

Interformer: An Interaction-Aware Model for Protein-Ligand Docking and Affinity Prediction

Houtim Lai

mosquitolkfo@gmail.com

Tencent

Longyue Wang

Tencent AI Lab

Ruiyuan Qian

Tencent

Geyan Ye

Tencent

Juhong Huang

Tencent

Fandi Wu

Tencent

Fang Wu

Westlake University <https://orcid.org/0000-0001-7240-3915>

Xiangxiang Zeng

Hunan University <https://orcid.org/0000-0003-1081-7658>

Wei Liu

Tencent AI lab, Tencent

Article

Keywords: Protein-ligand docking, Protein-ligand affinity prediction, AI based energy model, Structure based drug design

Posted Date: March 26th, 2024

DOI: <https://doi.org/10.21203/rs.3.rs-3995849/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: There is **NO** Competing Interest.

Version of Record: A version of this preprint was published at Nature Communications on November 25th, 2024. See the published version at <https://doi.org/10.1038/s41467-024-54440-6>.

Interformer: An Interaction-Aware Model for Protein-Ligand Docking and Affinity Prediction

Houtim Lai¹, Longyue Wang^{1*}, Ruiyuan Qian¹, Geyan Ye¹,
Juhong Huang¹, Fandi Wu¹, Fang Wu³, Xiangxiang Zeng²,
Wei Liu¹

¹AI Lab, Tencent, Shenzhen, China.

²Department of Computer Science, Hunan University, Hunan, China.

³Institute for AI Industry Research, Tsinghua University, Beijing, China.

*Corresponding author(s). E-mail(s): revoli@tencent.com,
vinnylywang@tencent.com;

Abstract

In recent years, there has been a growing interest in using deep learning models for protein-ligand docking and affinity prediction, both vital for structure based drug design. However, many of these models overlook the intricate modeling of interactions between ligand and protein atoms, thereby constraining their capabilities in generalization and interpretability. In this paper, we introduce INTERFORMER, a unified model built upon the Graph-Transformer architecture, which specially crafted to capture non-covalent interactions through the interaction-aware mixture density network. Besides, we incorporated a new strategy that utilizes negative samples, effective interaction distribution correction for affinity prediction. Experimental results on widely-used and our in-house datasets demonstrate the effectiveness and universality of the proposed approach. Extensive analyses confirm our claim that our approach improves performance by modelling protein-ligand specific interactions. Encouragingly, our approach propels the SOTA performance docking tasks forward. We intend to make our code publicly available, hoping to facilitate future research in this field.

Keywords: Protein-ligand docking, Protein-ligand affinity prediction, AI based energy model, Structure based drug design

1 Introduction

As a crucial task in optimization of drug molecular structures, *protein-ligand docking* is to predict the position and orientation of a ligand (a small molecule) when it binds to a protein receptor or enzyme. Leveraging accurate binding pose (protein-ligand binding complex conformations), the *affinity prediction* task offers a computational estimation of the binding strength between a ligand and its target protein, facilitating the screening of ligand with potential activity. In the intricate journey of drug development, both tasks have been important components of the drug discovery process for years [1, 2].

Recent years have witnessed a surge of interests of using deep learning (DL) approaches for molecular modeling [3, 4]. Treating docking as a generative modeling problem, [5] introduced DIFFDOCK, a graph neural network (GNN) based model that has established new benchmark in binding pose generation. However, existing DL models often overlook the modeling of *non-covalent interactions between ligand and protein atoms*, which is essential for interpretability and generalization. As illustrated in Fig.1 Left, the docking conformations produced by DIFFDOCK remain closely mirroring crystal structures without capturing the non-covalent interactions. Additionally, while traditional affinity prediction methods excel with crystal structures, they falter with less precise binding pose, presenting hurdles for real-world applications [6–8].

In this work, we introduce INTERFORMER, a novel model designed to alleviate the *interaction-aware problems* in boosting protein-ligand docking and affinity prediction. First, we proposed an interaction-aware mixture density network to model non-covalent interaction, specifically focusing on the *hydrogen bonds* and *hydrophobic interactions* present in the crystal structure. As illustrated in Fig.1 Right, Interformer can accurately produce specific interactions in the binding pose. Second, we introduce a pseudo-Huber loss function, harnessing the power of contrastive learning to guide the model to discriminate between favorable and unfavorable binding poses. These innovations are built upon a Graph-Transformer framework [9, 10], which has demonstrated superior performance compared to GNN-based models in various graph representation learning tasks [11]. An additional advantage of INTERFORMER is to interpret the internal mechanisms of protein-ligand interactions by examining the fusion coefficient of the mixture destiny network (MDN).

We validated our approach on both widely-used and our in-house datasets. Experiments on PDBBind timesplit benchmark [5, 12] show that INTERFORMER achieves a TOP-1 prediction performance of 54.9% with L-RMSD less than 2Å, a 16.7% improvement over DIFFDOCK. The improvements come from the better ability to capture non-covalent interactions between ligand and protein, which leads to generate less ambiguous conformation, is essential for successful performance in downstream tasks. Moreover, results of affinity prediction demonstrate that INTERFORMER can achieve comparable performance with DL-based models on CASF-2016 [13] benchmark while the best among all types of models on ours in-house benchmark. This confirms that the model exhibits pose-sensitive and strong generalization capabilities, demonstrates practical value and can be employed in real-world pharmaceutical research projects. In summary, the **main contributions** of this work are:

- Our study reveals the side effect of overlooking the protein-ligand non-covalent interaction modelling, which not only makes the existing DL-based models sub-optimal but also presents obstacles to their interpretability and generalization.
- To accomplish this goal, we introduce a novel interaction-aware mixture density network for small molecule docking. This proposed mechanism assists in making the binding pose more plausible, in addition to improving interpretation and generalization. It also demonstrating substantial improvements and setting a new benchmark in the domain.
- Furthermore, we pioneer the application of contrastive learning in the affinity task. Our model can predict a plausible affinity even with less accurate binding poses. This pose-sensitive capability offers a more robust and high-performing affinity prediction in real-world drug development scenarios.
- At the end, We have attempted to apply the Interformer in a real pharmaceutical pipeline, and have successfully obtained small molecules with activity levels of 1nM and 16nM in two separate projects.

2 Results

2.1 The Interformer model

Interformer is a deep learning model trained on crystal structure data and corresponding affinity values. Its architecture is inspired by the Graph-Transformer, originally proposed for graph representation learning tasks. Our proposed model is designed to perform two tasks: protein-ligand docking and protein-ligand affinity prediction. Initially, the model takes a single ligand 3D conformation and protein binding site structure as input. Graph representations, as illustrated in Fig.2a, are widely utilized in various methods to illustrate ligand and protein. In theses representations, nodes represent atoms, and edges to indicate proximity between two atoms. We employ pharmacophore atom types [14] as node features, and use the Euclidean distance between two atoms as edge feature. Theses Pharmacophore atom types provide essential chemical information, thus enabling the model to better comprehend specific interactions such as *hydrogen bonding* or *hydrophobic interactions*. A more detailed description of these features is provided in the Appendix.A.1.

In the second stage, the docking pipeline, as illustrated in Fig.2b, processes the node feature and edge feature from the protein and ligand through Intra-blocks. They are designed to capture intra-interactions within the molecule, thus producing updated node features for each atom of the protein and ligand. These updated node features are then fed into Inter-blocks, which are designed to capture inter-interactions between protein and ligand atom pairs, leading to further updated node and edge features. The edge output layer subsequently combines these two sets of features to generate an Inter-representation for each protein-ligand atom pair. Subsequently, the Inter-representation is processed by an Interaction-aware Mixture Density Network (MDN). This network predicts the parameters of four Gaussian function for each protein-ligand atom pair, each constrained by different possible specific interactions. The first two Gaussian functions encapsulate all types of pair interactions, while the third one signifies the hydrophobic interaction, and the fourth one denotes the hydrogen bond

interaction. By integrating these four Gaussian functions, we derive a Mixture Density Function (MDF), representing the conditional probability density function of distance for any given protein-ligand atom pair. This MDF can serve as an energy function to estimate the most probable distance between the protein atom and its corresponding ligand atom. The individual modeling each specific interaction distribution is crucial, as the third and fourth terms, which represent specific interactions, constitute 50% of our MDF. Therefore, the docking pose produced by our MDF are inherently depends on specific interactions. In contrast, other methods like [15, 16] utilize ten MDFs to model all types of protein-ligand pairs, thereby mixing all types of pairs and overlook the importance of specific interactions. Alternatively, methods like DiffDock solely minimize the Root Mean Square Deviation (RMSD) as the loss function, resulting in the docking process merely approximating the crystal structure, rather than capturing critical specific interactions. Finally, the MDF of all protein-ligand pairs is aggregated into a sum of energy functions, which is then introduced into a Monte Carlo (MC) sampling method for generating top candidate ligand conformations relative to its protein target. The MC sampling initially positions the ligand in various locations within the protein binding sites, and assign random torsion angles, then seeks to minimize the given energy function with respect to the ligand conformation. By aggregating all candidates sorted by energy value, we can obtain top candidate docking poses. A more detailed description of the sampling method is provided in the Appendix.B.2.

In the third stage, PoseScore and affinity prediction pipeline, as illustrated in Fig.2c. A new edge feature is updated by the new protein-ligand atom distance and specific interaction edge features derived from the generated docking pose. The node and edge features are then processed through Intra and Inter blocks to yield implicit interactions. A virtual node collects the entire binding pose information through self-attention mechanism. Finally, the binding embedding of virtual node is fed into the affinity and pose layer to predict the binding affinity value and the confidence pose score for the corresponding docking pose. By incorporating poor poses, a contrastive pseudo Huber loss function is utilized to guide the model in discerning whether a pose is good or poor. The training objective ensures that the model predicts a lower value for a poor pose and a higher value for a good pose. The primary distinction between a good and poor pose lies in their interactions. This strategy assists the model in learning crucial interactions rather than artificial features, a property we refer to as *pose-sensitive*, which demonstrates superior performance in real-world drug development projects.

2.2 Performance of Molecule Docking

Visualization of Energy Score Function

Typically, in drug development such as small molecule optimization, it is crucial to consider specific protein-ligand interactions in screening or downstream calculation scenarios. As shown in Fig.3a, demonstrated the predicted gaussian confusion coefficient α value among all potential hydrogen bond and hydrophobic pairs. α controlling the weight of hydrogen bond and hydrophobic term. It can be seen that Interformer correctly predicts the existence of hydrogen bond between ligand atom 20 and protein

atoms 90 and 134, and hydrophobic interactions between ligand atom 26 and protein atom 129 and 154. As shown in Fig.3b, demonstrated the predicted five MDNs. The two predicted MDNs with hydrogen bond pairs are noticeably on the left, indicating that the probability reaches 0.14 and 0.18 when the two atoms are close, and the mean μ of the hydrophobic MDN is at the position of $d = 0.8$, with the maximum probability being 0.23. The other two MDNs for non-interactions pairs, with d at 2.5 and 2.9, indicating that the two atoms should be relatively far apart during sampling. This analysis indicates that the Interformer can effectively predict reasonable MDNs. As shown in Appendix Fig.A3, Interformer generates four examples of binding poses. It can be seen that the generated binding poses directly exists hydrogen bonds or hydrophobic interactions, and theses binding poses can be forwarded seamlessly to CADD software without any optimization.

Interaction Statistic

The primary objectives of most deep learning methods aim to minimize the RMSD between the docked ligand and the crystal ligand. However, the generated docking poses tend to closely resemble the crystal ligand rather than adhering to physical principles such as non-covalent interaction. We have previously provided a sample that demonstrates the ability of our proposed method to predict a reasonable energy function. For a quantitative analysis, we identified the number of hydrogen and hydrophobic pairs in the crystal structure that were also present in the docking poses. This was achieved using PLIP [17], an open source tool that determines the formation of specific interactions based on physical rules. Upon evaluation of the PDBBind time-split testset, Diffdock could only capable of recovering an average of 22.76% of hydrogen bonds and 9.39% of hydrophobic interactions. In contrast, Interformer could recover an average of **51.7%** of hydrogen bonds and **59.4%** of hydrophobic interactions. For a more comprehensive view of these statistic, Fig.3c displays a histogram that outlines the count of different hit rates ranges. This further confirms that Interformer relies on non-covalent interactions rather than a simplistic data-driven strategy.

Docking accuracy

We employed the PDBBind time-split testset to calculate the success rate of docking poses with an RMSD of less than 2Å. We compared our results with those obtained from other methods as described in the Diffdock paper, as illustrated in Table 1. Our proposed method significantly surpassed all previous methods, achieving a top-1 success rate of 43.8% without the use of the pose score model. With the inclusion of a pose score model, the top-1 success rate rose to **54.9%**, which is 16.7% higher than the state-of-the-art method, DiffDock. Furthermore, the top-5 success rate reached an impressive **70.7%**, indicating that our proposed method generates a greater diversity and quality of binding poses among the top-ranked. The inference running time of the model is approximately one second, with the MC sampling algorithm being the most time-consuming component, resulting in a time cost similar to that of Gnina[18].

2.3 Performance on Affinity Prediction Task

Negative samples enhance pose-sensitive

Most methods performs well in predicting affinity based on crystal structure. However, in practical drug development scenarios, obtaining such realistic conformation is a challenge, predictions on poor binding pose often leading to substantial overfitting. We introduce a comparison between Interformer models that solely utilize crystal structure and those that also incorporate negatives samples. On PDDBind time-split testset, the affinity model using only crystal structure presents Pearson correlation coefficient $R=0.003$ between predicted affinity value and RMSD. However, when negative samples are used, the affinity model reaches a correlation of -0.545, and the PoseScore model achieves the highest correlation of -0.669. Additionally, in Fig.3d demonstrates that the affinity model without negatives samples keeps consistent predictions for good and poor binding poses, suggesting that the model does not leverage any non-covalent interaction features. In contrast, with negatives samples, both affinity and PoseScore models can predict lower values for binding pose with larger RMSD, demonstrating Interformer capacity to discern between good and poor poses.

Affinity Evaluation

In the small molecule development, not only is the accurate prediction of binding pose essential, but so is the prediction of affinity based on these poses as it directly impacts the performance of virtual screening or small molecule optimization. As the CASF2016 open benchmark often be used for comparison, many models have shown a tendency to overfit on this testset, leading us to question its persuasiveness. Therefore, we consider to use our internal testset for affinity performance comparison as follows:

- **CheEMBL-Kinase:** we curated data solely from the kinase family available on ChEMBL, resulting in 45 distinct protein targets and 3443 data points. The final assessment was based on the average affinity correlation across each target.
- **LSD1 Inner:** This is our in-house pharmaceutical pipeline, guided by Interformer model. It focuses on optimizing small molecules for the LSD1 target. In total, the project produced 54 novel small molecules, with the most potent achieving an affinity level of 0.7nM.
- **COVID-19 Test:** We sourced data on small molecules involving four types of covalent bonds from patents for COVID19 mPro target, yielding 142 data points. The final evaluation was based on the average affinity correlation for each covalent bond type.
- **COVID-19 Inner:** This is our in-house pharmaceutical pipeline, guided by Interformer model. It focuses on optimizing covalent type small molecules for the Covid19 mPro target. The project produced 18 novel small molecules, with the most potent achieving an affinity level of 16nM.

Table2 presents the performance of various models on these four real-world internal test sets, with all conformations in these test sets generated and predicted for affinity by the respective methods. On average, the affinity module that does not utilize negative samples training only achieves a correlation of 0.01, while the affinity module

incorporating negative samples reaches a correlation of 0.461. In the kinase dataset, it surpasses the SOTA Gnina model by 3.7% in correlation. In the LSD1 internal test set, it outperforms Gnina by a correlation of 2.9%. In the COVID-19 test, as Gnina can not handle docking with covalent bonds, it was not tested, and our method outperforms glide covalent by 8.5% in correlation. In the COVID-19 inner test set, it significantly outperforms by 19.3%. Due to pose-sensitivity, the model significantly outperforms models trained only on crystal structures in terms of generalization. Furthermore, on the Public Benchmark CASF2016, as shown in Appendix Table A4, the affinity model also demonstrates strong predictive capabilities, achieving a performance of $R=0.784$ when predicting on crystal structures.

Real world drug design application

In early 2022, guided by Interformer as the predictive model for affinity, we initiated an in-house small molecule optimization project targeting LSD1, a potential treatment for cancer. We developed two series of small molecules, resulting in 54 novel compounds, with the most potent reaching an affinity of 0.7nM. This compound (Cpd 27 pg 64) was later tested in oral and intravenous mouse trials, with half-lives ($t_{1/2}$) of 5.86 and 8.33 hours, and AUC(0-t) of 6528 and 11502h.ng/mL respectively, as shown in Fig. 4a. Amidst the pandemic in late 2022, we also attempted to target the popular COVID-19 viral target mPro, experimenting with macrocyclic modifications of small molecules. We designed 18 small molecules, with the most potent reaching an activity of 16nM, as shown in Fig. 4b. The correlation of Interformer in the LSD1 and COVID-19 projects was 0.511 and 0.672 respectively, as shown in Fig. 4c and 4d. Fig. 4e and f show the correlation of activity prediction models trained only on crystal structures, which was only 0.277 and -0.211. This real-world pharmaceutical scenario validated that models trained only on crystal structures perform very poorly when the actual docking pose is not necessarily accurate. The docking poses generated by Interformer greatly influenced the design of small molecules in the project, as shown in Fig. 4g. The small molecule maintains the key hydrogen bond interaction with LYS-661 of the reference molecule, while modifying other functional groups to form two additional hydrogen bond interactions, improving the metabolic stability of the newly designed small molecule while maintaining an activity level of 1nM, similar to the reference molecule.

3 Discussion

In conclusion, we have presented Interformer, a novel deep learning generative model designed to perform protein-ligand docking and affinity prediction. Our model, built on graph-Transformer backbone, offers robust embedding capabilities and is specifically tailored to model non-covalent interactions, thereby facilitating the effective learning of various atomic interaction types. By incorporating negative samples and utilizing contrastive learning for affinity prediction, Interformer achieves an impressive 54.9% TOP-1 success rate of molecule docking, significant improvement over the state-of-the-art models on the PDBBind time split benchmark dataset, and suppress all SOTA models in real-world internal test sets in affinity prediction task.

These advancements in performance are primarily due to the ability of Interformer to capture essential non-covalent interactions in binding pose, which subsequently enabling improved performance in downstream tasks. Furthermore, Interformer demonstrates strong generalizability in affinity prediction tasks, suppress all notable SOTA methods, showcasing its potential to impact computational biology and accelerate drug discovery efforts. As a future direction, we envision integrating this model into a broader pipeline of computational biology applications and further optimizing its performance for more diverse protein-ligand complexes.

4 Methods

4.1 Graph-Transformer

Graph Representation in Protein-Ligand Complex

Graph representation is well-suited for depicting the structure of protein-ligand complex, shown in Fig.2a. Here, $G = (V, E)$ denote a graph represents the complex, where $V = \{x_1, \dots, x_n\}$ be a set of nodes(atoms) within the complex, n is the number of nodes and x is node feature. Furthermore $E = \{e\}$ represents a set of edges e of each nodes pairs in G , where e is determined by the condition $d < 6\text{\AA}$ ¹, where d denotes the Euclidean distance between a pair of atoms.

Graph Transformer

The graph transformer encoder leveraging a powerful self-attention mechanism [9] excels at learning node relationships. It has achieved superior performance in multiple graph representation learning tasks [11]. It adds a bias term to the Multi-Head Self-Attention(MHA) of the transformer to obtain graph structure information. It operates on two inputs, the node features x and edge features e : the query $Q \in \mathbb{R}^{n \times d}$, key $K \in \mathbb{R}^{n \times d}$, and value $V \in \mathbb{R}^{n \times d}$ are linear transformations of x and can be described as follows:

$$\text{Attn}(x, e) = \text{softmax}(A)V \quad (1)$$

$$A = \frac{QK^T}{\sqrt{d}} + z(e) \quad (2)$$

where d denotes the dimension of the model. The bias term, denotes as $z \in \mathbb{R}^{n \times n \times d}$, functions as a linear layer that transform edge feature e from the dimension of edge feature d_e into model dimension d when two nodes are connect; otherwise, it is set to $-\text{inf}$. $A \in \mathbb{R}^{n \times n \times d}$ denotes self-attention weight, which captures the similarity between the query Q and the key K , then A is used to update the query representation through a weighted sum of the values V .

¹The maximum distance for all specific interactions within the PDDBank, which is statistically analyzed using PLIP[17]

4.2 Interformer

Masked Self-Attention (MSA)

Traditional self-attention mechanism permit each node to attend to all other nodes through attention weights. We propose MASKED SELF-ATTENTION that restricts certain nodes to attend only to specific designated nodes by a mask M , shown in Fig.5a.

$$\text{MaskAttn}(x, e, M) = \text{softmax}(A \odot M)V \quad (3)$$

where M is applied to A through element-wise multiplication denoted by $\odot$.

Intra-blocks

To better capture essential feature, we suggest the encoding of intra-interactions, leading us to propose Intra-blocks as our initial module. This module takes the node feature x and edge features e as inputs. To exclusively concentrate on the internal interaction within molecule, we formulate an Intra-mask M' for MSA as follows:

$$h = \text{MaskAttn}(x, e, M') \quad (4)$$

$$M'_{ij} = \begin{cases} 1 & i \in L, j \in L \text{ or } i \in P, j \in P \\ -inf & \text{otherwise} \end{cases} \quad (5)$$

where L and P represents a set of ligand and protein atoms indices respectively, lastly we get the updated node feature $h \in \mathbb{R}^{n \times d}$.

Inter-blocks

To capture the inter-interaction and predict a new relationship between ligand and protein atoms, we employ Inter-mask M'' that ensures atoms can attend to others atoms within a different molecule. This process can be described as follows:

$$h' = \text{MaskAttn}(h, e', M'') \quad (6)$$

$$e' = A + \text{FFN}(\text{LN}(A)) \quad (7)$$

$$M''_{ij} = \begin{cases} 1 & i \in L, j \in P \\ -inf & \text{otherwise} \end{cases} \quad (8)$$

where a new edge feature e' can be updated by a residual of A , and both layer normalization (LN) and feed forward network (FFN) are applied to A in each block, then e' is applied to MSA to update node feature h' .

Edge Output Layer for Protein-Ligand Docking

To predict the energy between protein-ligand atom pairs, the fusion of node and edge features can provide a more comprehensive inter-feature. As shown in Fig.5b, and can be described as follows:

$$h^a = \text{FFN}(h'_i h'_j + \frac{e_{ij}^o + e_{ji}^o}{2}), i \in L, j \in P \quad (9)$$

where e^o is the output edge feature obtained by applying a linear weight W and LN to the edge feature e' . Initially, we compute the mean the semi-positive matrix e^o and then add pair-wise node features. Finally, we employ a FFN to learn the protein-ligand inter-feature $h^a \in \mathbb{R}^{L \times P \times d}$.

4.3 Training

The training procedure consists of two stages. Initially, the energy model is trained using the crystal structure, which is subsequently employed to generate negative sample poses. Following this, both positive and negative sample poses are used to train a the pose score and affinity prediction model. The training objectives of the three modules will be elaborated in greater detail below:

Interaction Aware Mixture Density Network

To model specific interactions and output an energy function for the sampling algorithm to generates binding poses, we predict the interaction aware mixture density network (MDN) of each pair of ligand $i \in L$ and protein $j \in P$ atoms using the inter-feature h^a . This MDN is formed by the weighted sum of four mixed Gaussian functions. The first two can be considered as regular interaction forces, the third represents hydrophobic interactions and is modeled only in hydrophobic atom pairs, and the fourth represents hydrogen bonding and is modeled only between hydrogen bond donor and acceptor. The details are as follows:

$$\begin{aligned} \ell_{MDN}(d) &= -\log P(d|h^a) \\ &= -\log (p_1 + p_2 + \gamma p_3 + \omega p_4) \end{aligned} \quad (10)$$

$$p_k(d) = \alpha_k \mathcal{N}(d|\mu_k, \sigma_k), \text{ where } k = 1, 2, 3, 4 \quad (11)$$

$$\gamma_{i,j} = \begin{cases} 1 & i \in \mathcal{H}, j \in \mathcal{H} \\ 0 & \text{otherwise} \end{cases} \quad (12)$$

$$\omega_{i,j} = \begin{cases} 1 & i \in \mathcal{A}, j \in \mathcal{D} \text{ or } j \in \mathcal{A}, i \in \mathcal{D} \\ 0 & \text{otherwise} \end{cases} \quad (13)$$

where fusion coefficient α , mean μ and variance σ are obtained by applying a linear weight W to inter-feature h^a , and they are learnable parameters of each gaussian distribution $\mathcal{N}$ in the MDN for each pair of atoms. Given the Van der Waals (VdW) radius distance d , we can compute probability p of each term in MDN, in the third

term p_3 , α is non-zero only on hydrophobic $\mathcal{H}$ pair, in the fourth term p_4 , α is non-zero only on hydrogen bond donor $\mathcal{D}$ and acceptor $\mathcal{A}$ pair. Then the final probability of each atom pair P is derived by summing four $\mathcal{N}$ multiplied by α . Lastly, the model can optimize model parameters through the negative log likelihood loss function ℓ_{MDN} .

Pose Score Module

MC sampling generates multiple candidates docking pose based on energy function P . However, the best pose may not necessarily be at the top of the list, necessitating an additional model to further re-rank the docking poses for enhancing the success rate of docking. In this work, we apply a FFN on the virtual node feature v to predict p , which indicates whether the input binding pose is correct. This is determined by calculating if the docking pose is less than 2Å the RMSD from the crystal pose. The training objective is a binary loss function.

Affinity Module

Docking algorithms typically rank the affinity value of ligand using an energy function. However, these energy functions merely fit the crystal structure rather than the actual affinity value. To address this, we employ a FFN on virtual node feature v to predict the experimental affinity value y^2 . The training objective is a constrative pseudo-Huber loss function, which ensures the model predicts a lower value for a poor pose and a higher value for a good pose. The primary difference between these two poses lies in their interactions, such strategy aids the model in learning essential interaction rather than artificial feature, we refer it as *pose-sensitivity* as follows:

$$\ell_{affinity}(p, y, \hat{y}) = \sigma^2 \left(\sqrt{1 + \frac{diff^2}{\sigma}} - 1 \right) \quad (14)$$

$$diff = \begin{cases} \hat{y} - y & p = 1 \\ \hat{y} - y & p = 0, \hat{y} > y \\ 0 & p = 0, \hat{y} < y \end{cases} \quad (15)$$

where $\hat{y}$ is the predicted affinity value, and we select huber loss hyperparameter σ to be 4.

Join training further enhance pose-sensitive by virtual node

We introduce an extra virtual node to represent the entire information of the binding pose, the node feature of v is learnable embedding, v is connected to all other nodes within the graph G via a mask M^v in Inter-blocks as follows:

²we use IC50/Kd/Ki as affinity unit, are normalized by -log, with a higher value indicating greater affinity.

507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552

$$M_{ij}^v = \begin{cases} 1 & i = v, j \in \{L, P\} \\ -inf & \text{otherwise} \end{cases} \quad (16)$$

PoseScore and our contrastive learning affinity module share the same purpose of determining the correctness of a binding pose. Given the strong complementarity between these two tasks, we choose a single model, based on the same virtual node v , to train and predict both the PoseScore and Affinity tasks simultaneously as follows:

$$\ell_{joint} = \ell_{affinity} * 0.1 + \ell_{pose} \quad (17)$$

Since the affinity value is approximately ten times greater than the binary value, the loss for training the affinity is reduced by a factor of ten.

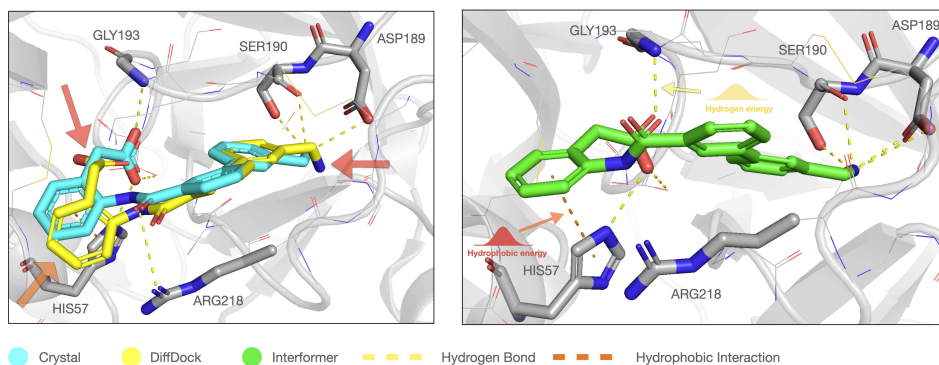


Fig. 1 *Left*: The incorrect case generated by DIFFDOCK. Pink and orange arrows indicates improperly formed hydrogen bond and hydrophobic interactions (RMSD: 1.13, PDB ID: 6qmt) *Right*: The correct binding pose generated by INTERFORMER, where the predicted interaction energy recovers almost all hydrogen bonds and hydrophobic interactions (RMSD: 0.67).

Appendix A Training Related

A.1 Node Feature and Edge Feature

We employ one-hot pharmacophore features to represent the node features of proteins and small molecules, which also include cofactors within the proteins. Water molecules are temporarily excluded in our system. These pharmacophore features were initially proposed by X-Score, comprising a total of 23 dimensions, as shown in Fig.A1. The first edge feature is the euclidean distance between two atoms, and when complex binding pose is available, the second edge feature is one-hot specific non-covalent interaction between two atoms, total 6 dimensions as shown in Fig.A2.

A.2 Batch Balance Sampling

Due to the long-tailed distribution of the dataset in the problem under consideration, certain proteins may have over four hundred PDBs (e.g., Carbonic Anhydrase2), shown in Fig.A1, while others may only have a few or even just one PDB. Additionally, negative samples are taken into account, leading us to modify the original sampling strategy as outlined in Algorithm1. In the proposed approach, we aim to optimize the model by including one protein and all of its ligands in a single batch, which helps balance the weight of each protein in a long-tailed distribution. The sampling of negative samples is restricted to having an equal number of positive samples in one batch, and the same ligand.

Models	TOP-1 RMSD		TOP-5 RMSD		Time (s) [↓]
	%<2Å [↑]	Median [↓]	%<2Å [↑]	Median [↓]	
<i>Traditional Docking Method</i>					
QVINA W	20.9	7.7	-	-	49
SMINA	18.7	7.1	29.3	4.6	126
GLIDE	21.8	9.3	-	-	1405
<i>AI-Based Docking Method</i>					
GNINA	22.9	7.7	32.9	4.5	127
EQUIBIND	5.5	6.2	-	-	0.04
TANKBIND	20.4	4.0	24.5	3.4	2.5
P2RANK+Gnina	28.8	5.5	38.3	3.4	127
EQUIBIND+Gnina	28.8	4.9	39.1	3.1	127
NEURALPLEXER	30.1	5.2	-	-	-
DIFFDOCK (40)	38.2	3.3	44.7	2.4	40
E3BIND	23.4	3.8	-	-	2.2
<i>Our Interformer</i>					
w/o Pose Score	43.8	2.4	62.5	1.4	105
w/ Pose Score	54.9	1.8	70.7	1.3	107

Table 1 Time-Based PDBBind docking accuracy benchmark(n=363). Shown is the percentage of predictions with RMSD less than 2Å, the median RMSD, and the method average runtime. DiffDock(40) refers to the best result obtained by sampling the top 40 candidates binding poses.

Methods	ChEMBL-Kinase		LSD1 Inner		Covid-19 Test		Covid-19 Inner	
	P	S	P	S	P	S	P	S
AUTO DOCK VINA	0.094	0.092	0.402	0.554	-	-	-	-
GLIDE SP	-	-	0.272	0.417	-	-	-	-
GLIDE XP	-	-	0.350	0.481	-	-	-	-
GNINA	0.188	0.184	0.437	0.516	-	-	-	-
GLIDE COVALENT	-	-	-	-	0.351	0.195	0.479	0.476
<i>Our Interformer-Affinity</i>								
w/o NEG	0.131	0.117	0.277	0.368	-0.137	-0.034	-0.211	-0.377
w/ NEG	0.225	0.210	0.511	0.619	0.436	0.376	0.672	0.535

Table 2 Performance of the affinity prediction of various methods on our internal test, in terms of Pearson (P) and Spearman (S) correlation coefficient. W/ NEG and W/O NEG denotes the use of negatives pose as training data or without.

Appendix B Inference Related

B.1 Docking

Firstly, we employ the Universal Force Field (UFF)[19] to generate the minimized energy pose of the ligand from its SMILES representation. The minimized pose and the pose of the protein pocket are fed input into to the energy model to predict energy function E for each pair of protein-ligand atoms as follows:

Algorithm 1 Batch Balance Strategy	645
1: Input: Pairs Training Data $\mathcal{D} = \{\langle \mathbf{x}^n, \mathbf{y}^n \rangle\}_{n=1}^N$; Interformer θ ;	647
2: $T =$ Shuffle targets based on uniprot id	648
3: $Z = []$	649
4: for t in T do	650
5: $x_{pos} \leftarrow$ Select all positive samples from t	651
6: $x_{neg} \leftarrow$ Retrieve the same amount of x_{pos} negatives samples with the same ligand from t	652
7: $Z.append(x_{pos} + x_{neg})$	654
8: end for	655
9: Split Z by batch_size into B	656
10: Training by MiniBatch SGD method	657
11: for b in B do	658
12: Update θ with batch loss $\mathbb{E}_{\langle \mathbf{x}, \mathbf{y} \rangle \sim b}$	659
13: end for	660
14: return θ	661

Pharmacophore Atom Type	662
AliphaticCarbonXSHydrophobe	663
AliphaticCarbonXSNonHydrophobe	664
AromaticCarbonXSHydrophobe	665
AromaticCarbonXSNonHydrophobe	666
Bromine	667
Calcium	668
Chlorine	669
Fluorine	670
Iodine	671
Iron	672
Magnesium	673
Nitrogen	674
NitrogenXSAcceptor	675
NitrogenXSDonnor	676
NitrogenXSDonorAcceptor	677
Oxygen	678
OxygenXSAcceptor	679
OxygenXSDonorAcceptor	680
Phosphorus	681
Sulfur	682
SulfurAcceptor	683
Zinc	684
IsProtein	685

Table A1 The node features employed in Interformer consist of 22 pharmacophore atom types, along with an additional feature indicating whether the atom belongs to the protein or the ligand.

Type	Description
Distance	Euclidean distance between two atoms.
None	There is no explicit interaction between two atoms.
HydrogenBond	Whenever two centers of hydrogen accetor and hydrogen donnor come with a distance of <i>HYDROGEN_DIST_MAX</i> , hydrogen bond is reported.
HalogenBond	Halogen bond is similar to Hydrogen bond but for one of the side is halogen element.
SaltBridge	Whenever two centers of opposite charges come with a distance of <i>SALT_BRIDGE_DIST_MAX</i> , a salt ridge is reported.
PiStack	π -stacking for two aromatic rings is reported whenever their centers are within a distance of <i>PISTACK_DIST_MAX</i> , the angle deviates no more than <i>PIS-TACK_ANG_DEV</i> from the optimal angle of 90 for T-stacking or 180 for P-stacking
PiCation	π -Cation interactions are reported for each pairing of a positive charge and an aromatic ring if the distance between the charge center and the aromatic ring center is less than <i>PICATION_DIST_MAX</i> and the offset between the ring center and the charge is no larger than <i>PISTACK_OFFSET_MAX</i> .
Hydrophobic	Hydrophobic interactions are nonpolar atoms they are close together and they do no interact with water molecules.

Table A2 The edge features utilized in Interformer comprise the Euclidean distance and six specific interactions.

Type	Description
AutoBoxSize	20Å*20Å*20Å
AutoBoxAdd	4Å
GridResolution	0.125Å
Exhaustiveness	64
StepsOfEachMC	2000

Table A3 Monte Carlo sampling parameters. The movement space of ligand is set by the larger value between AutoBoxSize a cubic space centered on the small molecule and AutoBoxAdd an additional 4Å to the extreme coordinates of ligand. GridResolution specifies the finest grid unit during sampling. Exhaustiveness delineates the count of Monte Carlo chains, while StepsOfEachMC signifies the optimization iterations per Monte Carlo chain.

Methods	CASF-2016	
	Pearson	Spearman
<i>Traditional Docking Method</i>		
AUTODOCK VINA	0.604	0.528
GLIDESCORE SP	0.513	0.419
<i>Machine Learning Method</i>		
$\Delta vinaRF_{20}$	0.816	0.750
X-SCORE	0.631	0.604
<i>Deep Learning Method</i>		
GNINA	0.790	-
DEEPDOCK	0.525	0.495
RTMScore	0.507	0.563
<i>Our Interformer</i>		
ENERGY $\dagger$	0.536	0.527
AFFINITY w/o NEG $\dagger$	0.772	0.754
AFFINITY w/ NEG $\dagger$	0.784	0.778
ENERGY $\ddagger$	0.530	0.519
AFFINITY w/o NEG $\ddagger$	0.746	0.739
AFFINITY w/ NEG $\ddagger$	0.711	0.734

Table A4 Performances of scoring functions on CASF-2016 coresets n=285, in terms of Pearson correlation coefficient, Spearman ranking coefficient. $\dagger$ means using the crystal structure. $\ddagger$ means using the docking pose generated by Interformer.

$$E_{(i,j,d)} = - \sum_{i=1}^L \sum_{j=1}^P P(d_{i,j} | h_i^a, h_j^a) + S(d_{i,j}) \quad (B1)$$

$$S(d) = \begin{cases} -d^2 & d < 0 \\ 0 & d > 0 \end{cases} \quad (B2)$$

where S is a penalty term that calculates the intersecting Van der Waals radius d of two atoms, in order to prevent steric clashes. Lastly E is utilized as the energy function in MC sampling algorithm B.2 to sample candidate poses.

B.2 Monte Carlo Sampling

In order to generate candidate binding poses, we require the Interformer to predict the energy functions E of each ligand-protein atom pair. Monte Carlo (MC) sampling, a global optimization procedure, minimizes the energy of the conformer while preserving its chemical plausibility. To ensure this plausibility, the movement of the ligand must be accounted for in a space of $6+K$ variables, where three are for translation, another three for rotation, and K for the torsion angles of rotatable bonds.

In our implementation, the sampling method is a variant of the AutoDock Vina procedure applied in our MC Sampling. This process unfolds in three stages. Initially,

the 6+K variables of the ligand conformation undergo randomization until all ligand atoms reside within the protein pocket autobox (a 20Å³ box centered on the reference ligand or a box 4Å extend the maximum or minimum coordinates of the reference ligand). Following this, an iterative optimization of the 6+K variables is executed. Each iteration comprises a ligand mutation and a subsequent BFGS local gradient descent minimization based on energy function E , typically iterating 2000 times per MC chain. The mutation of the ligand involves one of three randomly selected operations: translating the entire ligand, rotating the entire ligand, or adjusting one of the K torsion angles. In the final stage, all energy-minimized conformers from all iterations are gathered, clustered, and sorted by energy function E , resulting in the selection of the top 10 candidate poses. Each MC chain can run independently in parallel, usually 64. The parameters of the MC sampling are presented in Table A3.

Contributions of MC sampling

We re-implemented the Monte Carlo sampling in C++, distinguishing our method from AutoDock Vina particularly in the computation of BFGS gradients. Instead of employing an approximation approach for differentiation, our method ensures precise derivation, thereby accelerating the docking speed up to 200 times compared to the PyTorch version. Additionally, we have modified the original grid map function module to operate at atom level rather than atom type level, thereby enabling it to adapt to the Interformer energy function or other potential functions like distance map. These modifications have led to significant improvements in both docking accuracy and speed.

Appendix C More Experiments

C.1 Docking Power

Docking Power is an evaluation method that measures whether the energy function can identify near-native conformations from a large number of decoys. We tested the Docking Power on the dataset from the RTMScore paper, using the CASF2016 coreset with only $< 2\text{\AA RMSD}$ conformations. Since native conformations are easily distinguishable, we used a test setting without native conformations to better reflect practical applications. Our method achieved a 95.7% success rate, a 12% improvement over traditional docking methods and a 2% improvement over RTMScore Figure A2.

Appendix D Future Works

There are several ideas that require further exploration to enhance the performance of our model. Here, we outline some directions for future work and potential improvements:

Inclusion of more interaction types: In our current model, we only considered hydrogen bonds, hydrophobic interactions. Other interactions, such as halogen bonds, ionic bonds and Pi-Pi stacking, have not been described. These additional interaction types can be incorporated into the model by adding atom-pair-specific Gaussian functions.

Refinement of energy function: We employed a single Gaussian function to represent individual interactions. Future work could focus on developing tailored probability functions for different interaction types to obtain a more accurate description of the energy landscape.

Water models: Our model does not account for the influence of water molecules explicitly. However, by including water molecules from the crystal structures that have relatively accurate positions, the model can account for water-filled pockets within the binding site during small molecule docking. This approach would enable the sampling of interactions involving water bridges.

Data augmentation with limited data: In this study, we adjusted the energy function using conformation data generated by the model. The data generated by the model, however, is limited. The development of a computationally efficient sampling method could potentially yield more data, resulting in higher accuracy during energy function adjustment.

References

- [1] Öztürk, H., Özgür, A., Ozkirimli, E.: Deepdta: deep drug–target binding affinity prediction. *Bioinformatics* **34**(17), 821–829 (2018)
- [2] Stanzione, F., Giangreco, I., Cole, J.C.: Use of molecular docking computational tools in drug discovery. *Progress in Medicinal Chemistry* **60**, 273–343 (2021)
- [3] Stärk, H., Ganea, O., Pattanaik, L., Barzilay, R., Jaakkola, T.: Equibind: Geometric deep learning for drug binding structure prediction. In: *International Conference on Machine Learning*, pp. 20503–20521 (2022). PMLR
- [4] Zhang, Y., Cai, H., Shi, C., Zhong, B., Tang, J.: E3Bind: An End-to-End Equivariant Network for Protein-Ligand Docking (2023)
- [5] Corso, G., Stärk, H., Jing, B., Barzilay, R., Jaakkola, T.: Diffdock: Diffusion steps, twists, and turns for molecular docking. *International Conference on Learning Representations (ICLR)* (2023)
- [6] Ballester, P.J., Mitchell, J.B.: A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking. *Bioinformatics* **26**(9), 1169–1175 (2010)
- [7] Wang, C., Zhang, Y.: Improving scoring-docking-screening powers of protein–ligand scoring functions using random forest. *Journal of computational chemistry* **38**(3), 169–177 (2017)
- [8] Karimi, M., Wu, D., Wang, Z., Shen, Y.: DeepAffinity: interpretable deep learning of compound–protein affinity through unified recurrent and convolutional neural networks. *Bioinformatics* **35**(18), 3329–3338 (2019) <https://doi.org/10.1093/bioinformatics/btz111> <https://academic.oup.com/bioinformatics/article-pdf/35/18/3329/48975803/bioinformatics.35.18.3329.pdf>

- 875 [9] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N.,
876 Kaiser, L., Polosukhin, I.: Attention Is All You Need. arXiv (2017). <https://doi.org/10.48550/ARXIV.1706.03762> . <https://arxiv.org/abs/1706.03762>
877
878
- 879 [10] Yun, S., Jeong, M., Kim, R., Kang, J., Kim, H.J.: Graph transformer networks.
880 Advances in neural information processing systems **32** (2019)
881
- 882 [11] Ying, C., Cai, T., Luo, S., Zheng, S., Ke, G., He, D., Shen, Y., Liu,
883 T.-Y.: Do transformers really perform badly for graph representation? In:
884 Thirty-Fifth Conference on Neural Information Processing Systems (2021).
885 <https://openreview.net/forum?id=OeWooOxFwDa>
886
- 887 [12] Liu, Z., Su, M., Han, L., Liu, J., Yang, Q., Li, Y., Wang, R.: Forging the basis
888 for developing protein–ligand interaction scoring functions. Accounts of chemical
889 research **50**(2), 302–309 (2017)
890
- 891 [13] Su, M., Yang, Q., Du, Y., Feng, G., Liu, Z., Li, Y., Wang, R.: Comparative assess-
892 ment of scoring functions: the casf-2016 update. Journal of chemical information
893 and modeling **59**(2), 895–913 (2018)
894
- 895 [14] Wang, R., Lai, L., Wang, S.: Further development and validation of empiri-
896 cal scoring functions for structure-based binding affinity prediction. Journal of
897 computer-aided molecular design **16**, 11–26 (2002)
898
- 899 [15] Zhang, X., Zhang, O., Shen, C., Qu, W., Chen, S., Cao, H., Kang, Y., Wang, Z.,
900 Wang, E., Zhang, J., *et al.*: Efficient and accurate large library ligand docking
901 with karmadock. Nature Computational Science **3**(9), 789–804 (2023)
902
- 903 [16] Méndez-Lucio, O., Ahmad, M., Rio-Chanona, E.A., Wegner, J.K.: A geometric
904 deep learning approach to predict binding conformations of bioactive molecules.
905 Nature Machine Intelligence **3**(12), 1033–1039 (2021)
906
- 907 [17] Adasme, M.F., Linnemann, K.L., Bolz, S.N., Kaiser, F., Salentin, S., Haupt, V.J.,
908 Schroeder, M.: PLIP 2021: expanding the scope of the protein-ligand interaction
909 profiler to DNA and RNA. Nucleic Acids Res **49**(W1), 530–534 (2021)
910
- 911 [18] Ragoza, M., Hochuli, J., Idrobo, E., Sunseri, J., Koes, D.R.: Protein–ligand scor-
912 ing with convolutional neural networks. Journal of Chemical Information and
913 Modeling **57**(4), 942–957 (2017) <https://doi.org/10.1021/acs.jcim.6b00740>
914
- 915 [19] Rappé, A.K., Casewit, C.J., Colwell, K., Goddard III, W.A., Skiff, W.M.: Uff,
916 a full periodic table force field for molecular mechanics and molecular dynam-
917 ics simulations. Journal of the American chemical society **114**(25), 10024–10035
918 (1992)
919
920

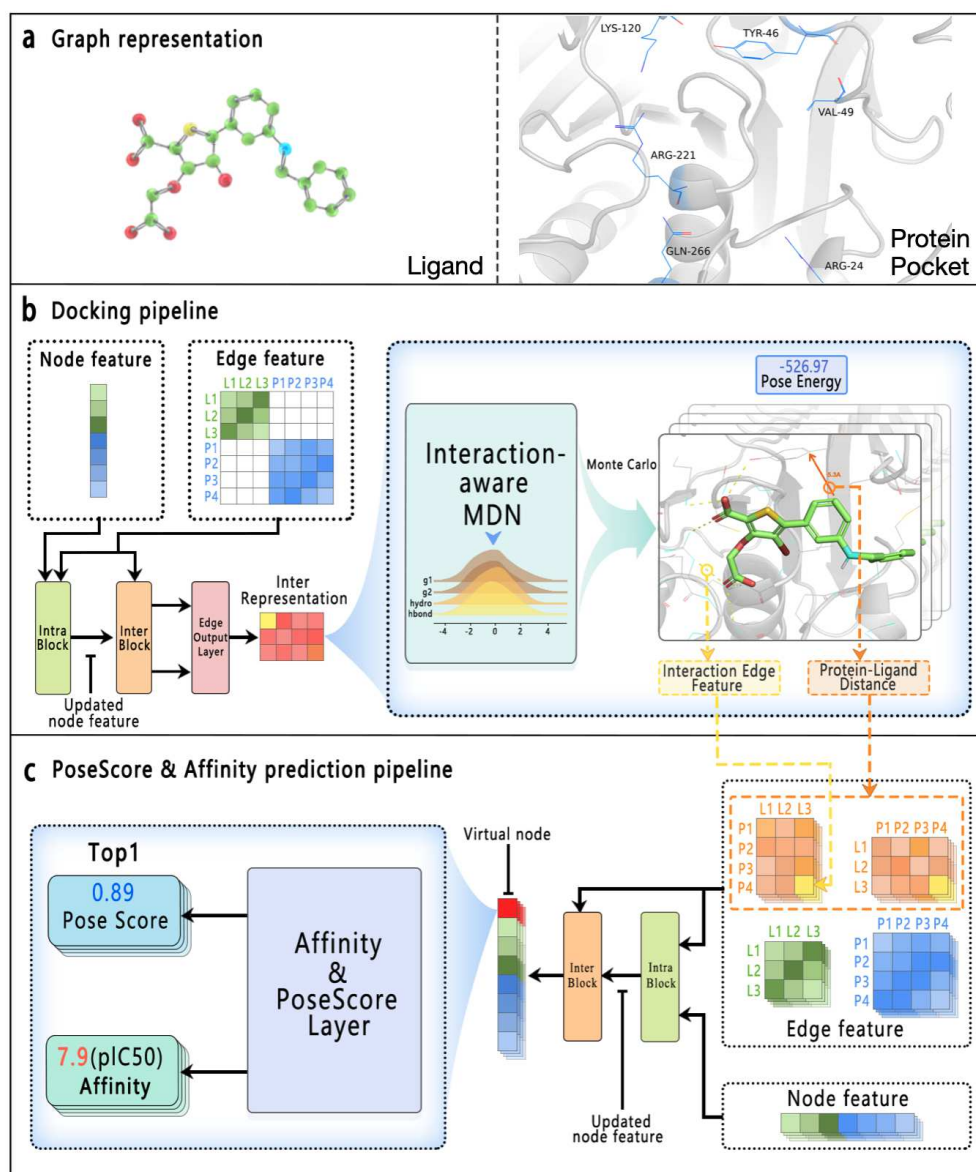


Fig. 2 An overview of Interformer model architecture. a) Graph representation, **b)** Docking pipeline, **c)** PoseScore and Affinity prediction pipeline.

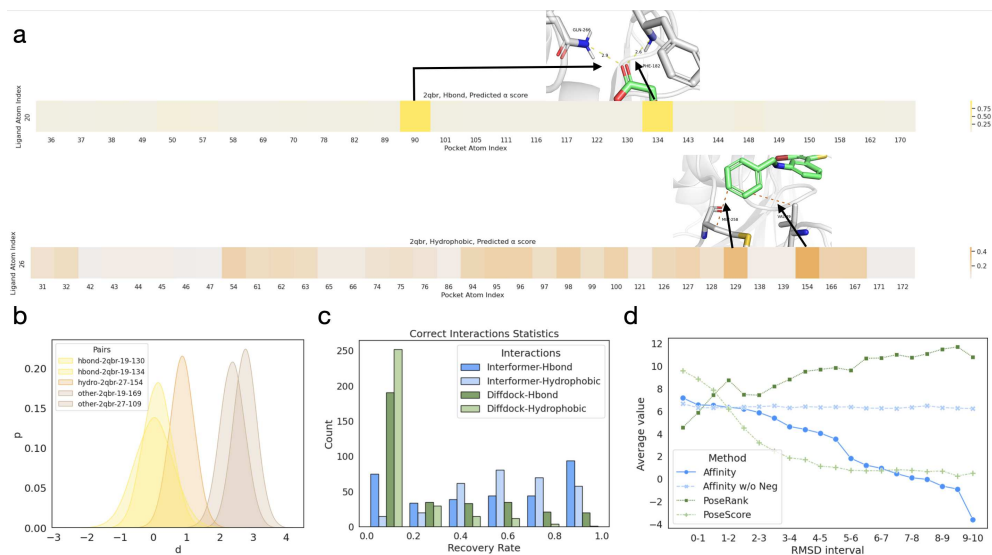


Fig. 3 Evaluations on protein ligand docking and affinity prediction tasks. **a)** The upper diagram illustrates the predicted α value for the hydrogen bond interaction between the 19th ligand atom (a hydrogen bond acceptor) and all other protein atom hydrogen bond donors. The model successfully predicts the highest alpha values for interactions 19-130 and 19-134, which correspond to the existence of a hydrogen bond in the crystal structure. A similar analysis applies to the lower diagram, but for a different hydrophobic pair. **b)** The predicted five MDN distributions, x axis is VdW radius distance of the atoms pair d and y axis is the corresponding probability. Yellow is hydrogen bond pairs, orange is hydrophobic pairs, and brown is no interaction pairs. **c)** The histogram depicts the recovery rate of hydrogen bonds and hydrophobic interactions from the crystal structure. The x-axis represents the percentage of correct interactions present in the docking poses, while the y-axis corresponds to the total number of poses within this range. **d)** We utilized RMSD intervals of 1 Å to calculate the average predictions of all conformations for various methods. PoseRank values were determined by sorting the energy from 1 to 10. To facilitate display, we multiplied the predicted values of PoseScore by 10.

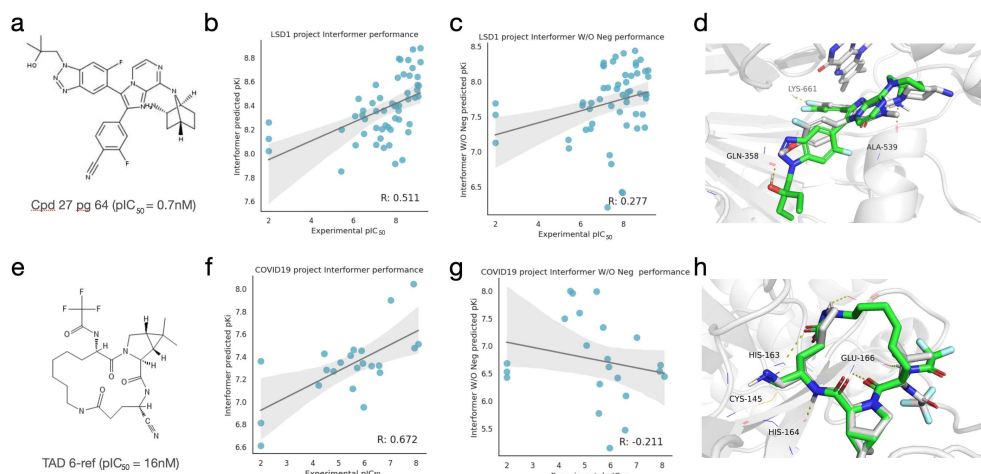


Fig. 4 Evaluation of two real-world drug design projects. **a)** The most potent compound, Cpd27 in the LSD1 project achieved an affinity of 0.7nM. **b)** The correlation plot between the Interformer predicted affinity value and experimental pIC₅₀ in LSD1 project. **c)** Similar to the setting of B), Interformer was trained without the use of any negatives sample data. **d)** The docking pose of Cpd27 aligns with the PDB 6w4k, it holds crucial hydrogen bond interaction with LYS-661 and creating two additional hydrogen bond interactions with GLN-358 and ALA-539. **e)** The most potent compound, TAD 6-ref, in the COVID-19 project achieved an affinity of 16nM. **f)** The correlation plot in the COVID-19 project. **g)** Similar to the setting of F), Interformer was trained without the use of any negatives sample data. **h)** The docking pose of TAD 6-ref align with the PDB 7rfs, and is designated to form a macrocyclic to stabilize the ligand.

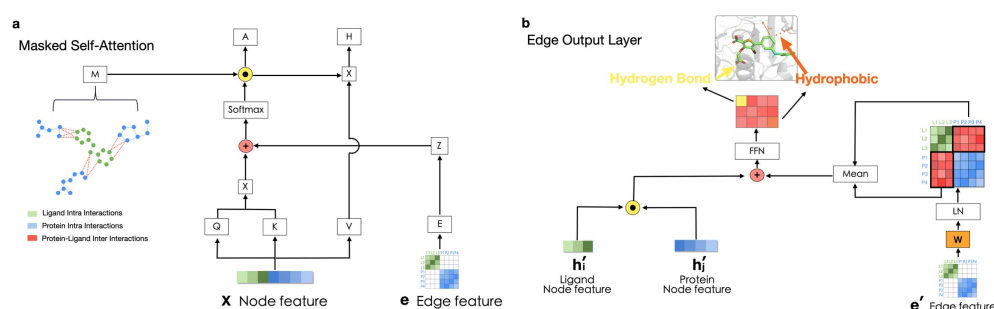


Fig. 5 Architecture of two modules. **a)** Masked Self-Attention, this is a modified self-attention mechanism that uses an attention mask M , to restrict nodes from attending to only certain other nodes. Besides it can incorporate edge features through element-wise addition. $\odot$ denotes as Hadamard product; $\oplus$ denotes element-wise summation; $\boxtimes$ denotes dot product. **b)** Edge Output Layer, a two-dimensional node fusion feature of the ligand and protein is obtained through dot product, which, when added to the edge feature learned through W , creates a new feature that captures the interaction between the ligand and ligand. This interaction feature is then learned through a feed-forward network(FFN), resulting in a comprehensive protein-ligand Inter representation.

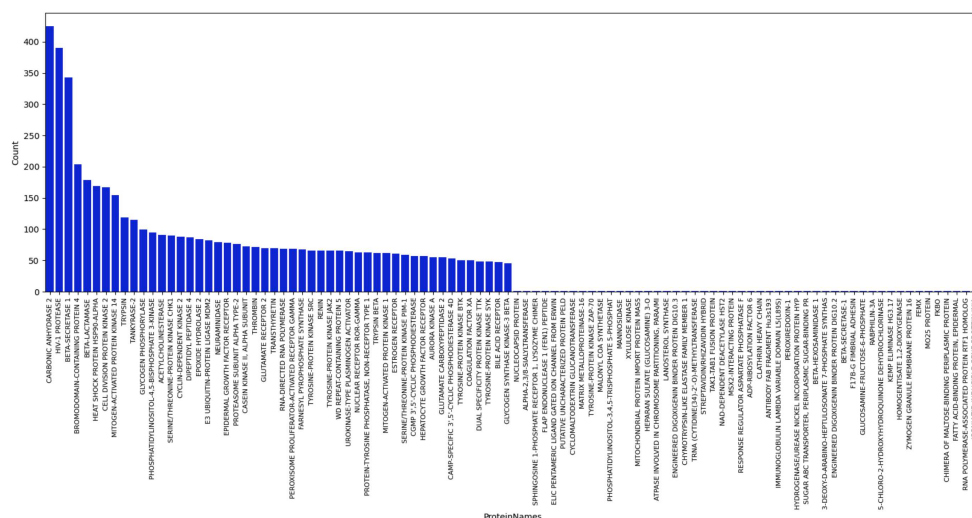


Fig. A1 This figure illustrates the distribution of PDB entries for each protein, sorted in descending order, with a focus on the top 50 and the bottom 50 proteins. Total proteins:4242

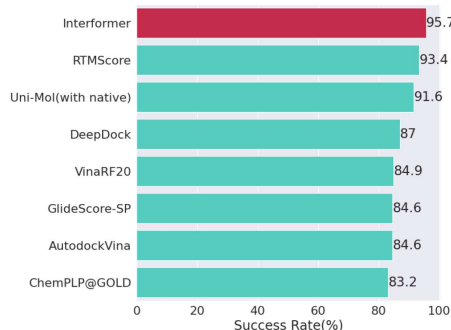


Fig. A2 Performance of top SOTA methods scoring functions on CASF-2016 benchmark, only docking powering. This success rate excluding crystal poses.

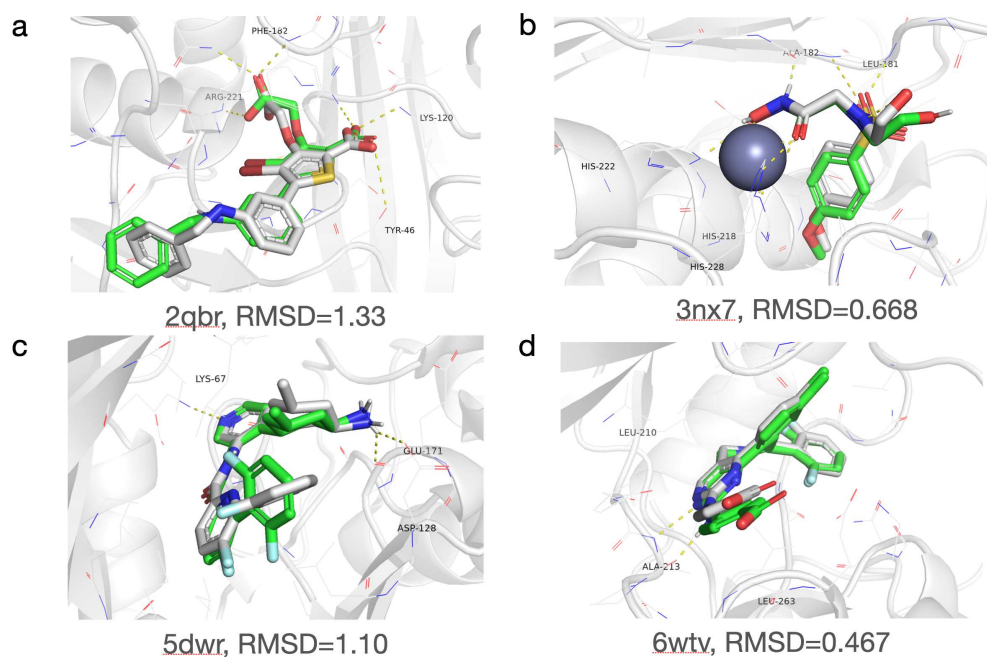


Fig. A3 This figure illustrates four docking results generated by Interformer in the testset of pdbbind coreset. The white represents the crystal structures, the green indicates the docking outcomes, and the yellow dashed lines signify the hydrogen bond interactions of the docked poses. As can be observed, the poses generated by Interformer are capable of directly forming hydrogen bond interactions.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [supplmentary.zip](#)