing all high-quality ligand-protein complexes irrespective of availability of the binding affinity data. The database includes 41,409 PDB structures with one or more ligands. We extracted all the combinations of PDB IDs and ligand Chemical IDs annotated by the database as “valid” and filtered them as described above.

It is important to note that the “PDB ID - ligand ID” pairs might be represented by more than one binding pocket and the corresponding ligand pose. This is caused by the existence of multiple protein chains in a single PDB file; multiple identical ligands in a single PDB file; different biological units/assemblies [17] provided by the Binding MOAD for a single PDB identifier. All of them were used for model training.

2.2.4. Data from molecular dynamics simulations

We performed massive MD simulations of about ~17,000 protein-ligand complexes extracted from PDBbind v.2020. The ligand topologies and charges were taken from the curated MISATO dataset [18] whenever possible. The rest of the ligands were curated and processed using the workflow similar to one implemented in MISATO to ensure compatibility. The protein structures were processed using the proprietary structure preparation module of Receptor.AI platform in order to reconstruct partially resolved residues, remove crystallographic agents, properly split the co-crystallized complexes with antibodies and engineered chimeric protein chains. All complexes were simulated with an amber03 force field as implemented in Gromacs [19]. Gromacs 2023.4 [20] was used with the recommended simulation parameters for the Amber force field. Simulations were run for a fixed amount of wall time for all complexes, which resulted in different simulation times depending on the system size. The distribution of the reached simulation times is shown in Fig. 1.

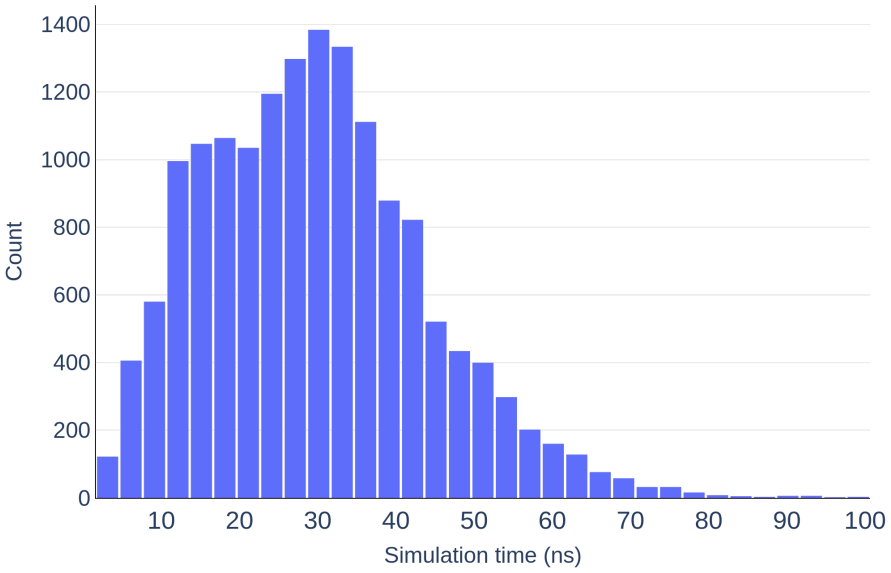

Figure 1. Distribution of the simulation times for massive MD simulations of ~17,000 protein ligand complexes.

Obtained trajectories were analyzed using custom scripts written with Pteros 2.5 molecular modeling library [21]. For each complex the probability map of amino acid residues to be in contact with the ligand was computed. Those residues, which are in contact with the ligand for more than 5% of time were considered to constitute the binding pocket. Then the binding pockets were aligned and clustered using the RMSD of their heavy atoms using k-means agglomerative clustering with ward linkage. The clusters which are at least 0.1 nm apart from each other were kept and the trajectory frames, which are the closest to cluster centroids, were determined. Resulting ensembles of pocket conformations were processed as described above and up to 10 frames were selected randomly among them as alternative pocket-ligand conformations for the training set augmentation.

2.2.5. Training data summary

Table 1 shows the number of data points in the training dataset obtained from different sources.

Table 1. Model training data sources.

Dataset	Unique PDB IDs	Unique Chemical IDs*	Unique PDB ID - Chemical ID pairs	Training samples
PDBbind	16,906	11,856	16,906	16,906
Binding MOAD	38,119	16,798	45,839	92,337
MD	13,955	10,353	13,955	130,962

* The oligomeric ligands annotated by multiple Chemical IDs are not accounted for in the column.

The intersection of entities from different data sources is shown in Figure 2.

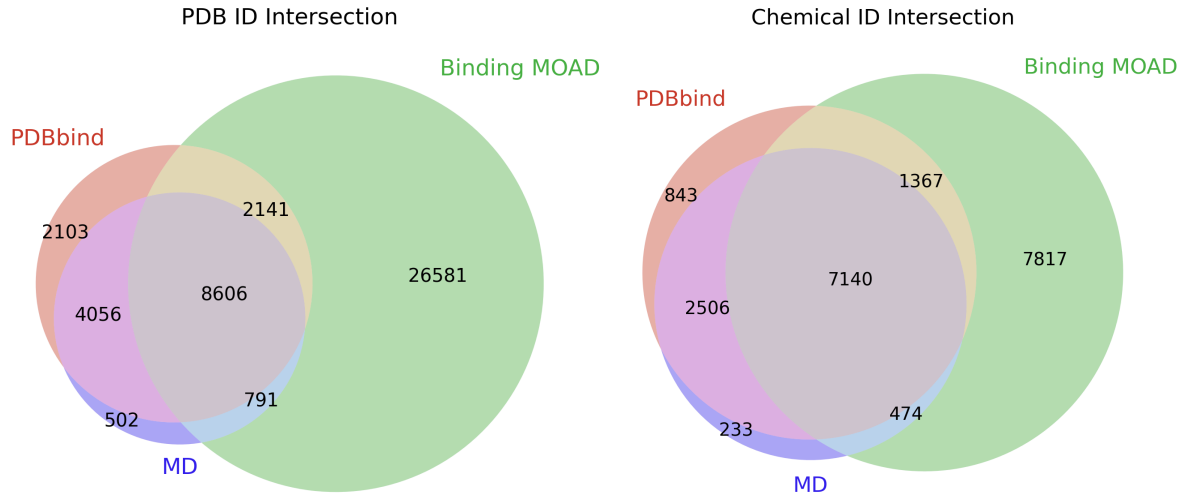

Figure 2. The intersection of PDB IDs and ligand Chemical IDs in the experimental and MD data sources.

2.2.6. Artificial binding pockets

In this study, we used an evolution of PocketCFDM approach [10] of augmenting the training set of the protein-ligand complexes with artificial data, which mimics real protein binding pockets in terms of statistical distributions of the non-bond interactions. This method allows generating artificial binding pockets for arbitrary small molecule conformers and was previously proven to produce high-quality augmentation data.

We used the “In-Stock” subset of the ZINC20 database of commercially available chemicals widely used for virtual screening [22] as a small molecule base for artificial pocket generation. The final dataset consisted of 960,000 compounds with 8 to 55 heavy atoms. To reduce the model bias towards the most frequent molecule sizes (the median is around 25 heavy atoms for both PDB and ZINC data), we enforced a uniform small molecule size distribution for the artificial data. Repetitions are possible in the training set if there are not enough compounds in a particular size range in ZINC to obtain a uniform distribution.

Additionally, we used the same compounds with 32-55 heavy atoms to generate an artificial pocket and then keep only a randomly chosen fragment of the ligand as an input. We assume that such fragment sampling enhances the model's ability to dock smaller ligands into the subpockets of extended binding pockets.

2.2.7. Data split

We employed the same train-validation-test split as used previously in EquiBind [4] and DiffDock [7], which makes the results of our model directly comparable with these two

reference techniques. The split consisted of 968 validation and 363 test PDB entries from PDBbind v.2020. After the low-quality complexes filtering, 915 validation complexes were retained. We removed from the training set all experimental and MD samples intersecting with validation or test PDB IDs.

2.3. Performance metrics

Traditionally, the primary metric for the docking techniques is RMSD between experimentally determined and predicted ligand poses in the given protein binding pocket. However, despite providing decent results in terms of RMSD, many ML-based approaches tend to output physically and chemically implausible ligand poses. Recently the PoseBusters approach was introduced to perform quality checks of a predicted ligand pose including chirality and stereochemistry preservation, bond length validity, internal energy, and intramolecular and intermolecular steric clashes [23].

The PoseBusters [23] dataset and the corresponding Python package was used for model performance benchmarking. Versions 1 and 3 of the PoseBusters were used in parallel because not all of the docking techniques, which we would like to include into comparison, are currently benchmarked against the latest PoseBusters version 3.

Thus, the following main performance metrics were used in this paper:

- a fraction of the predicted ligand poses with RMSD < 2 Å from experimental crystal structures;
- a fraction of the predicted ligand poses with RMSD < 2 Å that pass all the PoseBusters quality checks.

In order to check the influence of RMSD cutoff on the model performance the cutoffs from 2 Å to 5 Å were used for ArtiDock and several selected competing techniques.

2.4. Feature extraction

The ligands were represented as the atom-level molecular graphs. The node features included one-hot encoded atom symbol, number of covalently bound heavy atoms and hydrogens, valence, charge, hybridization, and aromaticity. The graph edges represented covalent bonds between heavy atoms and the one-hot encoded covalent bond type.

Protein node features were extracted for each heavy atom and included scalar (one-hot encoded residue and atom names) and vector (distance from a pocket centroid to an atom) components. The graph edges were formed between a node and its 30 closest neighbors. The edge scalar features represented positional embedding calculated by the distance radial basis function [24]. The vectors between the nodes connected by edges were considered edge vector features.

2.5. Model architecture, training, and validation

ArtiDock is based on a proprietary model architecture inspired by the lightweight Trigonometry-Aware Neural Networks [3]. The model was built with an open-source machine learning framework PyTorch [25] v.2.0.0.
The training objective was based on the minimization of the difference between predicted and reference pocket-ligand intermolecular distance matrices. We trained the model for 500 epochs on an NVIDIA GeForce RTX 4090 GPUs using the MSE loss and Adam optimizer. A batch size of 2 was used. Subsequently, the Exponential Moving Average (EMA) technique was applied to smooth the noise in the training process and to improve the generalization of the model.

At each epoch, the model was trained on:

- PDBbind training split (15,646 samples);
- Binding MOAD training split (44,888 samples, one per unique PDB ID - ligand ID pair);
- MD training split (13,046 samples, one per unique PDB ID);
- artificial pockets generated for randomly sampled 10,000 small molecules;
- artificial pockets generated for randomly sampled 5,000 small molecules split into the segments.

The model training, which is a GPU-intensive task, and pocket generation, which is a CPU-intensive task, were separated into distinct parallel workflows. On average, the generated data (10,000 pocket-ligand complexes and 5,000 pocket-ligand fragment complexes) was fully updated by newly generated samples each 3 model training epochs.

After each epoch, we tracked median model loss on the 915 experimental complexes from the validation data split and selected the best model checkpoint based on the validation performance.

2.6. Model inference

The model outputs the pocket-ligand intermolecular distance matrix DM^{pc} for any arbitrary pocket and small molecule. Thus, an additional distance matrix-to-pose algorithm is needed to convert the matrix into ligand atom coordinates (the actual binding pose). ArtiDock utilizes an algorithm that first infers a 3D point cloud from the distance matrix and then aligns a ligand conformer to the generated cloud.

For the point cloud generation we combined and improved the ligand pose generation approaches from TankBind [3,26,27] and Uni-Mol [28]. Given the model predicted DM^{pc}, we randomly initialized the 3D point cloud $C' \in \mathbb{R}^{c\times 3}$, where c - number of ligand atoms, and applied back-propagation to optimize the C' by Adam optimizer. The optimization stopped as soon as the loss reached a plateau. The loss function was calculated as the weighted sum of intramolecular and intermolecular distance and steric clash contributions.

Despite accounting for interatomic distance constraints during the distance matrix to point cloud transformation, the resulting point cloud C' is still subject to artifacts and violations of

geometric quality criteria. These issues are well known and were reported for the majority of ML-based docking methods [23].

We addressed the problem by applying the custom distance matrix-to-pose algorithm, which is based on differential evolution (DE) technique with additional enforced penalization of steric clashes. The DE is a stochastic approach that does not use gradient methods to find the minimum and can search large areas of the conformational space [29]. We utilized a customly modified version of DE for the ligand position optimization [30]. Given a point cloud C' and a random input ligand conformer, we set the DE objective of RMSD minimization between C' and the conformer coordinates by modifying conformer rotatable bonds and aligning it to the C'. As a result we obtain an optimized ligand conformer with atomic coordinates close to C', which has significantly improved tetrahedral chirality, double bond stereochemistry, covalent bond lengths and angles, aromatic ring planarity, and double bound planarity.

To address the issue with steric clashes, we included additional adjustable contributions to the DE objective function accounting for inter- and intramolecular clashes. The conformer coordinates are used instead of C' for defining the clashes. The amount of these extra constraints is adjusted empirically to achieve an optimal balance between the accuracy in terms of RMSD and the amount of remaining clashes.

2.6.1 Parameters tuning

The inference point cloud generation and DE objective function weights together with DE parameters were tuned using the PoseBusters [23] quality checks on 915 experimental complexes from the validation data split. The tuning was performed for 200 trials with the help of parameters optimization software Optuna [31] v.3.5.0 using the tree-structured Parzen estimator algorithm. The best parameters were selected based on the highest fraction of the predicted ligand poses with RMSD < 2 Å and passing all the PoseBusters quality checks.

2.7. Benchmarking

2.7.1 Testing data

We used the PoseBusters Benchmark set [23] that consists of 308 crystal complexes from the PDB each representing a unique protein and ligand. All the complexes have been released since 2021. Since all the experimental and MD data used for training is based on the database versions of the year 2020, the PoseBusters Benchmark set can be safely used as the time-split test dataset for the model performance assessment.

2.7.2. Comparison to other docking techniques

We compared the performance of our model to other docking approaches reported in PoseBusters publication [23], namely AutoDock Vina [32], Gold [33], DiffDock [7], Uni-Mol [28], TankBind [3], EquiBind [4], DeepDock [30]. We recalculated Uni-Mol results on the more recent version (source code accessed 2023-12-06). Also, Glide [34] v10.1 was included into comparison with default preprocessing and docking parameters except the docking box size,

which was set to 25 Å to be the same as in other classical docking methods in PoseBusters. The AutoDock Vina, Gold, and Glide represent well-established classical docking approaches. The rest of the techniques are ML-based methods performing either blind docking (DiffDock, TankBind, EquiBind) or docking into a predetermined protein binding pocket (Uni-Mol, DeepDock, ArtiDock).

We have also included AlphaFold-latest into the comparison although it is formally a protein-ligand co-folding method rather than a ligand pose prediction one. It is by far the most architecturally advanced and heavyweight ML model, which is expected to show the best accuracy among all other techniques, which do not use data augmentation. Although the AlphaFold-latest performance is only reported for PoseBusters v1 we still decided to include it into comparison with a necessary caution because of the importance of this model for the community.

We performed all the benchmarking using a workstation with NVIDIA GeForce RTX 3060 GPU and AMD Ryzen 5 5600X 6-Core CPU.

2.7.3. Tested ArtiDock versions

In this work we compared four ArtiDock versions, which differ by the training dataset and the post-processing of predicted ligand-protein distance matrices. All versions share the same general architecture with minor improvements in the later ones.

- ArtiDock v1.0 is trained on the PBDbind dataset only.
- ArtiDock v1.5 is trained on the PBDbind dataset and the augmented data from artificial binding pockets.
- ArtiDock v1.8 is trained on the PBDbind and the Binding MOAD datasets with addition of augmented data from artificial binding pockets and MD simulations.
- ArtiDock v2.0 additionally uses enforced penalization of the steric clashes in differential evolution algorithm for transforming the ligand-protein distance matrix to the ligand coordinates.

3. Results

It is clearly seen that the expansion of the training dataset leads to notable improvement of the ArtiDock performance. Augmentation with artificial binding pockets leads to the increase of accuracy by ~3% (v1.5 vs v1.0). Further inclusion of the Binding MOAD data and augmentation from MD simulations leads to a much more pronounced increase of accuracy by ~10-12% (v1.5 vs v1.8).

The RMSD evaluation and PoseBusters quality checks reveal the superior performance of ArtiDock v1.8 over all other techniques included in comparison (Fig. 3). ArtiDock 1.8 predicts 78% of ligand binding poses within 2 Å RMSD from experimental ones outperforming the closest rival by 14%. It demonstrates good generalization of the DL model over experimental training data and its ability to predict intramolecular distance matrix with high precision. Due to the good performance in terms of RMSD, the fraction of samples passing both RMSD and all PoseBusters checks is also the highest.

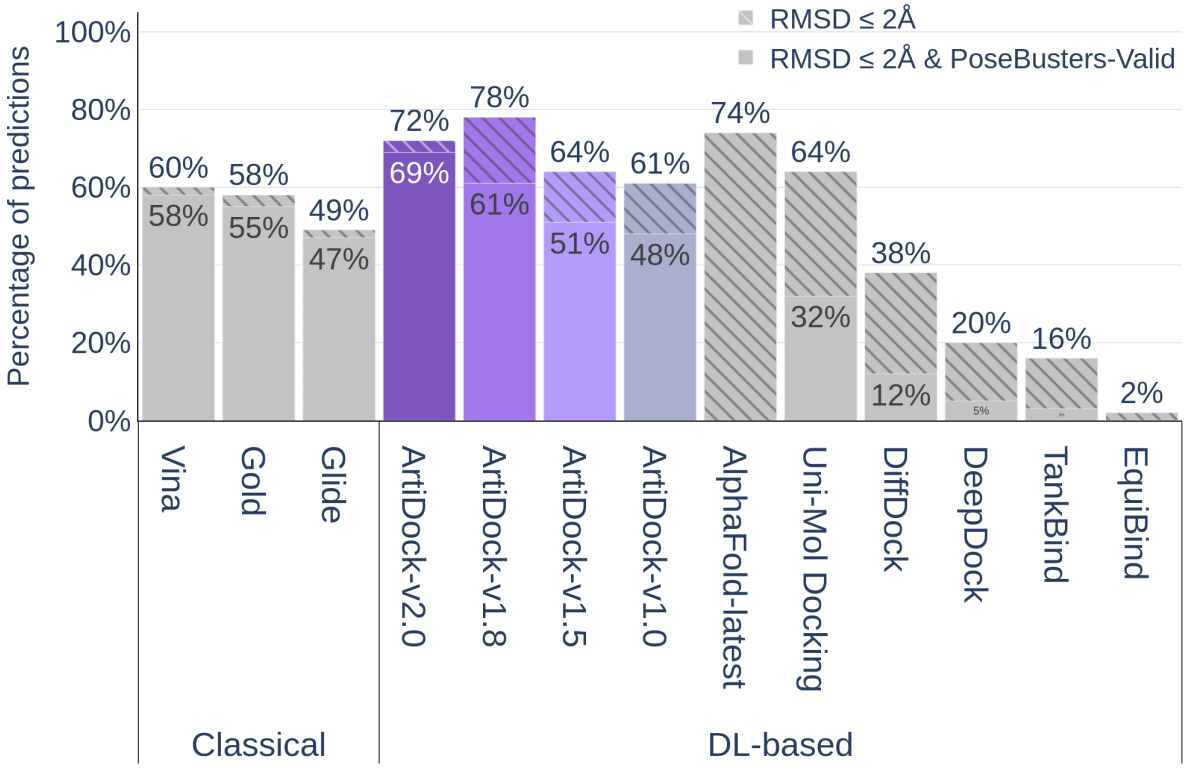

Figure 3. Performance of the docking methods on the PoseBusters v3 dataset (N=308). For AlphaFold-latest only RMSD is reported in the literature without additional PoseBusters metrics and the PoseBusters v1 is used (N=428), so the comparison should be treated with caution.

It is evident that all ML docking techniques, including ArtiDock up to v2.0, demonstrate a significant gap between RMSD of all predicted poses and RMSD of the PoseBusters-valid poses. For ArtiDock v1.8 this gap is ~17%, which is substantially less than for UniMol (~32%) and DiffDock (~26%), but still a lot more than for classical docking methods (2-3% for Gold, Vina and Glide). This indicates that a considerable fraction of predicted ligand poses in ML docking techniques is still inferior in terms of the structural quality metrics.

The main quality concerns are minimal interatomic distance violations and the van der Waals volume overlap as shown in Figure 4. About 25% of the protein-ligand complexes are predicted with minor protein-ligand steric clashes. Additionally, 4% of predicted ligand conformations have high energies as calculated using the Universal force field [35].

This issue could be mitigated by improving the distance matrix-to-pose transformation algorithm. In the ArtiDock v2.0 we introduced additional penalization of the steric clashes in the differential evolution algorithm for transforming the ligand-protein distance matrix to the ligand coordinates. This results in the drastic decrease of the RMSD gap from 17% in v1.8 to 3% in v2.0 (Fig. 3). The minimal interatomic distance violations decrease impressively from ~23% in v1.8 to ~5% in v2.0 and the volume overlap with protein and the violation of distances with organic cofactors almost vanishes in v2.0 (Fig.4).

This impressive improvement of the ligand poses quality comes at the cost of decreased overall precision in terms of RMSD<2.0 Å from ~78% in v1.8 to ~72% in v2.0. However, the percentage of correctly predicted poses with RMSD<2.0 Å that are simultaneously PoseBusters-valid increases from 61% to 69%. We believe that the drastic increase of the overall ligand pose quality and RMSD of the valid subset of poses is more important than getting higher accuracy in terms of RMSD only, thus we consider ArtiDock v2.0 as the best model for practical applications.

It is important to note that we didn't account explicitly for the non-protein cofactors in the binding pockets in the current version of ArtiDock, thus the presence of steric clashes with organic and inorganic cofactors is an expected behavior. Even though the cofactor-related PoseBusters metrics are already surprisingly good (especially in ArtiDock v2.0), they could be further improved by including the cofactors into the training data.

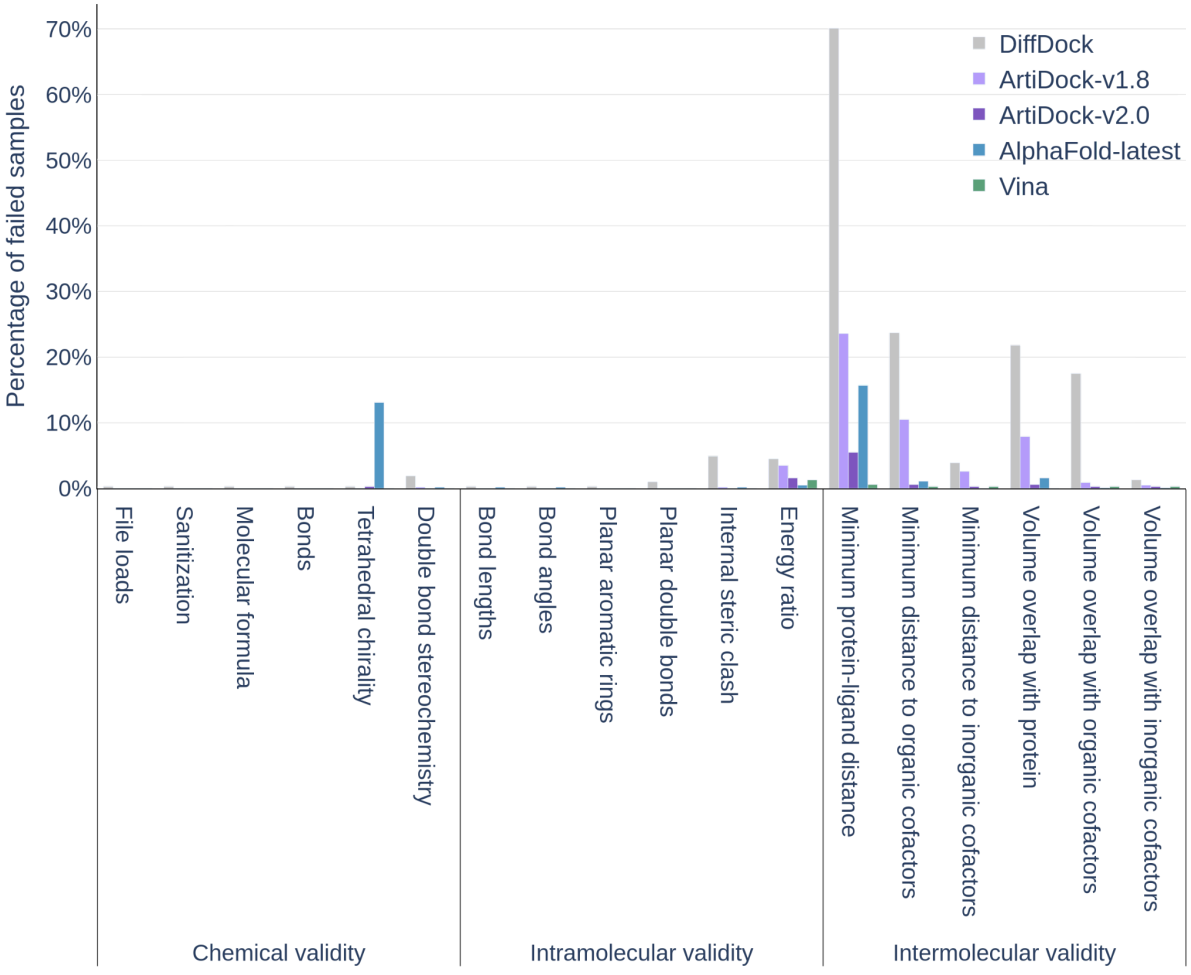

Figure 4. PoseBusters quality metrics of the predicted ligand poses using PoseBusters v3 dataset (N=308). The percentage of failures is reported (the lower the better). AlphaFold-latest results are only reported in the literature for the PoseBusters v1 (N=428), so the comparison should be treated with caution.

Figure 5 demonstrates the performance of ArtiDock at different RMSD cut-offs. Expectibly, the number of poses that pass all PoseBuster quality checks increases with the increase of RMSD cut-off. The ArtiDock v2.0 performs significantly better than v1.8 in the whole range of cut-offs and reaches an impressive 93% of correct ligand poses without RMSD constraints. For the most practically used range of cut-offs from 1.5 Å to 5 Å ArtiDock v2.0 outperforms Vina, Gold and Glide significantly, which makes it the method of choice in those applications where the ligand pose quality is a priority.

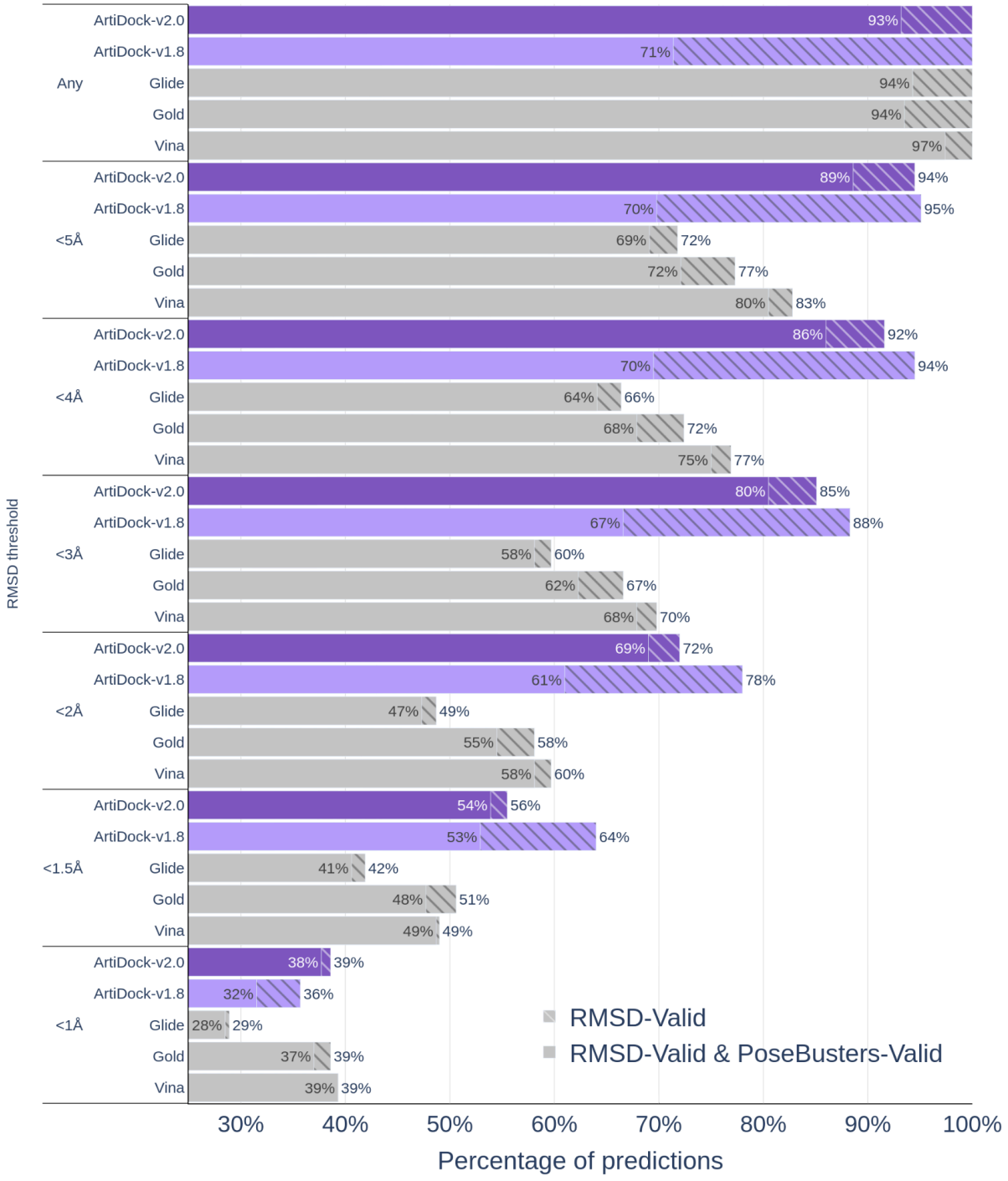

Figure 4. Performance of ArtiDock and classical docking methods on the PoseBusters v3 Benchmark set (N=308) at different RMSD thresholds.

One of the key aspects of ArtiDock in terms of its practical applicability is the fast inference. On average, it takes around 1.5 seconds to retrieve a single docked ligand conformer with the ArtiDock on our testing workstation (Figure 5). The inference speed is the same for all ArtiDock versions within the measurement error, thus a single value is reported. The classical docking methods are one or more orders of magnitude slower than our approach, which makes ArtiDock the most efficient when performing virtual screening of the big chemical libraries.

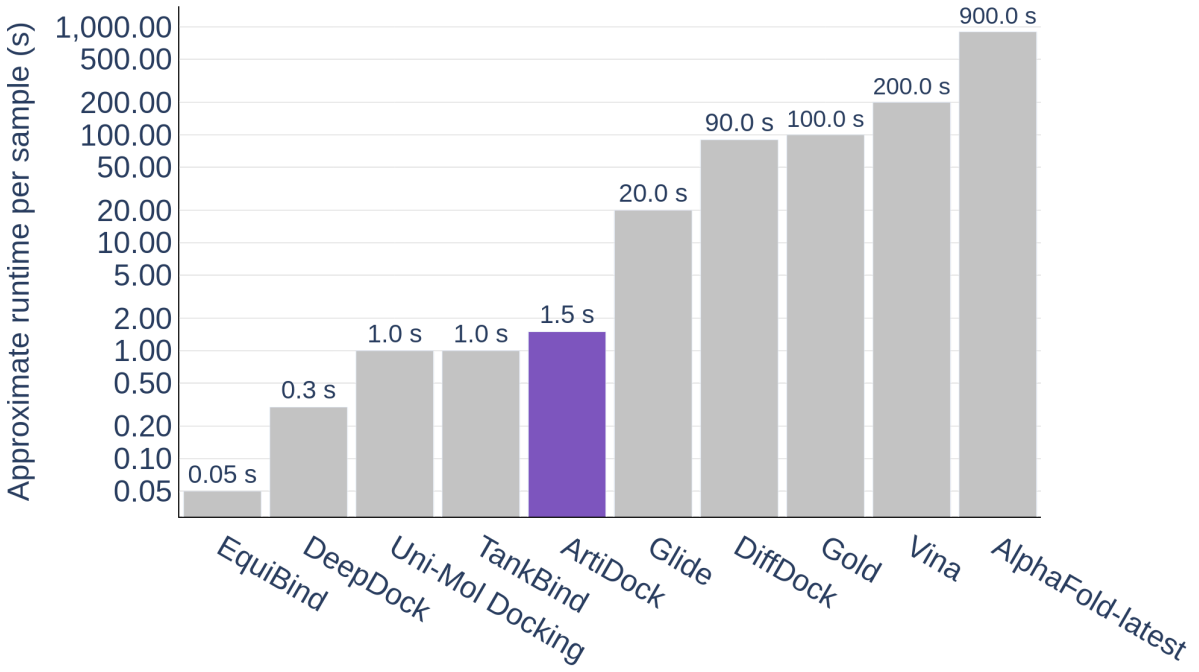

Figure 5. Approximate binding pose inference time of the docking methods. Note the log scale of the vertical axis.

4. Discussion

The ML approaches to protein-ligand docking are severely hampered by the inherently limited amount of high quality training data. The set of protein-ligand complexes in PDB is growing too slowly to keep up with the demands of rapidly advancing ML techniques. Following the enormous success of AlphaFold, recent development in the field of ML docking was biased toward more complex architectures and larger model sizes, which could, to some extent, balance limited amounts of data.

In this work we demonstrate that the opposite approach, with providing a deliberately simple and fast model with large amounts of augmented data, results in better prediction accuracy without compromising the inference speed. ArtiDock, which follows this approach, significantly outperforms not only all competing ML docking techniques but also conventional docking methods including the industry-standard Glide, Gold and the most widely used open

source Vina. This superior performance comes with no inference speed trade-off - ArtiDock is at least one order of magnitude faster than its closest competitors.

One of the most important limitations of existing ML docking models is the inferior quality of the ligand poses in comparison with conventional physics-based docking techniques. Due to the absence of explicit protein-ligand interactions it is challenging to ML models to avoid minor steric clashes and subtle but physically relevant distortions of the ligands' structure.

This issue is clearly visible in Fig. 3 as a performance gap between RMSD of all predicted poses and RMSD of the PoseBusters-valid poses, which is evident for all ML techniques and absent for conventional docking. For ArtiDock 1.8 this gap is already 1.5-2 times smaller than for two most recent ML docking models (UniMol and DiffDock), which indicates that larger and more diverse augmented training set helps to reduce the amount of steric clashes and structural imperfections. However, ArtiDock 2.0 decreases this gap to the values, which are observed for conventional docking, by introducing an additional penalization of the steric clashes at the stage of inferring the ligand coordinates from the predicted protein-ligand distance matrix.

Fig. 3 clearly shows that the steric clashes (minimal ligand-protein distances) remain the most challenging for all compared ML techniques. However, ArtiDock v2.0 manages to get rid of most of them, performing even better than the much more sophisticated AlphaFold-latest model and approaching the range of accuracy of conventional docking techniques. The rest of PoseBusters metrics are predicted with nearly perfect accuracy, which is indistinguishable from the conventional docking.

The ArtiDock 2.0 is at least one order of magnitude faster than conventional docking programs, at least two orders of magnitude faster than DiffDock and three orders of magnitude faster than AlphaFold-latest. Combined with RMSD superior to and the ligand pose quality approaching one of the conventional docking, this makes our technique an attractive choice for high-throughput applications in drug discovery.

4.1. Limitations

ArtiDock shares some common limitations with other ML docking techniques. Since there is no scoring function, which is assessed during the pose prediction, ArtiDock is not suitable for ranking alternative poses of the same ligand or comparing different ligands in terms of the binding strength, like the conventional docking techniques do. The poses generated by ArtiDock are optimal for the given ligand and given conformation of the binding pocket, but their scoring should be performed externally by either algorithmic scoring functions or ML scoring models. The latter is advantageous because such models allow attributing binding poses not only to the binding affinity, but also to the biological activity of compounds. Development and benchmarking of such rescoring models is out of scope of the current work.

ArtiDock is designed to perform docking into the predefined binding pocket, thus it is neither directly competitive nor apples-to-apples comparable to the blind docking techniques. In principle, the concept of multimodal data augmentation, which is used in ArtiDock, could be transferred to the blind docking ML techniques. However, while the data of MD simulations is straightforward to use in the blind docking scenario, the artificial pocket generation requires

significant rethinking since the binding pocket is not known in advance. The model itself is also likely to become larger and significantly slower due to the necessity to encode the whole protein structure. Development of the blind pocket-agnostic version of ArtiDock is an interesting future direction of research.

4.2. Future directions

There are clear directions of improvement which should eliminate remaining accuracy issues of ArtiDock, which are common for ML-based docking in general. First, data augmentation with artificial binding pockets allows easy model tuning towards better representations of structures with optimal ligand-protein distances by careful generation of pockets with desired ranges of distances. Second, remaining minor steric clashes could be removed by further optimization of the post-processing stage and including additional terms directly in the loss function. Last, but not least, organic and inorganic cofactors could be included explicitly into the model training. The latter is especially promising because it could give ML models an additional advantage over conventional docking, which traditionally struggles with metals and ions.

This work shows that inclusion of dynamic data from MD simulations leads to a large boost of model accuracy and robustness. Although the majority of used MD trajectories do not exceed 20-30 ns in length, they contribute to up to 15% performance difference between ArtiDock versions 1.5 and 2.0. Longer MD simulations and/or the usage of enhanced sampling is expected to provide an even larger performance boost.

5. Conclusions

ArtiDock is the ML docking model, which represents a new generation of the ligand pose prediction tools. It is based on the approach of utilizing lightweight and fast model architecture in conjunction with the training dataset augmented with artificial protein binding pockets and ensembles of representative conformations from massive MD simulations of existing protein-ligand complexes. This allows the model to achieve accuracy, which is superior to all major ML docking techniques and conventional docking programs on PoseBusters v3 dataset, while being from one to three orders of magnitude faster. Thus ArtiDock is an appealing solution for high throughput ligand-protein docking at the time of writing. ArtiDock v2.0 is currently integrated into the production virtual screening pipeline of the Receptor.AI drug discovery platform. It is also being prepared to be deployed on the Nvidia BioNeMo cloud platform for drug development.

6. Author contributions

TV developed the model and performed its tuning and performance assessment. SS and AN designed the study and controlled its progress. SY coordinated the work, performed MD simulations and participated in results interpretation. TV, IK, VB and RS researched and assessed existing methodologies of machine-learning-based molecular docking. LP, ZO, and RZ participated in developing strategies for efficient model training and tuning. IK, PH, and VV performed technical assistance and supervision of the tuning, training and testing process. The manuscript was written by TV and SY.

7. Conflicts of interest

All authors but VB are employees of Receptor. AI INC. SS, AN and SY have shares in Receptor.AI INC.

8. Availability

ArtiDock is available as a part of the Receptor.AI commercial drug discovery workflow and is being deployed on the Nvidia BioNeMo cloud platform.

9. References

(1) Li, X.; Li, Y.; Cheng, T.; Liu, Z.; Wang, R. Evaluation of the Performance of Four Molecular Docking Programs on a Diverse Set of Protein-Ligand Complexes. *J. Comput. Chem.* **2010**, *31* (11), 2109–2125. https://doi.org/10.1002/jcc.21498.

(2) Ghasemi, J. B.; Abdolmaleki, A.; Shiri, F. Molecular Docking Challenges and Limitations. In *Pharmaceutical Sciences: Breakthroughs in Research and Practice*; IGI Global, 2017; pp 770–794. https://doi.org/10.4018/978-1-5225-1762-7.ch030.

(3) Lu, W.; Wu, Q.; Zhang, J.; Rao, J.; Li, C.; Zheng, S. *TANKBind: Trigonometry-Aware Neural NetworKs for Drug-Protein Binding Structure Prediction*; preprint; Biophysics, 2022. https://doi.org/10.1101/2022.06.06.495043.

(4) Stärk, H.; Ganea, O.-E.; Pattanaik, L.; Barzilay, R.; Jaakkola, T. EquiBind: Geometric Deep Learning for Drug Binding Structure Prediction. arXiv June 4, 2022. https://doi.org/10.48550/arXiv.2202.05146.

(5) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2.

(6) Yang, Z.; Zeng, X.; Zhao, Y.; Chen, R. AlphaFold2 and Its Applications in the Fields of Biology and Medicine. *Signal Transduct. Target. Ther.* **2023**, *8* (1), 1–14. https://doi.org/10.1038/s41392-023-01381-z.

(7) Corso, G.; Stärk, H.; Jing, B.; Barzilay, R.; Jaakkola, T. DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking. **2022**. https://doi.org/10.48550/ARXIV.2210.01776.

(8) Dptech-Corp/Uni-Mol, 2024. https://github.com/dptech-corp/Uni-Mol (accessed 2024-02-25).

(9) *A glimpse of the next generation of AlphaFold*. Google DeepMind. https://deepmind.google/discover/blog/a-glimpse-of-the-next-generation-of-alphafold/ (accessed 2024-02-25).

(10) Voitsitskyi, T.; Bdzhola, V.; Stratiichuk, R.; Koleiev, I.; Ostrovsky, Z.; Vozniak, V.; Khropachov, I.; Henitsoi, P.; Popryho, L.; Zhytar, R. Augmenting a Training Dataset of the Generative Diffusion Model for Molecular Docking with Artificial Binding Pockets. *RSC Adv.* **2024**, *14* (2), 1341–1353.

(11) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An Open Chemical Toolbox. *J. Cheminformatics* **2011**, *3* (1), 33. https://doi.org/10.1186/1758-2946-3-33.

(12) Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind Database: Collection of Binding

Affinities for Protein−Ligand Complexes with Known Three-Dimensional Structures. *J. Med. Chem.* **2004**, *47* (12), 2977–2980. https://doi.org/10.1021/jm030580l.

(13) Liu, Z.; Li, Y.; Han, L.; Li, J.; Liu, J.; Zhao, Z.; Nie, W.; Liu, Y.; Wang, R. PDB-Wide Collection of Binding Data: Current Status of the PDBbind Database. *Bioinformatics* **2015**, *31* (3), 405–412. https://doi.org/10.1093/bioinformatics/btu626.

(14) Ahmed, A.; Smith, R. D.; Clark, J. J.; Dunbar, J. B., Jr; Carlson, H. A. Recent Improvements to Binding MOAD: A Resource for Protein–Ligand Binding Affinities and Structures. *Nucleic Acids Res.* **2015**, *43* (D1), D465–D469. https://doi.org/10.1093/nar/gku1088.

(15) Hu, L.; Benson, M. L.; Smith, R. D.; Lerner, M. G.; Carlson, H. A. Binding MOAD (Mother Of All Databases). *Proteins Struct. Funct. Bioinforma.* **2005**, *60* (3), 333–340. https://doi.org/10.1002/prot.20512.

(16) Smith, R. D.; Clark, J. J.; Ahmed, A.; Orban, Z. J.; Dunbar, J. B.; Carlson, H. A. Updates to Binding MOAD (Mother of All Databases): Polypharmacology Tools and Their Utility in Drug Repurposing. *J. Mol. Biol.* **2019**, *431* (13), 2423–2433. https://doi.org/10.1016/j.jmb.2019.05.024.

(17) Xu, Q.; Dunbrack, R. L. Principles and Characteristics of Biological Assemblies in Experimentally Determined Protein Structures. *Curr. Opin. Struct. Biol.* **2019**, *55*, 34–49. https://doi.org/10.1016/j.sbi.2019.03.006.

(18) Siebenmorgen, T.; Menezes, F.; Benassou, S.; Merdivan, E.; Kesselheim, S.; Piraud, M.; Theis, F. J.; Sattler, M.; Popowicz, G. M. MISATO - Machine Learning Dataset of Protein-Ligand Complexes for Structure-Based Drug Discovery. bioRxiv May 28, 2023, p 2023.05.24.542082. https://doi.org/10.1101/2023.05.24.542082.

(19) Sorin, E. J.; Pande, V. S. Exploring the Helix-Coil Transition via All-Atom Equilibrium Ensemble Simulations. *Biophys. J.* **2005**, *88* (4), 2472–2493. https://doi.org/10.1529/biophysj.104.051938.

(20) Abraham, M.; Alekseenko, A.; Bergh, C.; Blau, C.; Briand, E.; Doijade, M.; Fleischmann, S.; Gapsys, V.; Garg, G.; Gorelov, S.; Gouaillardet, G.; Gray, A.; Irrgang, M. E.; Jalalypour, F.; Jordan, J.; Junghans, C.; Kanduri, P.; Keller, S.; Kutzner, C.; Lemkul, J. A.; Lundborg, M.; Merz, P.; Miletić, V.; Morozov, D.; Páll, S.; Schulz, R.; Shirts, M.; Shvetsov, A.; Soproni, B.; Spoel, D. van der; Turner, P.; Uphoff, C.; Villa, A.; Wingbermühle, S.; Zhmurov, A.; Bauer, P.; Hess, B.; Lindahl, E. GROMACS 2023.4 Manual. **2024**. https://doi.org/10.5281/zenodo.10560024.

(21) Yesylevskyy, S. O. *Pteros 2.0: Evolution of the Fast Parallel Molecular Analysis Library for C++ and Python*; 2015.

(22) Irwin, J. J.; Tang, K. G.; Young, J.; Dandarchuluun, C.; Wong, B. R.; Khurelbaatar, M.; Moroz, Y. S.; Mayfield, J.; Sayle, R. A. ZINC20—A Free Ultralarge-Scale Chemical Database for Ligand Discovery. *J. Chem. Inf. Model.* **2020**, *60* (12), 6065–6073. https://doi.org/10.1021/acs.jcim.0c00675.

(23) Buttenschoen, M.; M. Morris, G.; M. Deane, C. PoseBusters: AI-Based Docking Methods Fail to Generate Physically Valid Poses or Generalise to Novel Sequences. *Chem. Sci.* **2024**, *15* (9), 3130–3139. https://doi.org/10.1039/D3SC04185A.

(24) Ingraham, J.; Garg, V.; Barzilay, R.; Jaakkola, T. Generative Models for Graph-Based Protein Design. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc., 2019; Vol. 32.

(25) Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Köpf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; Chintala, S. PyTorch: An Imperative Style, High-Performance Deep Learning Library. arXiv December 3, 2019. https://doi.org/10.48550/arXiv.1912.01703.

(26) Masters, M. R.; Mahmoud, A. H.; Wei, Y.; Lill, M. A. Deep Learning Model for Efficient Protein–Ligand Docking with Implicit Side-Chain Flexibility. *J. Chem. Inf. Model.* **2023**, *63* (6), 1695–1707. https://doi.org/10.1021/acs.jcim.2c01436.

(27) Zsoldos, Z.; Reid, D.; Simon, A.; Sadjad, S. B.; Johnson, A. P. eHiTS: A New Fast, Exhaustive Flexible Ligand Docking System. *J. Mol. Graph. Model.* **2007**, *26* (1),

198–212. https://doi.org/10.1016/j.jmgm.2006.06.002.
(28) Zhou, G.; Gao, Z.; Ding, Q.; Zheng, H.; Xu, H.; Wei, Z.; Zhang, L.; Ke, G. Uni-Mol: A Universal 3D Molecular Representation Learning Framework; 2022.
(29) Storn, R.; Price, K. Differential Evolution – A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces. *J. Glob. Optim.* **1997**, *11* (4), 341–359. https://doi.org/10.1023/A:1008202821328.
(30) Méndez-Lucio, O.; Ahmad, M.; del Rio-Chanona, E. A.; Wegner, J. K. A Geometric Deep Learning Approach to Predict Binding Conformations of Bioactive Molecules. *Nat. Mach. Intell.* **2021**, *3* (12), 1033–1039. https://doi.org/10.1038/s42256-021-00409-9.
(31) Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A Next-Generation Hyperparameter Optimization Framework. arXiv July 25, 2019. https://doi.org/10.48550/arXiv.1907.10902.
(32) Trott, O.; Olson, A. J. AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *J. Comput. Chem.* **2010**, *31* (2), 455–461. https://doi.org/10.1002/jcc.21334.
(33) Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. Improved Protein–Ligand Docking Using GOLD. *Proteins Struct. Funct. Bioinforma.* **2003**, *52* (4), 609–623. https://doi.org/10.1002/prot.10465.
(34) *Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy | Journal of Medicinal Chemistry*. https://pubs.acs.org/doi/full/10.1021/jm0306430 (accessed 2024-03-05).
(35) Rappe, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A.; Skiff, W. M. UFF, a Full Periodic Table Force Field for Molecular Mechanics and Molecular Dynamics Simulations. *J. Am. Chem. Soc.* **1992**, *114* (25), 10024–10035. https://doi.org/10.1021/ja00051a040.