Chemical
Science

Accepted Manuscript

This article can be cited before page numbers have been issued, to do this please use: T. Thaingtamtanha, J. Preto and F. Gentle, *Chem. Sci.*, 2025, DOI: 10.1039/D4SC07875F.

ROYAL SOCIETY
OF CHEMISTRY

rsc.li/chemical-science

View Article Online
DOI: 10.1039/D4SC07875F

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 K1N 6N5, 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 K1H 8M5, Canada*

^*equal contribution*

**email: fgentile@uottawa.ca*

Abstract. Predicting the geometry and strength governing small molecule-protein interactions remains a paramount challenge in drug discovery due to their complex and dynamic nature. Several 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 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 framework via 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, compared with state-of-the-art physics- and ML-based tools. Our results encourage further exploration of ML models learning from conformational ensembles to accurately model small molecule-protein interactions and thermodynamics. The DBX2 code is available at https://github.com/jp43/DockBox2.

View Article Online
DOI: 10.1039/D4SC07875F

Introduction

Drugs exert their therapeutic effects by binding to specific biomolecular targets, typically proteins or nucleic acids, and modulating their function, thereby inhibiting or restoring processes related to various diseases. The initial step in the drug discovery pipeline involves identifying molecules binding to the target of interest with high affinity and specificity [1], hence making the accurate prediction of both a crucial aspect for therapeutic 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 the estimation of ligand-protein affinities and interactions is essential, significant challenges arise due to the dynamic nature of these complexes. Molecular dynamics (MD) simulations can provide valuable insights into the nature of these interactions, *e.g.,* by considering an ensemble of bound conformations to compute thermodynamically accurate energies [8]. This is usually done by simulating the complexes in their thermodynamic equilibrium and considering 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]. 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 consistent 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. However, both standard and biased MD methods require significant computational power that makes these techniques unsuited for high-throughput screening purposes. Consequently, faster and less accurate methods such as

View Article Online
DOI: 10.1039/D4SC07875F

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, both to identify the most probable pose and to estimate the binding affinity [15]. Despite its simplicity, docking has shown great potential for the identification active molecules from vast backgrounds of inactive compounds [17], [18], with its impact extending across numerous therapeutic areas. Manglik et al., for example, docked over 3 million molecules against the μ-opioid receptor (μOR), leading to the discovery of PZM21, a G protein-biased μOR agonist [19]. Zernov et al. discovered a compound targeting the transient receptor potential cation channel 6 as a potential starting point to develop anti-Alzheimer's therapies, with *in-vitro* studies confirming its efficacy, stability, and target specificity without adverse effects [20]. Stein et al. employed docking to screen over 150 million molecules targeting melatonin receptor 1 (MT1) in the search for therapeutics addressing sleep disorders and depression, reporting a novel chemotype with experimentally validated, selective MT1 agonist activity [21]. Fink et al. utilized large-scale docking to identify novelα2A-adrenergic receptor (α2AAR) agonists with fewer adverse effects compared to earlier treatments, as new starting points to develop nonopioid analgesics [22]. These and many other studies underscore the important role of docking in advancing drug discovery.

However, several limitations remain in docking, mainly due to the approximative nature of scoring functions and the neglection of flexibility [15], [23]. Thus, ML methods 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 [24]. Several models have been proposed, such as CurvAGN [25], PIGNet [26], GenScore [27] and SS-GNN [28], reporting strong correlations between predicted and experimental affinities [24], [29], [30]. Additionally, GNNs have been applied in generative settings to replace physics-based sampling in generate and scoring ligand-protein poses, such as in DiffDock [31] and MedusaGraph [32]. 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 phycial

View Article Online
DOI: 10.1039/D4SC07875F

chemistry of the interactions [24], [30]. Moreover, these methods generally map single pose graphs to binding affinities, thus neglecting full thermodynamic profile and dynamics of ligand-protein interactions that depends on multiple conformations [24]. Notably, recent efforts have been made to consider multiple conformations in training GNNs for binding affinity predictions, such as Dynaformer, a method that encode each MD-derived binding conformation into a graph within a framework to provide better affinity estimates [33]. Notably, Dynaformer still relies on mapping each conformation to a single affinity value, and requires the use of costly simulations, hence limiting its scalability.

In this work, we introduce DockBox2 (DBX2), a GNNs framework enabling to encode multiple ligand-protein conformations derived from docking within individual graph neural networks in order to leverage ensemble representations for jointly predicting pose likelihood at the node level and binding affinities at the graph level. 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 [34]. The refined set of PDBbind v2016 (4,057 complexes) [35] was used to train the model. PDBbind is a comprehensive and widely adopted benchmark for protein-ligand binding, and several widely used benchmark datasets, such as CASF-2016 [36], are derived from this refined set. The PDBbind v2019-based hold-out test set built by Volkov et al. [30] and the Runs N' Pose database from Škrinjar et al. [37], consisting of 3,393 and 2,600 complexes respectively, were used as external test sets. Volkov's dataset is curated to mitigate latent biases, such as structural patterns in ligands or proteins, which can favor GNN memorization rather than protein-ligand interaction learning. As highlighted in previous studies [24], [30], this memorization often arises from significant redundancies

View Article Online
DOI: 10.1039/D4SC07875F

between training and test sets, resulting in data leakage. The Runs N' Poses dataset is a recently developed dataset containing high-resolution protein-ligand systems released after the publication of PDBbind v2020 and the training date cutoff of several protein-ligand co-folding models (e.g., AlphaFold3 [38], Chai-1 [39], Protenix [40], and Boltz-1 [41]). A subset of the LIT-PCBA database [42] was used to perform retrospective VS experiments.

Protein and ligand preparation

Complexes from PDBbind were prepared following the same procedure of our previous work [43]. For retrospective VS, dominant protonation and tautomerization states of small molecules were computed from the SMILES using Openeye 's QUACPAC [44] and converted into low-energy 3D conformations (mol2 format) using Openeye 's OMEGA tool [44]. The target proteins were prepared by removing redundant protein chains, along with non-essential ions, waters, and heteroatoms. The resulting protein structures were prepared using the Molecular Operating Environment (MOE) QuickPrep tool [45], to automatically add missing loops and assign reasonable conformations to the residues with alternate orientation. Subsequently, protonation states were generated 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) [43] was utilized to generate binding poses with AutoDock [46], Vina [47] and DOCK 6 (DOCK) [48], and rescore with their scoring function in addition to Gnina [49] and DSX [50]. The DBX configuration file used for this purpose on 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 [51] 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. All other parameters were left as default. 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

View Article Online
DOI: 10.1039/D4SC07875F

methods. Energy minimization was performed using AmberTools 17 [52] to prevent structural clashes and ensure appropriate rescoring with different programs. Rescoring was then conducted with AutoDock, Vina, DOCK, Gnina 's CNNScore, and DSX scoring functions.

Dockbox2 architecture

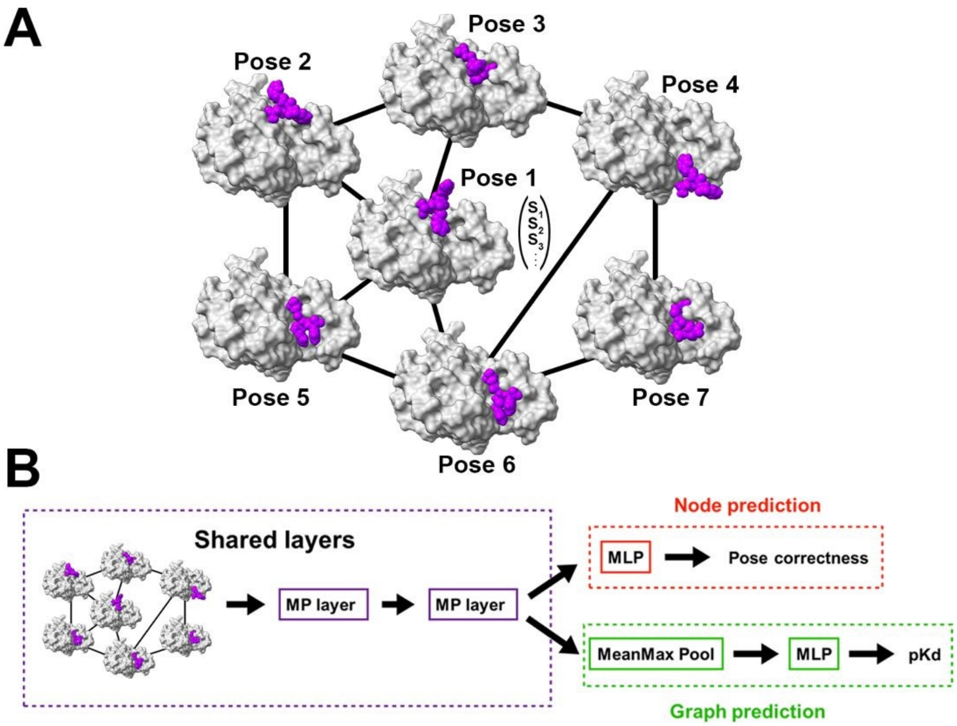

***Figure 1**: Architecture of DBX2. (A) Binding poses are represented as nodes. Two pose nodes are connected by an edge based on the root mean square deviation (RMSD) between them. Docking-derived energies and categorical features of each binding pose, here referred as s_1, s_2, s_3…, are used as node features. (B) Schematic of the DBX2 architecture; pose correctness and pKd are jointly learned as node- and graph-level tasks, respectively.*

DBX2 architecture is based on the GraphSAGE model [53] as shown in **Figure 1**. The ensemble of poses generated by docking a given ligand-protein pair is used to construct

View Article Online
DOI: 10.1039/D4SC07875F

a graph (**Figure 1A)**, with each node encoding an individual binding pose represented by categorical and energetic features, listed in **Table 1**.

Table 1. *DBX2 node features*.

Features	Description
Instance	Docking software utilized to generate the binding pose
Score	Docking score from original docking program
Rescoring score (AutoDock, Vina, Dock, DSX, Gnina)	Docking score obtained by rescoring the pose with another scoring function
Gaussian Terms (gauss1_inter, gauss2_inter, gauss1_intra, gauss2_intra)	Gaussian terms of the binding pose, as provided by Vina [47]
Hydrophobic interactions (hydrophobic_inter, hydrophobic_intra)	Hydrophobic terms evaluated by Vina [47]
Hydrogen bonding (hydrogrenbonding_inter, hydrogenbonding_intra)	Hydrogen bond terms evaluated by Vina [47]
Repulsion (repulsion_inter, repulsion_intra)	Repulsive Lennard-Jones energies from Vina [47]

All available scoring terms provided by Vina [47] were included as node features, with the exception of the entropy term which is determined solely by the ligand structure and therefore remains constant across different poses of the same ligand. In the constructed graph, pairwise root mean square deviation (RMSD) values are calculated between all poses. Two nodes are connected by an edge if the RMSD between the two poses is below a predefined threshold while the RMSD value is kept as edge feature. Graphs were generated using the *create_graphs* script available in the DBX2 package. In the shared layers, the DBX2 model uses the message passing (MP) framework [54], *i.e.*, for each node *i*, information from its neighbors $j \in \mathcal{N}(i)$ is gathered and aggregated using the symmetric mean (symmean) aggregation:

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

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

View Article Online
DOI: 10.1039/D4SC07875F

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:

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

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

Model training and evaluation

The total loss function of DBX2 consists of three components $Loss_n$, $Loss_g$, and $Loss_{reg}$

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

$Loss_n$ is the loss function for node-level task, where the binary focal cross entropy [55] is used as loss function applied to each node in the batch and averaged:

$$Loss_n = \frac{1}{N}\sum_{i=1}^{N} -\alpha_t^{(i)} \cdot \left(1 - p_t^{(i)}\right)^{\gamma} \cdot \log\left(p_t^{(i)}\right) \tag{4}$$

Where *N* is the number of nodes in the batch, γ is the focusing parameter (set to 1.0 in this study), and $\alpha_t^{(i)}$ is the weighting factor for each i-th sample:

$$\alpha_t^{(i)} = \begin{cases} \alpha & if\ y^{(i)} = 1 \\ 1-\alpha & if\ y^{(i)} = 0 \end{cases} \tag{5}$$

View Article Online
DOI: 10.1039/D4SC07875F

Where *α* is computed as:

$$\alpha = \frac{1}{1 + \frac{1}{G_t}\sum_{i=1}^{G_t}\frac{C_i}{I_i}} \tag{6}$$

Where G_t is the number of graphs in the training set, and C_i and I_i are the number of correct poses and incorrect poses in the i-th graph, respectively. A pose was considered as correct if it was 2 Å or less of RMSD from the experimental one. $p_t^{(i)}$ is the predicted probability output by the model for the correct class label of each i-th node:

$$p_t^{(i)} = \begin{cases} p^{(i)} & if\ y^{(i)} = 1 \\ 1 - p^{(i)} & if\ y^{(i)} = 0 \end{cases} \tag{7}$$

Where $p^{(i)}$ is the model output for each pose.

$Loss_g$ and w_1 are the loss function for the graph-level task and its weight, respectively. The optimal value of w_1 was determined through hyperparameter optimization (**Table S1**). $Loss_g$ corresponds to the root mean square error (RMSE) [56]:

$$Loss_g = \sqrt{\frac{1}{G}\sum_{i=1}^{G}(y_i - \hat{y}_i)^2} \tag{8}$$

Here G denotes the number of ligand-protein complexes in the batch, 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 $Loss_g$ contributes to correctly predicting the ligand-protein affinity, in which all poses within a graph are processed through message passing and readout, then used to predict the binding affinity. $Loss_{reg}$ is the regularization loss, while L2 regularization loss [57] was here used to prevent overfitting of model:

$$Loss_{reg} = \frac{1}{2}\sum_{i=1}^{n}\theta_i^2 \tag{9}$$

where θ_i represent the model parameter, n is the number of model parameter. The model was trained using the *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

View Article Online
DOI: 10.1039/D4SC07875F

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 the following hyperparameters: RMSD cutoff value to define an edge (RMSD cutoff), number of adjacent nodes to randomly sample for aggregation (nrof-neigh), and graph loss weight (w_1), for a total of 30 combinations (**Table S1**). Training and validation sets were prepared using the *split_train_val_dbx2* routine of the DBX2 package. The generated graphs were split for stratified 5-fold cross-validation, keeping a consistent distribution of protein families across all folds. Node and edge features for each graph were standardized using scikit-learn's StandardScaler [58]. For node-level predictions, success rate, accuracy, and area under the curve (AUC) were used as evaluation metrics. For graph-level predictions, RMSE was used.

Model testing

Models were compared for docking and scoring tasks with other methods on the hold-out and Runs N' Poses test sets. To evaluate docking power, the success rate was computed as the ratio of top-ranked poses with an RMSD equal or lower than a predefined threshold with respect to the experimental pose. Five different thresholds were tested, 1, 1.5, 2, 2.5 and 3 Å. For DBX2, the success rate was evaluated by considering the top-ranked poses from node-level predictions.

Next, the scoring power was assessed to evaluate the model's ability to predict experimental binding affinities 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. 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 combinations of scoring functions, as described in our previous study [43].

Scoring power was also evaluated using Pearson correlation coefficient and the predictive index (PI) as before [43]. Proposed by Pearlman et al. [59], PI measures the reliability of

View Article Online
DOI: 10.1039/D4SC07875F

a scoring function in identifying the most potent binder between two compounds. It is calculated as follows:

$$PI = \sum_{j>i} \sum_{i} w_{ij} C_{ij} \tag{10}$$

With

$$w_{ij} = |E_j - E_i| \tag{11}$$

$$C_{ij} = \begin{cases} 1 & if \quad \dfrac{E_j - E_i}{S_j - S_i} < 0 \\ -1 & if \quad \dfrac{E_j - E_i}{S_j - S_i} > 0 \\ 0 & if \quad S_j - S_i = 0 \end{cases} \tag{12}$$

Where E_i is the experimental binding affinity of compound i, and 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. w_{ij} is the weighting term which underscores the accurate ranking of compounds exhibiting substantial disparities in experimental binding affinities.

Retrospective virtual screening

VS experiments were conducted on the three target proteins from the LIT-PCBA database [42] that were not present in the DBX2 training set: Flap structure-specific Endonuclease 1 (FEN1, PDB id: 5FV7) [60], Glucocerebrosidase (GBA, PDB id: 2XWE) [61], and Mammalian Target of Rapamycin Complex 1 (MTORC1, PDB id: 5GPG) [62]. Initially, Vina was used to screen active-inactive sets derived from LIT-PCBA against each corresponding structure. The top 20,000 compounds based on the Vina ranking were then docked also with AutoDock to their respective targets. 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 (considering the CNNAffinity of the pose with the highest CNNScore) [49]. VS performances were evaluated by computing the logarithmic area under the curve (logAUC) [63], enrichment

View Article Online
DOI: 10.1039/D4SC07875F

factors (EF) and Boltzmann-enhanced discrimination of receiver operating characteristic (BEDROC) with adjust parameter (α) values of 20 and 80.5 using the CROC Python package [64], [65], [66].

The logAUC quantifies the performance of a VS method by assessing its ability to distinguish active compounds from decoys across the ranked list. By applying a logarithmic scale to the false positive rate axis, it places greater emphasis on the early retrieval of active compounds, which is critical in VS.

EF measures how effectively a VS method identifies active compounds within a specific fraction of the ranked list [67]. EF at a given cutoff $(\boldsymbol{x})$ is calculated from the ratio of true active compounds in the top $\boldsymbol{x}$ ranked compounds in relation to the ratio of true active compounds in the entire dataset:

$$EF(x) = \frac{TP/(TP+FP)}{[(TP+FN)/(TP+TN+FP+FN)]} = \frac{N \times n_s}{n \times N_s} \qquad (13)$$

Where TP and TN are true positives and true negatives, FP and FN are false positives and false negatives. N is a total number of compounds in the entire dataset, N_s is a total number of predicted active compounds in the selection set (x), n is a total number of true active compounds in the entire dataset, n_s is the number of true active compounds in the selection set (x). EF was computed by considering the top 2% of the ranked compounds for each scoring functions and for both graph-level and node-level predictions in DBX2 (EF2).

Normalized enrichment factor (NEF) rescales EF values into a range from 0 (bad prediction) to 1 (perfect prediction) [68], with the goal of standardizing comparison across different datasets. NEF is calculated as follow:

$$NEF(x) = \frac{EF(x)}{EF(x)_{max}} \qquad (14)$$

With

$$EF(x)_{max} = \frac{\mathrm{mi\,n}\{n_s,\ N \times x\}}{n \times x} \qquad (15)$$

View Article Online
DOI: 10.1039/D4SC07875F

Where $EF(x)_{max}$ denotes the maximum enrichment factor achievable within a selection set (x). n_s is the number of true active compounds in the selection set (x), N is the number of compounds in the entire dataset.

BEDROC metric emphasizes the concentration of active compounds at several range of ranked data sets [65], [68] through a scaling function (α). This metric is defined as:

$$BEDROC = \frac{RIE - RIE_{min}}{RIE_{min} - RIE_{max}} \quad (16)$$

With

$$RIE_{min} = \frac{1 - e^{\alpha R_{\alpha}}}{R_{\alpha}(1 - e^{\alpha})} \quad (17)$$

$$RIE_{max} = \frac{1 - e^{-\alpha R_{\alpha}}}{R_{\alpha}(1 - e^{-\alpha})} \quad (18)$$

$$RIE = \frac{\frac{1}{n}\sum_{i=1}^{n} e^{\alpha x_i}}{\frac{1}{n}\left(\frac{1 - e^{\alpha}}{{}^{e^{\alpha}}/_{N_{-1}}}\right)} \quad (19)$$

the Robust Initial Enhancement proposed by Sheridan et al [69], x_i is a relative ranking of active compound *i*. R_{α} is the fraction of active compound ($R_{\alpha} = \frac{n}{N}$), α is the scaling function.

We also investigated the potential of DBX2 to improve VS performance of individual docking programs (rather than on pose pools deriving from different software) and by using different docking setups to generate poses. The top 20,000 LIT-PCBA compounds docked and scored with Vina against FEN1 were redocked with AutoDock and Vina using several combinations of docking parameters for each program (**Table S2**). The resulting poses were subsequently subjected to DBX2, using the same settings used in the retrospective VS experiments. The same metrics were calculated to assess the effectiveness of DBX2 in this specific scenario.

View Article Online
DOI: 10.1039/D4SC07875F

Baseline models

We compared DBX2 model with other methods, including docking and rescoring tools either physics- or ML-based, using the following protocol:

- AutoDock, Vina, DOCK, Gnina, KarmaDock [70], RTMscore [71] and DBX2 were compared both in terms of docking and scoring power, as well as for retrospective VS (CarsiDock was excluded from the VS experiments due to the computational cost)
- CarsiDock [72], DSX and DBX2 were compared for docking and scoring power

Default settings were used for all programs. To evaluate the docking, scoring and VS capabilities of RTMscore and Gnina on the hold-out and Runs N' Poses sets, the binding poses used in DBX2 were also utilized for rescoring with these tools. For Gnina, the success rate was evaluated using CNNScore, and the scoring power was evaluated using the CNNAffinity and Minimized Affinity scores of the pose with the best CNNscore for each system. KarmaDock and CarsiDock, both generative models, automatically generated their own protein-ligand poses and associated scores.

Results and Discussion

Hyperparameter optimization

The results of hyperparameter optimization for the DBX2 model are summarized in **Table S3**. The best performing set of hyperparameters included a RMSD cutoff of 10 Å to define edges, a nrof-neigh of 30, and a graph-level loss weight (w_1) of 0.02, yielding an average success rate of 60% on 5-fold cross validation. The model with the highest performance was then retained and used in subsequent testing.

Docking and scoring power

We compared the success rate of DBX2 and other physics-based methods for the docking and rescoring tasks on the hold-out test set, as described in the Material and Methods section (**Figure 2A**). As expected, rescoring ensembles of docking poses with

View Article Online
DOI: 10.1039/D4SC07875F

different scoring functions led to significantly improved performance due to enhanced pose sampling, as observed in previous studies [43]. Noticeably, the node-level pose classification method implemented in DBX2 significantly outperformed all docking and rescoring schemes at all the tested RMSD thresholds. 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 indipendently. **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) [73]. Additionally, we compared DBX2 against four ML-based docking methods, Gnina, KarmaDock, CarsiDock, and RTMscore, using a 2 Å cutoff on the hold-out dataset (**Figure S4A**) and Runs N’s Poses dataset (**Figure 2C**). Unsurprisingly, KarmaDock, CarsiDock, and RTMscore outperformed both DBX2 and Gnina on the PDBbind v2019-based hold-out test set, which was part of the PDBbind v2020 general set used to train these models [70], [71], [72]. Nevertheless, DBX2 displayed encouraging performance despite the limited size of the training set (4,057 complexes) compared with the other methods. Next, we performed the same comparison on the Runs N’s Poses dataset, which was completely unseen by all five investigated methods during training. Moreover, we investigated the performance of docking before and after removing the Runs N’s Poses protein families that overlapped with v2016 and v2020. Notably, DBX2 demonstrated superior performance compared to all other models on the Runs N’s Poses dataset, followed by Gnina, both before and after the removal of overlapping protein families (**Figure 2C**). Interestingly, upon overlap removal, the success rates for RTMscore, DBX2, and Gnina experienced a slight increase. In contrast, the success rates for KarmaDock and CarsiDock slightly declined. Moreover, the impact of node count per graph on DBX2 prediction performance was further examined by generating additional graphs from the PDBbind v2016 and the hold-out set with reduced node counts: 70 nodes (30 poses from AutoDock and DOCK, 10 from Vina) and 35 nodes (15 poses from AutoDock and DOCK, 5 from Vina). For each setting, the model was retrained and revaluated on the hold-out test set. The success rate was then compared to the default 140-node configuration. While DBX2 achieved its highest performance with the default

View Article Online
DOI: 10.1039/D4SC07875F

setting, the prediction accuracy did not decline dramatically with fewer nodes (**Figure S4B**). These results suggest that when generating or training with a large number of poses is challenging, DBX2 can still achieve reasonable performance using ensembles of limited size.

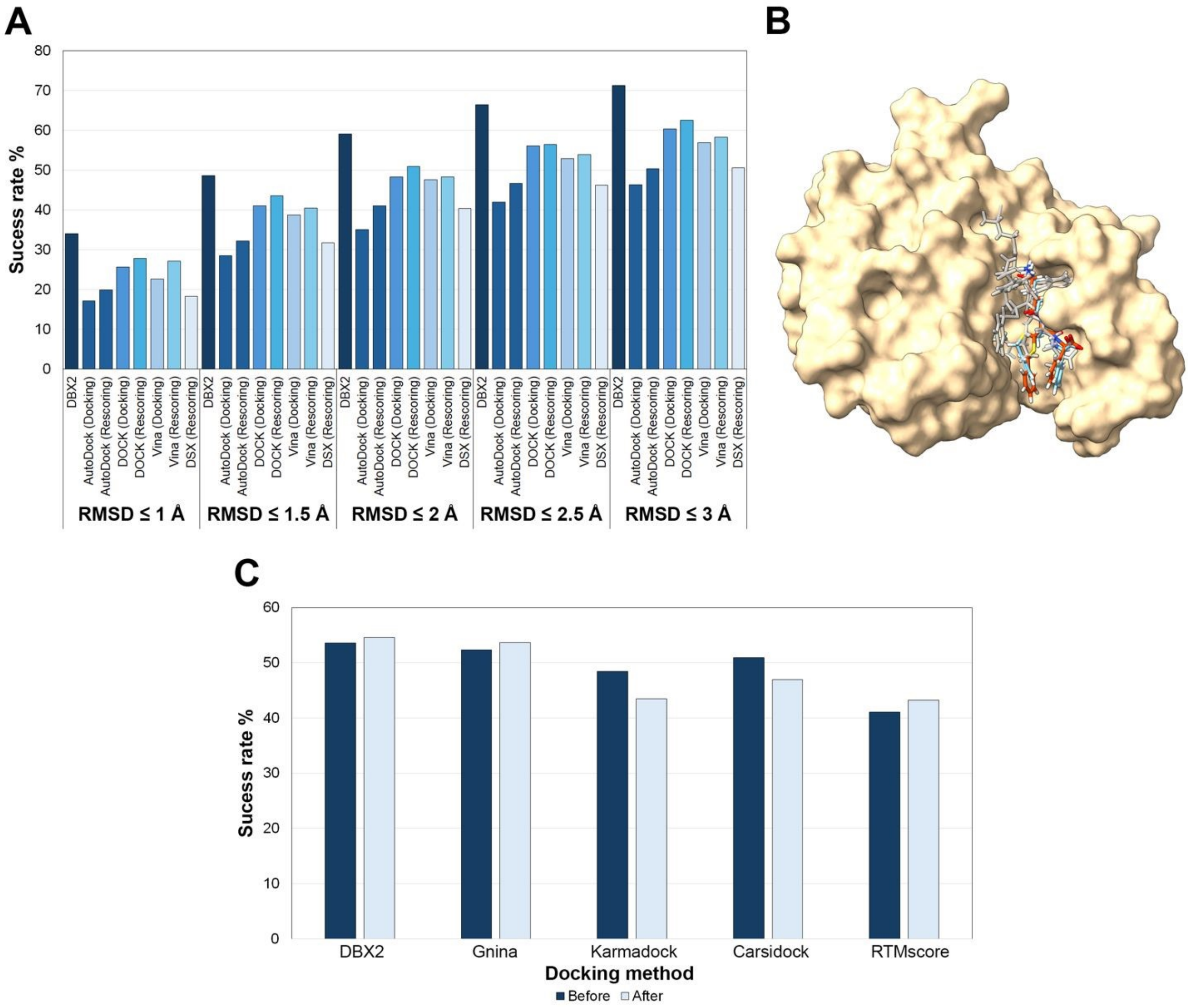

Figure 2: (A) Comparison of success rates of identification of the correct pose on hold-out test set between AutoDock, DOCK, Vina, DSX, and DBX2, comparing docking and rescoring strategies. Rescoring improved the performance of each docking program compared to standard docking, emphasizing the advantage 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

View Article Online
DOI: 10.1039/D4SC07875F

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). (C) Success rate of identification of the pose correctness on Runs N's Poses dataset before (light) and after (dark) removing overlapping protein families with PDBbind v2020 for DBX2, Gnina, KarmaDock, CarsiDock, and RTMscore.

Next, we evaluated the ability of the scoring functions to reproduce experimentally determined binding affinities in the hold-out test set (**Table 2**). Notably, 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. Thus, 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, similarly to our previous work [43]. Thus, linear regression was performed to compare binding affinities from the hold-out dataset with the scores of the best DOCK poses using different scoring functions and their linear combinations [43]. 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 pose ensembles via docking rather than relying on a single pose.

Table 2: *R^2, Pearson correlation coefficients and predictive index values between experimental binding affinities and the scores provided by tested scoring functions. Best values are indicated in bold.*

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
1	KarmaDock	0.03	0.18	-0.79
1	CarsiDock	0.03	0.17	-0.68

View Article Online
DOI: 10.1039/D4SC07875F

1	RTMscore	0.22	0.46	-0.36
1	Gnina CNNAffinity	0.36	**0.61**	0.55
1	Gnina MinimizedAffinity	0.25	0.44	0.18
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

Interestingly, DBX2 exhibited the highest correlation with experimental binding affinities on the hold-out dataset, outperforming other tested 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 [74]. Nevertheless, the R^2 value, while indicating positive correlation as well as an improvement compared with other methods, remained low (0.38), underscoring remaining challenges in accurate thermodynamics predictions via docking-based sampling. Indeed, while our results suggest that docking poses ensembles appear to be 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 necessary to ensure the high throughput required in docking. Additionally, DBX2 also outperforms other ML models (KarmaDock, CarsiDock, and RTMscore) in this task, despite being trained on fewer protein-ligand complexes, highlighting the challenges that these methods may face in VS due to the neglection of experimental affinities in their training [70], [72]. Correlation plots between experimental and computational affinities are shown in **Figure S5**.

Moreover, the DBX2 scoring power on the hold-out set was compared with established 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 [30]

View Article Online
DOI: 10.1039/D4SC07875F

and Pafnucy model from Stepniewska-Dziubinska et al [75]. The first class of models are GNNs mapping protein- (P), ligand- (L) and protein-ligand interactions (I) graph representations to ligand-protein affinities. The Pafnucy model is a 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 used in DBX2 (PDBbind v2016 dataset and the hold-out test set, respectively) as previously reported [30]. The comparison of Pearson coefficient and RMSE is summarized in **Table S4**. Even in this case, DBX2 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. [30] and Pafnucy model [75], as evident from the Pearson coefficient and RMSE values, and comparable performances with GNN models that included protein and ligand structural information explicitly, while being based entirely on energetic representations without taking into account any structural information. 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. [30], while significantly outperforming the success rate of pure interaction models.

Retrospective virtual screening

To test the VS power of DBX2 in realistic scenarios, we focused on the three LIT-PCBA targets that were not present in our training set: FEN1, GBA, and MTORC1. LIT-PCBA is a small molecule bioactivity dataset to mitigate biases and avoid overestimating VS performances. Derived from bioassays, it mimics experimental active and potency distributions within screening libraries, spans diverse protein targets, and has been validated across multiple screening methods, making it suitable for both structure- and ligand-based VS retrospective experiments [42]. 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

View Article Online
DOI: 10.1039/D4SC07875F

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 when compared to other scoring functions, on the three target proteins. Surprisingly, DBX2's node-level predictions, which assess the likelihood of each binding pose to be the correct one within a specific graph, consistently matched the screening power of graph-level predictions of binding affinities. Gnina, a ML-based tool that recently demonstrated state-of-the-art performance in prospective drug discovery challenges [76], [77], and the other ML-based tools (KarmaDock and RTMscore) also performed well, further validating the potential of data-driven models in VS tasks. Additionally, logAUC (**Figure 6D, 6E, 6F**) and BEDROC (**Table S6**) were calculated to further assess each scoring functions' ability to distinguish between active and inactive compounds. DBX2 demonstrates superior performance across both these metrics as well, suggesting a robust efficacy in prioritizing active compounds throughout top and broad ranks of compounds. Node-level predictions showed the highest performance, followed by graph-level predictions, KarmaDock, CarsiDock, and Gnina's CNNAffinity scoring function.

View Article Online
DOI: 10.1039/D4SC07875F

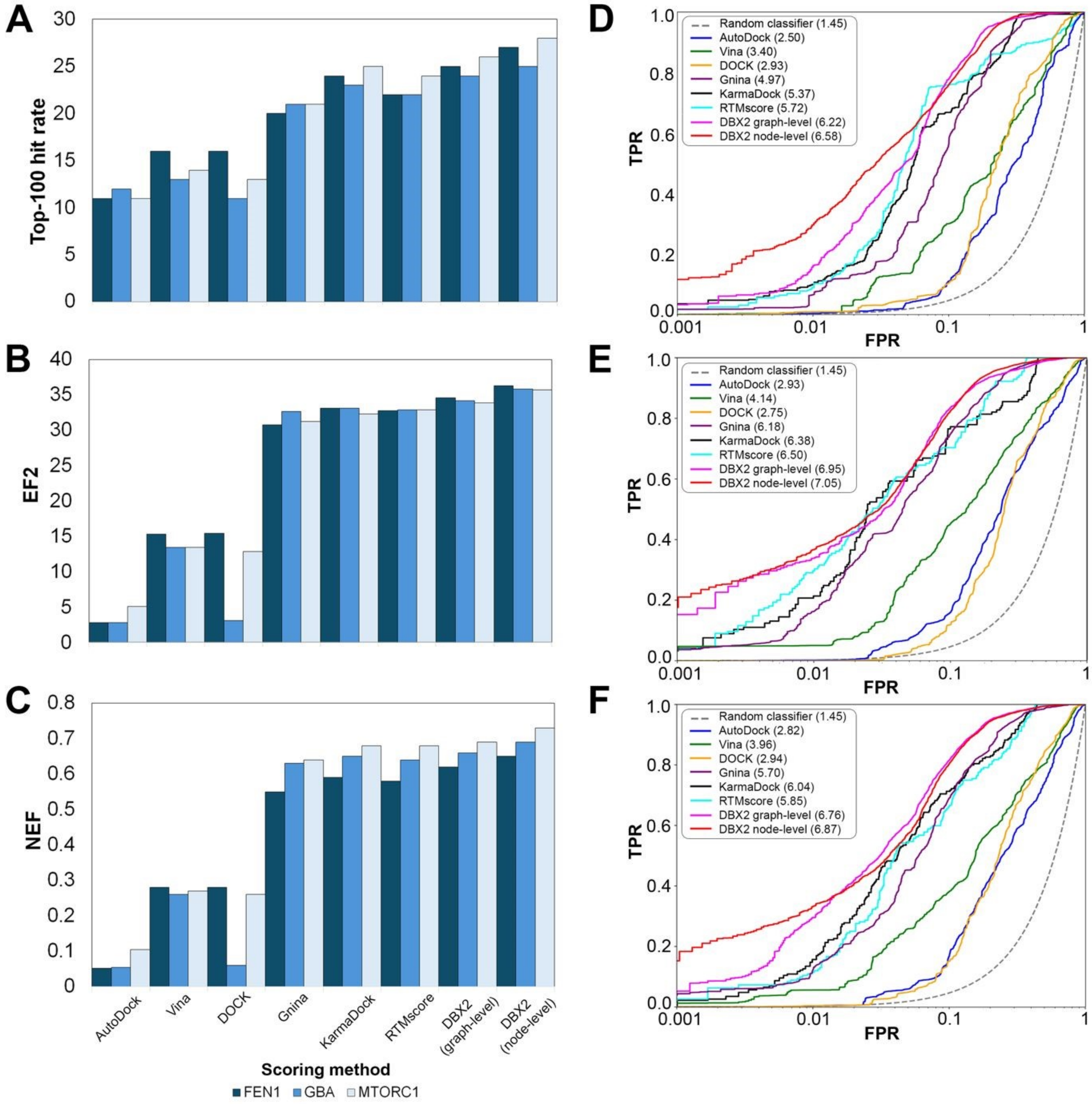

Figure 3: *Retrospective VS results of different scoring functions on three proteins from the LIT-PCBA database, (A) top-100 hit rate (B) EF2 (C) NEF, illustrating the significant performances of DBX2 node- and graph-level scores across different targets. LogAUC plots computed for (D) Flap structure-specific Endonuclease 1 (FEN1), (E) Glucocerebrosidase (GBA), and (F) Mechanistic Target of Rapamycin (MTORC1) confirmed the promising performance of the two DBX2 scores.*

View Article Online
DOI: 10.1039/D4SC07875F

Lastly, since the use of multiple programs may result computationally expensive in large-scale screens, we investigated the effect of DBX2 in enhancing the VS performance of single docking programs, Focusing on FEN1 as the target, we used DBX2 to rescore the top 20,000 Vina-scored molecules from LIT-PCBA, computing top-100 hit rate, EF2 and NEF metrics as well as logAUC before and after the application of DBX2 (**Figure S6, S7**). The results clearly indicated that also in this case, DBX2 significantly improved upon both AutoDock and Vina outcomes across different sets of docking parameters.

Conclusions

We introduced DBX2, a novel GNN framework that enables to represent computational ensembles of small molecule-protein conformations as single graphs to jointly predict binding modes and affinities. The model relies solely on simple energetic features derived directly from docking, thus without requiring additional costly sampling steps. We comprehensively evaluated DBX2 across various metrics for docking and VS tasks, underscoring its effectiveness as a robust tool with superior performances compared to conventional scoring functions and ML models relying on single pose. At the same time, some caveats associated with the newly proposed ensemble-based method emerged, especially reflected in the relatively poor correlation between graph-level prediction 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 estimating the binding energy contributions that are used as features. Nevertheless, the performances observed for DBX2 not only advocate for its adoption in prospective VS campaigns relying on high throughput VS but encourages also further exploration of ML models learning from computationally generated ensembles that can represent the thermodynamics of binding better 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 protein flexibility, induced fit effect, solvation, and overall equilibrium ensembles.

View Article Online
DOI: 10.1039/D4SC07875F

Author contributions

TT: methodology, data curation, investigation, formal analysis, writing - original draft, writing - review & editing. JP: conceptualization, data curation, software, writing - original draft. FG: conceptualization, funding acquisition, supervision, writing - original draft, writing - review & editing.

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) and a uOttawa start-up grant awarded to FG. Computations were performed on the resources of the Digital Research Alliance of Canada (RRG ID 4879 awarded to FG) and the University of Ottawa's Wooki supercomputing cluster. We thank Cadence Molecular Sciences for providing an academic license for Openeye suite.

View Article Online
DOI: 10.1039/D4SC07875F

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, Sept. 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

View Article Online
DOI: 10.1039/D4SC07875F

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, June 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, Sept. 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, Sept. 2016, doi: 10.1038/nature19112.

[20] 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.

[21] 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.

View Article Online
DOI: 10.1039/D4SC07875F

[22] 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, Sept. 2022, doi: 10.1126/science.abn7065.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

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

[29] 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.

[30] 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, June 2022, doi: 10.1021/acs.jmedchem.2c00487.

[31] 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.

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

[33] 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.

View Article Online
DOI: 10.1039/D4SC07875F

[34] 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, June 2004, doi: 10.1021/jm030580l.

[35] 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.

[36] M. Su *et al.*, “Comparative Assessment of Scoring Functions: The CASF-2016 Update,” *J. Chem. Inf. Model.*, vol. 59, no. 2, pp. 895–913, Feb. 2019, doi: 10.1021/acs.jcim.8b00545.

[37] P. Škrinjar, J. Eberhardt, J. Durairaj, and T. Schwede, “Have protein-ligand co-folding methods moved beyond memorisation?,” Feb. 07, 2025, *Bioinformatics*. doi: 10.1101/2025.02.03.636309.

[38] J. Abramson *et al.*, “Accurate structure prediction of biomolecular interactions with AlphaFold 3,” *Nature*, vol. 630, no. 8016, pp. 493–500, June 2024, doi: 10.1038/s41586-024-07487-w.

[39] Chai Discovery *et al.*, “Chai-1: Decoding the molecular interactions of life,” Oct. 11, 2024, *Synthetic Biology*. doi: 10.1101/2024.10.10.615955.

[40] ByteDance AML AI4Science Team *et al.*, “Protenix - Advancing Structure Prediction Through a Comprehensive AlphaFold3 Reproduction,” Jan. 11, 2025, *Bioinformatics*. doi: 10.1101/2025.01.08.631967.

[41] J. Wohlwend *et al.*, “Boltz-1: Democratizing Biomolecular Interaction Modeling,” Nov. 20, 2024, *Biophysics*. doi: 10.1101/2024.11.19.624167.

[42] 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, Sept. 2020, doi: 10.1021/acs.jcim.0c00155.

[43] 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, Sept. 2019, doi: 10.1007/s10822-019-00227-7.

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

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

View Article Online
DOI: 10.1039/D4SC07875F

[46] 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.

[47] 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.

[48] 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, July 2011, doi: 10.1002/jcc.21814.

[49] 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.

[50] 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.

[51] 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%3C1639::AID-JCC10%3E3.0.CO;2-B.

[52] 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.

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

[54] 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

[55] 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

[56] 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.

View Article Online
DOI: 10.1039/D4SC07875F

[57] C. Cortes, M. Mohri, and A. Rostamizadeh, “L2 Regularization for Learning Kernels,” May 09, 2012, *arXiv*: arXiv:1205.2653. Accessed: Oct. 20, 2024. [Online]. Available: http://arxiv.org/abs/1205.2653

[58] F. Pedregosa *et al.*, “Scikit-learn: Machine Learning in Python,” *J Mach Learn Res*, vol. 12, no. null, pp. 2825–2830, Nov. 2011.

[59] 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.

[60] 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.

[61] 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.

[62] 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.

[63] 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.

[64] 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.

[65] 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.

[66] 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.

[67] 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

View Article Online
DOI: 10.1039/D4SC07875F

screening applications with early recovery capability," *J. Cheminformatics*, vol. 9, no. 1, p. 7, Dec. 2017, doi: 10.1186/s13321-016-0189-4.

[68] 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.

[69] 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, Sept. 2001, doi: 10.1021/ci0100144.

[70] X. Zhang *et al.*, "Efficient and accurate large library ligand docking with KarmaDock," *Nat. Comput. Sci.*, vol. 3, no. 9, pp. 789–804, Sept. 2023, doi: 10.1038/s43588-023-00511-5.

[71] C. Shen *et al.*, "Boosting Protein–Ligand Binding Pose Prediction and Virtual Screening Based on Residue–Atom Distance Likelihood Potential and Graph Transformer," *J. Med. Chem.*, vol. 65, no. 15, pp. 10691–10706, Aug. 2022, doi: 10.1021/acs.jmedchem.2c00991.

[72] H. Cai *et al.*, "CarsiDock: a deep learning paradigm for accurate protein–ligand docking and screening based on large-scale pre-training," *Chem. Sci.*, vol. 15, no. 4, pp. 1449–1471, 2024, doi: 10.1039/D3SC05552C.

[73] 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.

[74] 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.

[75] 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.

[76] F. Li *et al.*, "CACHE Challenge #1: targeting the WDR domain of LRRK2, a Parkinson's Disease associated protein," July 18, 2024, *Biochemistry*. doi: 10.1101/2024.07.18.603797.

[77] I. Dunn, S. Pirhadi, Y. Wang, S. Ravindran, C. Concepcion, and D. R. Koes, "CACHE Challenge #1: Docking with GNINA Is All You Need," *J. Chem. Inf. Model.*, vol. 64, no. 24, pp. 9388–9396, Dec. 2024, doi: 10.1021/acs.jcim.4c01429.

View Article Online
DOI: 10.1039/D4SC07875F

View Article Online
DOI: 10.1039/D4SC07875F

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.