

Multi-level, multi-body atomic interaction graphs for machine learning-based prediction of protein-ligand binding energies

Tram TH. Le¹, Huy Vo¹, Nhat HM. Nguyen¹, Binh T. Nguyen^{1*},
Duc Duy Nguyen^{2*}

¹University of Science, Vietnam National University, Ho Chi Minh City,
Vietnam.

²University of Tennessee, Knoxville, Tennessee, USA.

*Corresponding author(s). E-mail(s): ngtbinh@hcmus.edu.vn;
ducnguyen@utk.edu;

Contributing authors: htrample.0310@gmail.com;
vghuy23@apcs.fitus.edu.vn; 24280090@student.hcmus.edu.vn;

Abstract

Accurate prediction of binding affinity is crucial for rational drug design and discovery. Traditional computational methods often rely on complex scoring functions that incorporate a multitude of physical and chemical descriptors, leading to high computational demands and sometimes limited generalizability. In this work, we propose a novel scoring function that models multi-level, multi-body atomic interactions using graph-based representations. Our method constructs comprehensive interaction graphs that incorporate both pairwise and triplet-wise atomic features that help capture cooperative spatial patterns essential for binding affinity prediction. By employing a feature fusion strategy, GMI-Score maintains model simplicity while enhancing accuracy. Extensive evaluation across multiple datasets, such as PDBbind v2013, PDBbind v2016, PDBbind v2020, CSAR-NRC-HiQ, and PDBbind-Redocked, demonstrates that our model consistently outperforms state-of-the-art scoring functions, achieving Pearson correlation coefficients up to 0.877. Furthermore, it retains strong predictive power under strict data leakage controls and realistic docking conditions to highlight its robustness and generalizability.

Scientific Contribution: In this study, we present a scoring methodology that systematically captures higher-order atomic interactions within a unified graph framework, making a conceptual shift in cheminformatics scoring functions.

Its consistent outperformances of existing methods and strong validity under redocked and withheld data scenarios demonstrate its utility for broad-scale molecular modeling applications and open cheminformatics workflows.

Keywords: Machine learning, Binding affinity, Drug discovery, Multi-body interactions, Graph-based learning, Virtual screening, Feature fusion

1 Introduction

Protein-ligand interactions are fundamental to nearly all biological processes, such as enzyme catalysis, signal transduction, and molecular recognition [3, 5, 16, 67]. These interactions typically depend on non-covalent forces, including hydrogen bonds, ionic interactions, hydrophobic effects, and van der Waals forces. These interactions provide critical insights into systems biology, revealing mechanisms such as signaling pathways [8], allosteric regulation [9], and feedback control that underlie complex cellular processes [17].

Given their central role in cellular function, quantifying protein-ligand interactions becomes essential, particularly in fields like drug discovery, where precise control over these interactions can determine therapeutic efficacy. The strength of these interactions is typically expressed through binding affinity, often quantified by the equilibrium dissociation constant K_d [2]. A lower K_d value indicates a higher affinity between the protein and ligand and vice versa. Accurate binding affinity predictions can be utilized in virtual screening, lead optimization, and cost-effective *de novo* drug design [53]. Consequently, developing robust computational methods to predict binding affinities has become a key objective in modern pharmacological research [57].

Traditionally, experimental methods such as isothermal titration calorimetry (ITC) or surface plasmon resonance (SPR) are used to determine protein-ligand binding affinity [12, 46]. However, these methods are usually expensive and time-consuming, making them unsuitable for large-scale drug screening [71]. As a result, there is a need for computational models that can effectively and accurately predict binding. In response, recent technological advances have led to the development of a variety of computational approaches for studying protein-ligand interactions. These include structure-based methods such as molecular docking, standalone scoring functions designed to predict binding affinity, and physics-based techniques like molecular dynamics simulations that analyze the stability and conformational dynamics of protein-ligand complexes [1].

In computational drug design, scoring functions are mathematical models used to estimate the protein-ligand binding affinity. They are commonly categorized into four types: physics-based, empirical, knowledge-based, and machine learning-based. Physics-based scoring functions focus on intermolecular interactions, such as van der Waals forces and electrostatic attractions [21, 30]. Empirical scoring functions use energy terms derived from molecular interactions and some empirical chemistry factors like hydrophobicity and metal-ligand interactions [47, 49]. Knowledge-based scoring functions rely on statistical analysis of known protein-ligand complexes to identify

favorable interaction patterns [60]. With the growth of data, machine learning-based functions are gaining much more attention. These methods learn patterns from large training data and have been shown to outperform traditional methods in binding affinity prediction [4, 22, 29, 44, 51, 61, 74]. Despite extensive research throughout the years, protein-ligand interactions are still far from fully understood. As the binding processes are influenced by many factors ranging from spatial arrangement to interactions between the molecules, feature selection remains a difficult task in machine learning-based methods [66]. Many existing works focus on the interactions between atom pairs in the protein-ligand complex. RF-score utilizes a feature vector of occurrence-specific atom type pairs within a cutoff distance [4]. ECIF extends this by considering atomic connectivity, valence, and hybridization [51]. At a more detailed level, GGL-ETA-Score constructs subgraphs based on atom types and computes features through summation of the generalized Lorentz or Exponential function over atom pairs [50]. A similar approach by Fangqiang Zhu et al. also uses feature vectors consisting of the summation of a neural network over atom pairs [74]. However, we hypothesize that pairwise interactions may not fully capture the characteristics of binding processes. We believe that multi-level interactions can provide a better representation for the protein-ligand complex and enhance the performance of machine learning-based scoring function models.

Models such as DimeNet [19], GemNet [20], and PaiNN [52] incorporate angular and dihedral information over bonded molecular graphs to improve expressiveness. However, these models were designed and benchmarked primarily on covalently bonded small-molecule systems, where connectivity is chemically prescribed and training datasets are relatively modest in scale. In contrast, the protein-ligand scoring problem presents a fundamentally different regime: interaction graphs are defined by spatial distance cutoffs rather than covalent bonds, labeled data is available at large scale, and inference efficiency is critical for practical screening workflows. Direct application of these architectures is therefore computationally prohibitive, as models such as GemNet-OC require quadratic or cubic operations through multi-hop directional message passing and process orders of magnitude fewer samples per GPU-second than simpler models at inference time.

In this work, we present GMI-Score (Graph geometric learning with Multi-body Interactions based scoring functions), a novel framework for modeling protein-ligand binding affinity. While prior models such as RI-Score and GGL-Score [44, 50] rely on graphs constructed from pairwise atomic interactions, GMI-Score introduces a more expressive representation by constructing multi-body interaction graphs that capture higher-order geometric motifs, including triplet-wise interactions. In addition to expanding the graph structure, GMI-Score incorporates a multi-level interaction fusion mechanism that jointly learns from different interaction orders. This design enables the model to capture cooperative spatial patterns that are often overlooked by pairwise methods. By learning from both fine-grained and collective structural features, GMI-Score provides a more faithful representation of the binding interface. This design consciously trades the full expressive power of equivariant architectures for practical scalability, providing the geometric inductive bias necessary to surpass pairwise

methods without incurring the computational overhead that makes geometric GNNs impractical at protein-ligand screening scale.

Evaluated across multiple standard and challenging benchmarks, including CASF-2013 [37], CASF-2016 [59], PDBbind v2020, CSAR-NRC-HiQ [13], and PDBbind-Redocked [54], our method achieves better results than other state-of-the-art models. The strong performance and efficient design of GMI-Score underscore its potential as a powerful tool for accelerating drug discovery in both academic and industrial settings.

2 Results and Discussion

In this section, we present the scoring power of our proposed multi-level model, GMI-Score, for the benchmark datasets.

2.1 Datasets

Table 1: Summary datasets used in our work.

Datasets	Training sets		Test set
	General set	Refined set	
PDBbind v2013	-	-	195
PDBbind v2016	12998	3772	285
PDBbind v2020	23496	5316	285
CSAR-NRC-HiQ	-	-	49
PDBbind-Redocked	-	-	285

To evaluate the predictive power of our proposed method, we utilize the PDBbind v2013, the PDBbind v2016, and the PDBbind v2020 datasets, which are standard datasets for the binding affinity prediction task. The PDBbind datasets are divided into three nested subsets: the general set, the refined set, and the core set, with each subset being contained within the previous one. The general and refined sets, after excluding the core set, are used as training data for our model, whereas the core set is used as the testing set. Despite being a well-acknowledged dataset, several studies have demonstrated concerns regarding the reliability of the PDBbind datasets as a benchmark for scoring functions. The protein-ligand complexes in the subsets are not consistent in terms of quality. Specifically, the core set contains protein-ligand complexes with higher resolution and binding affinity compared to complexes in the general and refined sets [38].

In order to provide a more comprehensive assessment of our method’s predictive ability, we conducted further evaluations on two external test sets: the CSAR-NRC-HiQ and PDBbind-Redocked datasets.

The CSAR-NRC-HiQ dataset [13] is a valuable resource for validating molecular docking and scoring models. It originally contains a total of 343 high-quality protein-ligand complexes, which are divided into two distinct subsets of 167 and 176 complexes.

To create an independent test set, we combined both subsets and then removed any complexes that overlapped with the PDBbind database. This filtering process resulted in a final, non-overlapping set of 49 complexes, which we used for external validation of our model’s performance.

The PDBbind-Redocked [54] is used for evaluating a scoring function’s ability to perform in real-world scenarios, where co-crystal structures aren’t available. Instead of using the experimentally determined poses from the PDBbind v2016 dataset, this dataset uses redocked structures generated by Surflex-Dock [24]. To assess a scoring function’s performance on this dataset, all complexes that are part of the CASF-2016 set are used as a test set.

All the datasets are shown in Table 1.

2.2 Hyperparameters optimization

To achieve the optimal hyperparameters for the binding affinity prediction task, we followed the hyperparameter settings from our previous work [50] for the pairwise interaction model, i.e., GMI-Score-2. We then optimized the GMI-Score-3 scoring function by evaluating its performance across a range of hyperparameters (t , k) and cutoff distances (c). We also hyperparameter-tune the model with grid search to identify optimal hyperparameters for our chosen machine learning models, aiming to maximize the correlation between predicted and experimental binding affinities. To balance thoroughness with computational feasibility, we performed GridSearchCV on the first training dataset (PDBBind v2016) to identify optimal hyperparameter configurations, and then applied these same configurations across all other training-test scenarios. This approach follows established practice in machine learning, where hyperparameters optimized on one representative dataset are transferred to related tasks to avoid overfitting to any specific dataset and to reduce computational overhead. The resulting best-performing models for each dataset served as the foundation for further analysis. The performances of various Gradient Boosting models on the CASF-2016 [59] benchmark trained on general set are presented in Table 2 and Table A1.

Table 2: Hyperparameter selection for the Gradient Boosting model of GMI-Score-3 on the CASF-2016 dataset. RMSE is in kcal/mol units.

Model	Train with Refined		Train with General	
	R_p	$RMSE$	R_p	$RMSE$
Model 1	0.836 ± 0.002	1.742 ± 0.004	0.860 ± 0.002	1.638 ± 0.007
Model 2	0.836 ± 0.003	1.703 ± 0.003	0.857 ± 0.002	1.623 ± 0.007
Model 3	0.824 ± 0.006	1.734 ± 0.021	0.850 ± 0.006	1.630 ± 0.025
Model 4	0.836 ± 0.002	1.709 ± 0.007	0.857 ± 0.001	1.626 ± 0.002

To explore the potential for improved predictive power, we investigated the fusion of features derived from GMI-Score-3 and GMI-Score-2. Integrated feature sets were

constructed by combining these two scoring functions at various ratios. For each benchmark dataset and each feature ratio, the optimized machine learning model was trained and evaluated with the five independent runs using various random seeds to ensure robust performance assessment.

2.3 PDBbind v2016 Benchmark

Our findings indicate that the GMI-Score-2, through its atom-specific feature representation, and the GMI-Score-3, with its atom-wise approach, capture different facets of the complex binding affinity between proteins and ligands. While each scoring function possesses inherent strengths, our analysis suggested that not all constituent features contribute equally to the predictive accuracy; indeed, some may have a negligible or even detrimental impact. This observation motivated the development of a fusion strategy focused on integrating the most relevant variables from both scoring paradigms. The primary objectives of this fusion were twofold: firstly, to achieve a more lightweight model by reducing the total number of input features, which in turn lessens computational cost and can enhance model interpretability; and secondly, to elevate the predictive performance beyond what is achievable by either scoring function in isolation.

Utilizing the PDBbind v2016 dataset for evaluation, models were trained on a combination of their general sets, and their baseline performances were subsequently tested on the core set. We first established baseline performances for the individual scoring functions, obtaining Pearson correlation coefficients of 0.870 ± 0.001 for the GMI-score-2 and 0.863 ± 0.002 for the GMI-Score-3 (Table 3). Specifically, the optimized fusion model achieved a Pearson correlation coefficient of 0.877 ± 0.001 . Paired t-test have been applied to see if the difference of nearly 0.02 of R_p meaningful between GMI-Score and the two individual models (Table A2). The outcomes of these experiments were encouraging and demonstrated that the resulting hybrid models statistically significantly better than both individual models. This enhanced performance, together with the advantages of feature reduction, validates the effectiveness of our feature fusion approach. The fusion model GMI-Score (fusion 7000–4000) is used for all evaluations; hereafter, it is referred to as GMI-Score.

Table 3: Performance comparison of individual (GMI-Score-2 and GMI-Score-3) and fusion models on the CASF-2016 dataset. The table reports the Pearson correlation coefficient (R_p) and RMSE in kcal/mol.

Model	Train with General	
	R_p	$RMSE$
GMI-Score-2	0.870 ± 0.001	1.582 ± 0.002
GMI-Score-3	0.863 ± 0.002	1.620 ± 0.009
GMI-Score (fusion 7000-4000)	0.877 ± 0.001	1.552 ± 0.005
GMI-Score (fusion 6000-5000)	0.877 ± 0.002	1.550 ± 0.007
GMI-Score (fusion 6000-6000)	0.877 ± 0.001	1.552 ± 0.005

“fusion a - b ” denotes a model constructed using a GMI-Score-3 features and b GMI-Score-2 features.

When benchmarked against other state-of-the-art models on the CASF v2016 dataset, our proposed GMI-Score performed competitively (Figure 1). We believe this improvement may be attributed to the synergistic combination of the GMI-Score-2’s capacity to capture overarching protein-ligand relationships and the GMI-Score-3’s detailed atom-wise insights. Our approach attempts to distill these strengths into a concise and computationally efficient feature set. Importantly, this fusion strategy appears to do more than simply preserve the predictive power of the individual components; it also seems to help mitigate their respective limitations. As a result, the proposed method is more robust and accurate for determining protein-ligand binding affinity than previous ones [4, 36, 50].

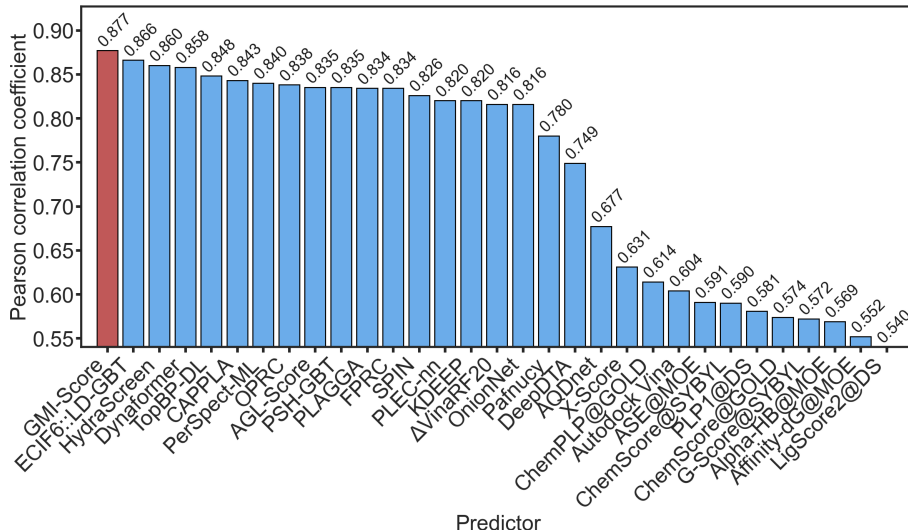


Fig. 1: Comparison of Pearson’s correlation scores of various scoring functions on the CASF-2016 benchmarks [11, 27, 33, 39, 48, 50, 56]. Our proposed *GMI-Score* is shown in red, while other methods are shown in blue.

2.4 PDBbind v2020 Benchmark

We extended its application to the PDBbind v2020 dataset in order to validate the generalizability of our GMI-Score. This dataset is a significantly expanded version of PDBbind v2016, encompassing 19,443 protein-ligand complexes. Notably, it retains the same core set as PDBbind v2016, facilitating a direct comparison of performance and allowing us to assess the model’s ability to handle a larger and more diverse dataset.

Adhering to the established PDBbind evaluation protocol, we trained the GMI-Score framework on general and refined sets of PDBbind v2020. To mitigate computational costs associated with extensive parameter tuning on this larger dataset, we transferred the optimized parameters obtained from our experiments with PDBbind v2016. This approach assumes a degree of parameter generalizability across the

two datasets, which is reasonable given the overlap in their core sets and the underlying physicochemical principles governing protein-ligand binding. Although fine-tuning the GMI-Score model on the v2020 dataset could potentially yield better performance, we chose to omit this step to highlight the feasibility and transferability of our method.

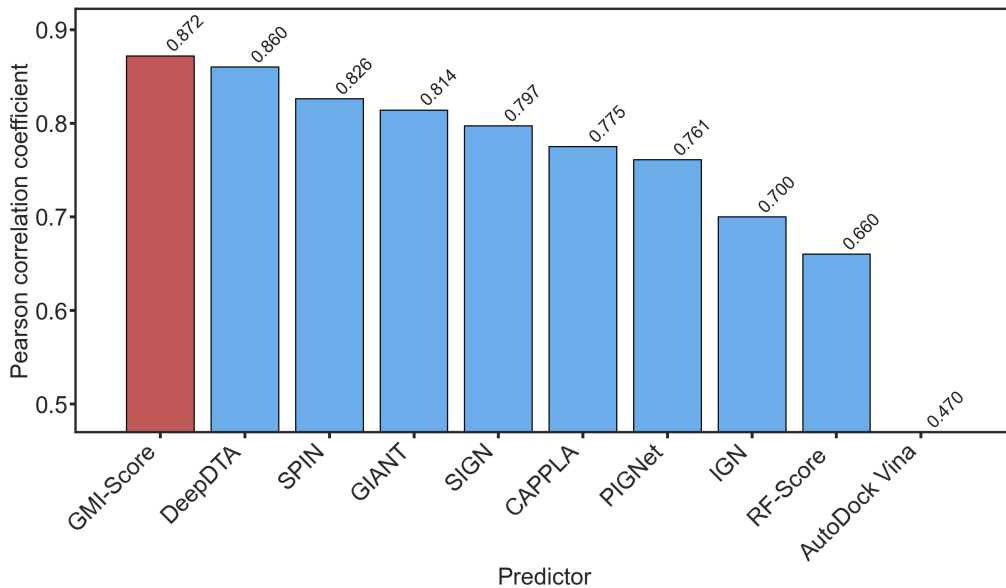


Fig. 2: Comparison of Pearson’s correlation of various scoring functions trained on the PDBbind v2020 general set [4, 11, 25, 27, 33–35]. Our proposed *GMI-Score* is shown in red, while other methods are shown in blue. The performances of other models are taken from [11]. Only models achieving high R_p are reported.

The trained GMI-Score was then tested on the PDBbind v2020 core set. As reported in Table A3, our method achieved a Pearson correlation coefficient (R_p) of 0.872 ± 0.001 . A strong linear correlation between predicted and experimental binding energies is further illustrated in the right panel of Fig. B1, which shows that most data points lie close to the diagonal line with RMSE=1.555 kcal/mol. The p-value of the t-test between the individual model and GMI-Score showed significance for both models trained with the refined and general sets (Table B4). This result is particularly noteworthy for several reasons. First, the high correlation coefficient indicates a strong linear relationship between the GMI-Score predictions and the experimentally determined binding affinities, demonstrating the model’s predictive accuracy. Second, when compared to the performance of separate scoring functions with fusion model on the same benchmark (0.862 ± 0.002 and 0.863 ± 0.002 , respectively), this result underscores the generalization capability of our GMI-Score. The ability to maintain a high level of accuracy on a significantly larger and more diverse dataset, without extensive

re-parameterization, suggests that GMI-Score has effectively captured the fundamental determinants of protein-ligand binding affinity. This enhanced performance is a critical advantage for practical applications in drug discovery, where models are often applied to novel protein-ligand systems not seen during training.

Further comparison with existing scoring functions is shown in Fig. 2, where GMI-Score achieves the highest Pearson correlation coefficient among all evaluated models. DeepDTA [45] and SPIN [11] show lower correlations (0.860 and 0.826, respectively), which demonstrates that GMI-Score provides better predictive accuracy and robustness on the PDBbind v2020 dataset over other state-of-the-art models. This pretrained model on this dataset is used for further evaluation on external test datasets and community use.

2.5 PDBbind v2013 Benchmark

We further evaluated the generalization ability of GMI-Score through a cross-dataset validation strategy. The model was trained separately on the PDBbind v2020 general set and the PDBbind v2013 refined set, and then tested on the CASF-2013 core set. This benchmark comprises 195 high-quality protein-ligand complexes that are structurally diverse and span a broad range of binding affinities. Since none of these complexes overlap with the training sets, the CASF-2013 set provides an independent and unbiased assessment of model performance. The comparative results are presented in Figure 3.

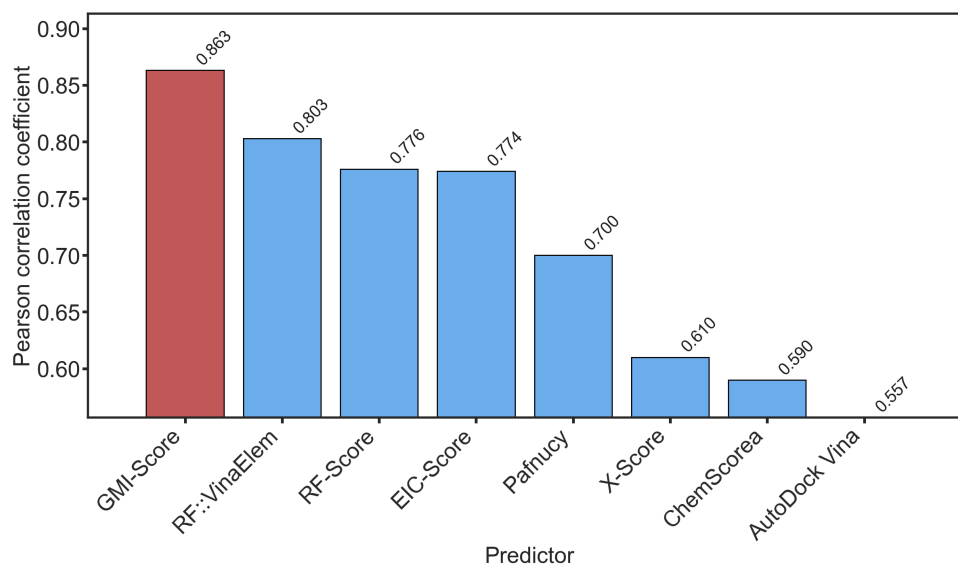


Fig. 3: Comparison of Pearson’s correlation scores of various scoring functions of CASF-2013 set. Our proposed *GMI-Score* is shown in red, while other methods are shown in blue. The performances of other models are taken from [32, 43, 64, 72].

The GMI-Score achieves strong predictive performance on the CASF-2013 benchmark. Our method yields a higher correlation of 0.863 ± 0.001 . This result demonstrates that GMI-Score maintains consistent accuracy across datasets of different sizes and compositions, which indicates its robust prediction to unseen protein-ligand systems.

2.6 CSAR-NRC-HiQ dataset

To further assess the generalizability of our GMI-Score, we conducted an external validation using a curated subset of the CSAR-HiQ dataset [13]. This dataset was filtered to remove any complexes that overlap with the PDBbind v2020 dataset (both general, refined and core sets), ensuring that our test set was truly independent. The final set comprised 49 unique complexes, which served as our independent validation set to evaluate the predictive power of the proposed model.

Predictions on the CSAR-NRC-HiQ dataset with the model trained on the PDBbind v2020 general set show excellent performance. Our model achieved a Pearson’s correlation coefficient of 0.796 ± 0.016 , with an associated *RMSE* of 2.321 ± 0.036 kcal/mol. It is worth noting that while GMI-Score achieves the highest R_p among all compared models, its *RMSE* (2.321 kcal/mol) is slightly higher than that of EGNA (2.094 kcal/mol). This seemingly contradictory result can be explained by the fact that R_p and *RMSE* capture complementary aspects of model performance. R_p reflects the model’s ability to correctly rank the relative binding affinities across complexes, whereas *RMSE* quantifies the absolute magnitude of prediction errors. A model can achieve strong ranking performance while exhibiting a degree of systematic offset or scale difference in its absolute predictions, particularly when optimization is targeted toward R_p as the primary metric. Furthermore, given the small size of the CSAR-NRC-HiQ set (49 complexes), *RMSE* values are more susceptible to the influence of individual outlier complexes. Taken together, the higher of GMI-Score suggests that it may provide a more reliable ordering of ligand binding strength, which is often the more practically relevant criterion in early-stage drug discovery workflows.

Table 4: Pearson correlation with predictions on CSAR-NRC-HiQ dataset. RMSE is in kcal/mol units. The performances of other models are taken from [68].

Model	R_p	RMSE
GMI-Score	0.796	2.321
EGNA [68]	0.750	2.094
KDEEP [26]	0.710	2.222
BAPA [53]	0.669	2.318
OnionNet [73]	0.661	2.337
Pafnucy [58]	0.620	2.591
RF-Score v3 [63]	0.532	2.613

all p-values from the t-test are < 0.01

2.7 PDBbind-Redocked

To further assess the robustness of our framework under realistic drug-design conditions, we used the PDBbind-Redocked dataset [54] for this study. Unlike crystal-structure-based evaluations, this dataset was created by re-docking ligands from the PDBbind v2016 refined and core sets with Surflex-Dock [24], which generated up to 20 poses for each complex and selected the pose with the lowest RMSD among the top-scoring candidates. The redocking process imitates practical virtual-screening situations where experimental 3D structures are unavailable, providing a stringent and realistic test of a scoring function’s ability to generalize beyond idealized crystal data.

Table 5: Performance comparison on the PDBbind-Redocked dataset.

Model	R_p
GMI-Score	0.829
GatedGCN_1.0 [7, 15]	0.810
GT_1.0 [14]	0.800
Δ Lin_F9XGB [70]	0.783
Pafnucy [58]	0.587
Autodock Vina [62]	0.521

The performances of other models are taken from [55].

In this setup, the refined set (3,767 complexes) served as the training data, while the core set (285 complexes) was reserved as an independent test set. As shown in Table 5, GMI-Score reached the highest Pearson correlation coefficient ($R_p = 0.829$) among all compared models and shows strong consistency with experimental binding affinities. GatedGCN_1.0 had slightly a lower correlation ($R_p = 0.810$). The relatively small gap between GMI-Score and GatedGCN_1.0 suggests that both models capture meaningful interaction patterns even from computationally generated poses. AutoDock Vina showed the weakest performance ($R_p = 0.521$). The considerably weaker performance of AutoDock Vina is consistent with its well-documented limitations in absolute affinity estimation [62], and aligns with findings from other external validation studies showing that physics-based scoring functions tend to struggle more than machine learning-based approaches when applied to non-native conformations [59, 70]. Overall, the high correlation of GMI-Score demonstrates that it provides a more reliable rank order of ligand binding strength, even for docking-generated conformations. This result highlights the model’s potential in real-world drug-design applications where crystal structures are often unavailable.

3 Methods

3.1 Multi-geometric subgraphs

Understanding the complex interactions within biomolecules often begins by representing them as graphs. Existing works have experimented on a variety of schemes to classify the atomic interactions. Early approaches [41, 44] classified the interactions

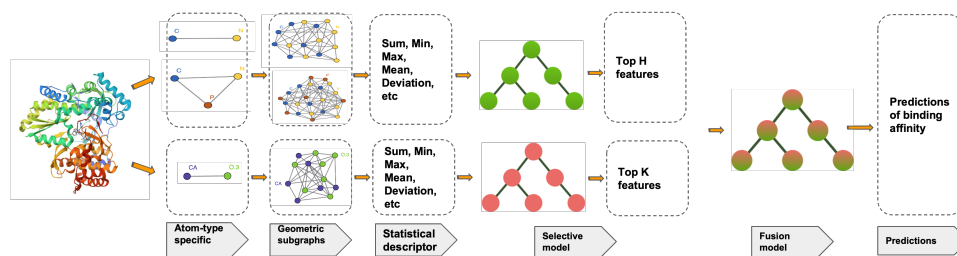


Fig. 4: Illustration of GMI strategy using 2GNW complex. In the second and third columns, element-specific clusters of GMI-Score-3 and GMI-Score-2 are presented, respectively, from top to bottom. The statistical descriptors (fourth column) are selected by a feature selection method using a machine learning model, then a fusion technique is applied before training and predicting with a gradient boosting trees model.

based on element symbols of the protein-ligand atoms, leading to coarse interaction types such as C-N or C-O. In more recent work [50], two additional schemes have been utilized to classify atomic interactions in protein-ligand complexes. In the GGL-Score scheme[50], protein atoms (excluding hydrogens) are identified by their atom names, and ligand atoms are described using SYBYL atom types, resulting in more detailed interaction pairs such as CD1-C.2, CG-C.ar, or the second scheme considers interaction between heavy atoms of each complex using a geometric subgraph of many clusters of pairwise and triplet-wise bonding space. Let \mathcal{T} be the set of all interested atom types of any of the schemes above.

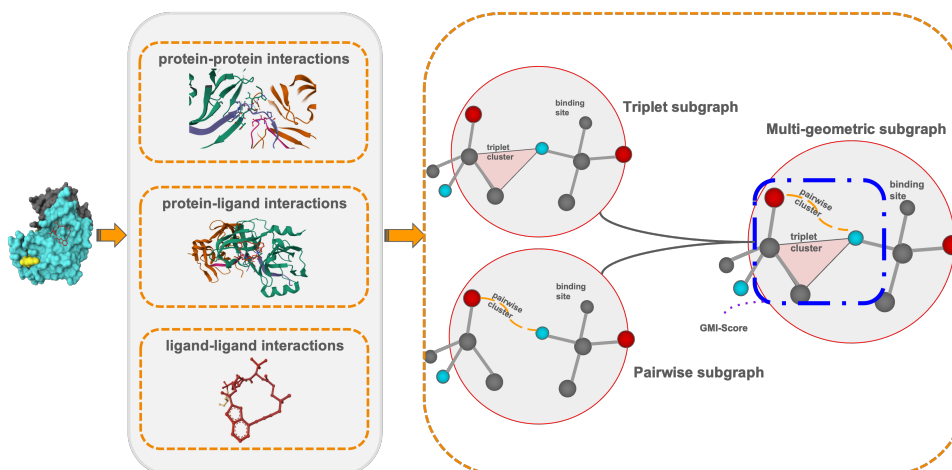


Fig. 5: Multi-geometrics subgraphs (multibody subgraphs) are the combination of triplet subgraphs and pairwise subgraphs.

3.1.1 Complete weighted molecular graph

We represent a protein-ligand complex as a complete weighted molecular graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, w)$. Each vertex corresponds to an atom in the complex. Let the entity labels be $\mathcal{S} = \{\text{P}, \text{L}\}$, denoting protein (P) and ligand (L) atoms. Each entity is associated with its own set of chemical atom types

$$\mathcal{T}_{\text{P}} = \{\mathcal{T}_1^{\text{P}}, \dots, \mathcal{T}_{M_{\text{P}}}^{\text{P}}\}, \quad \mathcal{T}_{\text{L}} = \{\mathcal{T}_1^{\text{L}}, \dots, \mathcal{T}_{M_{\text{L}}}^{\text{L}}\}, \quad (1)$$

where M_{P} and M_{L} denote the numbers of distinct atom types in the protein and ligand, respectively. The augmented type set is the disjoint union

$$\tilde{\mathcal{T}} = \mathcal{T}_{\text{P}} \sqcup \mathcal{T}_{\text{L}}, \quad (2)$$

so that each atom is uniquely characterized by both its chemical type and its entity label.

The vertex set is

$$\mathcal{V} = \{v_i = (\mathbf{r}_i, \alpha_i, r_i) \mid i = 1, \dots, N\}, \quad (3)$$

where N is the number of considered atoms in the protein-ligand complex, $\mathbf{r}_i \in \mathbb{R}^3$ is the coordinate of atom i , $\alpha_i \in \tilde{\mathcal{T}}$ is its augmented type, and r_i is its van der Waals radius. The edge set includes all unordered pairs of distinct atoms:

$$\mathcal{E} = \{(v_i, v_j) \mid i \neq j, i, j \in \{1, \dots, N\}\}. \quad (4)$$

Thus, the molecular graph is complete, and every atom pair is connected. Each edge (v_i, v_j) is assigned a type-dependent weight

$$w(v_i, v_j) = \Phi(\|\mathbf{r}_i - \mathbf{r}_j\|; \eta_{\alpha_i \alpha_j}), \quad (5)$$

where Φ is a kernel function parameterized by a characteristic distance $\eta_{\alpha_i \alpha_j}$, we place the following conditions on the weight function Φ as previous work [44]

$$\Phi(d; \eta) \rightarrow 1 \quad \text{as } d \rightarrow 0, \quad \Phi(d; \eta) \rightarrow 0 \quad \text{as } d \rightarrow \infty. \quad (6)$$

Evidence from previous works [44, 50] have shown that the generalized exponential function:

$$\Phi_E(d; \eta) = e^{-(d/\eta)^\kappa}, \quad \kappa > 0, \quad (7)$$

and the generalized Lorentz function

$$\Phi_L(d; \eta) = \frac{1}{1 + (d/\eta)^\kappa}, \quad \kappa > 0 \quad (8)$$

are good candidates for the non-covalent interactions.

This complete weighted molecular graph captures the full atomistic structure of the protein-ligand complex and serves as the foundation for defining type-specific subgraphs and multi-body interaction scores.

3.1.2 Multi-body type-specific subgraphs

While the complete weighted molecular graph \mathcal{G} captures all pairwise interactions in the protein-ligand complex, practical analysis often requires focusing on subsets of atom types to isolate physically meaningful multi-body interactions.

Let

$$\boldsymbol{\alpha}^l = (\alpha_{i_1}, \alpha_{i_2}, \dots, \alpha_{i_l}) \subset \tilde{\mathcal{T}}, \quad (9)$$

denote an ordered collection of l augmented atom types drawn from the combined set $\tilde{\mathcal{T}}$. The vertex set associated with $\boldsymbol{\alpha}^l$ is

$$\mathcal{V}_{\boldsymbol{\alpha}^l} = \bigcup_{m=1}^l \{v_j \in \mathcal{V} : \alpha_j = \alpha_{i_m}\}, \quad (10)$$

and the edge set is

$$\mathcal{E}^{(\boldsymbol{\alpha}^l)} = \{(v_i, v_j) \in \mathcal{E} : \alpha_i, \alpha_j \in \boldsymbol{\alpha}^l, r_i + r_j + \sigma \leq \|\mathbf{r}_i - \mathbf{r}_j\| \leq c_{\alpha_i \alpha_j}\}, \quad (11)$$

where $c_{\alpha_i \alpha_j}$ is a type-pair-specific cutoff distance. This parameter restricts edges to atom pairs that are spatially close enough to form chemically relevant interactions. The cutoff may be chosen as a fixed constant for each type pair or adaptively defined from empirical distributions of interaction distances. σ is the mean value of the standard deviation of the radii in the dataset.

If one is interested in considering only the intermolecular interactions between protein and ligand, the edge set will be defined as follows

$$\mathcal{E}^{(\boldsymbol{\alpha}^l)} = \{(v_i, v_j) \in \mathcal{E} : \alpha_i \in \boldsymbol{\alpha}^l \cap \mathcal{T}_P, \alpha_j \in \boldsymbol{\alpha}^l \cap \mathcal{T}_L, r_i + r_j + \sigma \leq \|\mathbf{r}_i - \mathbf{r}_j\| \leq c_{\alpha_i \alpha_j}\}, \quad (12)$$

The resulting multi-body type-specific subgraph for l atom types is

$$\mathcal{G}^{(\boldsymbol{\alpha}^l)} = (\mathcal{V}_{\boldsymbol{\alpha}^l}, \mathcal{E}^{(\boldsymbol{\alpha}^l)}, w). \quad (13)$$

This framework accommodates arbitrary selections of protein and ligand atom types, giving rise to the construction of subgraphs that capture intermolecular interactions (protein-ligand), intramolecular interactions (protein-protein, ligand-ligand), as well as higher-order cooperative subgraphs involving multiple atom types simultaneously. When $l = 2$, we obtain the pairwise atom type-based models as mentioned in our previous work [44, 50].

3.1.3 Multi-body interaction scores

The construction of multi-body type-specific subgraphs provides a natural framework for quantifying cooperative interactions among selected atom types in the protein-ligand complex. While pairwise kernels $w(v_i, v_j)$ capture local atomic affinities, they describe only isolated atom-atom contacts and do not scale naturally to reflect the collective behavior of larger subgraphs. Thus, we propose multi-body interaction scores,

which aggregate edge weights across l -atom types subgraphs to measure the strength of cooperative interactions within a chosen set of atom types. We have two different kinds of the multi-body (MB) scores:

Cyclic interaction Score:

For an ordered l -tuple $\alpha^l = (\alpha_{i_1}, \alpha_{i_2}, \dots, \alpha_{i_l})$ with cyclic closure $\alpha_{i_{l+1}} = \alpha_{i_1}$, the cyclic score is

$$\mu_{\text{MB,cyc}}^{(\alpha^l)} = \sum_{p=1}^l \sum_{i:\alpha_i=\alpha_{i_p}} \sum_{j:\alpha_j=\alpha_{i_{p+1}}} \Phi(\|\mathbf{r}_i - \mathbf{r}_j\|; \eta_{\alpha_{i_p}\alpha_{i_{p+1}}}) \mathbf{1}\left((v_i, v_j) \in \mathcal{E}(\alpha^l)\right). \quad (14)$$

Clique interaction score:

Alternatively, the clique score aggregates all pairwise couplings among the chosen types

$$\mu_{\text{MB,clq}}^{(\alpha^l)} = \sum_{\beta, \gamma \in \alpha^l} \sum_{i:\alpha_i=\beta} \sum_{j:\alpha_j=\gamma} \Phi(\|\mathbf{r}_i - \mathbf{r}_j\|; \eta_{\beta\gamma}) \mathbf{1}\left((v_i, v_j) \in \mathcal{E}(\alpha^l)\right). \quad (15)$$

The operator $\mathbf{1}\left((v_i, v_j) \in \mathcal{E}(\alpha^l)\right)$ is an indicator function

$$\mathbf{1}(A) = \begin{cases} 1, & \text{if the condition } A \text{ is true} \\ 0, & \text{otherwise} \end{cases} \quad (16)$$

Since this work does not consider the ordering of atom types, only the clique interaction score is employed in all experiments. Moreover, to avoid redundant feature representations, we restrict attention to the pairwise multi-body model (GMI-Score-2) with $l = 2$, and the triplet multi-body model (GMI-Score-3) with $l = 3$.

3.2 Fusion learning

Many existing works demonstrate that combining various descriptors can enhance the prediction of protein-ligand binding affinities [28, 31, 65, 69]. Moreover, the importance of data quality over quantity, specifically, the relevance and informativeness of descriptors are shown to have a substantial impact on model performance and efficiency [41]. To obtain richer representations of the protein-ligand complexes, we fuse the most informative features from both GMI-Score-2 and GMI-Score-3. Particularly, we focus on subsets of the pair and triplet atom types from both feature sets, which not only preserves key biochemical interactions but also eliminates non-contributory features, thereby reducing the computational cost. GMI-Score-2 provides the interaction strengths between atom pair types, whereas GMI-Score-3 captures both pairwise and triplet-wise energy terms across multi-body interactions, together forming multi-level representations of protein-ligand interactions.

To capitalize on the complementary strengths of both methods, we adopted a fusion strategy. The rationale behind combining different scoring functions or models

is well-established in computational drug discovery [10]. Ensemble methods, such as those employing combinatorial fusion analysis [23], often lead to improved accuracy and robustness by mitigating the individual limitations and biases of single models [40]. Here, we developed a new SF model by integrating GMI-Score-2 and GL-MB-Score-3 via a fusion process. The process can be done through methods like weighted averaging, linear regression, or more sophisticated ensemble techniques such as stacking or boosting. Such methods focus on leveraging the individual predictive power of each model at the decision stage. In our work, we applied an ensemble learning strategy that integrates their outputs at the decision stage. Specifically, we adopted a rank-based aggregation method: ensemble models ranked all candidate ligand–protein complexes according to their predicted binding affinity score. The final ensemble ranking was then obtained by aggregating these individual ranks, such that the relative order of complexes reflected the consensus between the two models.

3.2.1 Feature selection

Gradient boosting is a machine learning method that builds an ensemble of weak prediction models (typically decision trees) in a stage-wise fashion, optimizing an arbitrary differentiable loss function using gradient descent [18]. At each iteration, the algorithm fits a new model to the pseudo-residuals of the current ensemble, gradually reducing prediction errors [42]. Formally, given training data $\{(x_i, y_i)\}_{i=1}^n$ and a differentiable loss function $L(y, F(x))$, the method builds an additive model $F^*(x) = \sum_{m=0}^M f_m(x)$ where each f_m is a weak learner that minimizes the expected loss [18].

$$F^* = \arg \min_F \mathbb{E}_{y, \mathbf{x}} L(y, F(\mathbf{x})) = \arg \min_F \mathbb{E}_{\mathbf{x}} [\mathbb{E}_y (L(y, F(\mathbf{x})) \mid \mathbf{x})]. \quad (17)$$

where $\mathcal{L}(y, F(\mathbf{x}))$ is the loss function. Decision trees are developed using the greedy CART algorithm. During this process, features are chosen selectively to maintain sparsity in order to minimize the total impurity of child nodes [6]. Given a node t and its probabilities $p(j|t)$, the impurity is calculated as follow:

$$i(t) = \phi(p(1|t), \dots, p(J|t)) \quad (18)$$

Gini index is then applied to rank the importance of variables can be written as:

$$i(t) = \sum_{j \neq i} p(j|t)p(i|t) = \left(\sum_j p(j|t) \right)^2 - \sum_j p^2(j|t) = 1 - \sum_j p^2(j|t) \quad (19)$$

3.2.2 Feature fusion

To effectively merge two distinct feature sets while ensuring we capture their most informative descriptors, we employed a strategic feature fusion technique. This method centers on assessing the importance of each feature, a widely recognized approach for pointing out influential predictors within data. Rather than simply combining

or averaging features, which can lead to redundancy or dilute vital information, our process utilizes the feature ranking of Gradient Boosting (GB) models.

Given n minimum node impurity of features from m datasets $D_m = \{(x^{(i)}, y^{(i)})\}_{i=1}^n$, the new variables are defined as:

$$K = \left\{ \left(x_j^{(i)}, y_j^{(i)} \right) \right\}_{i=1, j=1}^{n, m} \quad (20)$$

This concatenation creates a richer representation by combining the most informative descriptors from both original sets, allowing the subsequent model to leverage a broader spectrum of predictive signals.

3.3 Benchmark evaluation

In this work, our proposed GMI-Score was evaluated for scoring power across several datasets. Following the standard CASF benchmark protocol [59] for scoring power, the Pearson’s correlation coefficient (R_p) between the predicted binding scores and the experimental binding constants (in logarithm units, log Ka) was computed as a quantitative indicator of scoring power:

$$R_p = \frac{\sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2 \sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2}} \quad (21)$$

To account for variability, each experiment was repeated at least 5 times with different random seeds, and the reported results represent the average across runs. RMSE values are optionally included where available to facilitate comparison with other methods.

4 Conclusions

In this work, we introduced GMI-Score, a novel scoring function for protein-ligand binding affinity prediction that leverages multi-level, multi-body atomic interaction graphs. Unlike traditional approaches that depend heavily on handcrafted physico-chemical descriptors, our method utilizes graph-based representations as core features, significantly reducing model complexity while maintaining high predictive accuracy. By integrating both pairwise and triplet-wise interactions through a feature fusion strategy, GMI-Score effectively captures the physical and chemical information governing the protein-ligand binding mechanism. Our extensive experiments on multiple benchmark datasets, including PDBbind v2016, v2020, CSAR-NRC-HiQ, and PDBbind-Redocked, demonstrate consistent improvements over state-of-the-art methods. Our proposed models achieve Pearson correlations up to 0.877. Among all evaluated ML-based models, GMI-Score exhibits the smallest performance degradation on the leak-proof partition, further supporting its robustness on unseen and diverse molecular systems.

Despite these promising results, this study focuses solely on the scoring power measured by the Pearson correlation coefficient following the PDBbind protocol. Other important evaluation criteria, including ranking power, docking power, and screening

power, which require additional dataset preparation and methodological extensions, were not assessed and remain the main direction for future work.

Author contributions

T.TH.L, H.V, N.HM.N prepared the manuscript; T.TH.L, B.T.N, D.D.N finalized the manuscript. T.TH.L collected data and performed analysis. B.T.N, D.D.N supervised the project.

Funding

D.D.N. was supported in part by funds from the National Science Foundation (NSF: # 2516126, # 2151802, and # 2534947).

Data availability

The PDBbind data used in this study are publicly available at <https://www.PDBbind-plus.org.cn>. The source code and pretrained model are available at the GitHub repository: <https://github.com/BinhMisfit/DrugWise-Implementation>.

Declarations

Competing interests

The authors declare that they have no competing interests.

Appendix A Supplementary Tables

Table A1: Model hyperparameters including number of estimators, feature selection method, tree depth, split criteria, learning rate, subsampling ratio, and loss function.

Model	Hyperparameters
Model 1	n_estimators=20000, max_features='sqrt', max_depth=8, min_samples_split=2, learning_rate=0.005, subsample=0.7, loss='squared_error'
Model 2	n_estimators=15000, max_features='sqrt', max_depth=7, min_samples_split=16, min_samples_leaf=10, learning_rate=0.008, subsample=0.4, loss='squared_error'
Model 3	n_estimators=24500, max_features='sqrt', max_depth=7, min_samples_split=3, min_samples_leaf=19, learning_rate=0.03, subsample=0.4, loss='squared_error'
Model 4	n_estimators=23500, max_features='sqrt', max_depth=6, min_samples_split=16, min_samples_leaf=8, learning_rate=0.006, subsample=0.7, loss='squared_error'

Table A2: Paired t-test on Pearson correlation: GMI-Score vs individual model on CASF v2016 dataset.

Model	Train with Refined			Train with General		
	t-stat	p-value	significant	t-stat	p-value	significance
GMI-Score-2	16.163	0.0001	***	13.191	0.0002	***
GMI-Score-3	10.090	0.0005	***	25.142	0.0000	***

Significance: ***: $p < 0.001$ | **: $p < 0.01$ | *: $p < 0.05$ | n.s. = not significant.

Table A3: Performance comparison of individual and fusion models on the PDBbind v2020 dataset. The table reports the Pearson correlation coefficient (R_p) and RMSE in kcal/mol.

Model	Train with Refined		Train with General	
	R_p	$RMSE$	R_p	$RMSE$
GMI-Score-2	0.816 ± 0.001	1.829 ± 0.003	0.862 ± 0.002	1.607 ± 0.008
GMI-Score-3	0.834 ± 0.002	1.769 ± 0.009	0.863 ± 0.001	1.622 ± 0.007
GMI-Score	0.842 ± 0.002	1.732 ± 0.008	0.872 ± 0.001	1.560 ± 0.003

Significance: ***: $p < 0.001$ | **: $p < 0.01$ | *: $p < 0.05$ | n.s. = not significant.

Table B4: Paired t-test on Pearson correlation: GMI-Score vs individual model on PDBbind v2020 dataset .

Model	Train with Refined			Train with General		
	t-stat	p-value	significant	t-stat	p-value	significance
GMI-Score-2	23.791	0.0000	***	13.870	0.0002	***
GMI-Score-3	5.396	0.0057	***	13.425	0.0002	***

Significance: ***: $p < 0.001$ | **: $p < 0.01$ | *: $p < 0.05$ | n.s. = not significant.

Appendix B Supplementary Figures

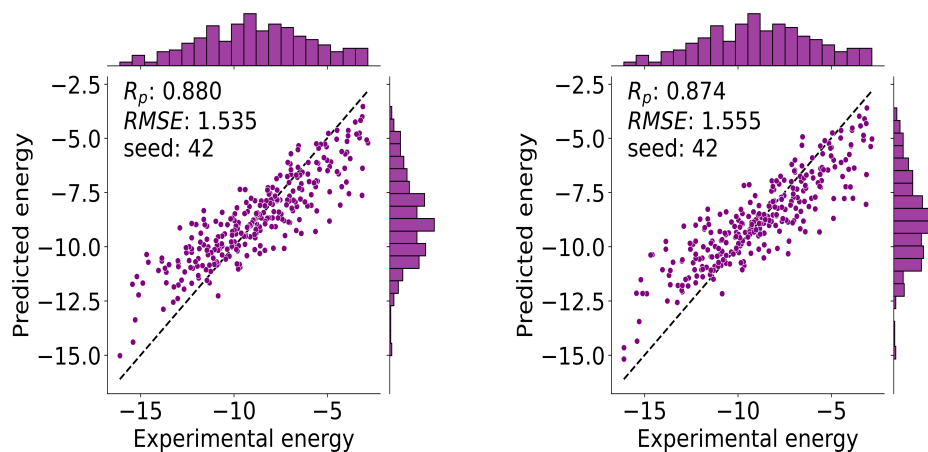


Fig. B1: Correlation between GMI-Score predictions and experimental data for various benchmarks. (Left) PDBbind v2016: Pearson correlation coefficient $R_p = 0.880$ and $RMSE = 1.535$ kcal/mol. (Right) PDBbind v2020: Pearson correlation coefficient $R_p = 0.874$ and $RMSE = 1.555$ kcal/mol.

Appendix C Predictions of PDBbind v2016 dataset

PDBID	actual	prediction	PDBID	actual	prediction	PDBID	actual	prediction
3ao4	2.07	3.82	2w66	4.05	5.11	3b27	5.16	5.79
3gv9	2.12	3.72	3kwa	4.08	6.13	4gkm	5.17	5.16
1uto	2.27	2.59	3g2n	4.09	4.49	4m0z	5.19	6.5
1ps3	2.28	4.68	4cr9	4.1	3.86	1w4o	5.22	4.48
4ddk	2.29	2.93	4ih5	4.11	5.1	3ueu	5.24	4.96
4jsz	2.3	5.59	4de2	4.12	5.31	4ih7	5.24	5.75
3g2z	2.36	3.05	3ozt	4.13	6.82	4jfs	5.27	5.91
3dxg	2.4	3.65	3f3a	4.19	5.17	3ozs	5.33	6.96
3l7b	2.4	3.49	1a30	4.3	5.93	3bv9	5.36	7.39
3gr2	2.52	3.53	3ivg	4.3	5.57	1gpk	5.37	6.45
3kgp	2.57	3.11	3u9q	4.38	5.13	1syi	5.44	6.4
3fcq	2.77	4.35	3rsx	4.41	4.49	2cbv	5.48	6.23
3lka	2.82	5.54	3pxf	4.43	4.88	1ydr	5.52	5.75
3zt2	2.84	3.4	2wbg	4.45	5.96	4de3	5.52	4.92
3udh	2.85	3.58	3rr4	4.55	5.82	3coz	5.57	6.16
3g31	2.89	4.14	4w9c	4.65	4.81	2wca	5.6	6.36
4llx	2.89	3.77	3mss	4.66	5.4	3u5j	5.61	5.77
4u4s	2.92	5.15	4agp	4.69	5.2	4dli	5.62	5.74
4owm	2.96	4.2	4mgd	4.69	6.88	1z9g	5.64	5.42
5aba	2.98	4.68	1vso	4.72	5.72	3arv	5.64	5.77
2xdl	3.1	5.3	4jxs	4.74	4.91	3n86	5.64	5.75
4kz6	3.1	4.12	1q8t	4.76	5.35	5c28	5.66	5.04
2ymd	3.16	5.92	3acw	4.76	6.41	4j28	5.7	6.65
3aru	3.22	4.87	4lzs	4.8	3.91	3jvr	5.72	5.62
1bcu	3.28	4.79	3r88	4.82	4.05	1o5b	5.77	6.15
3zsx	3.28	3.39	4ciw	4.82	4.65	2y5h	5.79	7.39
4ddh	3.3	3.7	2w4x	4.85	4.98	3qqs	5.82	5.33
4eky	3.52	4.46	2brb	4.86	5.98	3wz8	5.82	6.36
4abg	3.57	4.62	1p1q	4.89	6.78	4dld	5.82	5.9
5a7b	3.57	5.47	3d4z	4.89	4.91	3ehy	5.85	6.65
3dx1	3.58	6.47	1bzc	4.92	5.82	3uev	5.89	5.72
4bkt	3.62	3.5	1nc3	5	5.76	3ebp	5.91	5.86
2v00	3.66	4.77	4agq	5.01	5.45	1o0h	5.92	5.53
4cig	3.67	4.41	4w9l	5.02	6.11	1q8u	5.96	6.36
3n7a	3.7	3.95	2yge	5.06	6.03	4de1	5.96	5.56
3d6q	3.76	3.86	2r9w	5.1	4.88	4w9i	5.96	5.56
2hb1	3.8	4.32	3gy4	5.1	4.19	3ary	6	5.33
3twp	3.92	3.87	3syr	5.1	4.73	3coy	6.02	6.37
4agn	3.97	4.71	3zso	5.12	4.53	3f3c	6.02	5.41
1c5z	4.01	4	2br1	5.14	6.51	2fxs	6.06	5.78
3nq9	4.03	4.51	1s38	5.15	5.63	4kzq	6.1	6.48

PDBID	actual	prediction	PDBID	actual	prediction	PDBID	actual	prediction
2qnq	6.11	6.56	3jya	6.89	5.4	4hge	7.92	7.42
1nc1	6.12	7.46	3gbb	6.9	7.25	1g2k	7.96	9
2wvt	6.12	5.95	3uex	6.92	6.21	1o3f	7.96	7.92
1yc1	6.17	6.18	4f9w	6.94	6.98	2qe4	7.96	7.46
3bgz	6.26	6.36	2wer	7.05	6.88	3uuu	7.96	6.99
4wiv	6.26	5.78	1lpg	7.09	8.03	3e92	8	7.95
3k5v	6.3	6.78	3zdg	7.1	5.39	3fur	8	7.28
4eor	6.3	7.3	1z95	7.12	7.93	2cet	8.02	7.85
3uew	6.31	5.93	1pxn	7.15	6.55	5tmn	8.04	7.13
2wnc	6.32	6.25	3arp	7.15	7.31	3ag9	8.05	6.47
2zb1	6.32	7.56	3f3d	7.16	6.51	3kr8	8.1	7.43
2qbr	6.33	6.67	3tsk	7.17	8.09	3nx7	8.1	7.87
3arq	6.4	6.24	2j7h	7.19	6.72	3fv2	8.11	7.93
2j78	6.42	6.18	2xii	7.2	7.48	4eo8	8.15	7.72
4ea2	6.44	6.56	4cra	7.22	7.92	3e5a	8.23	7.53
1r5y	6.46	6.21	4gfm	7.22	6.01	1nvq	8.25	8.17
4m0y	6.46	6.88	1oyt	7.24	7.42	2v7a	8.3	7.96
1gpn	6.48	6.75	3gc5	7.26	6.63	4x6p	8.3	8.55
2weg	6.5	7.03	3p5o	7.26	6.9	1h23	8.35	7.34
4kzu	6.5	6.52	2vvn	7.3	6.22	4e6q	8.36	7.6
4mme	6.5	6.4	1qf1	7.32	6.77	2al5	8.4	7.08
3cj4	6.51	5.78	1ydt	7.32	6.45	2qbp	8.4	6.64
3uo4	6.52	7.43	3pww	7.32	7.17	2zda	8.4	7.45
3wtj	6.53	6.92	1owh	7.4	6.81	3b68	8.4	8.48
3jvs	6.54	6.42	2zy1	7.4	6.37	2xbv	8.43	8.12
1kli	6.58	6.43	3up2	7.4	7.28	3b1m	8.48	6.59
2yfe	6.63	6.57	4j21	7.41	6.94	2fvd	8.52	7.26
4k77	6.63	6.52	2xys	7.42	7.32	2vw5	8.52	6.41
2xj7	6.66	5.7	2qbq	7.44	7.17	2wn9	8.52	6.12
2iwx	6.68	6.62	3oe4	7.47	6.39	3ejr	8.57	7.97
4f09	6.7	7	3rlr	7.52	7.66	4qd6	8.64	6.88
4djv	6.72	6.59	2xb8	7.59	6.59	3u8k	8.66	7.18
4w9h	6.73	5.97	2c3i	7.6	7.41	3ge7	8.7	7.4
4ogj	6.79	6.65	4e5w	7.66	7.75	4crc	8.72	8.34
1p1n	6.8	6.15	3f3e	7.7	6.9	4ivb	8.72	8.39
3dx2	6.82	6.68	1u1b	7.8	6.17	2vkm	8.74	7.74
2xnb	6.83	7.64	3qgy	7.8	6.97	2wtv	8.74	8.31
3n76	6.85	5.84	3ryj	7.8	7.61	3b5r	8.77	7.99
3pyy	6.86	6.94	4j3l	7.8	7.31	2zcq	8.82	6.58
2zcr	6.87	6.94	3prs	7.82	7.24	3e93	8.85	7.84
3oe5	6.88	6.98	4pcs	7.85	6.58	4k18	8.96	8.06

PDBID	actual	prediction
2p4y	9	7.03
3dd0	9	8.57
3nw9	9	7.53
3ui7	9	7.41
3uri	9	6.55
1qkt	9.04	8.04
1h22	9.1	7.33
3gnw	9.1	7.8
1sqa	9.21	7.54
4jia	9.22	7.37
3b65	9.27	8.18
3fv1	9.3	8.77
4qac	9.4	8.47
2yki	9.46	6.69
3g0w	9.52	7.9
4ivd	9.52	8.63
4ty7	9.52	8
2pog	9.54	7.81
4gr0	9.55	9.08
1eby	9.7	8.7
1z6e	9.72	9.21
1e66	9.89	7.94
4ivc	10	8.56
4twp	10	8.19
4rfm	10.05	8.06
1y6r	10.11	9.45
3u8n	10.17	7.82
4tmn	10.17	7.72
2p15	10.3	8.14
3myg	10.7	7.86
4gid	10.77	7.6
3utu	10.92	8.56
5c2h	11.09	8.96
1mq6	11.15	8.34
5dwr	11.22	9.81
4f2w	11.3	10.56
2x00	11.33	8.61
3o9i	11.82	11.02
4f3c	11.82	11.02

Appendix D Predictions of PDBbind v2020 dataset

PDBID	actual	prediction	PDBID	actual	prediction	PDBID	actual	prediction
1a30	4.3	6.06	1ydt	7.32	6.43	2xbv	8.43	8.27
1bcu	3.28	4.49	1z6e	9.72	9.22	2xb8	7.59	6.5
1bzc	4.92	5.81	1z95	7.12	7.84	2xdl	3.1	5.31
1c5z	4.01	4.18	1z9g	5.64	5.66	2xii	7.2	7.24
1e66	9.89	7.86	2al5	8.4	7.21	2xj7	6.66	5.61
1eby	9.7	8.65	2br1	5.14	6.46	2xnb	6.83	7.55
1g2k	7.96	8.92	2brb	4.86	6.14	2xys	7.42	7.45
1gpn	6.48	6.72	2c3i	7.6	7.51	2y5h	5.79	7.39
1gpk	5.37	6.49	2cbv	5.48	6.07	2yfe	6.63	6.41
1h23	8.35	7.15	2cet	8.02	7.57	2yge	5.06	6.03
1h22	9.1	7.25	2fvd	8.52	7.44	2yki	9.46	6.64
1kli	6.58	6.47	2fxs	6.06	5.87	2ymd	3.16	5.7
1lpg	7.09	7.97	2hb1	3.8	4.74	2zb1	6.32	7.71
1mq6	11.15	8.44	2iwx	6.68	6.45	2zcq	8.82	6.4
1nc1	6.12	7.48	2j78	6.42	6.29	2zcr	6.87	6.74
1nc3	5	5.67	2j7h	7.19	6.54	2zda	8.4	7.67
1nvq	8.25	7.99	2p15	10.3	8.11	2zy1	7.4	6.31
1o0h	5.92	5.4	2p4y	9	7.15	3acw	4.76	6.47
1o3f	7.96	7.84	2pog	9.54	7.81	3ag9	8.05	6.81
1o5b	5.77	6.03	2qbp	8.4	6.85	3ao4	2.07	3.69
1owh	7.4	6.85	2qbp	7.44	7.01	3arp	7.15	7.14
1oyt	7.24	7.3	2qbr	6.33	6.5	3arq	6.4	6.34
1p1n	6.8	6.52	2qe4	7.96	7.49	3aru	3.22	4.96
1p1q	4.89	6.72	2qnq	6.11	6.62	3arv	5.64	5.73
1ps3	2.28	4.89	2r9w	5.1	4.81	3ary	6	5.1
1pxn	7.15	6.55	2v00	3.66	3.98	3b1m	8.48	6.58
1q8t	4.76	5.44	2v7a	8.3	8.13	3b27	5.16	5.55
1q8u	5.96	6.26	2vkm	8.74	7.72	3b5r	8.77	8.11
1qf1	7.32	6.92	2vvn	7.3	7.53	3b65	9.27	8.39
1qkt	9.04	8.11	2vw5	8.52	6.39	3b68	8.4	8.13
1r5y	6.46	6.28	2w4x	4.85	5.11	3bgz	6.26	6.32
1s38	5.15	5.67	2w66	4.05	5.09	3bv9	5.36	7.57
1sqa	9.21	7.42	2wca	5.6	6.47	3cj4	6.51	5.94
1syi	5.44	6.43	2weg	6.5	6.89	3coy	6.02	6.47
1ulb	7.8	6.15	2wer	7.05	6.85	3coz	5.57	6.38
1vso	4.72	5.8	2wbg	4.45	5.7	3d4z	4.89	4.95
1w4o	5.22	4.55	2wnc	6.32	6.54	3d6q	3.76	3.85
1uto	2.27	2.64	2wn9	8.52	6.07	3dd0	9	8.55
1y6r	10.11	9.17	2wtv	8.74	8	3dxg	2.4	3.72
1yc1	6.17	6.31	2wvt	6.12	5.96	3dx1	3.58	6.38
1ydr	5.52	5.6	2x00	11.33	8.91	3dx2	6.82	6.69

PDBID	actual	prediction	PDBID	actual	prediction	PDBID	actual	prediction
3e5a	8.23	7.32	3nw9	9	7.73	3zt2	2.84	3.32
3e93	8.85	7.75	3nx7	8.1	7.86	4abg	3.57	4.4
3e92	8	7.96	3o9i	11.82	11.13	3zdg	7.1	5.26
3ehy	5.85	6.61	3oe4	7.47	6.64	4agn	3.97	4.73
3ebp	5.91	5.78	3oe5	6.88	6.82	4agp	4.69	5.26
3ejr	8.57	8.03	3ozt	4.13	6.83	4agq	5.01	5.38
3f3a	4.19	5.17	3ozs	5.33	7.14	4bkt	3.62	3.49
3f3e	7.7	6.86	3p5o	7.26	7.03	4cig	3.67	4.26
3f3c	6.02	5.45	3prs	7.82	6.93	4ciw	4.82	4.6
3f3d	7.16	6.21	3pww	7.32	7.22	4cra	7.22	7.88
3fcq	2.77	4.67	3pxf	4.43	4.8	4cr9	4.1	3.93
3fur	8	7.12	3pyy	6.86	6.91	4crc	8.72	8.1
3fv2	8.11	8	3qgy	7.8	7.63	4ddh	3.3	3.66
3fv1	9.3	8.66	3qqs	5.82	5.11	4ddk	2.29	3.16
3g0w	9.52	8.86	3r88	4.82	3.91	4de1	5.96	5.29
3g2z	2.36	3.01	3rlr	7.52	7.58	4de2	4.12	5.3
3g2n	4.09	4.69	3rsx	4.41	4.33	4de3	5.52	5.08
3g31	2.89	3.92	3rr4	4.55	5.98	4djv	6.72	6.5
3gbb	6.9	7.23	3ryj	7.8	7.89	4dli	5.62	5.72
3gc5	7.26	6.67	3syr	5.1	4.75	4dld	5.82	6.25
3ge7	8.7	7.33	3tsk	7.17	8.11	4e6q	8.36	7.68
3gr2	2.52	3.25	3twp	3.92	3.75	4e5w	7.66	7.92
3gnw	9.1	7.87	3u5j	5.61	6	4ea2	6.44	6.55
3gv9	2.12	3.94	3u8k	8.66	6.73	4eky	3.52	4.63
3gy4	5.1	4.4	3u9q	4.38	5.01	4eo8	8.15	7.65
3ivg	4.3	5.66	3u8n	10.17	7.42	4eor	6.3	7.27
3jvs	6.54	6.26	3udh	2.85	3.4	4f09	6.7	6.83
3jvr	5.72	5.21	3ueu	5.24	5.13	4f2w	11.3	10.53
3jya	6.89	5.24	3uew	6.31	6.24	4f3c	11.82	10.75
3k5v	6.3	6.87	3uev	5.89	5.99	4f9w	6.94	7
3kgp	2.57	2.92	3uex	6.92	6.31	4gfm	7.22	6.08
3kr8	8.1	7.39	3ui7	9	7.59	4gid	10.77	7.61
3kwa	4.08	5.32	3uo4	6.52	7.41	4gkm	5.17	5.35
3l7b	2.4	3.57	3up2	7.4	7.38	4gr0	9.55	8.81
3lka	2.82	5.69	3uri	9	6.61	4hge	7.92	7.77
3mss	4.66	5.35	3utu	10.92	8.92	4ih5	4.11	4.88
3myg	10.7	7.61	3uuo	7.96	7.09	4ih7	5.24	5.8
3n7a	3.7	4	3wtj	6.53	6.96	4ivb	8.72	8.35
3n76	6.85	5.83	3wz8	5.82	6.29	4ivc	10	8.62
3n86	5.64	5.74	3zso	5.12	4.77	4ivd	9.52	8.65
3nq9	4.03	4.66	3zsx	3.28	3.4	4j21	7.41	6.99

PDBID	actual	prediction
4j28	5.7	6.81
4j3l	7.8	7.2
4jia	9.22	7.22
4jfs	5.27	5.92
4jsz	2.3	5.9
4jxs	4.74	5.01
4k18	8.96	8.32
4k77	6.63	6.76
4kz6	3.1	4.39
4kzq	6.1	6.44
4kzu	6.5	6.55
4llx	2.89	3.88
4lzs	4.8	3.62
4m0y	6.46	6.91
4m0z	5.19	6.52
4mgd	4.69	6.82
4mme	6.5	6.22
4ogj	6.79	6.58
4owm	2.96	4.22
4pcs	7.85	6.87
4qd6	8.64	6.69
4qac	9.4	8.44
4rfm	10.05	8.3
4tmn	10.17	7.15
4twp	10	8.14
4ty7	9.52	7.78
4u4s	2.92	5.31
4w9c	4.65	4.91
4w9h	6.73	6.15
4w9i	5.96	5.8
4w9l	5.02	6.3
4wiv	6.26	6.42
4x6p	8.3	8.43
5a7b	3.57	5.7
5aba	2.98	4.72
5c28	5.66	4.95
5c2h	11.09	8.91
5dwr	11.22	9.87
5tmn	8.04	6.24

References

- [1] Adediwura VA, Koirala K, Do HN, et al (2024) Understanding the impact of binding free energy and kinetics calculations in modern drug discovery. Expert

- [2] Alberts B, Heald R, Johnson A, et al (2022) Molecular biology of the cell: seventh international student edition with registration card. WW Norton & Company
- [3] Bacon F (1965) On the nature of allosteric transitions: a plausible model. *J mol biol* 12(1):88–118
- [4] Ballester PJ, Mitchell JB (2010) A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking. *Bioinformatics* 26(9):1169–1175
- [5] Boehr DD, Nussinov R, Wright PE (2009) The role of dynamic conformational ensembles in biomolecular recognition. *Nature chemical biology* 5(11):789–796
- [6] Breiman L, Friedman J, Olshen RA, et al (2017) Classification and regression trees. Chapman and Hall/CRC
- [7] Bresson X, Laurent T (2017) Residual gated graph convnets. arXiv preprint arXiv:171107553
- [8] Brivanlou AH, Darnell Jr JE (2002) Signal transduction and the control of gene expression. *Science* 295(5556):813–818
- [9] Changeux JP, Edelstein SJ (2005) Allosteric mechanisms of signal transduction. *Science* 308(5727):1424–1428
- [10] Cheng T, Li X, Li Y, et al (2009) Comparative assessment of scoring functions on a diverse test set. *Journal of chemical information and modeling* 49(4):1079–1093
- [11] Choi S, Seo S, Park S (2024) Spin: Se (3)-invariant physics informed network for binding affinity prediction. In: ECAI 2024. IOS Press, p 169–176
- [12] Damian L (2013) Isothermal titration calorimetry for studying protein–ligand interactions. *Protein-Ligand Interactions: Methods and Applications* pp 103–118
- [13] Dunbar Jr JB, Smith RD, Damm-Ganamet KL, et al (2013) Csar data set release 2012: ligands, affinities, complexes, and docking decoys. *Journal of chemical information and modeling* 53(8):1842–1852
- [14] Dwivedi V, Bresson X (2020) A generalization of transformer networks to graphs. arxiv preprint arxiv: 201209699
- [15] Dwivedi VP, Joshi CK, Luu AT, et al (2023) Benchmarking graph neural networks. *Journal of Machine Learning Research* 24(43):1–48
- [16] Fang Y (2012) Ligand–receptor interaction platforms and their applications for drug discovery. *Expert opinion on drug discovery* 7(10):969–988

- [17] Ferrell Jr JE (2002) Self-perpetuating states in signal transduction: positive feedback, double-negative feedback and bistability. *Current opinion in cell biology* 14(2):140–148
- [18] Friedman JH (2001) Greedy function approximation: a gradient boosting machine. *Annals of statistics* pp 1189–1232
- [19] Gasteiger J, Groß J, Günnemann S (2020) Directional message passing for molecular graphs. arXiv preprint arXiv:200303123
- [20] Gasteiger J, Becker F, Günnemann S (2021) Gemnet: Universal directional graph neural networks for molecules. *Advances in neural information processing systems* 34:6790–6802
- [21] Guterres H, Im W (2020) Improving protein-ligand docking results with high-throughput molecular dynamics simulations. *Journal of chemical information and modeling* 60(4):2189–2198
- [22] Harren T, Gutermuth T, Grebner C, et al (2024) Modern machine-learning for binding affinity estimation of protein–ligand complexes: Progress, opportunities, and challenges. *Wiley Interdisciplinary Reviews: Computational Molecular Science* 14(3):e1716
- [23] Hsu DF, Chung YS, Kristal BS (2006) Combinatorial fusion analysis: methods and practices of combining multiple scoring systems. In: *Advanced data mining technologies in bioinformatics*. IGI Global Scientific Publishing, p 32–62
- [24] Jain AN (2007) Surflex-dock 2.1: robust performance from ligand energetic modeling, ring flexibility, and knowledge-based search. *Journal of computer-aided molecular design* 21(5):281–306
- [25] Jiang D, Hsieh CY, Wu Z, et al (2021) Interactiongraphnet: a novel and efficient deep graph representation learning framework for accurate protein–ligand interaction predictions. *Journal of medicinal chemistry* 64(24):18209–18232
- [26] Jiménez J, Skalic M, Martinez-Rosell G, et al (2018) K deep: protein–ligand absolute binding affinity prediction via 3d-convolutional neural networks. *Journal of chemical information and modeling* 58(2):287–296
- [27] Jin Z, Wu T, Chen T, et al (2023) Capla: improved prediction of protein–ligand binding affinity by a deep learning approach based on a cross-attention mechanism. *Bioinformatics* 39(2):btad049
- [28] Jones D, Kim H, Zhang X, et al (2021) Improved protein–ligand binding affinity prediction with structure-based deep fusion inference. *Journal of Chemical Information and Modeling* 61(4):1583–1592. <https://doi.org/10.1021/acs.jcim.0c01306>, URL <https://doi.org/10.1021/acs.jcim.0c01306>, pMID: 33754707,

<https://doi.org/10.1021/acs.jcim.0c01306>

- [29] Karlov DS, Sosnin S, Fedorov MV, et al (2020) graphdelta: Mpnn scoring function for the affinity prediction of protein–ligand complexes. *ACS omega* 5(10):5150–5159
- [30] Kuhn M, Firth-Clark S, Tosco P, et al (2020) Assessment of binding affinity via alchemical free-energy calculations. *Journal of Chemical Information and Modeling* 60(6):3120–3130
- [31] Kundu I, Paul G, Banerjee R (2018) A machine learning approach towards the prediction of protein–ligand binding affinity based on fundamental molecular properties. *RSC Adv* 8:12127–12137. <https://doi.org/10.1039/C8RA00003D>, URL <http://dx.doi.org/10.1039/C8RA00003D>
- [32] Li H, Leung KS, Wong MH, et al (2015) Improving autodock vina using random forest: the growing accuracy of binding affinity prediction by the effective exploitation of larger data sets. *Molecular informatics* 34(2-3):115–126
- [33] Li J, Guan X, Zhang O, et al (2024) Leak proof pdbind: a reorganized dataset of protein-ligand complexes for more generalizable binding affinity prediction. *ArXiv pp arXiv–2308*
- [34] Li S, Zhou J, Xu T, et al (2021) Structure-aware interactive graph neural networks for the prediction of protein-ligand binding affinity. In: *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pp 975–985
- [35] Li S, Zhou J, Xu T, et al (2023) Giant: Protein-ligand binding affinity prediction via geometry-aware interactive graph neural network. *IEEE Transactions on Knowledge and Data Engineering* 36(5):1991–2008
- [36] Li Y, Han L, Liu Z, et al (2014) Comparative assessment of scoring functions on an updated benchmark: 2. evaluation methods and general results. *Journal of chemical information and modeling* 54(6):1717–1736
- [37] Li Y, Su M, Liu Z, et al (2018) Assessing protein–ligand interaction scoring functions with the casf-2013 benchmark. *Nature protocols* 13(4):666–680
- [38] Liu Z, Li Y, Han L, et al (2014) Pdb-wide collection of binding data: current status of the pdbind database. *Bioinformatics* 31(3):405–412. <https://doi.org/10.1093/bioinformatics/btu626>, URL <https://doi.org/10.1093/bioinformatics/btu626>, https://academic.oup.com/bioinformatics/article-pdf/31/3/405/49012002/bioinformatics_31_3_405.pdf
- [39] Min Y, Wei Y, Wang P, et al (2024) From static to dynamic structures: Improving binding affinity prediction with graph-based deep learning. *Advanced Science* 11(40):2405404

- [40] Mohamed Abdul Cader J, Newton MH, Rahman J, et al (2024) Ensembling methods for protein-ligand binding affinity prediction. *Scientific Reports* 14(1):24447
- [41] Mukta FT, Rana MM, Meyer A, et al (2025) The algebraic extended atom-type graph-based model for precise ligand-receptor binding affinity prediction. *Journal of Cheminformatics* 17(1):10
- [42] Natekin A, Knoll A (2013) Gradient boosting machines, a tutorial. *Frontiers in neurorobotics* 7:21
- [43] Nguyen DD, Wei GW (2019) Dg-gl: Differential geometry-based geometric learning of molecular datasets. *International journal for numerical methods in biomedical engineering* 35(3):e3179
- [44] Nguyen DD, Xiao T, Wang M, et al (2017) Rigidity strengthening: A mechanism for protein-ligand binding. *Journal of chemical information and modeling* 57(7):1715–1721
- [45] Öztürk H, Özgür A, Ozkirimli E (2018) Deepdta: deep drug-target binding affinity prediction. *Bioinformatics* 34(17):i821–i829
- [46] O’Connell N (2021) Protein ligand interactions using surface plasmon resonance. *Targeted Protein Degradation: Methods and Protocols* pp 3–20
- [47] Park H, Bradley P, Greisen Jr P, et al (2016) Simultaneous optimization of biomolecular energy functions on features from small molecules and macromolecules. *Journal of chemical theory and computation* 12(12):6201–6212
- [48] Prat A, Abdel Aty H, Bastas O, et al (2024) Hydrascreen: a generalizable structure-based deep learning approach to drug discovery. *Journal of Chemical Information and Modeling* 64(15):5817–5831
- [49] Quiroga R, Villarreal MA (2016) Vinardo: A scoring function based on autodock vina improves scoring, docking, and virtual screening. *PloS one* 11(5):e0155183
- [50] Rana MM, Nguyen DD (2023) Geometric graph learning with extended atom-types features for protein-ligand binding affinity prediction. *Computers in Biology and Medicine* 164:107250
- [51] Sánchez-Cruz N, Medina-Franco JL, Mestres J, et al (2021) Extended connectivity interaction features: improving binding affinity prediction through chemical description. *Bioinformatics* 37(10):1376–1382
- [52] Schütt K, Unke O, Gastegger M (2021) Equivariant message passing for the prediction of tensorial properties and molecular spectra. In: *International conference on machine learning*, PMLR, pp 9377–9388

- [53] Seo S, Choi J, Park S, et al (2021) Binding affinity prediction for protein–ligand complex using deep attention mechanism based on intermolecular interactions. *BMC bioinformatics* 22:1–15
- [54] Shen C, Hu X, Gao J, et al (2021) The impact of cross-docked poses on performance of machine learning classifier for protein–ligand binding pose prediction. *Journal of cheminformatics* 13(1):81
- [55] Shen C, Zhang X, Hsieh CY, et al (2023) A generalized protein–ligand scoring framework with balanced scoring, docking, ranking and screening powers. *Chemical Science* 14(30):8129–8146
- [56] Shi MH, Zhang SW, Zhang QQ, et al (2025) Plagca: Predicting protein–ligand binding affinity with the graph cross-attention mechanism. *Journal of Biomedical Informatics* 165:104816
- [57] Sliwoski G, Kothiwale S, Meiler J, et al (2014) Computational methods in drug discovery. *Pharmacological reviews* 66(1):334–395
- [58] Stepniewska-Dziubinska MM, Zielenkiewicz P, Siedlecki P (2018) Development and evaluation of a deep learning model for protein–ligand binding affinity prediction. *Bioinformatics* 34(21):3666–3674
- [59] Su M, Yang Q, Du Y, et al (2018) Comparative assessment of scoring functions: the casf-2016 update. *Journal of chemical information and modeling* 59(2):895–913
- [60] Su Y, Zhou A, Xia X, et al (2009) Quantitative prediction of protein–protein binding affinity with a potential of mean force considering volume correction. *Protein Science* 18(12):2550–2558
- [61] Tan H, Wang Z, Hu G (2024) Gaabind: a geometry-aware attention-based network for accurate protein–ligand binding pose and binding affinity prediction. *Briefings in Bioinformatics* 25(1):bbad462
- [62] Trott O, Olson AJ (2010) Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry* 31(2):455–461
- [63] Wang B, Zhao Z, Nguyen DD, et al (2017) Feature functional theory–binding predictor (fft–bp) for the blind prediction of binding free energies. *Theoretical Chemistry Accounts* 136(4):55
- [64] Wang C, Zhang Y (2017) Improving scoring-docking-screening powers of protein–ligand scoring functions using random forest. *Journal of computational chemistry* 38(3):169–177

- [65] Wang G, Zhang H, Shao M, et al (2024) Deeptgin: a novel hybrid multimodal approach using transformers and graph isomorphism networks for protein-ligand binding affinity prediction. *Journal of Cheminformatics* 16(1):147
- [66] Wang H (2024) Prediction of protein–ligand binding affinity via deep learning models. *Briefings in Bioinformatics* 25(2):bbae081
- [67] Wang W, Donini O, Reyes CM, et al (2001) Biomolecular simulations: recent developments in force fields, simulations of enzyme catalysis, protein-ligand, protein-protein, and protein-nucleic acid noncovalent interactions. *Annual review of biophysics and biomolecular structure* 30(1):211–243
- [68] Xia C, Feng SH, Xia Y, et al (2023) Leveraging scaffold information to predict protein–ligand binding affinity with an empirical graph neural network. *Briefings in bioinformatics* 24(1):bbac603
- [69] Xia K, Opron K, Wei GW (2013) Multiscale multiphysics and multidomain models—flexibility and rigidity. *The Journal of chemical physics* 139(19)
- [70] Yang C, Zhang Y (2022) Delta machine learning to improve scoring-ranking-screening performances of protein–ligand scoring functions. *Journal of chemical information and modeling* 62(11):2696–2712
- [71] Yang R, Zhang L, Bu F, et al (2024) Ai-based prediction of protein–ligand binding affinity and discovery of potential natural product inhibitors against erk2. *BMC chemistry* 18(1):108
- [72] Zhang H, Liao L, Saravanan KM, et al (2019) Deepbindrg: a deep learning based method for estimating effective protein–ligand affinity. *PeerJ* 7:e7362
- [73] Zheng L, Fan J, Mu Y (2019) Onionnet: a multiple-layer intermolecular-contact-based convolutional neural network for protein–ligand binding affinity prediction. *ACS omega* 4(14):15956–15965
- [74] Zhu F, Zhang X, Allen JE, et al (2020) Binding affinity prediction by pairwise function based on neural network. *Journal of chemical information and modeling* 60(6):2766–2772