DEEP LEARNING MODEL FOR FLEXIBLE AND EFFICIENT PROTEIN-LIGAND DOCKING

Matthew R. Masters, Amr H. Mahmoud, Yao Wei & Markus A. Lill
Department of Pharmaceutical Sciences
University of Basel
Klingelbergstrasse 50, 4056
Basel, Switzerland
{matthew.masters, amr.abdallah, yao.wei, markus.lill}@unibas.ch

ABSTRACT

Protein-ligand docking is an essential tool in structure-based drug design with applications ranging from virtual high-throughput screening to pose prediction for lead optimization. Most docking programs for pose prediction are optimized for re-docking to an existing co-crystalized protein structure ignoring protein flexibility. In real-world drug design applications, however, protein flexibility is an essential feature of the ligand-binding process. Here we present a deep learning model for flexible protein-ligand docking based on the prediction of an intermolecular Euclidean distance matrix (EDM), making the typical use of search algorithms obsolete. Our method introduces a new approach for the reconstruction of ligand poses in Cartesian coordinates, utilizing EDM completion and restrained energy-based optimization. The model was trained on a large-scale dataset of protein-ligand complexes and evaluated on standardized test sets. Our model generates high quality poses for a diverse set of protein and ligand structures and outperforms comparable docking methods.

1 INTRODUCTION

Protein-ligand docking is widely used in structure-based drug design and throughout early drug development. Molecular docking enables the identification of key binding interactions which can be optimized for affinity or selectivity. Furthermore, docking can be applied to virtual screening (VS) protocols, wherein a large library of compounds are screened for potential target binding (Pereira et al., 2016; Fan et al., 2019). Docking can also be applied in the context of inverse (or reverse) docking in which one compound is screened against a large set of potential targets. This technique is particularly useful in drug repurposing, polypharmacology and side-effect prediction (Kharkar et al., 2014).

Despite the broad applicability of docking, several challenges still exist with traditional docking methods. One challenge is the computational cost of the sampling or search algorithms. Docking programs often generate millions of potential poses and attempt to rank the strongest binding ones towards the top. This creates a burdensome computational requirement and often makes molecular docking of large chemical libraries infeasible or requires the use of simplified and less accurate scoring metrics. Another limitation of many docking methods is the use of a rigid receptor, which neglects the induced-fit effect known to play a critical role in protein-ligand recognition and binding (Koshland Jr, 1958; Savir & Tlusty, 2007). Simplified scoring functions which neglect protein flexibility have detrimental effects on the performance of docking algorithms. Finally, there are still protein systems where state-of-the-art docking algorithms fail to generate any correct pose.

1.1 RELATED WORK

Over recent years, several different methods employing deep learning models have been applied to molecular docking. These models can be divided into two predominant approaches: reranking and generation. In the reranking approach, an ensemble of docked poses are first generated using

a traditional docking method. Then the deep learning model is trained to rerank the ensemble such that the top-ranked poses have the lowest RMSD to the native pose obtained from experiment. A number of neural network architectures have been applied to this approach including convolutional neural networks(Ragoza et al., 2017; Mahmoud et al., 2020b; Zheng et al., 2019) and graph neural networks (Wang et al., 2021). While this approach has proven to be powerful in improving the ranking of poses, they do not address the sampling problem.

In the second approach, deep learning has been employed to generate the docked poses directly. This approach has been less extensively studied with only a few methods developed in recent years (McNutt et al., 2021; Mahmoud et al., 2020a; Stärk et al., 2022). Our work falls into the later of these categories.

Our contributions to the existing methods are two-fold:

- We present an equivariant graph neural network (EGNN) model for the accurate prediction of protein-ligand distance matrices.
- We provide a method for the reconstruction of the binding poses in Cartesian space based on the predicted Euclidean distance matrix which respects the physical constraints and energy of the pose.

2 METHODS

2.1 MODEL

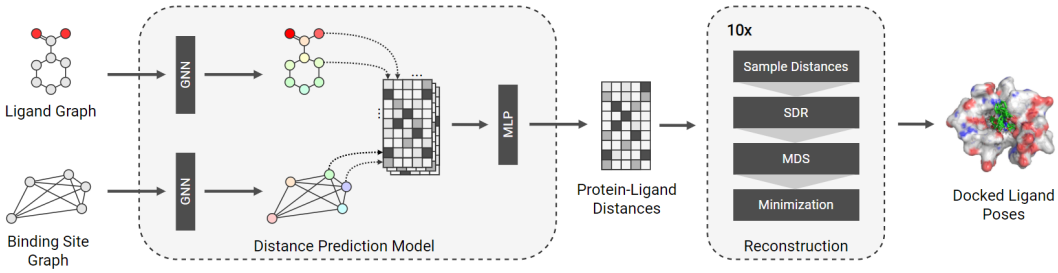

Figure 1: Overview of the model architecture and reconstruction process. The deep neural network is capable of predicting protein-ligand intermolecular distances. This distance matrix is used as an intermediate representation of the ligand pose in the binding site. In the second phase, the distances are used to reconstruct an ensemble of poses in Cartesian coordinates.

In this section, we propose a model architecture for the prediction of protein-ligand distance matrices. The model, depicted in figure 1, consists of two independent graph neural networks (GNNs) (one for ligand, one for binding site) and stack of fully connected layers. The model predicts distance matrices as an intermediate representation due to it's inherent invariance to rigid transformations such as translation and rotation. The GNN model used in this work, the Equivariant Graph Neural Network (EGNN) from Satorras et al. (2021) is also unaffected by rigid transformations. The EGNN was chosen for its ability to learn information about a nodes environment and 3D geometry.

The nodes of the binding site graph are defined by the C_α atoms of residues with any atom within 8Å of the co-crystallized ligand. Therefore, each node represents an entire residue centered on the C_α atom. As a result, the model is inherently flexible as it disregards the side chain conformation entirely. The nodes of the ligand are defined by the ligand heavy atoms and input coordinates are given from a random conformation generated by RDKit (Landrum, 2021).

Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with nodes $v_i \in \mathcal{V}$ and edges $e_{ij} \in \mathcal{E}$ an EGNN can be defined as a set of equations which updates $\mathcal{V}$ in successive layers. Each node v_i is associated with a set of node features $\mathbf{h}_i \in \mathbb{R}^f$ where f is the number of features, and a set of n-dimensional coordinates $\mathbf{x}_i \in \mathbb{R}^n$ (here: $n = 3$). Each EGNN layer l is provided with these node features $\mathbf{h}^l$ (with $\mathbf{h}_i^0 = \mathbf{h}_i$), coordinates $\mathbf{x}^l$ (with $\mathbf{x}_i^0 = \mathbf{x}_i$), and edge information $\mathcal{E}$ and outputs updated $\mathbf{h}^{l+1}$ and $\mathbf{x}^{l+1}$. The

equations defining this update are as follows:

$$\mathbf{m}_{ij} = \phi_e\left(\mathbf{h}_i^l, \mathbf{h}_j^l, \left\|\mathbf{x}_i^l - \mathbf{x}_j^l\right\|, e_{ij}\right) \tag{1}$$

$$\mathbf{x}_i^{l+1} = \mathbf{x}_i^l + C\sum_{j\neq i}(\mathbf{x}_i^l - \mathbf{x}_j^l)\phi_x(\mathbf{m}_{ij}) \tag{2}$$

$$\mathbf{m}_i = \sum_{j\neq i}\mathbf{m}_{ij} \tag{3}$$

$$\mathbf{h}_i^{l+1} = \phi_h(\mathbf{h}_i^l, \mathbf{m}_i) \tag{4}$$

Both EGNNs have the same structure, each with stacks of three layers which successively transform the input features into feature embeddings containing chemical and geometric information. These feature embeddings are denoted as $\mathbf{h}^{lig}$ and $\mathbf{h}^{bs}$ for the ligand and binding site graph respectively. These two sets of feature embeddings are concatenated for each pair of protein and ligand nodes to form a tensor $\mathbf{B} \in \mathbb{R}^{N\times M\times f}$ with

$$\mathbf{B}_{ij} = concat(\mathbf{h}_i^{lig}, \mathbf{h}_j^{bs}) \tag{5}$$

for $i \in \{1, 2, ..., N\}$ and $j \in \{1, 2, ..., M\}$ where N is the number of ligand nodes and M is the number of binding site nodes.

This tensor $\mathbf{B} \in \mathbb{R}^{N\times M\times f}$ is finally transformed into the distance matrix $\mathbf{D} \in \mathbb{R}^{N\times M}$ between ligand and protein atoms using a multi-layer perceptron (MLP) Φ_d

$$\mathbf{D} = \Phi_d(\mathbf{B}) \tag{6}$$

The distance matrix $\mathbf{D}$ can then be used to generate an ensemble of poses in Cartesian coordinates by the reconstruction process described in section 2.5.

2.2 Datasets

2.2.1 Training and Validation

During early experimentation, a significant improvement in performance was observed with increasing training set size. Therefore, we chose the large-scale BioLiP dataset of protein-ligand structures for training and validation (Yang et al., 2012). The complete set contains nearly 300k protein-ligand structures. However, many ligands are non-drug-like and had to be excluded from the training set. The refinement process resulted in about 53k protein-ligand structures used for training and validation. The complete refinement process is described in figure 7 in the supplementary information.

The validation set was then created from the refined BioLiP dataset. The importance of having an unbiased validation set has been demonstrated numerous times in the literature (Davis et al., 2020; Wu et al., 2018; Wallach & Heifets, 2018; Tran-Nguyen et al., 2020). Unfortunately, most of these recent unbiasing approaches rely on activity data, which is not available for the BioLiP dataset. Therefore, we decided to apply the Kennard-Stone (KS) algorithm to generate a robust validation set (Kennard & Stone, 1969; Xu & Goodacre, 2018). The KS algorithm works by selecting the two most distant samples first. Then selecting the third sample based on the distance from the first two, and so on. KS produces a uniform selection of samples across the dataset, including samples from the boundary of the dataset. RDKit fingerprints were generated for every ligand in the dataset and a sample-wise distance matrix was created using the Tanimoto similarity metric. KS was performed using this matrix and the size of the desired validation set, 5,233, which represents 10% of the full dataset.

2.2.2 Independent Test Sets

The PDBbind core set is a subset of 285 high resolution, manually curated protein-ligand structures used to validate docking methods (Su et al., 2018). We chose to use the latest 2016 PDBbind core set to evaluate our method for the task of re-docking. Because these structures are also contained in the larger BioLiP dataset, we excluded them from the training and validation sets. In addition to re-docking, our method was also evaluated on the task of cross-docking, wherein a ligand is docked to an independent protein structure which hasn't been influenced by the ligand. Cross-docking

is more difficult but also more rigorous in evaluating docking methods with regards to real-world applications. To this end, the DISCO cross-docking dataset was employed as a second test set (Wierbowski et al., 2020). The dataset features 94 targets with an average of 50 ligands per target. As none of these complexes were included in the BioLiP dataset, nothing had to be removed from the training or validation sets.

2.3 FEATURIZATION

As the ligand and binding site graphs are processed by the model separately, and are have different levels of granularity, different featurization schemes were used for each. A combination of embeddings generated using pre-trained, self-supervised neural networks and hand-crafted features were used. The ligand embeddings were generated using a self-supervised graph transformer known as GROVER (Rong et al., 2020). Hand-crafted features for the ligand describe the physical and chemical nature of each atom. The binding site embeddings were generated using a state-of-the-art protein sequence transformer model known as ESM1b (Rives et al., 2021). Protein hand-crafted features describe the sequence and structure of each residue. All hand-crafted features were encoded and normalized as needed. A complete description of the featurization is included in section 6.2 of the supplementary information.

2.4 MODEL TRAINING

The model was trained to minimize the mean squared error between the predicted and known distances. Training was done using the Adadelta optimizer (Zeiler, 2012) with a learning rate of 0.01. The Adadelta optimizer was found to have better stability and performance compared to Stochastic Gradient Descent or Adam optimizers. The model was trained on the full training set (47,095 structures) for a total of 100 epochs. The validation loss was calculated following each epoch and was used for the selection of the best model weights. The training and validation loss improved significantly during training, and good correlation between the two indicates there was no overfitting of the training data.

2.5 RECONSTRUCTION

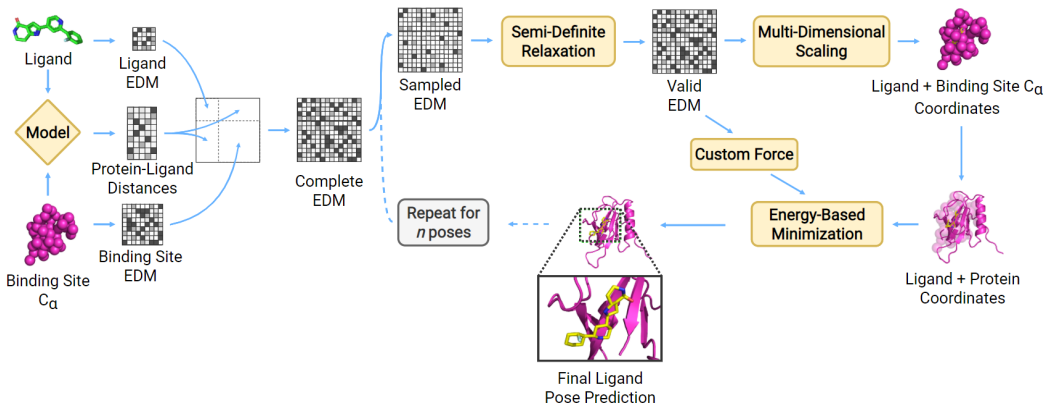

Figure 2: Diagram of the reconstruction process.

In order to reconstruct the Cartesian coordinates, first the predicted distance matrix, that contains only intermolecular protein-ligand distances, must be completed into the full square EDM. This is accomplished through tiling the ligand distance matrix D_{lig}, binding site distance matrix D_{bs}, and the predicted protein-ligand distance matrix D_{pred} as depicted in equation 7.

$$D_{complete} = \begin{bmatrix} D_{lig} & D_{pred}^T \\ D_{pred} & D_{bs} \end{bmatrix} \tag{7}$$

The exact distance matrix of the binding site is computed from the original X-ray structure, as the backbone atoms (including C_α atoms) are constrained during pose generation. However, the exact distance matrix of the ligand is not known prior to having the docked pose. Using the native pose would bias the reconstruction towards the correct result and would show higher performance than actually achieved. Therefore, the ligand distance matrix used in reconstruction is generated from an unbiased conformation generated with RDKit.

In order to generate an ensemble of different poses, our method relies on randomly sampling from the predicted protein-ligand distances and only using these samples to reconstruct the pose. Sampling 50% of distances was found to be a good balance between increased diversity while maintaining accuracy. Although, this sampling introduces a new problem; The EDM is now sparse and the missing values will lead to a deformed structure. Thus, an EDM completion algorithm called Semi-Definite Relaxation (SDR) was used to complete the matrix, resulting in a valid EDM.

Given a precise EDM, Multi-Dimensional Scaling (MDS) is capable of reconstructing the Cartesian coordinates exactly. MDS uses an intermediate representation called a Gram matrix in order to accomplish this conversion (Hoffmann & Noé, 2019). The relationship between an EDM D and its corresponding Gram matrix G is defined in equation 8.

$$G = \frac{1}{2}(D_{1j} + D_{i1} - D_{ij}) \tag{8}$$

Because the Gram matrix G is real symmetric matrix, it can be decomposed into eigenvalues and eigenvectors denoted in equation 9 as S and U respectively.

$$G = USU^T = (U\sqrt{S})(U\sqrt{S})^T \tag{9}$$

The Cartesian coordinates $X \in \mathbb{R}^{3N}$ can then be reconstructed from U and S using equations 10 and 11 where N is the total number of protein and ligand nodes.

$$V = U \cdot diag(\sqrt{S}) \tag{10}$$

$$X = \begin{bmatrix} V_{N-2} \\ V_{N-1} \\ V_N \end{bmatrix} \tag{11}$$

where the last three row vector of matrix V are used.

However, even when small levels of noise are present in the distance matrix, as is the case with distances predicted by a neural network, reconstructing the coordinates without distortion becomes considerably harder. MDS can attempt to reconstruct the coordinates, but will often lead to unrealistic molecular conformations with extremely unfavorable bonds and angles. Therefore, we chose to develop a new approach which uses an energy-based model to refine the atomic coordinates into a reasonable docked pose which satisfies both the predicted protein-ligand distances as well as physical constraints described by a molecular force field.

The approach is to utilize the predicted distances as an additional harmonic restraint force between atoms of the protein and atoms of the ligand during energy minimization. This allows us to perform a correction on the initial reconstruction so that both the predicted distances are satisfied and the potential energy of the protein-ligand complex is minimized. This can be formalized with the following equations, where U represents the potential energy as a combination of U_{ff}, the potential energy of the protein-ligand system described by the ff14SB (Maier et al., 2015) and gaff-2.11 (Wang et al., 2004) force fields, and U_{pred}, the harmonic restraints imposed by the predicted distances.

$$U = U_{ff} + U_{pred} \tag{12}$$

$$U_{pred} = \sum_{ij} \begin{cases} \frac{k}{D_{ij}} * (R_{ij} - D_{ij})^2, & D_{ij} <= C \\ 0, & D_{ij} > C \end{cases} \tag{13}$$

U_{pred} is a harmonic potential based on the difference between current distances R_{ij} (during energy minimization) and the predicted distances D_{ij}. Additionally, a cutoff C is introduced so that restraints are only introduced for distances smaller than a prescribed value. Typically, harmonic forces have a constant spring coefficient k defined by the two bonded atom types. However, in our case we chose a spring constant that is inversely proportional to the predicted distance. This signifies that small distances represent important, strong interactions between protein and ligand, e.g. hydrogen bonds. Those interaction distances should be strongly confined to the predicted values and therefore have a larger force constant than larger distances. The C_α atoms of the protein were harmonically restrained with a spring constant of 2 kJ mol^{-1} Å^{-2} during minimization. This prevented large changes to the backbone, while sidechains were allowed to move freely. Minimization was performed until a tolerance of 0.5 kJ/mol was reached. The hyperparameters C and k were optimized using a grid search with a smaller subset of the training set, and were set to be 10.0 Å and 2.0 kJ mol^{-1} Å^{-2} respectively. Finally, the ensemble of poses were ranked according to the final potential energy U.

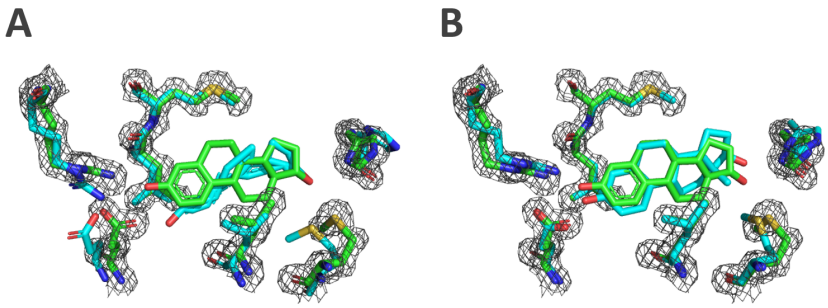

Figure 3: Example of a ligand reconstruction before (A) and after (B) minimization procedure. The docked ligand structure is shown in cyan and the known crystal structure is shown in green. The electron density map is also shown as a gray mesh to highlight how the residues are initially deformed. In the unminimized structure, the position of the ligand is generally correct. However some bonds and angles are unrealistic. The minimized ligand has corrected bond lengths and angles, and is in better agreement with the known pose. Additionally, some of the nearby residues adopt a conformation similar to the known ligand-bound structure.

2.6 Benchmark Comparisons

In order to compare our results to traditional docking methods, we employed two different docking software packages for benchmarking. GeauxDock is a high-performance docking software developed to be used on CPU or GPU. It is based on a Monte Carlo search algorithm and has a scoring function composed of physical-based energy terms and knowledge-based potentials (Fang et al., 2016). Importantly, GeauxDock only considers C_α atoms of the receptor, thereby making it implicitly flexible like our method.

Autodock Vina is another docking method which is widely used in the context of protein-ligand docking(Koes et al., 2013; Pagadala et al., 2017). For this work, a fork of AutoDock Vina called Smina was used to run the protocol for its ease-of-use and performance improvements over the original program(Koes et al., 2013). As opposed to our method and GeauxDock, Autodock Vina uses all atoms of the receptor. The search space for AutoDock Vina was defined using the same protocol as the original paper (Koes et al., 2013). The box was centered on the binding site residues with sides extending 8.0Å from the native ligand with a minimum length of 22.5Å per side. No flexible residues were used and all other settings were left as their default value.

In the context of re-docking, using all atoms of the receptor imposes a bias towards the final pose and can lead to deceptively good results (Jain, 2008; 2009; Jain & Nicholls, 2008). Therefore, AutoDock Vina is not a suitable tool for comparison for the re-docking task. This bias is not an issue when using methods which use C_α atoms alone to represent the receptor, as the induced side chain conformation is discarded entirely. For this reason, GeauxDock was selected as the primary

benchmark for the re-docking task. In the context of cross-docking, ligands are docked to a protein structure resolved independently from the ligand. This provides a more realistic benchmark, better aligned to the real-world use of docking programs in structure-based drug design. Moreover, cross-docking lessens the bias imposed by using all atoms of the receptor. For this reason, AutoDock Vina was selected as the primary benchmark for the cross-docking task.

3 RESULTS

3.1 RE-DOCKING

First, we summarize the results for the re-docking task. A random sample of docked poses from the PDBbind core test set is shown in figure 4 superimposed with their native pose. In comparison to GeauxDock, our model had superior performance with respect to RMSD and rate of success among the top-1, top-3, and top-5 ranked poses (Figure 5). In terms of RMSD, our method outperformed GeauxDock with an average improvement of 1.3Å 1.1Å and 0.6Å for the top-1, top-3, and top-5 poses respectively. In terms of rate of success, our method had a 2.3, 1.7, and 1.3-fold improvement over GeauxDock for the top-1, top-3, and top-5 poses respectively. Success for the re-docking task was defined as having a docked pose with less than 2.0Å RMSD from the native pose. Another comparison using a higher cutoff of 3.5Å is provided in figure 9 of the supplementary information.

Although AutoDock Vina is not a suitable comparison for the re-docking task, it was still run for the PDBbind core test set. The results exceeded our method in several categories. This test, however, is not a fair comparison due to the artificial enhancement of AutoDock Vina results by using all receptor atoms in their native conformation in a re-docking setting. Nonetheless, there were about 50 systems where our method outperformed Autodock Vina in the top-1. While the improvement in most of these cases was modest, a handful of examples showed significant improvement of 7Å or more. These examples, shown in figure 10 in the supplementary information, failed entirely using Vina but succeeded with our method.

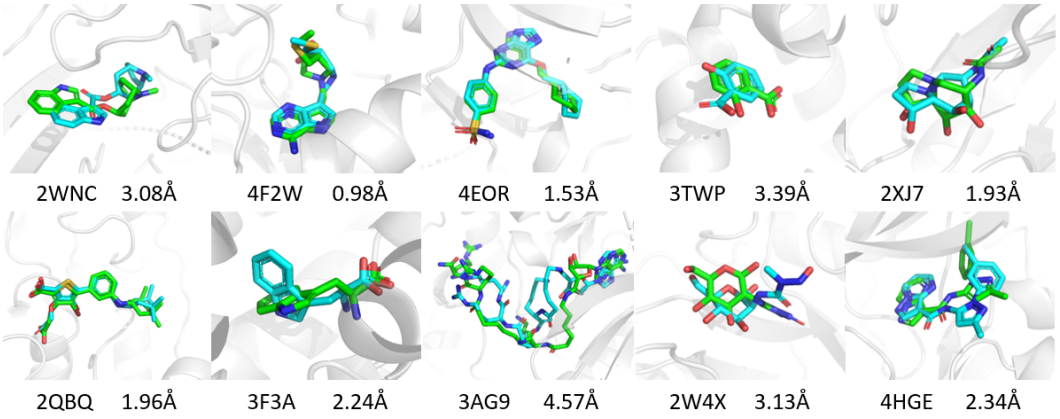

Figure 4: Non-cherry-picked docking results from test set. Only the top-ranked pose is shown. Docked poses are in cyan and native poses are in green. The corresponding PDB ID and RMSD to the native pose is listed below each frame.

3.2 CROSS-DOCKING

Next, we summarize the results for the cross-docking task. As with re-docking, performance was assessed in terms of RMSD and rate of success among the top-1, top-3, and top-5 ranked poses. Among the top-1 poses, our method was successful in 52% of systems, an improvement from 40% with AutoDock Vina. Improvement among the top-3 and top-5 poses was more modest. In terms of RMSD, our method performed exceptionally well in the top-1 with around 1.0 Å improvement in mean RMSD. The performance was also examined for each system. Some difficult targets, such as Adenosine A2a receptor, Farnesyltransferase, HIV integrase, and Histone deacetylase 8 all had

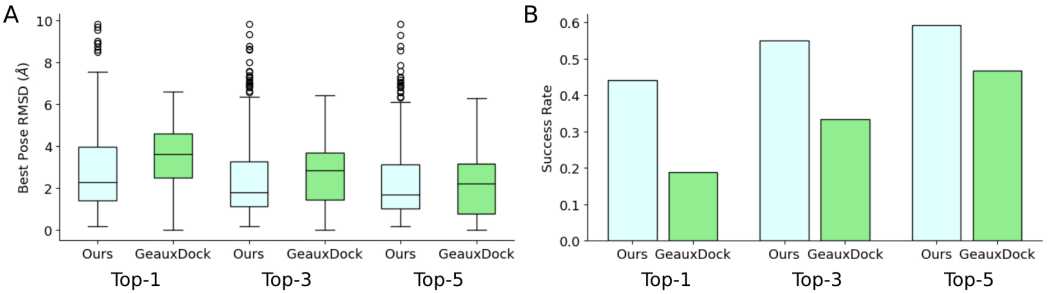

Figure 5: Results from the re-docking task using our method versus GeauxDock. A) Best pose RMSD (Å) among the top-1, top-3, and top-5 ranked poses. B) Success rate among the top-1, top-3, and top-5 ranked poses. Success is defined as having a pose less than 2.0Å.

greater than 2-fold improvements in RMSD. A full overview of the cross-docking results by system is included in tables 3 and 4 of the supplementary information. GeauxDock was not run for the cross-docking dataset.

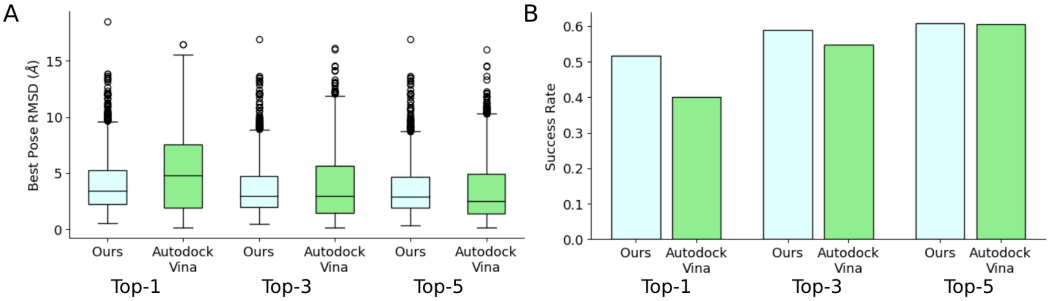

Figure 6: Results from the cross-docking task using our method versus AutoDock Vina. A) Best pose RMSD (Å) among the top-1, top-3, and top-5 ranked poses. B) Success rate among the top-1, top-3, and top-5 ranked poses. Success is defined as having a pose less than 3.5Å.

4 Discussion

Despite the excellent results, there are still a couple limitations to our method that should be pointed out. First, the diversity of the generated poses is not as high as traditional docking methods. The distance prediction model only generates a single distance matrix and thus the reconstruction procedure often leads to very similar poses. As such, our method performs exceptionally well in the top-1 rank but only improves modestly when considering lower ranked poses. Another limitation is the ranking procedure. The ranking used in this study is simple and involves no additional scoring function. The results do show some improvement when considering the lower ranked poses. Therefore, there is an opportunity to boost performance even further by adding an additional scoring function, either based on traditional scoring functions or a deep neural network. We hope that future work can address these limitations and develop our method further.

5 Conclusion

In this study, we proposed a novel docking method capable of generating high quality poses for a diverse set of protein and ligand structures. The method is based on a combination of two independent equivariant graph neural networks for protein and ligand, combined by a multi-layer perceptron model to predict the EDM between the two entities. Pose generation is based on EDM completion, reconstruction based on MDS and energy-based models. The method was rigorously evaluated on two independent test sets, covering both re-docking and cross-docking tasks. Moreover, two independent docking programs, GeauxDock for re-docking and AutoDock Vina for cross-docking were

selected as comparison. Our method showed superior performance in both tasks in terms of RMSD and rates of success. Additionally, our methods performance is uncoupled from a time-consuming search algorithm and the need to enumerate many possible docked poses. Together, these factors show that deep learning models are capable of flexible and efficient protein-ligand docking. This approach is a powerful new paradigm which will be investigated further and used to accelerate modern structure-based drug discovery.

REFERENCES

Huali Cao, Jingxue Wang, Liping He, Yifei Qi, and John Z Zhang. Deepddg: predicting the stability change of protein point mutations using neural networks. *Journal of chemical information and modeling*, 59(4):1508–1514, 2019.

Brian Davis, Kevin Mcloughlin, Jonathan Allen, and Sally Ellingson. Split optimization for protein/ligand binding models. *arXiv preprint arXiv:2001.03207*, 2020.

Jiyu Fan, Ailing Fu, and Le Zhang. Progress in molecular docking. *Quantitative Biology*, pp. 1–7, 2019.

Ye Fang, Yun Ding, Wei P Feinstein, David M Koppelman, Juana Moreno, Mark Jarrell, J Ramanujam, and Michal Brylinski. Geauxdock: accelerating structure-based virtual screening with heterogeneous computing. *PloS one*, 11(7):e0158898, 2016.

Moritz Hoffmann and Frank Noé. Generating valid euclidean distance matrices. *arXiv preprint arXiv:1910.03131*, 2019.

Ajay N Jain. Bias, reporting, and sharing: computational evaluations of docking methods. *Journal of computer-aided molecular design*, 22(3):201–212, 2008.

Ajay N Jain. Effects of protein conformation in docking: improved pose prediction through protein pocket adaptation. *Journal of computer-aided molecular design*, 23(6):355–374, 2009.

Ajay N Jain and Anthony Nicholls. Recommendations for evaluation of computational methods. *Journal of computer-aided molecular design*, 22(3):133–139, 2008.

Xiaoyang Jing and Jinbo Xu. Fast and effective protein model refinement using deep graph neural networks. *Nature Computational Science*, 1(7):462–469, 2021.

Wolfgang Kabsch and Christian Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers: Original Research on Biomolecules*, 22(12):2577–2637, 1983.

Ronald W Kennard and Larry A Stone. Computer aided design of experiments. *Technometrics*, 11 (1):137–148, 1969.

Prashant S Kharkar, Sona Warrier, and Ram S Gaud. Reverse docking: a powerful tool for drug repositioning and drug rescue. *Future medicinal chemistry*, 6(3):333–342, 2014.

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.

DE Koshland Jr. Application of a theory of enzyme specificity to protein synthesis. *Proceedings of the National Academy of Sciences of the United States of America*, 44(2):98, 1958.

Gred Landrum. Rdkit. `https://github.com/rdkit/rdkit`, 2021.

Amr H Mahmoud, Jonas F Lill, and Markus A Lill. Graph-convolution neural network-based flexible docking utilizing coarse-grained distance matrix. *arXiv preprint arXiv:2008.12027*, 2020a.

Amr H Mahmoud, Matthew R Masters, Ying Yang, and Markus A Lill. Elucidating the multiple roles of hydration for accurate protein-ligand binding prediction via deep learning. *Communications Chemistry*, 3(1):1–13, 2020b.

James A Maier, Carmenza Martinez, Koushik Kasavajhala, Lauren Wickstrom, Kevin E Hauser, and Carlos Simmerling. ff14sb: improving the accuracy of protein side chain and backbone parameters from ff99sb. *Journal of chemical theory and computation*, 11(8):3696–3713, 2015.

Ian K McDonald and Janet M Thornton. Satisfying hydrogen bonding potential in proteins. *Journal of molecular biology*, 238(5):777–793, 1994.

Andrew T McNutt, Paul Francoeur, Rishal Aggarwal, Tomohide Masuda, Rocco Meli, Matthew Ragoza, Jocelyn Sunseri, and David Ryan Koes. Gnina 1.0: molecular docking with deep learning. *Journal of cheminformatics*, 13(1):1–20, 2021.

Nataraj S Pagadala, Khajamohiddin Syed, and Jack Tuszynski. Software for molecular docking: a review. *Biophysical reviews*, 9(2):91–102, 2017.

Janaina Cruz Pereira, Ernesto Raul Caffarena, and Cicero Nogueira Dos Santos. Boosting docking-based virtual screening with deep learning. *Journal of chemical information and modeling*, 56 (12):2495–2506, 2016.

Matthew Ragoza, Joshua Hochuli, Elisa Idrobo, Jocelyn Sunseri, and David Ryan Koes. Protein–ligand scoring with convolutional neural networks. *Journal of chemical information and modeling*, 57(4):942–957, 2017.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15), 2021.

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.

Victor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E (n) equivariant graph neural networks. In *International Conference on Machine Learning*, pp. 9323–9332. PMLR, 2021.

Yonatan Savir and Tsvi Tlusty. Conformational proofreading: the impact of conformational changes on the specificity of molecular recognition. *PloS one*, 2(5):e468, 2007.

Hannes Stärk, Octavian-Eugen Ganea, Lagnajit Pattanaik, Regina Barzilay, and Tommi Jaakkola. Equibind: Geometric deep learning for drug binding structure prediction. *arXiv preprint arXiv:2202.05146*, 2022.

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, 2018.

Viet-Khoa Tran-Nguyen, Célien Jacquemard, and Didier Rognan. Lit-pcba: An unbiased data set for machine learning and virtual screening. *Journal of chemical information and modeling*, 60 (9):4263–4273, 2020.

Izhar Wallach and Abraham Heifets. Most ligand-based classification benchmarks reward memorization rather than generalization. *Journal of chemical information and modeling*, 58(5):916–932, 2018.

Junmei Wang, Romain M Wolf, James W Caldwell, Peter A Kollman, and David A Case. Development and testing of a general amber force field. *Journal of computational chemistry*, 25(9): 1157–1174, 2004.

Xiao Wang, Sean T Flannery, and Daisuke Kihara. Protein docking model evaluation by graph neural networks. *Frontiers in Molecular Biosciences*, 8:402, 2021.

Shayne D Wierbowski, Bentley M Wingert, Jim Zheng, and Carlos J Camacho. Cross-docking benchmark for automated pose and ranking prediction of ligand binding. *Protein Science*, 29(1): 298–305, 2020.

Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.

Yun Xu and Royston Goodacre. On splitting training and validation set: a comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. *Journal of Analysis and Testing*, 2(3):249–262, 2018.

Jianyi Yang, Ambrish Roy, and Yang Zhang. Biolip: a semi-manually curated database for biologically relevant ligand–protein interactions. *Nucleic acids research*, 41(D1):D1096–D1103, 2012.

Matthew D Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

Liangzhen Zheng, Jingrong Fan, and Yuguang Mu. Onionnet: a multiple-layer intermolecular-contact-based convolutional neural network for protein–ligand binding affinity prediction. *ACS omega*, 4(14):15956–15965, 2019.

6 SUPPLEMENTARY INFORMATION

6.1 DATASET PREPARATION

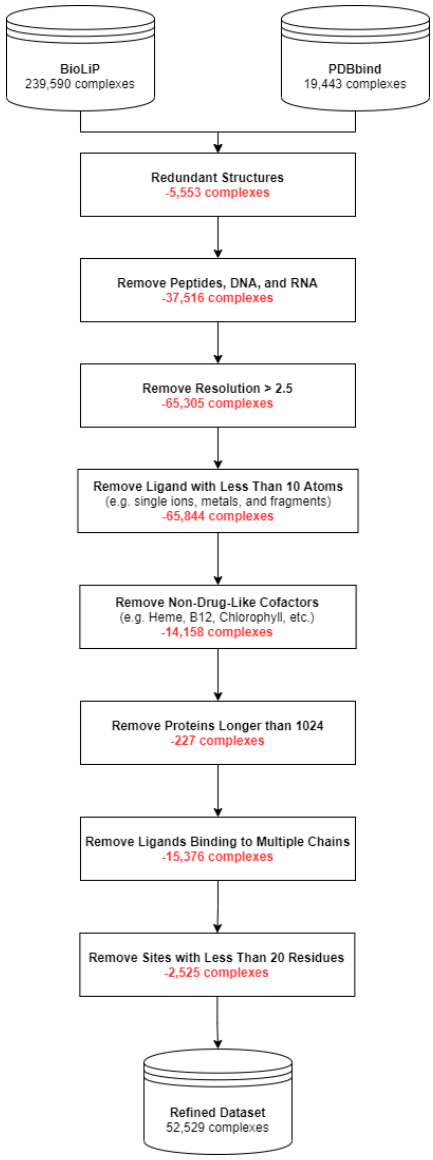

Figure 7: Diagram of the dataset creation and refinement process. Eight different filters were applied to the original dataset to filter out complexes with non-drug-like ligands. First, structures from BioLiP and PDBbind were combined and redundant structures were removed. Then, ligands marked as peptide, DNA, or RNA were removed to focus the dataset on small molecules. In order to have confidence in the protein structure and native ligand pose, structures with resolution greater than 2.5 Å were removed. Then, ligands with less than 10 atoms, such as single ions, metals, and small fragments were removed. Followed by non-drug-like cofactors such as hemes, B12, and chlorophyll. Proteins with a sequence longer than 1024 were removed as they could not be processed by the ESM model. Additionally, ligands in contact with multiple protein chains were removed as our method is only setup to handle a single protein target. And finally, binding sites with less than 20 residues were removed as these ligands bind are highly solvent exposed and do not have a well defined native conformation.

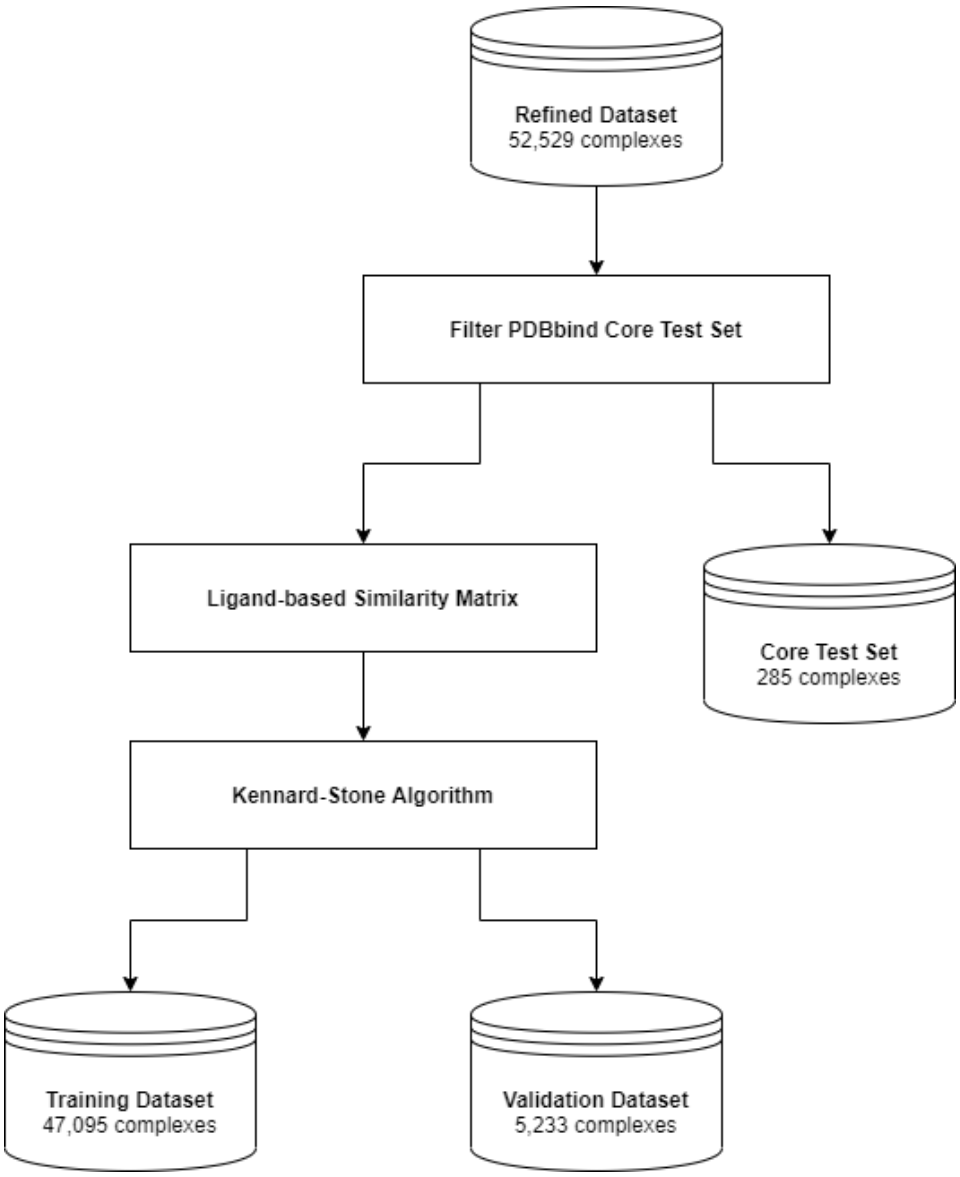

Figure 8: Diagram of the dataset splitting workflow. The refined BioLiP dataset contained 53k protein-ligand complexes. As there was some overlap with the PDBbind core set, redundant structures were removed from the refine set. Then, a ligand-based similarity matrix was constructed using molecular fingerprints and the Tanimoto similarity metric. This matrix was used by the Kennard-Stone algorithm to split the data and create a robust validation set representing 10% of the full dataset size.

6.2 FEATURIZATION DETAILS

6.2.1 BINDING SITE FEATURES

We employed two methods for featurizing the binding site nodes from different perspectives. In the first approach, we obtained sequence-based embeddings generated by the Evolutionary Scale Modeling (ESM) deep neural network (Rives et al., 2021). In the second approach, we collected several handcrafted features which describe the geometry and chemical environment of each residue.

The ESM-1b model was used to generate the sequence embeddings. This model contains roughly 650M parameters in 33 transformer layers. The model was trained on a diverse dataset of 27M unique protein sequences. It is considered a general-purpose protein language model and has shown state-of-the-art performance when applied to structure prediction tasks (Rives et al., 2021). To generate the embeddings, we provide the complete protein sequence to the ESM-1b model. The sequence is transformed through the stack of layers and is retrieved after the last hidden layer. The resulting embeddings have a rank of 1280 and are generated for each letter in the sequence. Embeddings specific to the binding site residues were then extracted.

The second featurization method combines several hand-crafted features which describe the geometric and chemical environment of each residue. The selected features were used in previous works on deep learning of protein structures (Jing & Xu, 2021; Cao et al., 2019). The residue type was one-hot encoded as one of the twenty standard amino acids with an additional category for non-standard residues. DSSP was used to assign the secondary structure and calculate Solvent Accessible Surface Area (SASA) for each residue (Kabsch & Sander, 1983). The secondary structure was one-hot encoded as one of eight types. Hydrogen bonding of the protein was calculated using HBPLUS (McDonald & Thornton, 1994). The hydrogen bonds were segmented into eight categories: one for each possible combination of donor/acceptor, backbone/sidechain, and occupied/accessible. The position of the residue within the full sequence is included and normalized by the sequence length. Finally, the backbone dihedral angles phi, psi, and omega are included. The periodicity of the angles are handled by transforming with sin and cos functions. This set of features is summarized in table 1.

6.2.2 LIGAND FEATURES

Similar to the binding site features, we also chose to explore multiple approaches for the featurization of ligand atoms. In the first method, we use another trained self-supervised transformer model, GROVER (Rong et al., 2020), to generate abstract atom-level embeddings. In the second method, we again include a collection of handcrafted features including several known to have a significant influence on protein-ligand binding.

The GROVER model is another large-scale transformer model. However, rather than working on a sequence representation, it works directly on the graph representation of molecules. The model contains roughly 100M parameters and was trained on 10M small, drug-like molecules.

The hand-crafted ligand features includes descriptors known to be important for binding. The atomic element was one-hot encoded into eight possible classes (C, O, N, P, B, S, Halogens, and Metals). Different halogens and metal elements were grouped together due to their sparsity in the dataset and similar chemical characteristics. Still, individual halogens and metal elements are able to be distinguished based on the other features such as atomic weight and radii. The hybridization type was also one-hot encoded into one of seven possible classes (S, SP, SP2, SP3, SP3D, SP3D2, and unknown). The aromaticity of ligand atoms were indicated by a single boolean value. Several other atom-level properties including atomic weight, atomic radii, partial charge, polar surface area (PSA), accessible surface area (ASA), and logP, were calculated using RDKit (Landrum, 2021). The hand-crafted ligand features are summarized in table 2.

Table 1: Hand-crafted features for protein C_α atoms.

Feature	**Encoding**	**Normalized**	**Size**
Residue	One-hot	N	21
Disulfide Bond	Boolean	N	1
Hydrogen Bond Counts	Integer	N	8
Solvent Accessible Surface Area	Float	Y	1
Position	Float	Y	1

Table 2: Hand-crafted features for ligand atoms.

Feature	**Encoding**	**Normalized**	**Size**
Element	One-hot	N	9
Hybridization	One-hot	N	7
Aromaticity	Boolean	N	1
Atomic Weight	Float	Y	1
Atomic Radii	Float	Y	1
Gasteiger Partial Charge	Float	Y	1
Topological Polar Surface Area	Float	Y	1
Accessible Surface Area	Float	Y	1
logP	Float	Y	2
Molar Refractivity	Float	Y	2

6.3 RESULTS

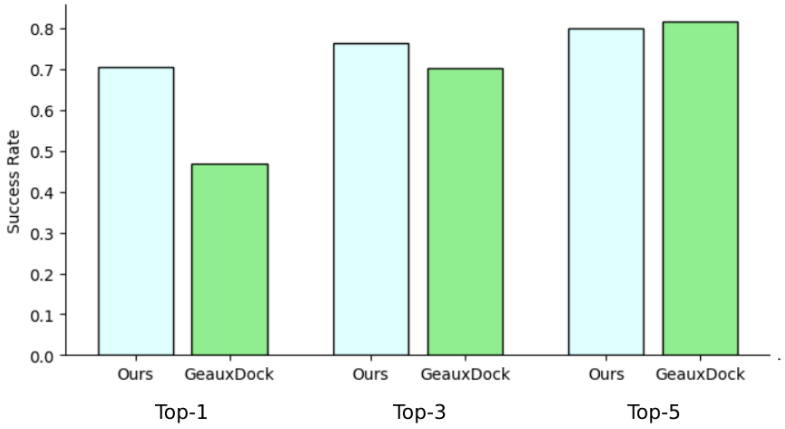

Figure 9: Success rate of the re-docking task using our method versus GeauxDock with success defined as having a pose less than 3.5Å.

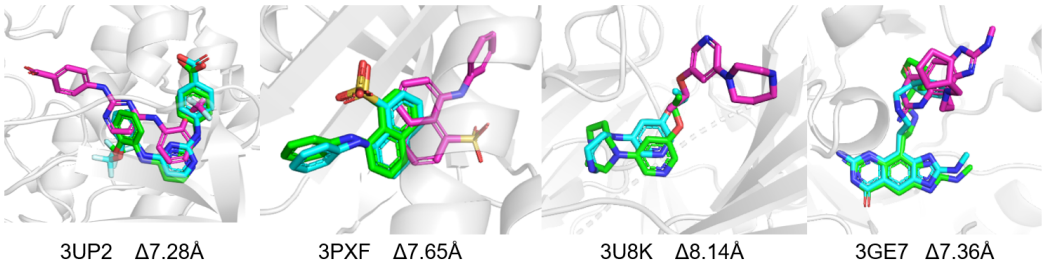

Figure 10: Cherry-picked examples from the re-docking test set which failed using AutoDock Vina but succeeded using our docking method. Autodock Vina poses are in magenta, our poses are in cyan, and the native poses are in green. The corresponding PDB ID and the improvement in RMSD to the native pose is listed below each frame.

Table 3: Cross-docking results by system.

Target	Reference Receptor	Num. of Ligands	AutoDock Vina Avg. RMSD (Å)	AutoDock Vina Success	Our Avg. RMSD (Å)	Our Success
AA2AR	3EML	2	10.24 ± 1.73	0.00%	4.06 ± 2.13	0.50%
ABL1	2HZI	38	8.33 ± 5.05	0.32%	4.69 ± 2.43	0.42%
ACE	3BKL	8	6.36 ± 2.67	0.12%	3.89 ± 1.98	0.50%
ACES	1E66	40	5.89 ± 2.99	0.20%	4.94 ± 2.65	0.40%
ADA	2E1W	15	3.84 ± 3.09	0.53%	2.60 ± 1.37	0.80%
ADA17	2OI0	19	5.39 ± 2.76	0.32%	3.65 ± 1.25	0.47%
ADRB1	2VT4	15	1.90 ± 1.01	0.87%	3.07 ± 1.16	0.73%
ADRB2	3NY8	11	2.97 ± 3.09	0.82%	3.08 ± 0.81	0.64%
AKT1	3CQW	10	5.05 ± 3.73	0.40%	2.62 ± 0.60	0.90%
AKT2	3D0E	7	6.27 ± 3.36	0.14%	3.01 ± 1.02	0.71%
ALDR	2HV5	1	1.15 ± 0.00	1.00%	1.99 ± 0.00	1.00%
AMPC	1L2S	40	4.72 ± 2.06	0.35%	3.56 ± 1.37	0.53%
ANDR	2AM9	94	2.17 ± 2.82	0.71%	2.17 ± 1.25	0.83%
AOFB	1S3B	1	1.97 ± 0.00	1.00%	2.05 ± 0.00	1.00%
BACE1	3L5D	281	6.43 ± 2.56	0.19%	3.42 ± 1.54	0.57%
BRAF	3D4Q	56	7.91 ± 3.47	0.14%	4.10 ± 2.53	0.48%
CAH2	1BCD	275	3.51 ± 2.40	0.54%	2.47 ± 1.25	0.85%
CASP3	2CNK	21	6.98 ± 2.73	0.10%	4.81 ± 1.73	0.29%
CDK2	1H00	306	5.27 ± 2.56	0.29%	3.64 ± 1.88	0.62%
CP2C9	1R9O	4	5.65 ± 3.08	0.25%	4.64 ± 2.21	0.25%
CP3A4	3NXU	1	5.76 ± 0.00	0.00%	4.85 ± 0.00	0.00%
CSF1R	3KRJ	12	6.26 ± 2.78	0.25%	4.59 ± 2.41	0.42%
CXCR4	3ODU	4	6.24 ± 0.50	0.00%	4.49 ± 0.54	0.00%
DEF	1LRU	10	4.60 ± 2.80	0.40%	4.32 ± 2.39	0.50%
DPP4	2I78	76	4.32 ± 3.50	0.54%	4.04 ± 1.32	0.43%
DYR	3NXO	14	4.75 ± 2.16	0.57%	2.40 ± 0.60	1.00%
EGFR	2RGP	90	6.13 ± 2.56	0.16%	4.03 ± 1.57	0.42%
ESR1	1SJ0	209	4.25 ± 3.00	0.50%	3.58 ± 1.49	0.53%
ESR2	2FSZ	33	2.36 ± 1.75	0.79%	3.47 ± 1.60	0.58%
FA10	3KL6	104	4.18 ± 3.53	0.62%	3.26 ± 1.14	0.70%
FA7	1W7X	44	4.10 ± 2.46	0.50%	4.03 ± 1.52	0.39%
FABP4	2NNQ	22	4.66 ± 1.94	0.18%	3.30 ± 1.31	0.50%
FAK1	3BZ3	18	5.04 ± 3.76	0.44%	5.15 ± 2.68	0.39%
FKB1A	1J4H	26	3.22 ± 3.29	0.69%	5.02 ± 2.16	0.19%
FNTA	3E37	14	9.33 ± 1.60	0.00%	4.17 ± 1.62	0.43%
GCR	3BQD	21	1.91 ± 1.79	0.81%	2.99 ± 1.66	0.62%
GLCM	2V3F	9	2.72 ± 1.46	0.89%	3.07 ± 1.23	0.44%
GRIA2	3KGC	83	4.71 ± 2.11	0.13%	2.14 ± 0.62	0.96%
GRIK1	1VSO	26	4.65 ± 1.50	0.23%	3.63 ± 0.93	0.42%
HDAC2	3MAX	4	2.61 ± 3.62	0.75%	1.46 ± 0.46	1.00%
HDAC8	3F07	8	7.29 ± 2.39	0.12%	3.12 ± 0.58	0.62%
HIVINT	3NF7	8	7.45 ± 1.39	0.00%	3.49 ± 1.61	0.75%
HIVPR	1XL2	415	7.13 ± 2.19	0.05%	7.02 ± 1.91	0.01%
HIVRT	3LAN	180	3.27 ± 2.62	0.68%	5.24 ± 1.23	0.07%
HMDH	3CCW	17	1.57 ± 0.64	0.94%	4.54 ± 1.45	0.29%
HS90A	1UYG	196	4.19 ± 2.57	0.45%	2.83 ± 1.52	0.77%
HXK4	3F9M	26	8.13 ± 2.90	0.12%	5.51 ± 2.62	0.31%
IGF1R	2OJ9	14	6.56 ± 2.46	0.07%	4.45 ± 2.22	0.29%
ITAL	2ICA	13	4.91 ± 2.48	0.31%	4.89 ± 1.68	0.15%
JAK2	3LPB	58	4.70 ± 3.17	0.45%	3.78 ± 1.66	0.53%

Table 4: Cross-docking results by system continued.

Target	Reference Receptor	Num. of Ligands	AutoDock Vina Avg. RMSD (Å)	AutoDock Vina Success	Our Avg. RMSD (Å)	Our Success
KIF11	3CJO	29	2.94 ± 2.03	0.55%	2.91 ± 0.97	0.79%
KIT	3G0E	5	4.00 ± 3.27	0.40%	3.91 ± 2.52	0.80%
KITH	2B8T	1	0.52 ± 0.00	1.00%	5.01 ± 0.00	0.00%
LCK	2OF2	31	5.59 ± 2.95	0.32%	3.90 ± 2.14	0.61%
LKHA4	3CHP	43	5.00 ± 2.75	0.47%	2.89 ± 1.52	0.72%
MAPK2	3M2W	13	2.62 ± 2.60	0.77%	3.47 ± 2.44	0.62%
MCR	2AA2	18	2.01 ± 2.61	0.78%	1.58 ± 0.46	1.00%
MET	3LQ8	60	6.34 ± 2.76	0.18%	3.67 ± 1.52	0.50%
MK01	2OJG	68	6.95 ± 2.30	0.04%	4.31 ± 1.86	0.38%
MK10	2ZDT	54	5.77 ± 3.15	0.31%	4.37 ± 1.99	0.46%
MK14	2QD9	186	5.28 ± 4.10	0.50%	5.61 ± 2.59	0.26%
MMP13	830C	33	5.56 ± 3.59	0.39%	3.66 ± 2.20	0.67%
MP2K1	3EQH	10	5.96 ± 3.54	0.30%	3.32 ± 1.67	0.80%
NRAM	1B9V	12	3.29 ± 2.12	0.42%	3.07 ± 1.07	0.58%
PA2GA	1KVO	8	3.07 ± 2.30	0.75%	3.66 ± 1.24	0.50%
PARP1	3L3M	27	2.55 ± 2.23	0.85%	1.82 ± 0.87	0.96%
PDE5A	1UDT	27	5.20 ± 2.82	0.37%	2.97 ± 1.11	0.81%
PGH1	2OYU	19	5.15 ± 2.17	0.26%	4.62 ± 1.52	0.26%
PGH2	3LN1	30	5.43 ± 3.07	0.27%	4.00 ± 1.96	0.47%
PLK1	2OWB	11	3.21 ± 2.48	0.55%	6.38 ± 2.17	0.09%
PNPH	3BGS	7	4.22 ± 2.99	0.43%	2.67 ± 1.50	0.71%
PPARA	2P54	17	7.39 ± 3.60	0.24%	3.47 ± 1.67	0.59%
PPARD	2ZNP	23	7.12 ± 3.29	0.26%	4.84 ± 1.95	0.30%
PPARG	2GTK	131	7.14 ± 2.95	0.17%	4.68 ± 2.31	0.42%
PRGR	3KBA	18	3.22 ± 2.17	0.61%	3.46 ± 1.21	0.56%
PTN1	2AZR	73	4.04 ± 3.06	0.58%	3.80 ± 2.41	0.53%
PUR2	1NJS	9	2.35 ± 2.76	0.89%	2.54 ± 0.54	0.89%
PYGM	1C8K	17	5.33 ± 3.40	0.35%	3.49 ± 1.13	0.47%
PYRD	1D3G	11	3.30 ± 3.00	0.64%	5.07 ± 1.64	0.27%
RENI	3G6Z	51	6.57 ± 2.71	0.16%	3.71 ± 1.07	0.47%
ROCK1	2ETR	14	4.56 ± 3.01	0.50%	2.50 ± 1.00	0.79%
RXRA	1MV9	44	2.58 ± 2.52	0.77%	3.25 ± 1.62	0.64%
SAHH	1LI4	2	7.74 ± 7.47	0.50%	3.35 ± 1.88	0.50%
SRC	3EL8	52	5.55 ± 3.43	0.33%	3.26 ± 1.73	0.65%
TGFR1	3HMM	21	4.38 ± 2.73	0.43%	3.43 ± 1.68	0.57%
THB	1Q4X	14	1.01 ± 0.74	1.00%	1.98 ± 0.79	0.93%
THRB	1YPE	207	2.93 ± 2.48	0.71%	3.21 ± 1.48	0.66%
TRY1	2AYW	172	4.13 ± 2.93	0.56%	3.16 ± 1.54	0.72%
TRYB1	2ZEC	11	2.95 ± 1.53	0.64%	7.30 ± 2.50	0.18%
TYSY	1SYN	11	6.59 ± 2.72	0.27%	3.96 ± 0.98	0.27%
UROK	1SQT	15	4.89 ± 3.30	0.53%	3.71 ± 2.15	0.53%
VGFR2	2P2I	24	5.19 ± 4.00	0.50%	4.03 ± 2.70	0.62%
WEE1	3BIZ	12	2.31 ± 2.90	0.83%	2.44 ± 0.81	0.92%
XIAP	3HL5	21	1.77 ± 1.57	0.90%	4.46 ± 2.02	0.33%