Pose Ensemble Graph Neural Networks to Improve Docking Performances

Thanawat Thaingtamtanha[1,⊥], Jordane Preto[2,⊥], Francesco Gentile[1,3*]

[1]Department of Chemistry and Biomolecular Sciences, University of Ottawa, Ottawa, ON, Canada

[2]Aix-Marseille University, Université de Toulon, CNRS, Centre de Physique Théorique UMR 7332, 13288 Marseille Cedex 09, France

[3]Ottawa Institute of Systems Biology, Ottawa, ON, Canada

⊥equal contribution

**email: fgentile@uottawa.ca*

Abstract. The prediction of the geometry and strength governing small molecule-protein interactions remains a paramount challenge in drug discovery due to their complex and dynamic nature. A number of machine learning (ML) methods have been proposed to complement and improve on physics-based tools such as molecular docking, usually by mapping three dimensional features of individual poses to their closeness to experimental structures and/or to binding affinities. Here, we introduce Dockbox2 (DBX2), a novel approach that encodes ensembles of computational poses within a graph neural network architecture via simple energy-based features derived from molecular docking. The model was jointly trained to predict binding pose likelihood as a node-level task and binding affinity as a graph-level task using the PDBbind dataset and demonstrated significant performance in comprehensive, retrospective docking and virtual screening experiments. Our results encourage further exploration of ML models based on conformational ensembles to provide more accurate estimates of small molecule-protein interactions and thermodynamics. The DBX2 code is available at https://github.com/jp43/DockBox2.

Introduction

Drugs exert their therapeutic effects by binding to specific biomolecular targets, typically proteins, and modulating their function, thereby inhibiting or restoring processes relevant for the treatment of various diseases. The initial step in the drug discovery pipeline involves identifying molecules binding to a target of interest with high affinity and specificity [1], making accurate prediction of both crucial for drug development [2]. Binding affinity, which reflects the strength of the interaction between a drug and its protein target, is commonly expressed in terms of dissociation constant (Kd), measurable via a plethora of experimental techniques [3]. However, these techniques are usually time-consuming and resource intensive [4], [5], especially at high throughput rates required to explore vast chemical spaces [6]. Consequently, *in-silico* screening methods have gained significant momentum, especially in the recent years [7].

Although accurate computational estimation of ligand-protein affinity and interactions is crucial, significant challenges arise due to the dynamic nature of these complexes. Molecular dynamics (MD) simulations provide valuable insights into the nature of these interactions, *e.g.,* by considering an ensemble of bound conformations to generate thermodynamically accurate estimates of various energy contributions [8]. This is usually done by calculating the statistical properties of systems in thermodynamic equilibrium and estimating the time spent in the various microstates. Therefore, MD has the potential to connect the chemical world to physical observables, aiding in the determination of state variables (free energy, enthalpy, entropy, …), kinetics, and the exploration of biomolecular mechanisms driven by rare events [9]. Numerous studies have illuminated the remarkable performance of MD simulations in predicting experimental outcomes, showcasing their transformative potential to accelerate and economize the drug discovery process. For instance, the ligand gaussian accelerated MD (LGMD) method, an enhanced sampling technique pioneered by Miao et al. [10], was employed to forecast the binding affinity of nirmatrelvir with the coronavirus 3C-like protease, yielding predictions in striking concordance with experimental observations [11], [12]. Likewise, Wolf et al. [13] harnessed the power of Langevin simulations an extended MD approach that delves into the intricate low-frequency motions governing large conformational shifts [14], to estimate the binding affinity of the benzamidine-trypsin complex, achieving results that closely mirrored experimental findings. However, standard and biased MD methods require significant computational power that render these techniques unsuited for high-throughput screening purposes. Consequently, faster and less accurate methods such as molecular docking and machine learning (ML) approaches have been proposed as alternatives.

Molecular docking methods generate bound conformations of a ligand within a rigid binding pocket and then rank the poses using a scoring function to estimate the binding affinity [15]. Despite its simplicity, docking has shown great potential to identify active molecules from vast backgrounds of inactive compounds [17], [18], with its impact extending across numerous therapeutic areas. A notable example is the work of Manglik et al., in which docking was used to screen over 3 million molecules against the μ-opioid receptor (μOR), leading to the discovery of PZM21, a G protein-biased μOR agonist [19]. This compound not only demonstrated remarkable analgesic efficacy but also lacked the severe side effects associated with traditional opioids, marking a significant milestone in pain management. Beyond its therapeutic promise, PZM21 exemplifies a new class of μOR agonists with enhanced specificity [20]. Zernov et al., for instance, discovered an anti-Alzheimer's compound targeting the transient receptor potential cation channel 6, with in-vitro studies confirming its efficacy, stability, and target specificity without adverse effects [21]. Docking has also been key in identifying treatments for infectious diseases. Agnihotri et al. identified potent inhibitors of γ-glutamylcysteine synthase for treating leishmaniasis, with four out of five candidates showing strong specificity and low toxicity in human cells [22]. Amid the global urgency of the COVID-19 pandemic, Wang et al. screened 2,467 compounds against the SARS-CoV-2 spike protein, yielding promising antiviral leads through docking [23]. Stein et al., for instance, employed docking to screen over 150 million molecules targeting melatonin receptor 1 (MT1) in the search for therapeutics addressing sleep disorders and depression. Despite numerous *in-vivo* studies aiming to identify selective MT1 ligands, few have demonstrated significant selectivity [24], [25]. Interestingly, docking identified a novel chemotype with selective MT1 agonist activity, later validated experimentally, underscoring the robustness of docking in discovering new chemical scaffolds for neurological disorders [26]. Additionally, Fink et al. identified promising α2A-adrenergic receptor (α2AAR) agonists with fewer adverse effects compared to earlier treatments. Screening over 300 million compounds via docking, their findings were corroborated through experimental validation, confirming both the efficacy and favorable pharmacokinetics of these compounds [27]. These studies underscore the vital role of docking in advancing drug discovery.

While molecular docking continues to be a transformative tool in drug discovery, several limitations remain due to the approximative nature of scoring functions and the neglection of flexibility, among others [15], [28]. Machine learning (ML) methods, on the other hand, have been introduced in the last decade to tackle molecular docking challenges [15]. For example, Graph Neural Networks (GNNs) have been widely explored to characterize ligand-protein interactions [29]. Several GNN models have been used for ligand-protein affinity prediction, such as CurvAGN [30], PIGNet [31], GenScore [32] and SS-GNN [33], reporting strong correlations between predicted and experimental affinities [29], [34], [35]. Additionally, GNNs have been applied in generative settings to replace physics-based sampling and generate and score potential ligand-protein poses, such as in DiffDock [36] and MedusaGraph [37]. Although these architectures have shown promising results, an increasing number of studies suggest that GNNs tend to memorize ligand and protein patterns instead of learning the true interactions between them [29], [35]. Moreover, single pose graphs are generally mapped to binding affinities, potentially missing the opportunity to capture the full thermodynamic profile and dynamics of ligand-protein interactions that depends on multiple conformations [29].

Recent efforts have been made to consider multiple conformations in training GNNs for binding affinity predictions, such as Dynaformer [38]. However, this method utilizes a data augmentation strategy that still relies on individual graphs for each binding conformation, derived from costly MD simulations, to predict affinities. In this work, we introduce DockBox2 (DBX2), a GNNs framework that enables to encode multiple ligand-protein conformations derived from docking within single graphs to leverage ensemble representations, for predicting simultaneously near-to-native binding poses and binding affinities. In a series of retrospective experiments, DBX2 demonstrated significant improved performances both for docking and virtual screening (VS) tasks compared with physics-based and ML methods, warrantying further investigation of ensemble-based ML models in computer-aided drug discovery.

Material and Methods

Datasets

The DBX2 model was trained and evaluated using the PDBbind database [39]. The refined set of PDBbind version 2016 (4,057 complexes) [40] was used to train the model. The hold-out test set from Volkov et al [35], which consists of 3,393 complexes, were used as test sets. A subset of the LIT-PCBA database [41] was used to perform retrospective VS experiments.

Protein and ligand preparation

Complexes from PDBbind were prepared following the same procedure of our previous work [42]. For retrospective VS, dominant protonation and tautomerization state was computed from the small molecule SMILES using Openeye 's QUACPAC [43]. Resulting SMILES strings were then converted into low-energy 3D conformations (mol2 format) using Openeye 's OMEGA tool [43]. The target proteins were prepared as follows: redundant protein chains, along with non-essential ions, waters, and heteroatoms, were removed. The resulting protein structures were prepared using the Molecular Operating Environment (MOE) QuickPrep tool [44], by automatically adding missing loops in the structure and assigning the proper conformation to the residues with alternate orientation. Subsequently, protonation states were generated and optimized using the Protonate 3D tool from MOE (at pH 7.4). Finally, the structures were energy-minimized using the AMBER10:EHT forcefield implemented in MOE , and saved in pdb format.

Molecular docking and rescoring

The first Dockbox package (DBX) [42] was utilized to generate binding poses with AutoDock [45], AutoDock Vina [46] and DOCK 6 (DOCK) [47], and rescore with their scoring function in addition to Gnina [48] and DSX [49]. The DBX configuration file used for generating binding pose from PDBbind v2016 and the test sets is illustrated in **Figure**

S1; a maximum of 140 binding poses were generated for each system, 60 from AutoDock, 20 from Vina, and 60 from DOCK. For AutoDock, grid spacing was set to 0.3 Å, and the Lamarckian genetic algorithm [50] was employed to generate poses. For Vina, the energy range for final poses was set to 3 kcal/mol. In DOCK, a grid-based scoring method was applied with a spacing of 0.3 Å. Docking with any of the above programs was followed by energy minimization, starting with 500 steps of the steepest descent method followed by 1,000 steps combining steepest descent and conjugate gradient methods. Energy minimization was performed using AmberTools 17 [51] to prevent structural clashes and ensure appropriate rescoring with different programs. Rescoring was then conducted with AutoDock, Vina, DOCK, and DSX scoring functions.

Dockbox2 architecture

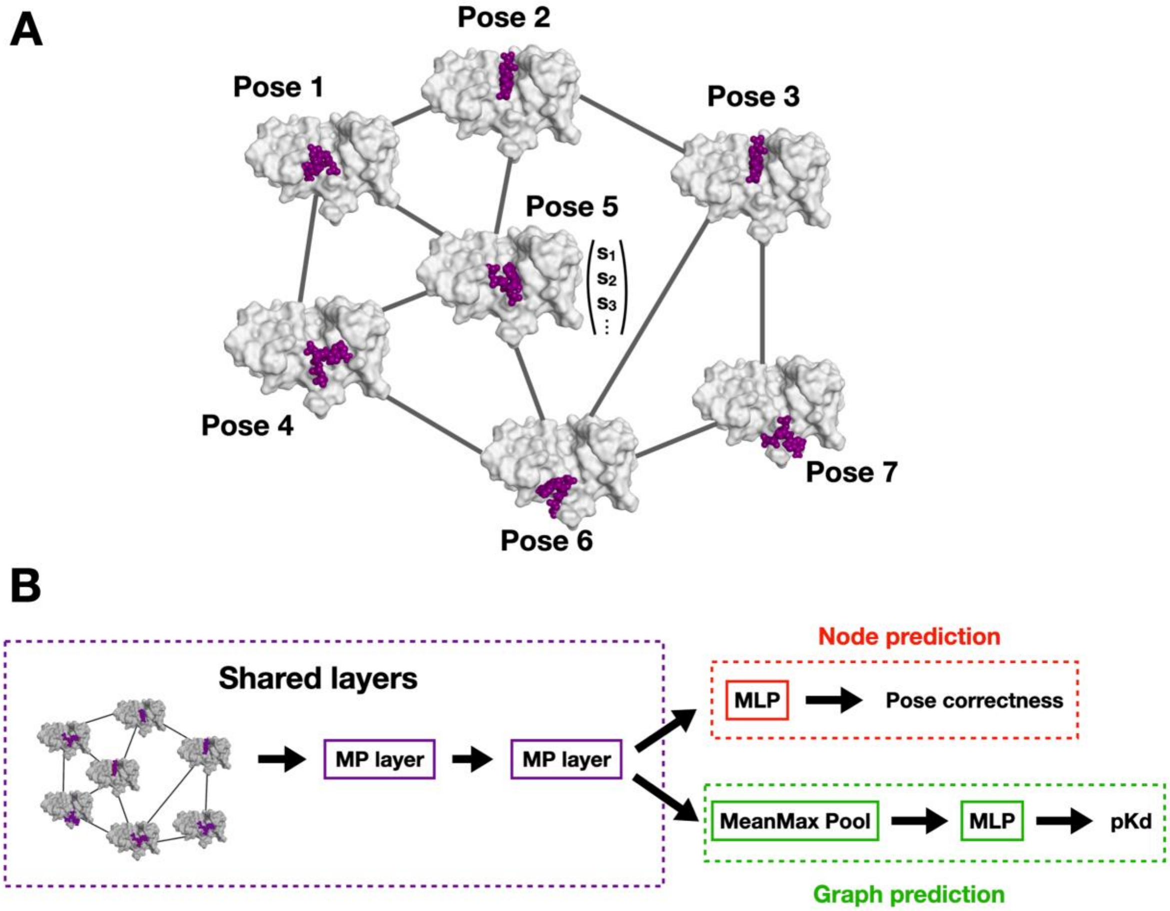

***Figure 1**: Architecture of DBX2. (A) Binding poses are represented as nodes. Two pose nodes are connected by an edge based on the root mean square deviation (RMSD)*

between them. Docking-derived energies and categorical features of each binding pose, here referred as s_1, s_2, s_3…, are used as node features. (B) DBX2 model showing the different layers involved. Pose correctness and pKd are jointly learned as node- and graph-level tasks, respectively.

DBX2 architecture is based on the GraphSAGE model [52] as shown in **Figure 1**. The ensemble of poses generated by docking a given ligand-protein pair is used to construct a graph (**Figure 1A)**, with each node encoding an individual binding pose represented by categorical and energetic features, listed in **Table S1**. Two nodes are connected by an edge if the root mean square deviation (RMSD) between the two poses is below a predefined threshold (usually 5Å or more) while the RMSD value is kept as edge feature. Graphs may be generated using the *create_graphs* script available in the DBX2 package. In the shared layers, the DBX2 model uses message passing (MP) [53], *i.e.*, for each node *i*, information from its neighbors $\boldsymbol{j} \in \boldsymbol{\mathcal{N}}(\boldsymbol{i})$ is gathered and aggregated using the symmetric mean (symmean) aggregation (capturing averaged features of node's neighborhood):

$$\boldsymbol{m}_{\mathcal{N}(i)}^{(k-1)} = SYMMEAN\{\boldsymbol{s}_j^{(k-1)} \oplus RMSD_{ij}, \forall j \in \mathcal{N}(i)\}, \quad (1)$$

where $\boldsymbol{m}_{\mathcal{N}(i)}^{(k-1)}$ is the aggregated message for node *i* from its neighbors, $\boldsymbol{s}_j^{(k-1)}$ is the feature vector of neighbor node *j*, $RMSD_{ij}$ is the RMSD between node *i* and *j*. The feature vector is concatenated with the RMSD between nodes *i* and *j*. The aggregation function then combines these concatenated vectors to produce a single aggregation message vector. The node feature vector is then updated:

$$\boldsymbol{s}_{\boldsymbol{i}}^{(\boldsymbol{k})} = \boldsymbol{\sigma}\left(\boldsymbol{W}_{\boldsymbol{self}}^{(\boldsymbol{k})}\boldsymbol{s}_{\boldsymbol{i}}^{(\boldsymbol{k-1})} \oplus \boldsymbol{W}_{\boldsymbol{neigh}}^{(\boldsymbol{k})}\boldsymbol{m}_{\boldsymbol{\mathcal{N}(i)}}^{(\boldsymbol{k-1})}\right), \quad (2)$$

where $s_i^{(k-1)}$ is the feature vector of node *i* at layer *k*. $\boldsymbol{s}_{\boldsymbol{i}}^{(\boldsymbol{k-1})}$ is the feature vector of node *i* from the previous layer *k-1*. $\boldsymbol{W}_{\boldsymbol{self}}^{(\boldsymbol{k})}$ and $\boldsymbol{W}_{\boldsymbol{neigh}}^{(\boldsymbol{k})}$ are learnable weight matrices that apply to the feature vector of the current node and to the aggregated message vector from neighbor nodes, respectively. $\boldsymbol{m}_{\boldsymbol{\mathcal{N}(i)}}^{(\boldsymbol{k-1})}$ is the aggregated message from the neighbors $\boldsymbol{\mathcal{N}}(\boldsymbol{i})$

of node *i*. The MP layers are followed by multilayer perceptron (MLP) layers to predict pose correctness (node-level task) and the pK_d/pK_i (graph-level task) as illustrated in **Fig. 1B**. For node-level predictions, aggregated information from the MP layers is passed to an MLP with Rectified Linear Unit (ReLU) and sigmoid activation function for hidden layers and final layer of MLP, respectively. For graph-level predictions, aggregated information is passed to a readout layer corresponding to a MeanMax pooling and then passed to a two-layers MLP, with ReLu activation function for the hidden layer and linear activation function for the output layer. This allows MLP to leverage energetic information from ensembles of binding poses for ligand-protein affinity predictions.

Model training and evaluation

The total loss function of DBX2 consists of three components $\boldsymbol{Loss_n}$, $\boldsymbol{Loss_g}$, and $\boldsymbol{Loss_{reg}}$ as in eqn (3):

$$\boldsymbol{Total\ loss = Loss_n + w_1\, Loss_g + w_2\, Loss_{reg}} \tag{3}$$

$\boldsymbol{Loss_n}$ is the loss function for node-level task, where the binary focal cross entropy [54] is used as loss function for node-level task:

$$\boldsymbol{Loss_n = -\alpha \cdot (1 - p_t)^{\gamma} \cdot log(p_t)} \tag{4}$$

where α is a weighting factor, γ is the focusing parameter and $\boldsymbol{p_t}$ is an estimate of the probability for the true class, typically given by the number of correct poses over the total number of poses in the training set. Minimizing $\boldsymbol{Loss_n}$ enables the model to correctly predict the likelihood of binding pose. $\boldsymbol{Loss_g}$ and $\boldsymbol{w_1}$ are the loss function and weight for graph-level task, respectively, where $\boldsymbol{Loss_g}$ corresponds to the root mean square error (RMSE) [55]:

$$\boldsymbol{Loss_g = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}} \tag{5}$$

Here $\boldsymbol{N}$ denotes the total number of ligand-protein complexes, $\boldsymbol{y_i}$ is the actual value of binding affinity for each complex and $\hat{y}_i$ is the predicted binding affinity for each ligand-protein complex. Minimizing $\boldsymbol{Loss_g}$ contributes to correctly predicting the ligand-protein affinity. $\boldsymbol{Loss_{reg}}$ and $\boldsymbol{w_2}$ are the regularization loss and weight, respectively, while L2 regularization loss [56] was here used to prevent overfitting of model:

$$\boldsymbol{Loss_{reg}} = \frac{1}{2}\sum_{i=1}^{n} \boldsymbol{t_i^2} \tag{6}$$

where $\boldsymbol{t_i}$ represent the model parameter, $\boldsymbol{n}$ is the number of model parameter. The model was trained by using *traindbx2* routine (example of a configuration file for *traindbx2* in the INI format is provided in **Figure S2**). Training was performed with a maximum of 200 epochs and early stopping was used by monitoring the total loss on the validation sets for 3 consecutive epochs. The model was trained with mini-batch gradient descent (batch size of 100) and the adaptive moment estimation (ADAM) optimizer with a learning rate of 5e-4 and a decay rate of 0.99.

Hyperparameter optimization was performed using a grid search, considering RMSD cutoff value to define an edge (RMSD cut-off), number of adjacent nodes to randomly sample for aggregation (nrof-neigh), and graph loss weight (lossg weight) as hyperparameters, for a total of 30 combinations (**Table S2**). Training and validation sets were prepared by using the *split_train_val_dbx2* routine of the DBX2 package. The graphs generated from PDBbind 2016 complexes were split as follows: the graph was created with the number of nodes per graph of 140. Then, the data was split for stratified 5-fold cross-validation (90% training, 10% validation), with each fold maintaining a consistent distribution of protein families (e.g. T4 lysozyme, β-galactosidase, etc.) across all folds. Node and edge features for each graph were normalized using standard scaler. For node-level predictions, success rate, accuracy, and area under the curve (AUC) were used as evaluation metrics. For graph-level predictions, RMSE, and R-squared (R^2) were used. The predictive power of DBX2 was further assessed by calculating the Pearson correlation coefficient (Rp) between experimental Kd and graph-level predicted Kd.

Model testing

Models were tested on the test sets and compared for docking and scoring tasks with other docking software. Several metrics were employed to evaluate performance. To evaluate docking power, the success rate was used to measure the likelihood that the top-ranked pose as determined by a given scoring function corresponds to the native pose. Specifically, the top-ranked pose was compared to the minimized experimental structure, with a pose deemed successful if its RMSD was less than 2 Å. For DBX2, the success rate was evaluated using top-ranked poses from node-level predictions.

Next, the scoring power was assessed to evaluate the model's ability to predict and reproduce experimental binding constants using linear and multiple linear regression. The correlation between experimental binding affinities and scores of the best-poses from different scoring functions was analyzed through linear regression, and the R^2 values were calculated to assess the quality of the fitting. For DBX2, graph-level predictions were utilized to evaluate the correlation with experimental binding affinities. Additionally, multiple linear regression was conducted to correlate experimental binding affinities with predicted values derived from various linear combinations of scoring functions, as described in a previous study [42].

Scoring power was also evaluated using the Pearson correlation coefficient (RP) and the predictive index, as described in a prior study [42]. Proposed by Pearlman et al. [57], the predictive index measures the reliability of a scoring function in accurately distinguishing the most potent binder between two compounds. It is calculated as follows:

$$\boldsymbol{PI = \sum_{j>i} \sum_{i} w_{ij} C_{ij}} \qquad (7)$$

With

$$\boldsymbol{w_{ij} = \left| E_j - E_i \right|}$$

$$\boldsymbol{C_{ij}} = \begin{cases} \boldsymbol{1} & \boldsymbol{if\ \ \dfrac{E_j - E_i}{S_j - S_i} < 0} \\ \boldsymbol{-1} & \boldsymbol{if\ \ \dfrac{E_j - E_i}{S_j - S_i} > 0} \\ \boldsymbol{0} & \boldsymbol{if\ \ S_j - S_i = 0} \end{cases}$$

Where $\boldsymbol{E_i}$ is the experimental binding affinity of compound i, and $\boldsymbol{S_i}$ is the score of compound i. Predictive index gives values in range from -1 (wrong prediction) to 1 (perfect prediction), with 0 being random prediction. $\boldsymbol{w_{ij}}$ is the weighting term which underscores the accurate ranking of compounds exhibiting substantial disparities in experimental binding affinities.

Retrospective virtual screening

The VS experiment was conducted on three target proteins from the LIT-PCBA database [41] that were not present in the training set: Flap structure-specific endonuclease 1 (FEN1, PDB id: 5FV7) [58], Glucocerebrosidase (GBA, PDB id: 2XWE) [59], and Mammalian Target of Rapamycin Complex 1 (MTORC1, PDB id: 5GPG) [60]. As a first step, Vina was used to screen active-inactive sets from LIT-PCBA against each corresponding structure. The top 20,000 compounds based on the Vina ranking were then docked also with AutoDock. 80 binding poses (60 from AutoDock and 20 from Vina) were generated for each ligand-protein complex (**Figure S3**). Rescoring was performed with AutoDock, Vina, DOCK, and Gnina (Gnina rescoring was done by selecting the best pose by CNNScore, then considering its CNNAffinity) [48]. VS performance was evaluated by calculating logarithmic area under the curve (logAUC) [61], enrichment factors (EF) and Boltzmann-Enhanced Discrimination of ROC (BEDROC) with adjust parameter (α) values of 20 and 80.5 using the CROC Python package [62], [63], [64].

The logAUC quantifies the overall performance of a virtual screening (VS) method by assessing its ability to distinguish active compounds from decoys across the ranked list. By applying a logarithmic scale to false positive rates, it places greater emphasis on the early retrieval of active compounds, which is critical for the efficiency of screening methods.

EF measures how effectively the VS method identifies active compounds within a specific fraction of the ranked list [65]. EF at a given cutoff $(\boldsymbol{x})$ is calculated from the proportion of true active compounds in the selection set in relation to the proportion of true active compounds in the entire dataset:

$$\boldsymbol{EF}(\boldsymbol{x}) = \frac{\boldsymbol{TP/TP+FP}}{\boldsymbol{TP+FN/TP+TN+FP+FN}} = \frac{N\times \boldsymbol{n_s}}{\boldsymbol{n}\times \boldsymbol{N_s}} \tag{8}$$

Where $\boldsymbol{TP}$ and $\boldsymbol{TN}$ are true positive and true negative, $\boldsymbol{FP}$ and $\boldsymbol{FN}$ are false positive and false negative. $\boldsymbol{N}$ is a total number of compounds in the entire dataset, $\boldsymbol{N_s}$ is a total number of predicted active compounds in the selection set $(\boldsymbol{x})$, $\boldsymbol{n}$ is a total number of true active compounds in the entire dataset, $\boldsymbol{n_s}$ is the number of true active compounds in the selection set $(\boldsymbol{x})$. The top 2% of the ranked compounds for each scoring functions and both graph-level and node-level predictions by DBX2 were calculated to assess EF (EF2).

Normalized enrichment factor (NEF) is calculated to rescale the EF values into a range from 0 (bad prediction) to 1 (perfect prediction) [66], with the goal of standardizing comparison across different datasets. NEF is calculated with following:

$$\boldsymbol{NEF(x)} = \frac{\boldsymbol{EF(x)}}{\boldsymbol{EF(x)_{max}}} \tag{9}$$

With

$$\boldsymbol{EF(x)_{max}} = \frac{\boldsymbol{min\{n_s,\ N \times x\}}}{\boldsymbol{n \times x}}$$

Where $\boldsymbol{EF(x)_{max}}$ denotes the maximum enrichment factor achievable within a selection set $(\boldsymbol{x})$. It serves as a quantitative measure of the highest potential efficiency of a virtual screening method in identifying active compounds from a selection set. $\boldsymbol{n_s}$ is the number of true active compounds in the selection set (x). $\boldsymbol{N}$ is a total number of compounds in the entire dataset.

BEDROC is used to emphasized the concentration of active compounds at several range of ranked data sets [63], [66] through a scaling function (α). This metric is defined as:

$$\boldsymbol{BEDROC} = \frac{\boldsymbol{RIE - RIE_{min}}}{\boldsymbol{RIE_{min} - RIE_{max}}} \tag{10}$$

With

$$\boldsymbol{RIE_{min}} = \frac{\boldsymbol{1 - e^{\alpha R_\alpha}}}{\boldsymbol{R_\alpha(1 - e^{\alpha})}}$$

$$\boldsymbol{RIE_{max}} = \frac{\boldsymbol{1 - e^{-\alpha R_\alpha}}}{\boldsymbol{R_\alpha(1 - e^{-\alpha})}}$$

$$\boldsymbol{RIE} = \frac{\boldsymbol{\frac{1}{n}\sum_{i=1}^{n} e^{\alpha x_i}}}{\boldsymbol{\frac{1}{n}\left(\frac{1 - e^{\alpha}}{e^{\alpha}/_{N_{-1}}}\right)}}$$

Where $\boldsymbol{RIE}$ is robust initial enhancement which proposed by Sheridan et al [67], $\boldsymbol{x_i}$ is a relative ranking of active compound *i*. $\boldsymbol{R_\alpha}$ is the fraction of active compound ($\boldsymbol{R_\alpha = n/N}$), $\boldsymbol{\alpha}$ is the scaling function.

Baseline models

The performance of our DBX2 model in predicting ligand-protein binding affinity and retrospective virtual screening was estimated using the following approach:

- AutoDock, Vina, DOCK6, and DBX2 were compared both in terms docking/scoring power and retrospective virtual screening.
- Gnina and DBX2 were compared only for retrospective virtual screening.
- DSX and DBX2 were compared only for docking/scoring power.

To demonstrate the accuracy of DBX2, docking and scoring performances were evaluated using a temporal split hold-out test set from Volkov et al [35]. This dataset was carefully curated to eliminate latent biases, such as patterns in ligands or proteins, which can lead neural networks to depend on memorization rather than genuine protein-ligand interaction learning. As highlighted in previous studies [29], [35], this memorization often arises from significant redundancies between training and test sets, resulting in data leakage.

Results and Discussion

Hyperparameter optimization

The results of hyperparameter optimization for the DBX2 model are summarized in **Table S3**. The best-performing hyperparameters were an RMSD cut-off of 10 Å, nrof-neigh of 30, and a loss graph weight of 0.02, yielding a success rate of around 60%. This underscores the significance of a higher RMSD cut-off and wider neighborhood size in enhancing model accuracy.

Docking and scoring power

To evaluate the effectiveness of predicting the correct binding pose in DBX2 and other docking programs, we compute the success rate on the hold-out test set as described in the Material and Methods section (**Figure 2A**). As expected, rescoring ensembles of docking poses with different scoring functions led to significantly improved performances for all the scoring functions likely due to enhanced pose sampling, as observed in previous studies [42]. Noticeably, the node-level pose classification method implemented in DBX2 significantly outperformed all docking and rescoring schemes while considering the same pool of poses. These findings suggest that by leveraging neighbor information via the GNN framework, DBX2 offers a significant advantage in accurately identifying native near-to-native ligand binding poses compared with docking methods that score each pose individually. **Figure 2B** illustrates an example of successful application of DBX2 for identifying the native pose of the potent TER-117 inhibitor bound to its target, the human Glutathione S-Transferase P1-1 (PDB id: 10gs) [68].

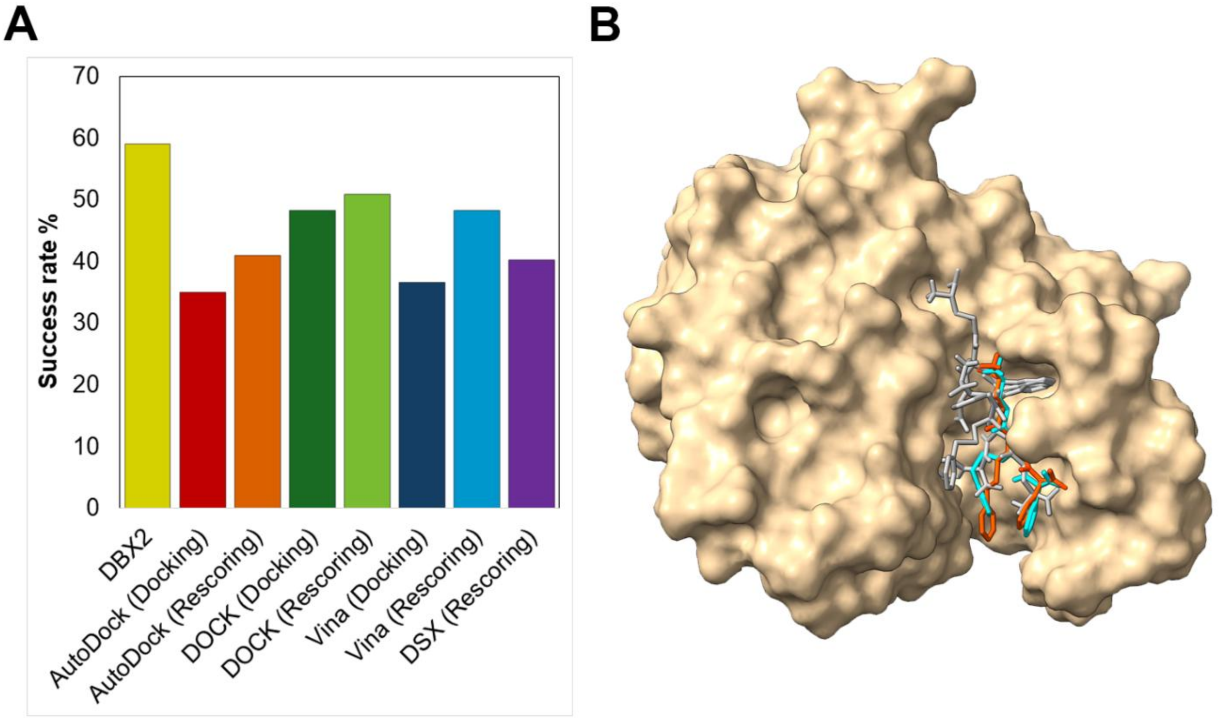

Figure 2: *(A) Success rate of identification of the pose correctness on hold-out dataset comparison between AutoDock, DOCK, Vina, DSX, and DBX2, comparing docking and rescoring strategies. Rescoring improved the performance of each docking program*

compared to standard docking alone, emphasizing the advantage of refining initial pose predictions by evaluating them with additional scoring functions. DBX2 node-level classification outperformed all the other tested methods (B) Crystal structure of human glutathione S-transferase (PDB id: 10gs) with bound TER117 inhibitor (cyan). The binding pose predicted by DBX2 (orange) aligns closely with the crystallographic structure, in contrast to the poses predicted as native by other docking software (grey).

Next, we evaluated the ability of the scoring functions to reproduce experimentally determined binding constants in the hold-out test set (**Table 1**). DBX2 directly computes the binding affinity from an ensemble of poses, so it does not require selecting a specific docking pose as input, unlike other scoring functions. For traditional scoring functions, since DOCK showed the best success rate among classical docking programs, we focused only on poses with the best DOCK scores (after rescoring) in order to compute binding affinities with docking scoring functions, similarly to our previous work [42]. Linear regression was performed to compare experimental binding affinities from the hold-out dataset with the scores of the best poses from DOCK using different scoring functions and their linear combinations [42]. For DBX2, the affinity values for each protein-ligand complex in the hold-out dataset were predicted as graph–level tasks, hence as readouts of ensembles of poses generated for a system rather than relying on a single pose.

Table 1: *R^2, Pearson correlation coefficients and predictive index values between experimental binding affinities and the scores provided by multiple scoring functions.*

Number of functions	Scoring function/combination	R^2	Pearson coefficient	Predictive index
1	**DBX2**	**0.38**	**0.61**	**0.79**
1	AutoDock	0.20	0.45	0.45
1	DOCK	0.16	0.41	0.42
1	Vina	0.25	0.52	0.48
1	DSX	0.22	0.47	0.46
2	AutoDock, Vina	0.25	0.50	0.49

3	AutoDock, Vina, DOCK	0.18	0.44	0.43
3	AutoDock, Vina, DSX	0.23	0.49	0.48
4	AutoDock, Vina, DSX, DOCK	0.22	0.47	0.47

Our results showed that DBX2 exhibited the highest correlation with experimental binding affinities on the hold-out dataset, outperforming other scoring functions. In contrast, DOCK, despite showing the best prediction of binding poses, had the lowest correlation ($R^2 = 0.16$). DBX2 scoring function also displayed a significantly higher predictive index (0.79) than other methods, indicating its potential suitability in ranking active molecules based on their binding affinities to a target of interest. Likewise, the Pearson coefficient of DBX2 (0.61) indicated a good predictive power based on pharmaceutical industry standards [69]. Nevertheless, the R^2 value, while indicating positive correlation as well as an improvement compared with physics-based methods, remained low (0.38). While our results suggest that docking poses ensembles are more suitable than single poses for binding affinity predictions, they likely fail to provide a comprehensive thermodynamic picture of binding processes, due to the approximations (especially, neglection of protein flexibility and water) necessary to ensure the high throughput required in VS. Correlations of experimental values versus computational scores are shown in **Figure S4**.

Moreover, the scoring power on the hold-out set of DBX2 was compared with published state-of-the-art methods that were trained and tested on the same splits or supersets of them. Thus, DBX2 was compared with GNN-MP neural network (MPNN) models from Volkov et al [35] and Pafnucy model from Stepniewska-Dziubinska et al [70]. The first class of models are GNNs with a customizable hidden size and a two-layer dense module, which map protein- (P), ligand- (L) and protein-ligand interactions (I) graph representations to ligand-protein affinities. The Pafnucy model is a state-of-the-art convolutional neural network utilizing 3D convolution to produce a feature map for protein and ligand atoms to predict ligand-protein affinity. Notably, these models were already trained and tested on the same datasets as used in DBX2 (PDBbind v2016 dataset and the hold-out test set, respectively) as reported in the previous studied [35]. The comparison of Rp and RMSE on all models is summarized in **Table S4**.

Even though the number of entries in the training set for DBX2 was lower than other models, it exhibited significantly improved performances in predicting binding affinity against hold-out set with respect to GNN-MPNN pure interaction (I) models from Volkov et al [35] and Pafnucy model [70], as evident from the Rp and RMSE values, and comparable performances with GNN models that include protein and ligand structural information explicitly. Importantly, DBX2 is entirely based on energetic ensemble representations that do not consider any structural information about ligand and/or protein structures, differently than the models from [35] and [70]. This observation suggests that DBX2 could (at least partially) overcome the hidden biases causing memorization of 2D molecular patterns that these models display, as described in the study by Volkov et al [35], while significantly outperforming the success rate of generalizable pure interaction models.

Retrospective virtual screening

LIT-PCBA is a chemical dataset designed to eliminate hidden chemical biases. Derived from bioassays, it mimics experimental screening decks, spans diverse protein targets, and has been validated across multiple screening methods, making it suitable for both structure- and ligand-based virtual screening experiments [41]. In order to test the VS power of DBX2 in realistic scenarios, we focused on three LIT-PCBA targets that were not present in our training set: FEN1, GBA, and MTORC1. The numbers of active and inactive compounds for each LIT-PCBA protein target at the beginning of the retrospective VS experiment and after the first round of Vina docking (with the top 20,000 molecules brought forward) are reported in **Table S5**.

After generating additional poses with AutoDock for molecules endowed by the Vina docking step, rescoring with different scoring functions (including DBX2) was performed and the result evaluated by computing top-100 hit rate, EF2, and NEF (**Figure 3A, 3B, and 3C**). DBX2 demonstrated superior performance across all metrics (EF2, NEF, and top-100 hit rate) when compared to other scoring functions for the three target proteins. DBX2's node-level predictions, which assess the likelihood of each binding pose within a specific graph, consistently matched the screening power of graph-level predictions of binding affinities. Interestingly, Gnina, another ML-based tool that recently demonstrated state-of-the-art performance in prospective drug discovery challenges [71], also performed well, further validating the potential of data-driven models in VS tasks.

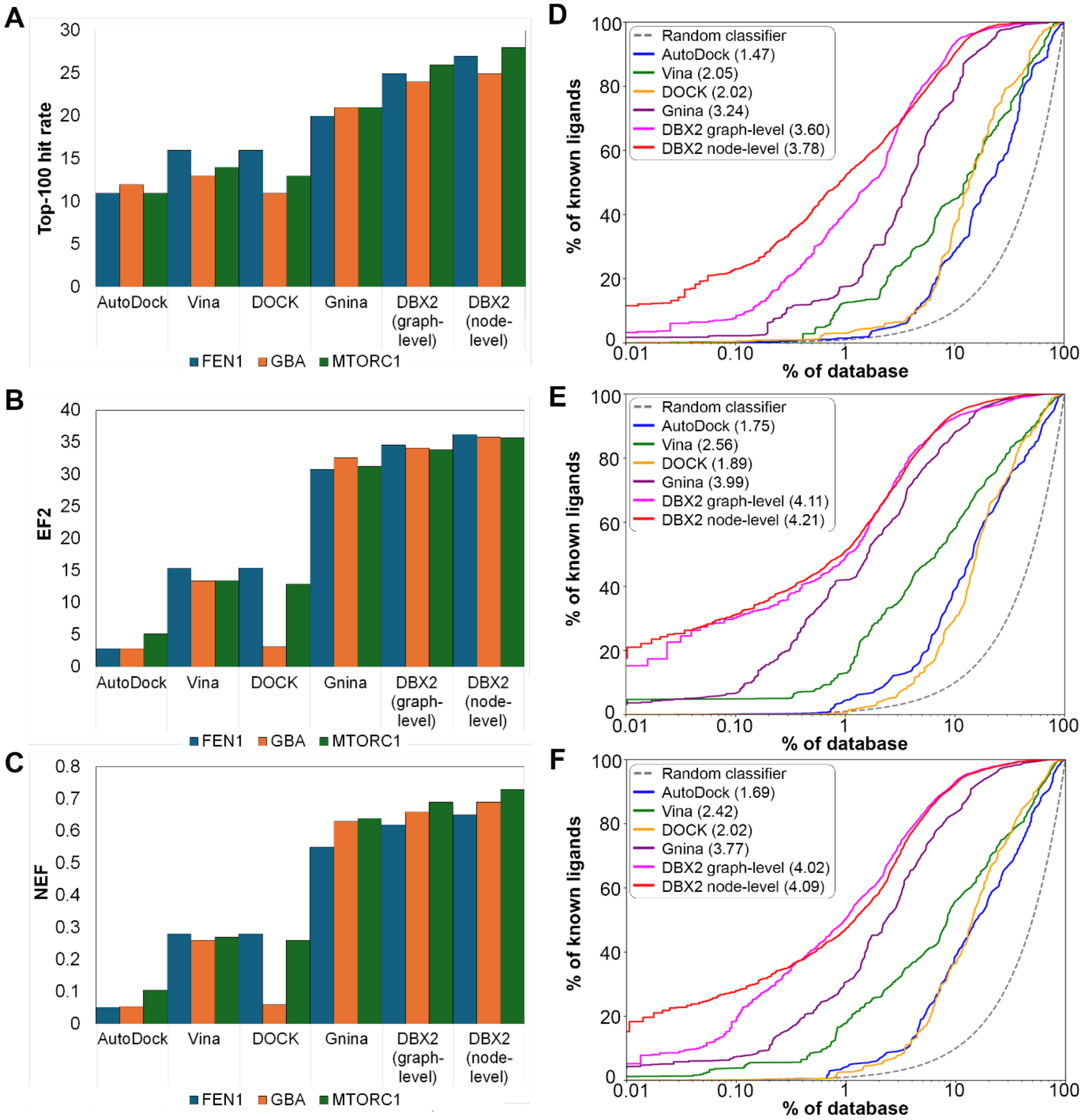

Figure 3: *Retrospective VS results of different scoring functions on LIT-PCBA database (A) top-100 hit rate (B) EF2 (C) NEF. Higher values in top-100 hit rate, EF2 and NEF corresponding to superior performance in identifying active compound at top-ranked.*

Enrichment plot comparison between DBX2 graph-level (magenta), DBX2 node-level (red), Gnina (purple), AutoDock (blue), Vina (green), and DOCK (yellow) on (A) FLAP Endonuclease (FEN1) protein, (B) Glucocerebrosidase (GBA) protein, and (C) Mechanistic target of rapamycin (MTORC1).

Additionally, logAUC was plotted (Figure 6D, 6E, 6F) and BEDROC were calculated (Table S6) to assess each scoring functions' ability to distinguish between active and inactive compounds. DBX2 demonstrates superior performance across both logAUC and BEDROC with the two scaling functions, suggesting a robust efficacy in prioritizing active compounds throughout top and broad ranks of compounds. Notably, node-level predictions show the highest performance, followed by graph-level predictions and Gnina's CNNAffinity scoring function.

Conclusions

We introduced DBX2, a novel GNN framework that enables to learn computational ensembles of small molecule-protein conformations as single graphs to predict binding modes and affinities. The model relies solely on simple energetic features derived from docking, without incorporating ligand and protein structural information that render conventional GNNs prone to memorization and consequently, poor generalization. We comprehensively evaluated DBX2 across various metrics for docking and VS tasks, underscoring its effectiveness as a robust tool for binding affinity prediction and virtual screening compared to conventional scoring functions and ML models based on single poses. At the same time, some caveats associated with the ensemble-based method emerged, especially reflected in the poor correlation between graph-level predicted and experimental binding affinities. We reasoned that these constraints can be ascribed to the limitations of the data generating process, i.e., docking, both in sampling the free energy landscape of binding and in quantitatively estimate binding energy contributions. Nevertheless, the significant performances observed for DBX2 not only advocate for its adoption in prospective drug discovery campaigns relying on high throughput VS but encourages also further exploration of ML models suitable for learning from computationally generated ensembles better representing binding thermodynamics than single poses. In this context, an exciting venue for further investigation could be the adaptation of the DBX2 architecture to MD-derived conformational ensembles of small molecule-protein complexes, to take into consideration also protein flexibility and induced fit as well as solvation.

Conflict of interest

The authors declare no conflict of interest.

Data availability

The DBX2 code is available at https://github.com/jp43/DockBox2. Trained models and training data are available at 10.5281/zenodo.14181651.

Acknowledgments

This work was supported by a Natural Sciences and Engineering Research Council of Canada Discovery Grant (RGPIN-2023-04129) awarded to F.G. Computations were performed on the resources of the Digital Research Alliance of Canada. We thank Cadence Molecular Sciences for providing an academic license for Openeye suite.

References

[1] S.-F. Zhou and W.-Z. Zhong, “Drug Design and Discovery: Principles and Applications,” *Molecules*, vol. 22, no. 2, p. 279, Feb. 2017, doi: 10.3390/molecules22020279.

[2] X. Zeng, S.-J. Li, S.-Q. Lv, M.-L. Wen, and Y. Li, “A comprehensive review of the recent advances on predicting drug-target affinity based on deep learning,” *Front. Pharmacol.*, vol. 15, Apr. 2024, doi: 10.3389/fphar.2024.1375522.

[3] X. Du *et al.*, “Insights into Protein–Ligand Interactions: Mechanisms, Models, and Methods,” *Int. J. Mol. Sci.*, vol. 17, no. 2, p. 144, Jan. 2016, doi: 10.3390/ijms17020144.

[4] D. J. Newman and G. M. Cragg, “Natural Products as Sources of New Drugs over the Nearly Four Decades from 01/1981 to 09/2019,” *J. Nat. Prod.*, vol. 83, no. 3, pp. 770–803, Mar. 2020, doi: 10.1021/acs.jnatprod.9b01285.

[5] T. Takebe, R. Imai, and S. Ono, “The Current Status of Drug Discovery and Development as Originated in UNITED STATES Academia: The Influence of Industrial and Academic Collaboration on Drug Discovery and Development,” *Clin. Transl. Sci.*, vol. 11, no. 6, pp. 597–606, Nov. 2018, doi: 10.1111/cts.12577.

[6] J. Kuan, M. Radaeva, A. Avenido, A. Cherkasov, and F. Gentile, “Keeping pace with the explosive growth of chemical libraries with structure-based virtual screening,” *WIREs Comput. Mol. Sci.*, vol. 13, no. 6, p. e1678, Nov. 2023, doi: 10.1002/wcms.1678.

[7] B. Shaker, S. Ahmad, J. Lee, C. Jung, and D. Na, “In silico methods and tools for drug discovery,” *Comput. Biol. Med.*, vol. 137, p. 104851, Oct. 2021, doi: 10.1016/j.compbiomed.2021.104851.

[8] M. De Vivo, M. Masetti, G. Bottegoni, and A. Cavalli, “Role of Molecular Dynamics and Related Methods in Drug Discovery,” *J. Med. Chem.*, vol. 59, no. 9, pp. 4035–4061, May 2016, doi: 10.1021/acs.jmedchem.5b01684.

[9] S. Decherchi and A. Cavalli, “Thermodynamics and Kinetics of Drug-Target Binding by Molecular Simulation,” *Chem. Rev.*, vol. 120, no. 23, pp. 12788–12833, Dec. 2020, doi: 10.1021/acs.chemrev.0c00534.

[10] Y. Miao, A. Bhattarai, and J. Wang, “Ligand Gaussian Accelerated Molecular Dynamics (LiGaMD): Characterization of Ligand Binding Thermodynamics and Kinetics,” *J. Chem. Theory Comput.*, vol. 16, no. 9, pp. 5526–5547, Sep. 2020, doi: 10.1021/acs.jctc.0c00395.

[11] Y.-T. Wang *et al.*, “Structural insights into Nirmatrelvir (PF-07321332)-3C-like SARS-CoV-2 protease complexation: a ligand Gaussian accelerated molecular

dynamics study,” *Phys. Chem. Chem. Phys.*, vol. 24, no. 37, pp. 22898–22904, 2022, doi: 10.1039/D2CP02882D.

[12] D. W. Kneller *et al.*, “Covalent narlaprevir- and boceprevir-derived hybrid inhibitors of SARS-CoV-2 main protease,” *Nat. Commun.*, vol. 13, no. 1, p. 2268, Apr. 2022, doi: 10.1038/s41467-022-29915-z.

[13] S. Wolf, B. Lickert, S. Bray, and G. Stock, “Multisecond ligand dissociation dynamics from atomistic simulations,” *Nat. Commun.*, vol. 11, no. 1, p. 2918, Jun. 2020, doi: 10.1038/s41467-020-16655-1.

[14] E. Paquet and H. L. Viktor, “Molecular Dynamics, Monte Carlo Simulations, and Langevin Dynamics: A Computational Review,” *BioMed Res. Int.*, vol. 2015, pp. 1–18, 2015, doi: 10.1155/2015/183918.

[15] K. Crampon, A. Giorkallos, M. Deldossi, S. Baud, and L. A. Steffenel, “Machine-learning methods for ligand–protein molecular docking,” *Drug Discov. Today*, vol. 27, no. 1, pp. 151–164, Jan. 2022, doi: 10.1016/j.drudis.2021.09.007.

[16] P. C. Agu *et al.*, “Molecular docking as a tool for the discovery of molecular targets of nutraceuticals in diseases management,” *Sci. Rep.*, vol. 13, no. 1, Aug. 2023, doi: 10.1038/s41598-023-40160-2.

[17] F. Liu *et al.*, “Large library docking identifies positive allosteric modulators of the calcium-sensing receptor,” *Science*, vol. 385, no. 6715, p. eado1868, Sep. 2024, doi: 10.1126/science.ado1868.

[18] J. Lyu *et al.*, “Ultra-large library docking for discovering new chemotypes,” *Nature*, vol. 566, no. 7743, pp. 224–229, Feb. 2019, doi: 10.1038/s41586-019-0917-9.

[19] A. Manglik *et al.*, “Structure-based discovery of opioid analgesics with reduced side effects,” *Nature*, vol. 537, no. 7619, pp. 185–190, Sep. 2016, doi: 10.1038/nature19112.

[20] H. Wang *et al.*, “Structure-Based Evolution of G Protein-Biased μ-Opioid Receptor Agonists,” *Angew. Chem. Int. Ed.*, vol. 61, no. 26, p. e202200269, Jun. 2022, doi: 10.1002/anie.202200269.

[21] N. Zernov, V. Ghamaryan, D. Melenteva, A. Makichyan, L. Hunanyan, and E. Popugaeva, “Discovery of a novel piperazine derivative, cmp2: a selective TRPC6 activator suitable for treatment of synaptic deficiency in Alzheimer’s disease hippocampal neurons,” *Sci. Rep.*, vol. 14, no. 1, p. 23512, Oct. 2024, doi: 10.1038/s41598-024-73849-z.

[22] P. Agnihotri, A. K. Mishra, S. Mishra, V. K. Sirohi, A. A. Sahasrabuddhe, and J. V. Pratap, “Identification of Novel Inhibitors of *Leishmania donovani* γ-Glutamylcysteine

Synthetase Using Structure-Based Virtual Screening, Docking, Molecular Dynamics Simulation, and in Vitro Studies," *J. Chem. Inf. Model.*, vol. 57, no. 4, pp. 815–825, Apr. 2017, doi: 10.1021/acs.jcim.6b00642.

[23] L. Wang *et al.*, "Discovery of potential small molecular SARS-CoV-2 entry blockers targeting the spike protein," *Acta Pharmacol. Sin.*, vol. 43, no. 4, pp. 788–796, Apr. 2022, doi: 10.1038/s41401-021-00735-z.

[24] R. Jockers *et al.*, "Update on melatonin receptors: IUPHAR Review 20," *Br. J. Pharmacol.*, vol. 173, no. 18, pp. 2702–2725, Sep. 2016, doi: 10.1111/bph.13536.

[25] D. P. Zlotos, N. M. Riad, M. B. Osman, B. R. Dodda, and P. A. Witt-Enderby, "Novel difluoroacetamide analogues of agomelatine and melatonin: probing the melatonin receptors for MT $_1$ selectivity," *MedChemComm*, vol. 6, no. 7, pp. 1340–1344, 2015, doi: 10.1039/C5MD00190K.

[26] R. M. Stein *et al.*, "Virtual discovery of melatonin receptor ligands to modulate circadian rhythms," *Nature*, vol. 579, no. 7800, pp. 609–614, Mar. 2020, doi: 10.1038/s41586-020-2027-0.

[27] E. A. Fink *et al.*, "Structure-based discovery of nonopioid analgesics acting through the α $_{2A}$ -adrenergic receptor," *Science*, vol. 377, no. 6614, p. eabn7065, Sep. 2022, doi: 10.1126/science.abn7065.

[28] K. M. Elokely and R. J. Doerksen, "Docking Challenge: Protein Sampling and Molecular Docking Performance," *J. Chem. Inf. Model.*, vol. 53, no. 8, pp. 1934–1945, Aug. 2013, doi: 10.1021/ci400040d.

[29] A. Mastropietro, G. Pasculli, and J. Bajorath, "Learning characteristics of graph neural networks predicting protein–ligand affinities," *Nat. Mach. Intell.*, vol. 5, no. 12, pp. 1427–1436, Nov. 2023, doi: 10.1038/s42256-023-00756-9.

[30] J. Wu, H. Chen, M. Cheng, and H. Xiong, "CurvAGN: Curvature-based Adaptive Graph Neural Networks for Predicting Protein-Ligand Binding Affinity," *BMC Bioinformatics*, vol. 24, no. 1, p. 378, Oct. 2023, doi: 10.1186/s12859-023-05503-w.

[31] S. Moon, W. Zhung, S. Yang, J. Lim, and W. Y. Kim, "PIGNet: a physics-informed deep learning model toward generalized drug–target interaction predictions," *Chem. Sci.*, vol. 13, no. 13, pp. 3661–3673, 2022, doi: 10.1039/D1SC06946B.

[32] C. Shen *et al.*, "A generalized protein–ligand scoring framework with balanced scoring, docking, ranking and screening powers," *Chem. Sci.*, vol. 14, no. 30, pp. 8129–8146, 2023, doi: 10.1039/D3SC02044D.

[33] S. Zhang *et al.*, “SS-GNN: A Simple-Structured Graph Neural Network for Affinity Prediction,” *ACS Omega*, vol. 8, no. 25, pp. 22496–22507, Jun. 2023, doi: 10.1021/acsomega.3c00085.

[34] H. Shen, Y. Zhang, C. Zheng, B. Wang, and P. Chen, “A Cascade Graph Convolutional Network for Predicting Protein–Ligand Binding Affinity,” *Int. J. Mol. Sci.*, vol. 22, no. 8, p. 4023, Apr. 2021, doi: 10.3390/ijms22084023.

[35] M. Volkov *et al.*, “On the Frustration to Predict Binding Affinities from Protein–Ligand Structures with Deep Neural Networks,” *J. Med. Chem.*, vol. 65, no. 11, pp. 7946–7958, Jun. 2022, doi: 10.1021/acs.jmedchem.2c00487.

[36] G. Corso, H. Stärk, B. Jing, R. Barzilay, and T. Jaakkola, “DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking,” 2022, *arXiv*. doi: 10.48550/ARXIV.2210.01776.

[37] H. Jiang *et al.*, “Predicting Protein–Ligand Docking Structure with Graph Neural Network,” *J. Chem. Inf. Model.*, vol. 62, no. 12, pp. 2923–2932, Jun. 2022, doi: 10.1021/acs.jcim.2c00127.

[38] Y. Min *et al.*, “From Static to Dynamic Structures: Improving Binding Affinity Prediction with Graph-Based Deep Learning,” 2022, doi: 10.48550/ARXIV.2208.10230.

[39] R. Wang, X. Fang, Y. Lu, and S. Wang, “The PDBbind Database: Collection of Binding Affinities for Protein−Ligand Complexes with Known Three-Dimensional Structures,” *J. Med. Chem.*, vol. 47, no. 12, pp. 2977–2980, Jun. 2004, doi: 10.1021/jm030580l.

[40] Z. Liu *et al.*, “Forging the Basis for Developing Protein–Ligand Interaction Scoring Functions,” *Acc. Chem. Res.*, vol. 50, no. 2, pp. 302–309, Feb. 2017, doi: 10.1021/acs.accounts.6b00491.

[41] V.-K. Tran-Nguyen, C. Jacquemard, and D. Rognan, “LIT-PCBA: An Unbiased Data Set for Machine Learning and Virtual Screening,” *J. Chem. Inf. Model.*, vol. 60, no. 9, pp. 4263–4273, Sep. 2020, doi: 10.1021/acs.jcim.0c00155.

[42] J. Preto and F. Gentile, “Assessing and improving the performance of consensus docking strategies using the DockBox package,” *J. Comput. Aided Mol. Des.*, vol. 33, no. 9, pp. 817–829, Sep. 2019, doi: 10.1007/s10822-019-00227-7.

[43] OpenEye, *OpenEye Toolkits*. Cadence Molecular Sciences, Santa Fe, NM. [Online]. Available: http://www.eyesopen.com

[44] *Molecular Operating Environment (MOE)*. Chemical Computing Group ULC, 910-1010 Sherbrooke St. W., Montreal, QC H3A 2R7.

[45] G. M. Morris *et al.*, “AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility,” *J. Comput. Chem.*, vol. 30, no. 16, pp. 2785–2791, Dec. 2009, doi: 10.1002/jcc.21256.

[46] O. Trott and A. J. Olson, “AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading,” *J. Comput. Chem.*, vol. 31, no. 2, pp. 455–461, Jan. 2010, doi: 10.1002/jcc.21334.

[47] T. E. Balius, S. Mukherjee, and R. C. Rizzo, “Implementation and evaluation of a docking-rescoring method using molecular footprint comparisons,” *J. Comput. Chem.*, vol. 32, no. 10, pp. 2273–2289, Jul. 2011, doi: 10.1002/jcc.21814.

[48] A. T. McNutt *et al.*, “GNINA 1.0: molecular docking with deep learning,” *J. Cheminformatics*, vol. 13, no. 1, p. 43, Dec. 2021, doi: 10.1186/s13321-021-00522-2.

[49] G. Neudert and G. Klebe, “*DSX* : A Knowledge-Based Scoring Function for the Assessment of Protein–Ligand Complexes,” *J. Chem. Inf. Model.*, vol. 51, no. 10, pp. 2731–2745, Oct. 2011, doi: 10.1021/ci200274q.

[50] G. M. Morris *et al.*, “Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function,” *J. Comput. Chem.*, vol. 19, no. 14, pp. 1639–1662, Nov. 1998, doi: 10.1002/(SICI)1096-987X(19981115)19:14<1639::AID-JCC10>3.0.CO;2-B.

[51] R. Salomon-Ferrer, D. A. Case, and R. C. Walker, “An overview of the Amber biomolecular simulation package,” *WIREs Comput. Mol. Sci.*, vol. 3, no. 2, pp. 198–210, Mar. 2013, doi: 10.1002/wcms.1121.

[52] W. L. Hamilton, R. Ying, and J. Leskovec, “Inductive Representation Learning on Large Graphs,” Sep. 10, 2018, *arXiv*: arXiv:1706.02216. Accessed: Oct. 01, 2024. [Online]. Available: http://arxiv.org/abs/1706.02216

[53] D. Duvenaud *et al.*, “Convolutional Networks on Graphs for Learning Molecular Fingerprints,” Nov. 03, 2015, *arXiv*: arXiv:1509.09292. Accessed: Oct. 01, 2024. [Online]. Available: http://arxiv.org/abs/1509.09292

[54] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal Loss for Dense Object Detection,” Feb. 07, 2018, *arXiv*: arXiv:1708.02002. Accessed: Oct. 19, 2024. [Online]. Available: http://arxiv.org/abs/1708.02002

[55] X. Zhang, Y. Li, J. Wang, G. Xu, and Y. Gu, “A Multi-perspective Model for Protein–Ligand-Binding Affinity Prediction,” *Interdiscip. Sci. Comput. Life Sci.*, vol. 15, no. 4, pp. 696–709, Dec. 2023, doi: 10.1007/s12539-023-00582-y.

[56] C. Cortes, M. Mohri, and A. Rostamizadeh, “L2 Regularization for Learning Kernels,” May 09, 2012, *arXiv*: arXiv:1205.2653. Accessed: Oct. 19, 2024. [Online]. Available: http://arxiv.org/abs/1205.2653

[57] D. A. Pearlman and P. S. Charifson, “Are Free Energy Calculations Useful in Practice? A Comparison with Rapid Scoring Functions for the p38 MAP Kinase Protein System,” *J. Med. Chem.*, vol. 44, no. 21, pp. 3417–3423, Oct. 2001, doi: 10.1021/jm0100279.

[58] J. C. Exell *et al.*, “Cellularly active N-hydroxyurea FEN1 inhibitors block substrate entry to the active site,” *Nat. Chem. Biol.*, vol. 12, no. 10, pp. 815–821, Oct. 2016, doi: 10.1038/nchembio.2148.

[59] B. Brumshtein *et al.*, “Cyclodextrin-mediated crystallization of acid β-glucosidase in complex with amphiphilic bicyclic nojirimycin analogues,” *Org. Biomol. Chem.*, vol. 9, no. 11, p. 4160, 2011, doi: 10.1039/c1ob05200d.

[60] S.-Y. Lee *et al.*, “Proximity-Directed Labeling Reveals a New Rapamycin-Induced Heterodimer of FKBP25 and FRB in Live Cells,” *ACS Cent. Sci.*, vol. 2, no. 8, pp. 506–516, Aug. 2016, doi: 10.1021/acscentsci.6b00137.

[61] K. Palacio-Rodríguez, I. Lans, C. N. Cavasotto, and P. Cossio, “Exponential consensus ranking improves the outcome in docking and receptor ensemble docking,” *Sci. Rep.*, vol. 9, no. 1, p. 5142, Mar. 2019, doi: 10.1038/s41598-019-41594-3.

[62] J.-F. Truchon and C. I. Bayly, “Evaluating Virtual Screening Methods: Good and Bad Metrics for the ‘Early Recognition’ Problem,” *J. Chem. Inf. Model.*, vol. 47, no. 2, pp. 488–508, Mar. 2007, doi: 10.1021/ci600426e.

[63] Y. Perez-Castillo *et al.*, “Fusing Docking Scoring Functions Improves the Virtual Screening Performance for Discovering Parkinson’s Disease Dual Target Ligands,” *Curr. Neuropharmacol.*, vol. 15, no. 8, Nov. 2017, doi: 10.2174/1570159X15666170109143757.

[64] G.-L. Xiong, W.-L. Ye, C. Shen, A.-P. Lu, T.-J. Hou, and D.-S. Cao, “Improving structure-based virtual screening performance via learning from scoring function components,” *Brief. Bioinform.*, vol. 22, no. 3, p. bbaa094, May 2021, doi: 10.1093/bib/bbaa094.

[65] J. C. D. Lopes, F. M. Dos Santos, A. Martins-José, K. Augustyns, and H. De Winter, “The power metric: a new statistically robust enrichment-type metric for virtual screening applications with early recovery capability,” *J. Cheminformatics*, vol. 9, no. 1, p. 7, Dec. 2017, doi: 10.1186/s13321-016-0189-4.

[66] S. Liu *et al.*, “Practical Model Selection for Prospective Virtual Screening,” *J. Chem. Inf. Model.*, vol. 59, no. 1, pp. 282–293, Jan. 2019, doi: 10.1021/acs.jcim.8b00363.

[67] R. P. Sheridan, S. B. Singh, E. M. Fluder, and S. K. Kearsley, “Protocols for Bridging the Peptide to Nonpeptide Gap in Topological Similarity Searches,” *J. Chem. Inf. Comput. Sci.*, vol. 41, no. 5, pp. 1395–1406, Sep. 2001, doi: 10.1021/ci0100144.

[68] A. J. Oakley *et al.*, “The structures of human glutathione transferase P1-1 in complex with glutathione and various inhibitors at high resolution,” *J. Mol. Biol.*, vol. 274, no. 1, pp. 84–100, Nov. 1997, doi: 10.1006/jmbi.1997.1364.

[69] E. J. Martin, V. R. Polyakov, X.-W. Zhu, L. Tian, P. Mukherjee, and X. Liu, “All-Assay-Max2 pQSAR: Activity Predictions as Accurate as Four-Concentration IC $_{50}$ s for 8558 Novartis Assays,” *J. Chem. Inf. Model.*, vol. 59, no. 10, pp. 4450–4459, Oct. 2019, doi: 10.1021/acs.jcim.9b00375.

[70] M. M. Stepniewska-Dziubinska, P. Zielenkiewicz, and P. Siedlecki, “Development and evaluation of a deep learning model for protein–ligand binding affinity prediction,” *Bioinformatics*, vol. 34, no. 21, pp. 3666–3674, Nov. 2018, doi: 10.1093/bioinformatics/bty374.

[71] F. Li *et al.*, “CACHE Challenge #1: targeting the WDR domain of LRRK2, a Parkinson’s Disease associated protein,” Jul. 18, 2024, *Biochemistry*. doi: 10.1101/2024.07.18.603797.