E3BIND: AN END-TO-END EQUIVARIANT NETWORK FOR PROTEIN-LIGAND DOCKING

Anonymous authors
Paper under double-blind review

ABSTRACT

In silico prediction of the ligand binding pose to a given protein target is a crucial but challenging task in drug discovery. This work focuses on blind flexible self-docking, where we aim to predict the positions, orientations and conformations of docked molecules. Traditional physics-based methods usually suffer from inaccurate scoring functions and high inference cost. Recently, data-driven methods based on deep learning techniques are attracting growing interest thanks to their efficiency during inference and promising performance. These methods usually either adopt a two-stage approach by first predicting the distances between proteins and ligands and then generating the final coordinates based on the predicted distances, or directly predicting the global roto-translation of ligands. In this paper, we take a different route. Inspired by the resounding success of AlphaFold2 for protein structure prediction, we propose E3Bind, an end-to-end equivariant network that iteratively updates the ligand pose. E3Bind models the protein-ligand interaction through careful consideration of the geometric constraints in docking and the local context of the binding site. Experiments on standard benchmark datasets demonstrate the superior performance of our end-to-end trainable model compared to traditional and recently-proposed deep learning methods.

1 INTRODUCTION

For nearly a century, small molecules, or organic compounds with small molecular weight, have been the major weapon of the pharmaceutical industry. They take effect by ligating (binding) to their target, usually a protein, to alter the molecular pathways of diseases. The structure of the protein-ligand interface holds the key to understanding the potency, mechanisms and potential side-effects of small molecule drugs. Despite huge efforts made for protein-ligand complex structure determination, there are by far only some 10^4 protein-ligand complex structures available in the protein data bank (PDB) (Berman et al., 2000), which dwarfs in front of the enormous combinatorial space of possible complexes between 10^{60} drug-like molecules (Hert et al., 2009; Reymond & Awale, 2012) and at least 20,000 human proteins (Gaudet et al., 2017; Consortium, 2019), highlighting the urgent need for *in silico* protein-ligand docking methods. Furthermore, a fast and accurate docking tool enables binding pose prediction for molecules that have yet to be synthesized, empowering mass-scale virtual screening (Lyu et al., 2019), which is a vital step in modern structure-based drug discovery (Ferreira et al., 2015). Compared with models that only output a scalar score (*e.g.* binding affinity) for each protein-ligand pair, docking software provides pharmaceutical scientists with an interpretable, information-rich result.

Being a crucial task, predicting the docked pose of a ligand is also a challenging problem. Traditional docking methods (Halgren et al., 2004; Morris et al., 1996; Trott & Olson, 2010; Coleman et al., 2013) rely on physics-inspired scoring functions and extensive conformation sampling to obtain the predicted binding pose. Some deep learning methods focus on learning a more accurate scoring function (McNutt et al., 2021; Méndez-Lucio et al., 2021), but often at the cost of even lower inference speed due to their adoption of the sampling-scoring framework. Distinct from the above methods, TankBind (Lu et al., 2022) drop the burden of conformation sampling by predicting the protein-ligand distance matrix, then converting the distance map to a docked pose using gradient descent. The optimization objective is the weighted sum of the protein-ligand distance error with respect to the predicted distance map and the intra-ligand distance error w.r.t. the reference ligand conformation. This two-stage approach might run into problems during the distance-to-coordinate

transformation, as the predicted distance map is, in many cases, not a valid Euclidean distance matrix (Liberti et al., 2014). Recently, Stärk et al. (2022) proposed EquiBind, an equivariant model that directly predicts coordinates of the docked pose. EquiBind refines the ligand conformation with a graph neural network, then roto-translates the refined ligand into the pocket using a key-point alignment mechanism. Compared to the popular docking baselines (Hassan et al., 2017; Koes et al., 2013), EquiBind enjoys significant speedup, but it's performance is not strong enough, underscoring the urgent need for capacity increase of the model. Overall, rapid generation of accurate binding pose remains an unsolved problem.

In this paper, we move one step forward in this important direction and propose E3Bind, the first end-to-end equivariant network that iteratively docks the ligand into the binding pocket. Inspired by AlphaFold2 (Jumper et al., 2021), our model comprises a Trioformer feature extractor and an iterative coordinate refinement module. The Trioformer encodes the protein and ligand graphs into three information-rich embeddings: the protein residue embeddings, the ligand atom embeddings and the protein-ligand pair embeddings, where the pair embeddings are fused with geometry awareness to enforce the implicit constraints in docking. Our iterative coordinate refinement module decodes the rich representations into E(3)-equivariant coordinate updates. We further propose a self-confidence predictor to select the final pose and evaluate the soundness of our pose predictions. E3Bind is trained end-to-end with loss directly defined on the output coordinates and can directly output docked ligand coordinates, relieving the burden of conformation sampling or distance-to-coordinate transformation. The iterative coordinate update scheme enables careful consideration of the local context, which increases model capacity.

Our contributions can be summarized as follows:

- We formulate the docking problem as an iterative refinement process, where the model updates the ligand coordinates based on the current context at each iteration.
- We propose an end-to-end E(3) equivariant network to produce the coordinate updates. The network comprises an expressive geometric-aware encoder and an equivariant context-aware coordinate update module.
- Quantitative results show that our method outperforms traditional docking software and recently proposed deep learning based models.

2 Related Works

Protein-ligand docking. Traditional approaches to protein-ligand docking (Morris et al., 1996; Halgren et al., 2004; Coleman et al., 2013) mainly adopt a sampling, scoring, ranking, and fine-tuning paradigm, with AutoDock Vina (Trott & Olson, 2010) being a popular example. Each part of the docking pipeline has been extensively studied in literature to increase both the accuracy and the speed (Durrant & McCammon, 2011; Liu et al., 2013; Hassan et al., 2017; Zhang et al., 2020). Multiple subsequent works use deep-learning on 3D voxels (Ragoza et al., 2017; Francoeur et al., 2020; McNutt et al., 2021; Bao et al., 2021) or graphs (Méndez-Lucio et al., 2021) to improve the scoring functions. Nevertheless, these methods are inefficient, and often takes minutes or even more to predict docking poses of a single protein-ligand pair, which hinders the accessibility of large-scale virtual screening experiments.

Recently, methods that directly model the distance geometry between protein-ligand pairs have been investigated (Masters et al., 2022; Lu et al., 2022; Zhou et al., 2022). They adopt a two-stage approach for docking, and generate docked poses using post-optimization algorithms based on predicted protein-ligand distance map. Advanced techniques in geometric deep learning, e.g. triangle attention (Jumper et al., 2021), have been leveraged to encourage the geometrical consistency of the distance map (Lu et al., 2022). To by pass the error prone two-stage framework, EquiBind (Stärk et al., 2022) propose a fully differentiable equivariant model, which directly predicts coordinates of docked poses with a novel attention-based key-point alignment mechanism (Ganea et al., 2021b). Despite being efficient, EquiBind fails to beat popular docking baselines without fine-tuning its predictions. This stress the importance for increasing the capacity of the model.

Molecular conformation generation. Deep learning has made great progress in predicting low-energy conformations given molecular graphs (Shi et al., 2021; Ganea et al., 2021a; Xu et al., 2022; Jing et al., 2022). State-of-the-art models are generally SE(3)-invariant or equivariant, and mainly

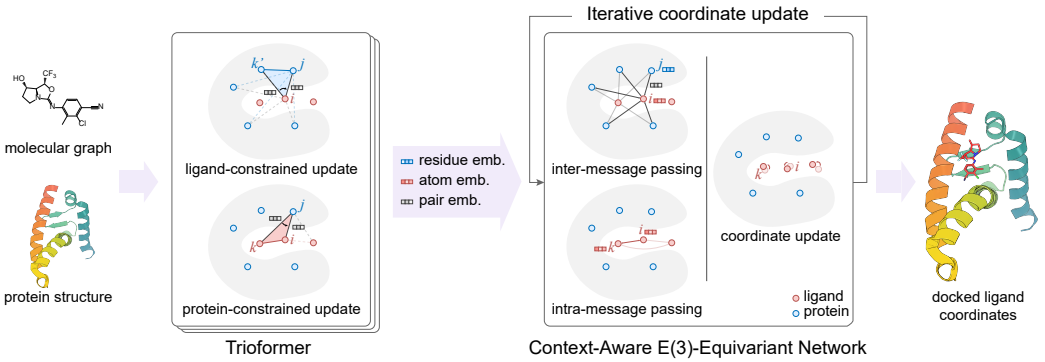


Figure 1: E3Bind model overview.

adopt the score-matching (Vincent, 2011; Song & Ermon, 2019) or diffusion (Ho et al., 2020) framework. In protein-ligand docking, we also aim to generate the conformation of the bound ligand. However it is not viable to directly apply these approaches as (1) there are much more atoms in the system and (2) the protein context must be carefully considered when generating the docked pose. Here we model the protein at the residue level and design our model to carefully capture the protein-ligand interactions.

Protein structure prediction. Predicting protein folds from sequences has long been a challenging task. Senior et al. (2020a); Wang et al. (2017); Senior et al. (2020b) use deep learning to predict the contact map, distance map or torsion angles between protein residues, and then convert them to coordinates using optimization-based methods. Recently, AlphaFold2 (Jumper et al., 2021) takes a leap forward by adopting an end-to-end approach with iterative coordinate refinement. It consists of an Evoformer to extract information from Multiple Sequence Alignments (MSAs) and structural templates, and a structure module to iteratively update the coordinates. Though this problem is very different from docking which models heterogeneous entities – a fixed protein structure and a ligand with unknown pose, in this paper we show that some ideas can be extended.

3 The E3Bind Model

E3Bind tackles the protein-ligand docking task with an encoder-decoder framework. Specifically, the protein and ligand graphs are first encoded by standard graph encoders. Pair embeddings are constructed between every protein residue - ligand atom pair. We then use a geometry-aware Trioformer to fully mix the protein, ligand and pair embeddings (Section 3.2). With the rich representations at hand, the iterative coordinate refinement scheme iteratively updates the ligand pose by a series of E(3)-equivariant networks to generate the final docked pose (Section 3.3). The model is trained end-to-end and capable of directly generating the docked pose (Section 3.4). An overview of E3Bind is shown in Figure 1.

3.1 Preliminaries

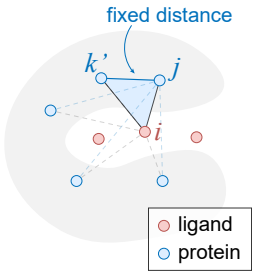
Notation and Input Feature Construction. The ligand is treated as an atom-level molecular graph $\mathcal{G}^{\mathrm{l}}$ where the edges denote chemical bonds. The ligand node features $\{\boldsymbol{h}_i^{\mathrm{l}}\}_{1\leq i\leq n_{\mathrm{l}}}$ are calculated by a TorchDrug (Zhu et al., 2022) implementation of the graph isomorphism network (GIN) (Xu et al., 2018), where n_{l} is the number of atoms in the ligand. Ligand coordinates are denoted as $\{\boldsymbol{x}_i^{\mathrm{l}}\}_{1\leq i\leq n_{\mathrm{l}}}$. Following Jing et al. (2020) and Lu et al. (2022), we represent the protein as a residue-level K-nearest neighbor graph $\mathcal{G}^{\mathrm{p}}$. Each protein node $j \in \{1,\ldots,n_{\mathrm{p}}\}$ has 3D coordinates $\boldsymbol{x}_j^{\mathrm{p}}$ (which corresponds to the position of the C_α atom of the residue), where n_{p} is the number of residues in the protein. The protein node features $\{\boldsymbol{h}_j^{\mathrm{p}}\}_{1\leq j\leq n_{\mathrm{p}}}$ are calculated with a geometric-vector-perceptron-based graph neural network (GVP-GNN). For each protein residue-ligand atom pair (i,j), we construct a pair embedding $\boldsymbol{z}_{ij}$ via the outer product module (OPM) which takes in protein and ligand embeddings: $\boldsymbol{z}_{ij} = \mathrm{Linear}\left(\mathrm{vec}\left(\mathrm{Linear}(\boldsymbol{h}_i^{\mathrm{l}}) \bigotimes \mathrm{Linear}(\boldsymbol{h}_j^{\mathrm{p}})\right)\right)$. For notation consistency, throughout this paper we will use i,k to index ligand nodes, and j,k' for protein nodes.

Problem Definition. Given a molecular graph of the ligand compound and a fixed protein structure, we aim to predict the binding (docked) pose of the ligand compound $\{\boldsymbol{x}_i^{\mathrm{l}*}\}_{1\leq i\leq n_{\mathrm{l}}}$. We focus on *blind* docking settings, where the binding pocket is not provided and has to be predicted by the model. In the *blind re-docking* setting, the docked ligand conformation is given, but its position and orientation relative to the protein is unknown. In the *flexible blind self-docking* setting, the docked ligand conformation needs to be inferred besides its position and orientation. The model is thus provided with an unbound ligand structure, which could be obtained by running the ETKDG algorithm (Riniker & Landrum, 2015) with RDKit (Landrum et al., 2013).

Equivariance. An important inductive bias in protein ligand docking is E(3) equivariance, *i.e.*, if the input protein coordinates are transformed by some E(3) transformation $g \cdot \{\boldsymbol{x}_j^{\mathrm{p}}\}_{1\leq j\leq n_{\mathrm{p}}} = \{\boldsymbol{R}\boldsymbol{x}_j^{\mathrm{p}} + \boldsymbol{t}\}_{1\leq j\leq n_{\mathrm{p}}}, \forall j = 1 \dots n_{\mathrm{p}}$, the predicted ligand coordinates should also be g-transformed: $F(g \cdot \{\boldsymbol{x}_j^{\mathrm{p}}\}_{1\leq j\leq n_{\mathrm{p}}}) = g \cdot F(\{\boldsymbol{x}_j^{\mathrm{p}}\}_{1\leq j\leq n_{\mathrm{p}}})$, where F is the coordinate prediction model. To inject such symmetry to our model, we adopt a variant of the equivariant graph neural network (EGNN) as the building block in our coordinate update steps (Satorras et al., 2021).

3.2 Extracting Geometry-Consistent Information with Trioformer

The E3Bind encoder extracts information-rich ligand atom embeddings $\boldsymbol{h}_i^{\mathrm{l}}$, protein residue embeddings $\boldsymbol{h}_j^{\mathrm{p}}$ and ligand-protein pair representation $\boldsymbol{z}_{ij}$ that captures the subtle protein-ligand interactions. Note that this is a non-trivial task since implicit geometric constraints must be incorporated in the representations. As shown in the figure on the right, the protein-ligand distance d_{ij} and $d_{ik'}$ can not be predicted independently. They are constrained by the fixed intra-protein distance $d_{jk'}$, as the protein structure is considered to be fixed during docking. In other words, by the triangle inequality, if residues j and k' are close, then ligand atom i cannot be both near j and far from k'. The same thing happens when we consider the given intra-ligand distance d_{ik}[1]. While using simple concatenation of protein and ligand embeddings as the final pair embeddings is common among previous methods (Méndez-Lucio et al., 2021; Masters et al., 2022), it dismisses the above geometry constraints and might result in geometrically inconsistent pose predictions.



Trioformer Overview. To tackle the above challenge, we use a deep stack of *Trioformer* blocks to intensively update the protein, ligand and pair embeddings. In each Trioformer block, information between the protein, ligand and pair embeddings are fully mixed. First, we update the protein and ligand node embeddings with multi-head cross-attention using the pair embeddings as attention bias. Next, we use protein and ligand embeddings to update the pair embeddings via the OPM: $\boldsymbol{z}_{ij} = \boldsymbol{z}_{ij} + \mathrm{OPM}\left(\boldsymbol{h}_i^{\mathrm{l}}, \boldsymbol{h}_j^{\mathrm{p}}\right)$. The pair embeddings then go through geometry-aware pair update modules described below to produce geometry-consistent representations. The final block outputs are the multi-layer perceptron (MLP)-transitioned protein, ligand and pair embeddings. Details of the Trioformer are described in Section B.

Geometry-Aware Pair Updates. To inject geometry awareness to the pair embeddings, we construct intra-ligand and intra-protein distance embeddings ($\boldsymbol{d}_{ik}^*$ and $\boldsymbol{d}_{jk'}^*$ respectively) and then use them to update the pair embeddings. Following (Jumper et al., 2021; Lu et al., 2022), the edge updates are arranged in the form of triangles comprising two ligand-protein edges and one intra-edge (intra-ligand or intra-protein edge) where the distance constraints apply.

In the protein-constrained attentive pair update module, each pair (i, j) attends to all "neighboring" pairs $\{(i, k')\}_{1\leq k'\leq n_{\mathrm{p}}}$ with a common ligand node i. We compute a geometry-informed attention weight $a_{ijk'}^{(h)}$ for each neighbor (i, k'), by adding an attention bias $t_{jk'}^{(h)} = \mathrm{Linear}^{(h)}\left(\boldsymbol{d}_{jk'}\right)$ computed from the intra-protein distance embeddings per attention head h. We then perform a standard multi-

[1]During docking, the bond lengths, bond angles and conformation of small rings are mostly unchanged, while torsion angles of rotatable bonds might change drastically (Trott & Olson, 2010; Méndez-Lucio et al., 2021; Stärk et al., 2022). In the flexible docking setting, we provide distance d_{ik}^* of atoms i and k in the *unbound* ligand structure predicted by ETKDG (Riniker & Landrum, 2015) to constrain the model if i, k are $\leq$ 2-hop neighbors or members of the same ring. Distance of all other ligand atom pairs are not provided.

head attention with H heads to aggregate information from neighboring pairs.

$$a_{ijk'}^{(h)} = \text{softmax}_{k'}\left(\frac{1}{\sqrt{c}}\boldsymbol{q}_{ij}^{(h)\top}\boldsymbol{k}_{ik'}^{(h)} + b_{ij}^{(h)} + t_{jk'}^{(h)}\right) \tag{1}$$

$$\boldsymbol{z}_{ij} = \boldsymbol{z}_{ij} + \text{Linear}\left(\text{concat}_{1\le h\le H}\left(\boldsymbol{g}_{ij}^{(h)} \odot \sum_{k'=1}^{n_\text{p}} a_{ijk'}^{(h)}\boldsymbol{v}_{ik'}^{(h)}\right)\right) \tag{2}$$

where $\boldsymbol{g}_{ij}^{(h)} = \sigma(\text{Linear}^{(h)}(\boldsymbol{z}_{ij}))$ is a per-head output gate, $\boldsymbol{q}_{ij}^{(h)}, \boldsymbol{k}_{ij}^{(h)}, \boldsymbol{v}_{ij}^{(h)}, b_{ij}^{(h)}$ are linear projections of the pair embedding $\boldsymbol{z}_{ij}$, and $t_{jk'}^{(h)}$ is a distance-based bias described above.

The ligand-constrained attentive pair update module is designed similarly. Here, for pair (i, j), the "neighboring" pairs $\{(k, j)\}_{1\le k\le n_\text{l}}$ are those sharing the same protein node j. Multi-head attention is performed across all such neighbors with constraints from intra-ligand distances $\boldsymbol{d}_{ik}$.

3.3 Iterative Coordinate Update with Context-Aware E(3)-Equivariant Layer

The coordinate update module iteratively adjust the current ligand structure $\{\boldsymbol{x}_i^\text{l}\}_{1\le i\le n_\text{l}}$ towards the docked pose based on the extracted representations. The module is designed to satisfy the following desiderata: (1) **protein context awareness**: the model must have an in-depth understanding of the interaction between the ligand and its protein context in order to produce a protein-ligand complex with optimized stability (*i.e.* lowest energy); (2) **self (ligand) context awareness**: the model should honor the basic geometric constraints of the ligand, so that the predicted ligand conformation is physically valid; (3) **E(3)-equivariance**: if the protein and the input ligand pose are roto-translated, the ligand should dock into the same pocket with the same pose, and its coordinates should be transformed with the same E(3) transformation. Compared with previous ont-shot method, iterative refinement allows the ligand to better sense the local environment as our ligand gradually move towards its right position.

We start from an unbound ligand structure[2]. At iteration step $t = 0, \ldots, T-1$, the module updates the protein, ligand and pair representations (summarized by $\boldsymbol{h}^{(t)}$) and the ligand coordinates $\{\boldsymbol{x}_i^{(t)}\}_{1\le i\le n_\text{l}}$ with a context-aware E(3)-equivariant layer.

$$\left(\boldsymbol{h}^{(t+1)}, \{\Delta\boldsymbol{x}_i^{(t+1)}\}_{1\le i\le n_\text{l}}\right) = \text{DecoderLayer}^{(t)}\left(\boldsymbol{h}^{(t)}, \{\boldsymbol{x}_i^{(t)}\}_{1\le i\le n_\text{l}}, \{\boldsymbol{x}_j\}_{1\le j\le n_\text{p}}\right) \tag{3}$$

$$\boldsymbol{x}_i^{(t+1)} = \boldsymbol{x}_i^{(t)} + \Delta\boldsymbol{x}_i^{(t+1)}, \qquad 0 \le t < T. \tag{4}$$

where $\{\boldsymbol{x}_j\}_{1\le j\le n_\text{p}}$ are the fixed protein coordinates. We now introduce $\text{DecoderLayer}^{(t)}$ in detail.

Context-Aware Message Passing. To ensure the coordinate update module captures both the protein and ligand context, we construct a heterogeneous context graph containing both protein and ligand nodes. In the graph, *inter-edges* connect protein and ligand nodes, while *intra-edges* connect two ligand nodes (Figure 1). We perform inter- and intra-edge message passing on the graph to explore the current context:

$$\boldsymbol{h}_i^{(t+1)} = \boldsymbol{h}_i^{(t)} + \sum_{j=1}^{n_\text{p}} \boldsymbol{m}_{ij}^{(t)} + \sum_{k=1}^{n_\text{l}} \boldsymbol{m}_{ik}^{(t)} \tag{5}$$

$$\boldsymbol{h}_j^{(t+1)} = \boldsymbol{h}_j^{(t)} + \sum_{i=1}^{n_\text{l}} \boldsymbol{m}_{ji}^{(t)} \tag{6}$$

Note that the three types of messages are generated using different sets of parameters.

Equivariant Coordinate Update. We use *equivariant graph convolution layer* (EGCL) to process current context geometry and update ligand coordinates in an E(3)-equivariant manner. Specifically, we compute messages from the node (and edge) representations and distance information using

[2]In the rigid docking setting, this structure has the same conformation as the docked ligand structure.

MLPs ϕ^m and φ^m.

$$\left(\boldsymbol{m}_{ij}^{(t)}, \boldsymbol{m}_{ji}^{(t)}\right) = \phi^m\left(\boldsymbol{z}_{ij}, \boldsymbol{h}_i^{(t)}, \boldsymbol{h}_j^{(t)}, \left\|\boldsymbol{x}_i^{(t)} - \boldsymbol{x}_j^{(t)}\right\|\right) \tag{7}$$

$$\boldsymbol{m}_{ik}^{(t)} = \varphi^m\left(\boldsymbol{h}_i^{(t)}, \boldsymbol{h}_k^{(t)}, \left\|\boldsymbol{x}_i^{(t)} - \boldsymbol{x}_k^{(t)}\right\|\right) \tag{8}$$

We generate equivariant coordinate updates using the following equation, where ϕ^x and φ^x are gated MLPs that evaluates the message importance. Ligand and protein node embeddings are updated using equations 5 and 6, respectively.

$$\Delta \boldsymbol{x}_i^{(t)} = \sum_{j=1}^{n_\text{p}} \frac{\boldsymbol{x}_j^{(t)} - \boldsymbol{x}_i^{(t)}}{\|\boldsymbol{x}_j^{(t)} - \boldsymbol{x}_i^{(t)}\|} \phi^x(\boldsymbol{m}_{ij}^{(t)}) + \sum_{k=1}^{n_\text{l}} \frac{\boldsymbol{x}_k^{(t)} - \boldsymbol{x}_i^{(t)}}{\|\boldsymbol{x}_k^{(t)} - \boldsymbol{x}_i^{(t)}\|} \varphi^x(\boldsymbol{m}_{ik}^{(t)}) \tag{9}$$

Note that the framework for our decoder is also compatible with other equivariant layers on graphs, such as the geometric vector perceptron (GVP) (Jing et al., 2020) or vector neuron (Deng et al., 2021). We choose EGCL as it is a powerful layer for molecular modeling and conformation generation (Satorras et al., 2021; Huang et al., 2022; Hoogeboom et al., 2022).

Self-Confidence Prediction. At the end of our decoder, an additional self-confidence module is added to predict the model's confidence for its predicted docked pose (Jumper et al., 2021). The confidence is a value from zero to one calculated by the equation $\hat{c} = \sigma(\text{MLP}\left(\sum_{i=1}^{n_\text{l}} \boldsymbol{h}_i^{(T)}\right))$.

3.4 Training and Inference

End-to-End Training. E3Bind directly generates the predicted ligand coordinates along with a confidence score. We define a simple coordinate loss and train the model end-to-end:

$$\mathcal{L}_\text{coord} = \sum_{i=1}^{n_\text{l}} \left(\boldsymbol{x}_i - \boldsymbol{x}_i^*\right)^2.$$

We train the self-confidence module with $\mathcal{L}_\text{confidence}$, a mean squared error (MSE) loss between the predicted self-confidence and its target value c^*. The target value is set based on the root-mean-square deviation (RMSD) of the predicted coordinates. Details are deferred to Section C.1. Intuitively, we want our model to predict a low c for high-RMSD predictions. The final training loss is the combination of the coordinate loss and the confidence loss $\mathcal{L} = \mathcal{L}_\text{coord} + \beta\mathcal{L}_\text{confidence}$, where β is a hyperparameter. Our loss aligns with the goal of docked pose prediction, thus avoids the time-consuming and potentially error-prone distance-to-coordinate transformation.

Inference Process. In practice, the target protein may contain multiple binding sites, or have an extremely large size where most parts are irrelevant for ligand binding. To solve this problem, we use P2Rank (Kriv\'ak & Hoksza, 2018) to segment protein to less than 10 functional block (defined as a 20Å graph around the block center) following TankBind (Lu et al., 2022). We then initialize the unbound ligand structure with random rotation and translation in each block, dock the ligand, and select the predicted docked pose with the highest self-confidence. Different from TankBind which selects the final pose through binding affinity estimation for all functional block's prediction, our pose selection is based on the self-confidence. As a result, our model does not require binding affinity data for training.

4 Experiments

4.1 Flexible Self Docking

As E3Bind is designed to model the flexibility of ligand conformation, it is natural to evaluate it in the blind flexible self-docking setting. We defer the blind re-docking results to the appendix (see Section A).

Data. We use the PDBbind v2020 dataset (Liu et al., 2017) for training and evaluation. We follow the time dataset split from (Stärk et al., 2022), where 363 complex structures uploaded later

than 2019 serve as test examples. After removing structures sharing ligands with the test set, the remaining 16739 structures are used for training and 968 structures are used for validation.

Baselines. We compared our model with recent deep leaning (DL) models and a series of traditional score-based docking methods. For recent deep learning models, TankBind (Lu et al., 2022) and EquiBind (Stärk et al., 2022) are included. For score-based method, QVina-W (Hassan et al., 2017), GNINA (McNutt et al., 2021), SMINA (Koes et al., 2013) and GLIDE (Halgren et al., 2004) are included.

We additionally distinguish between the uncorrected and corrected versions of the recent deep learning models following (Stärk et al., 2022). The corrected versions adopt post optimization methods (e.g. gradient descent or von Mises distributions optimization) to further enforce the geometry constraints when generating the predicted structure. The uncorrected versions, which do not use post-optimization, reflect the model's own capability for pose prediction. These uncorrected versions are suffixed with -U, where EquiBind-U and E3Bind-U outputs are directly generated by the models and TankBind-U is the variant of TankBind which optimize the finally coordinates without the ligand configuration loss. For fair comparison, in E3Bind results we refine the E3Bind-predicted coordinates by post-optimization method of TankBind.

Metric. We evaluate the quality of generated ligand pose by the metrics following (Stärk et al., 2022): (1) **Ligand RMSD** measures how well the model is able to find the docked pose including conformation and its rigid transformation. It is calculated as the root-mean-square deviation of ligand's Cartesian coordinates. (2) **Centroid Distance** reflects the model's capacity to find the binding site. It is defined as the distance between the average coordinate of predicted and ground-truth ligand structure. All metrics are calculated after hydrogen atoms are removed following previous work.

	Ligand RMSD						Centroid Distance					
	Percentiles ↓				% Below ↑		Percentiles ↓				% Below ↑	
Methods	25%	50%	75%	Mean	2Å	5Å	25%	50%	75%	Mean	2Å	5Å
QVina-W	2.5	7.7	23.7	13.6	20.9	40.2	0.9	3.7	22.9	11.9	41.0	54.6
GNINA	2.8	8.7	22.1	13.3	21.2	37.1	1.0	4.5	21.2	11.5	36.0	52.0
SMINA	3.8	8.1	17.9	12.1	13.5	33.9	1.3	3.7	16.2	9.8	38.0	55.9
GLIDE	2.6	9.3	28.1	16.2	21.8	33.6	0.8	5.6	26.9	14.4	36.1	48.7
Vina	5.7	10.7	21.4	14.7	5.5	21.2	1.9	6.2	20.1	12.1	26.5	47.1
EquiBind	3.8	6.2	10.3	8.2	5.5	39.1	1.3	2.6	7.4	5.6	40.0	67.5
TankBind	2.6	4.2	**7.6**	7.8	17.6	57.8	**0.8**	1.7	4.3	5.9	55.0	77.8
E3Bind	**2.1**	**3.8**	7.8	**7.2**	**23.4**	**60.0**	**0.8**	**1.5**	**4.0**	**5.1**	**60.0**	**78.8**
EquiBind-U	3.3	5.7	9.7	7.8	7.2	42.4	1.3	2.6	7.4	5.6	40.0	67.5
TankBind-U	3.9	7.7	13.6	10.5	8.0	34.7	1.3	3.0	8.2	6.6	40.5	66.4
E3Bind-U	**2.0**	**3.8**	**7.7**	**7.2**	**25.6**	**60.6**	**0.8**	**1.5**	**4.0**	**5.1**	**59.0**	**78.8**

Table 1: Blind self-docking performance

Performance in Flexible Self Docking. We evaluate the ligand RMSD and centroid distance for the generated pose as shown in Table 1. The quantitative results show that our model achieves state-of-the-art in most metrics. Specifically, E3Bind shows exceptional power in finding ligand poses with high resolution, where it exceeds state-of-the-art DL based model TankBind by 33% in finding qualified pose (below Threshold 2 Å). For the 25-th percentile ligand RMSD, E3Bind achieves 2.1 Å, ourperforming all previous methods by a large margin. These results verify that our model is able to better capture the protein-ligand interactions in the local context.

Notably, compared with uncorrected deep learning models, E3Bind-U enjoys more significant performance improvement, showcasing its low dependency for additional post optimization. Besides, E3Bind-U outperforms traditional docking softwares by orders of magnitude in inference speed. This speed-up demonstrates the model's potential for high-throughput virtual screening.

Performance in Flexible Self Docking for Unseen Protein. We further evaluate our model's capacity for unseen proteins, which form a subset of the time split based test set containing 144 complexs. As depicted in Table 2, we find that E3Bind show better generalization ability than other DL based model. Commercial software GLIDE show superior performance than all DL based model with respect to finding high quality docking pose (below Threshold 2Å). It's probably due to the data

	LIGAND RMSD						CENTROID DISTANCE					
	Percentiles ↓				% Below ↑		Percentiles ↓				% Below ↑	
Methods	25%	50%	75%	Mean	2Å	5Å	25%	50%	75%	Mean	2Å	5Å
QVina-W	3.4	10.3	28.1	16.9	15.3	31.9	1.3	6.5	26.8	15.2	35.4	47.9
GNINA	4.5	13.4	27.8	16.7	13.9	27.8	2.0	10.1	27.0	15.1	25.7	39.5
SMINA	4.8	10.9	26.0	15.7	9.0	25.7	1.6	6.5	25.7	13.6	29.9	41.7
GLIDE	3.4	18.0	31.4	19.6	**19.6**	28.7	**1.1**	17.6	29.1	18.1	29.4	40.6
Vina	7.9	16.6	27.1	18.7	1.4	12.0	2.4	15.7	26.2	16.1	20.4	37.3
EquiBind	5.9	9.1	14.3	11.3	0.7	18.8	2.6	6.3	12.9	8.9	16.7	43.8
TankBind	3.4	**5.7**	10.8	10.5	3.5	**43.7**	1.2	2.6	8.4	8.2	40.9	**70.8**
E3Bind	**3.0**	6.1	**10.2**	**10.1**	6.3	38.9	1.2	**2.3**	**7.0**	**7.6**	**43.8**	66.0
EquiBind-U	5.7	8.8	14.1	11.0	1.4	21.5	2.6	6.3	12.9	8.9	16.7	43.8
TankBind-U	4.0	7.9	14.9	8.3	3.5	34.0	1.4	3.3	10.9	8.3	35.4	65.2
E3Bind-U	**3.1**	**6.0**	**10.6**	**10.1**	**5.6**	**41.0**	**1.2**	**2.3**	**7.8**	**7.7**	**42.4**	**65.3**

Table 2: Blind self-docking performance on unseen receptors.

sparsity problem of DL based models. However, we observe that our model produces much fewer predictions that are far away from the ground-truth pose in general.

Benefit of Iterative Refinement. In Figure 2, we plot the RMSD and self-confidence score versus the number of coordinate refinement iterations for the successfully docked test examples. As the ligand go through more rounds of coordinate updates, its RMSD tends to decrease and the model self-confidence score tends to increase. This shows the benefit of refining ligand pose for multiple iterations over directly predicting its final pose. Figure 7 visualizes the iterative refinement process. After 4 iterations, E3Bind have already found the correct binding site. Subsequent iterations further refine the ligand conformation to maximize interactions with the protein. The final pose is close to the ground truth with an RMSD of 1.42 Å. In contrast, Equibind, which refines the ligand conformation multiple times but docks the ligand into the pocket with a one-shot global roto-translation, generates a pose with an RMSD of 4.18 Å.

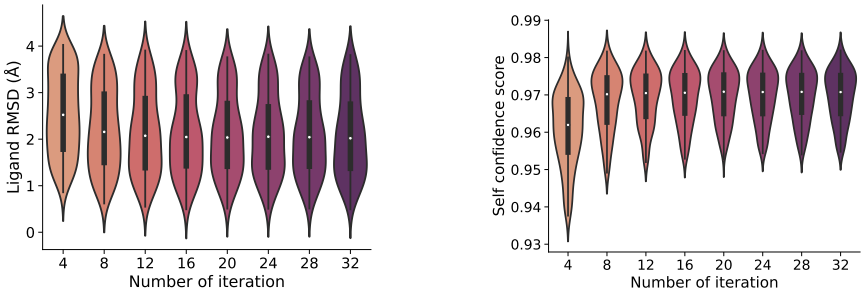


Figure 2: Interpreting effect of iterative refinement. Here we show the trend of top50% ligand RMSD (left) and self confidence score (right).

4.2 ABLATION STUDY

A series of ablation study is done to investigate different factors influencing the performance as depicted in E.First, we investigate the performance with fewer iteration. The results with 4 iteration decrease significantly especially in the fraction below threshold 2Å, validating the benefit of the iterative refinement in the docking problem. Second, we evaluate the model's performance without the Trioformer block. Simple fusion operation without Trioformer's edge-level message passing hurt the model's performance. Third, we test the performance without intra message passing, it slightly reduce the model's expressiveness. Lastly, we investigate the performance without P2Rank. Protein is segmented into 30 random blocks for docking without P2Rank. The result show that our model is not very sensitive to P2Rank's segmentation.

4.3 CASE STUDY

E3Bind correctly identifies the binding site in an unseen large protein. Figure 4a shows a representative case, which contains a large protein target unseen in training. This is a challenging example in blind docking, where the models have zero knowledge of the binding site. While QVina-

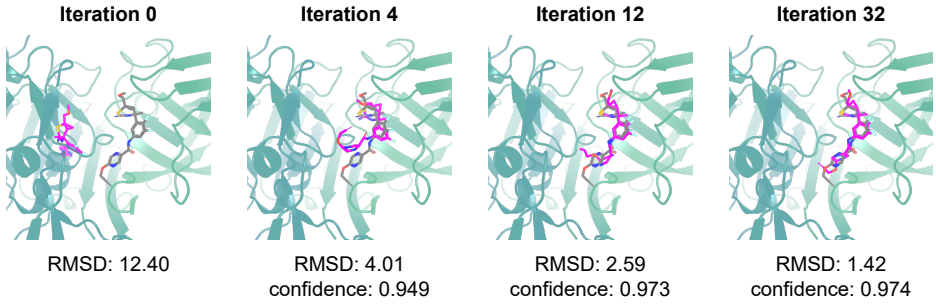


Figure 3: E3Bind coordinate refinement trajectory for ligand in PDB 6PZ4. In each figure, the ground-truth docked ligand pose is shown in gray and the predicted structures in magenta. RMSD and model confidence are written below the figures.

W, EquiBind and TankBind dock to three distinct binding sites away from the ground truth, E3Bind (shown in magenta) correctly identifies the native binding site and generates a pose with a low RMSD of 4.2 Å. Further examination on the self-confidence score shows this pose is the only one with confidence above 0.91, while poses at other binding sites all have confidence below 0.86 (Figure 5).

E3Bind strikes a balance between modeling protein-ligand interaction and structure validity. Figure 4b presents the docking result of a small protein. All methods except TankBind identify the correct binding site, with E3Bind prediction aligning better to the ground truth protein-ligand interaction pattern. Interestingly, TankBind produces a knotted structure which is invalid in nature. This shows that the two-stage approach might generate invalid protein-ligand distance maps that can not be transformed into plausible structures in the distance-to-coordinate optimization stage.

(a) PDB 6N94 (b) PDB 6QLR

Figure 4: Case studies. Predicted pose from E3Bind (magenta), TankBind (orange), EquiBind (red) and QVina-W (blue) are placed together with the target protein. RMSD from ground truth ligand pose (grey) are shown on the figure. (a) E3Bind correctly identifies the binding site from the large protein, while the other methods are off-site. (b) Among three models that identify the correct binding site, E3Bind predicts the binding pose most accurately. TankBind generates an invalid structure with rings knotted together.

5 Conclusion

Fast and accurate docking methods are vital tools for small molecule drug research and discovery. This work propose E3Bind, an end-to-end equivariant network for protein-ligand docking. E3Bind predicts the docked ligand pose through a feature extraction – coordinate refinement pipeline. Geometry-consistent protein, ligand and pair representations are first extracted by Trioformer. Then the ligand coordinates are iteratively updated by a context-aware E(3)-equivariant network. Empirical experiments show that E3Bind is competitive against state-of-the-art blind docking methods, especially so when ligand pose post-optimization is not applied. Interesting future directions include: modeling the protein backbone/side chain dynamics to better capture the drug-target interaction, and exploring better ways of feature extraction, geometry constraint incorporation and coordinate refinement.

REFERENCES

Jingxiao Bao, Xiao He, and John ZH Zhang. Deepbsp—a machine learning method for accurate prediction of protein–ligand docking structures. *Journal of Chemical Information and Modeling*, 61(5):2231–2240, 2021.

Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1): 235–242, 2000.

Ryan G Coleman, Michael Carchia, Teague Sterling, John J Irwin, and Brian K Shoichet. Ligand pose and orientational sampling in molecular docking. *PloS one*, 8(10):e75992, 2013.

UniProt Consortium. Uniprot: a worldwide hub of protein knowledge. *Nucleic acids research*, 47 (D1):D506–D515, 2019.

Congyue Deng, Or Litany, Yueqi Duan, Adrien Poulenard, Andrea Tagliasacchi, and Leonidas J Guibas. Vector neurons: A general framework for so (3)-equivariant networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12200–12209, 2021.

Jacob D Durrant and J Andrew McCammon. Nnscore 2.0: a neural-network receptor–ligand scoring function. *Journal of chemical information and modeling*, 51(11):2897–2903, 2011.

Leonardo G Ferreira, Ricardo N Dos Santos, Glaucius Oliva, and Adriano D Andricopulo. Molecular docking and structure-based drug design strategies. *Molecules*, 20(7):13384–13421, 2015.

Paul G Francoeur, Tomohide Masuda, Jocelyn Sunseri, Andrew Jia, Richard B Iovanisci, Ian Snyder, and David R Koes. Three-dimensional convolutional neural networks and a cross-docked data set for structure-based drug design. *Journal of chemical information and modeling*, 60(9):4200–4215, 2020.

Octavian Ganea, Lagnajit Pattanaik, Connor Coley, Regina Barzilay, Klavs Jensen, William Green, and Tommi Jaakkola. Geomol: Torsional geometric generation of molecular 3d conformer ensembles. *Advances in Neural Information Processing Systems*, 34:13757–13769, 2021a.

Octavian-Eugen Ganea, Xinyuan Huang, Charlotte Bunne, Yatao Bian, Regina Barzilay, Tommi Jaakkola, and Andreas Krause. Independent se (3)-equivariant models for end-to-end rigid protein docking. *arXiv preprint arXiv:2111.07786*, 2021b.

Pascale Gaudet, Pierre-André Michel, Monique Zahn-Zabal, Aurore Britan, Isabelle Cusin, Marcin Domagalski, Paula D Duek, Alain Gateau, Anne Gleizes, Valérie Hinard, et al. The nextprot knowledgebase on human proteins: 2017 update. *Nucleic acids research*, 45(D1):D177–D182, 2017.

Thomas A Halgren, Robert B Murphy, Richard A Friesner, Hege S Beard, Leah L Frye, W Thomas Pollard, and Jay L Banks. Glide: a new approach for rapid, accurate docking and scoring. 2. enrichment factors in database screening. *Journal of medicinal chemistry*, 47(7):1750–1759, 2004.

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

Jérôme Hert, John J Irwin, Christian Laggner, Michael J Keiser, and Brian K Shoichet. Quantifying biogenic bias in screening libraries. *Nature chemical biology*, 5(7):479–483, 2009.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Emiel Hoogeboom, Vıćtor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3d. In *International Conference on Machine Learning*, pp. 8867–8887. PMLR, 2022.

Wenbing Huang, Jiaqi Han, Yu Rong, Tingyang Xu, Fuchun Sun, and Junzhou Huang. Equivariant graph mechanics networks with constraints. *arXiv preprint arXiv:2203.06442*, 2022.

Bowen Jing, Stephan Eismann, Patricia Suriana, Raphael JL Townshend, and Ron Dror. Learning from protein structure with geometric vector perceptrons. *arXiv preprint arXiv:2009.01411*, 2020.

Bowen Jing, Gabriele Corso, Jeffrey Chang, Regina Barzilay, and Tommi Jaakkola. Torsional diffusion for molecular conformer generation. *arXiv preprint arXiv:2206.01729*, 2022.

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

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

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

Greg Landrum et al. Rdkit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum*, 2013.

Leo Liberti, Carlile Lavor, Nelson Maculan, and Antonio Mucherino. Euclidean distance geometry and applications. *SIAM review*, 56(1):3–69, 2014.

Yu Liu, Lei Zhao, Wentao Li, Dongyu Zhao, Miao Song, and Yongliang Yang. Fipsdock: a new molecular docking technique driven by fully informed swarm optimization algorithm. *Journal of computational chemistry*, 34(1):67–75, 2013.

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

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

Jiankun Lyu, Sheng Wang, Trent E Balius, Isha Singh, Anat Levit, Yurii S Moroz, Matthew J O'Meara, Tao Che, Enkhjargal Algaa, Kateryna Tolmachova, et al. Ultra-large library docking for discovering new chemotypes. *Nature*, 566(7743):224–229, 2019.

Matthew Masters, Amr H Mahmoud, Yao Wei, and Markus Alexander Lill. Deep learning model for flexible and efficient protein-ligand docking. In *ICLR2022 Machine Learning for Drug Discovery*, 2022.

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.

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

Garrett M Morris, David S Goodsell, Ruth Huey, and Arthur J Olson. Distributed automated docking of flexible ligands to proteins: parallel applications of autodock 2.4. *Journal of computer-aided molecular design*, 10(4):293–304, 1996.

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.

Jean-Louis Reymond and Mahendra Awale. Exploring chemical space for drug discovery using the chemical universe database. *ACS chemical neuroscience*, 3(9):649–657, 2012.

Sereina Riniker and Gregory A Landrum. Better informed distance geometry: using what we know to improve conformation generation. *Journal of chemical information and modeling*, 55(12): 2562–2574, 2015.

Vıctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E (n) equivariant graph neural networks. In *International conference on machine learning*, pp. 9323–9332. PMLR, 2021.

Andrew W Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander WR Nelson, Alex Bridgland, et al. Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710, 2020a.

Andrew W Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander WR Nelson, Alex Bridgland, et al. Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710, 2020b.

Chence Shi, Shitong Luo, Minkai Xu, and Jian Tang. Learning gradient fields for molecular conformation generation. In *International Conference on Machine Learning*, pp. 9558–9568. PMLR, 2021.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019.

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

Oleg Trott and Arthur J Olson. Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*, 31(2):455–461, 2010.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011.

Sheng Wang, Siqi Sun, Zhen Li, Renyu Zhang, and Jinbo Xu. Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS computational biology*, 13(1):e1005324, 2017.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. Geodiff: A geometric diffusion model for molecular conformation generation. *arXiv preprint arXiv:2203.02923*, 2022.

Wenyi Zhang, Eric W Bell, Minghao Yin, and Yang Zhang. Edock: blind protein–ligand docking by replica-exchange monte carlo simulation. *Journal of cheminformatics*, 12(1):1–17, 2020.

Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-mol: A universal 3d molecular representation learning framework. *ChemRxiv*, 2022. doi: 10.26434/chemrxiv-2022-jjm0j-v3.

Zhaocheng Zhu, Chence Shi, Zuobai Zhang, Shengchao Liu, Minghao Xu, Xinyu Yuan, Yangtian Zhang, Junkun Chen, Huiyu Cai, Jiarui Lu, et al. Torchdrug: A powerful and flexible machine learning platform for drug discovery. *arXiv preprint arXiv:2202.08320*, 2022.