JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

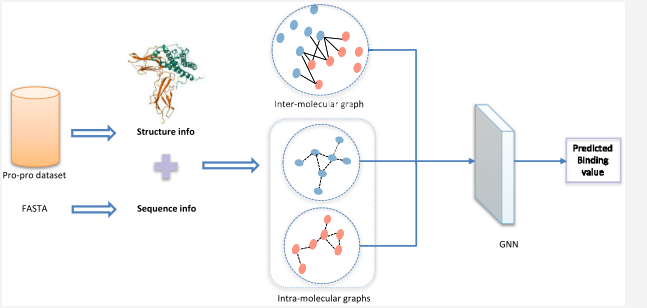
ProAffinity-GNN: A Novel Approach to Structure-Based Protein−Protein Binding Affinity Prediction via a Curated Data Set and Graph Neural Networks

Zhiyuan Zhou, Yueming Yin, Hao Han, Yiping Jia, Jun Hong Koh, Adams Wai-Kin Kong,* and Yuguang Mu*

Cite This: J. Chem. Inf. Model. 2024, 64, 8796−8808 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Protein−protein interactions (PPIs) are crucial for understanding biological processes and disease mechanisms, contributing significantly to advances in protein engineering and drug discovery. The accurate determination of binding affinities, essential for decoding PPIs, faces challenges due to the substantial time and financial costs involved in experimental and theoretical methods. This situation underscores the urgent need for more effective and precise methodologies for predicting binding affinity. Despite the abundance of research on PPI modeling, the field of quantitative binding affinity prediction remains underexplored, mainly due to a lack of comprehensive data. This study seeks to address these needs by manually curating pairwise interaction labels on available 3D structures of protein complexes, with experimentally determined binding affinities, creating the largest data set for structure-based pairwise protein interaction with binding affinity to date. Subsequently, we introduce ProAffinity-GNN, a novel deep learning framework using protein language model and graph neural network (GNN) to improve the accuracy of prediction of structure-based protein−protein binding affinities. The evaluation results across several benchmark test sets and an additional case study demonstrate that ProAffinity-GNN not only outperforms existing models in terms of accuracy but also shows strong generalization capabilities.



Downloaded via UPPSALA UNIV on October 28, 2025 at 09:09:30 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

INTRODUCTION

Protein−protein interactions (PPIs) serve as the cornerstone of nearly all biological processes, dictating the dynamics of signaling pathways and structural frameworks essential for cellular functionality.[1] These interactions are central to dissecting the molecular basis of diseases, thus forming a critical focus in the quest for novel therapeutic strategies.[2,3] The proliferation of advanced experimental techniques, including X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy, has significantly expanded the repository of 3D structural data on PPIs. This wealth of structural information has set the stage for the integration of artificial intelligence (AI) methods,[4−12] potentially providing a cost-effective and efficient alternative to traditional experimental approaches.

Beyond the structural delineation of PPIs, binding affinity emerges as a critical determinant, dictating the formation and specificity of protein complexes.[13] For instance, this parameter is particularly crucial in drug optimization phases of therapeutic development.[14,15] Nonetheless, the accurate prediction of protein−protein binding affinities remains a formidable challenge. While traditional experimental methods for affinity determination are known for their resource intensity,[16] computational approaches, such as molecular dynamics simulations and empirical energy functions, face hurdles in computational demand and accuracy.[17−22]

In the realm of machine learning, recent endeavors have shown promise for structure-based protein−protein binding affinity prediction.[23−27] Notably, the PRODIGY predictor harnesses a linear model focusing on inter-residue contacts, noninteracting surface, and buried surface area which are some elements crucial to PPIs.[24] Similarly, PPI-Affinity employs a Support Vector Machine (SVM) strategy, leveraging selected molecular descriptors as input features.[28] Despite these advancements, the complex nature of PPIs and the scarcity of comprehensive data sets have hampered the full realization of machine learning's potential in this field.[29]

PDBbind protein−protein complexes database (version 2020),[30] subsequently referred to as PDBbind, is the current

Received: October 9, 2024
Revised: November 10, 2024
Accepted: November 12, 2024
Published: November 19, 2024

ACS Publications © 2024 American Chemical Society

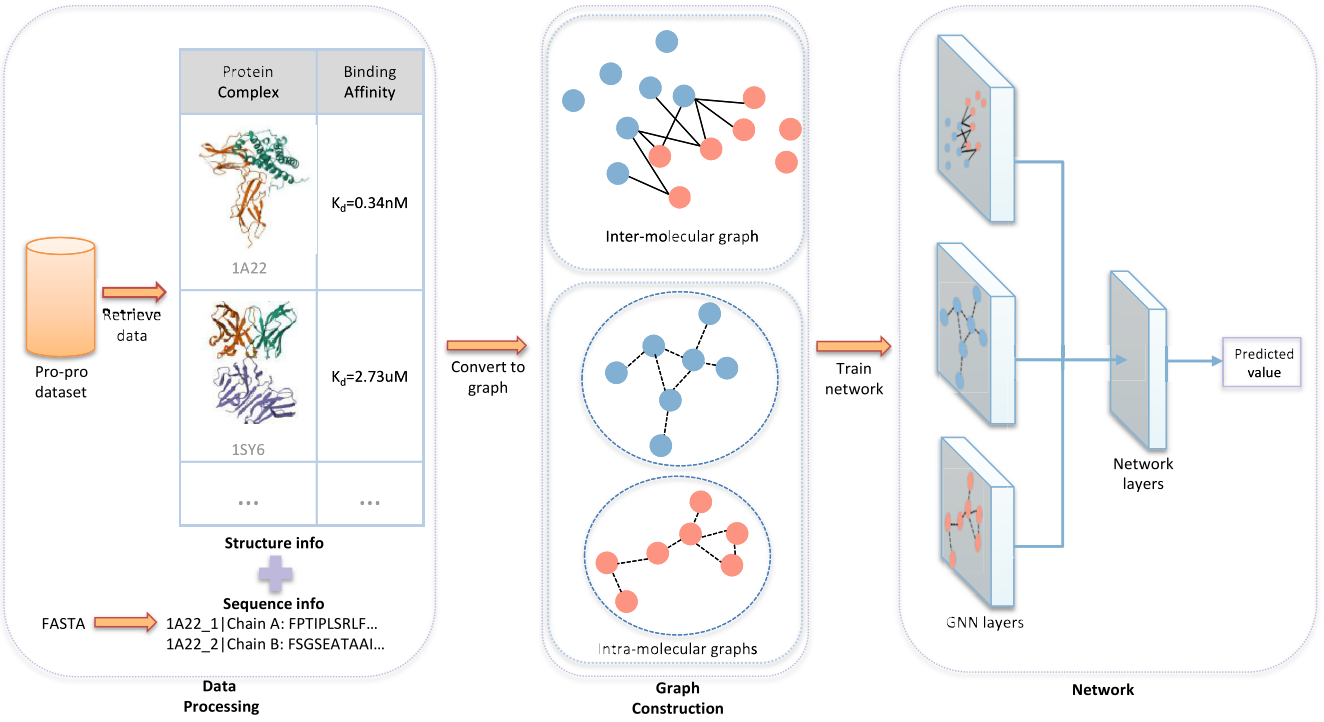


Figure 1. Pipeline of ProAffinity-GNN. First, 3D structure and sequence data of the protein−protein complexes with corresponding binding affinities are retrieved. Then one intermolecular and two intramolecular graphs are constructed based on the data. Finally, a network is trained, with the graphs as the input.

largest available PPI structure-based resource, with experimental binding affinities. To map the binding affinity onto the structure features, usually two-body or two-chain of interacting proteins are considered. PDBbind hosts 2,852 protein complexes, however, only about 590 of them contain only two chains of proteins which is easy for data preparation, most of the data points contain more than two chains of proteins. This complexity significantly diminishes the original data set's applicability for detailed protein−protein binding analysis. In contrast, a dedicated structure-based benchmark database for protein−protein binding affinity,[31] which can offer a granular view by clearly delineating chain components within pairwise interactions, has only 144 protein−protein pairs, highly limiting the potential for comprehensive analysis and development within this research topic.

In this work, our first endeavor is to manually curate the PPI complexes in the structure database to identify the interacting two chains or two groups of proteins based on the detailed structural features and related published papers. Leveraging the comprehensive coverage of PDBbind and the precise focus of the structure-based benchmark, we have created a refined data set with 2,283 unique protein−protein pairs with their binding affinity. The protein−protein pairs in this data set indicate the components involved in pairwise interactions, providing an accurate and extensive foundation for detailed structure-based benchmarking of protein−protein binding affinities prediction. Further details on the construction of this data set will be elaborated upon in the subsequent sections of this paper.

In addition, building upon our refined data set, we introduce ProAffinity-GNN, an innovative deep learning framework designed to advance the modeling of protein−protein complexes. Leveraging the fusion of protein language model[32,33] and graph neural networks (GNNs),[34] ProAffinity-GNN encapsulates both structural and sequence information within residue-level graphs, offering a comprehensive portrayal of protein−protein interactions. This model also enhances performance by collaboratively combining intra- and intermolecular graphs, providing a detailed and holistic view of protein−protein complexes. This approach enables deeper insights and more precise predictions of binding affinities.

ProAffinity-GNN represents a significant advancement over traditional machine learning models in structure-based protein−protein binding affinity prediction. Contrary to traditional models that primarily focus on two-chain interactions and depend heavily on the extraction of complex physicochemical properties,[23,27,28] ProAffinity-GNN introduces a purely deep learning-based approach. This method capitalizes on the entire complex structure, facilitating the prediction of protein−protein binding affinities without the need for intricate feature extraction. Furthermore, ProAffinity-GNN has demonstrated improved accuracy, broad generalizability, and enhanced usability across diverse benchmark test sets and an additional case study.

Overall, this study improves the field of protein−protein binding affinity prediction through both a carefully curated data set and an advanced deep learning method. By enhancing the understanding of protein−protein interactions, these contributions may offer valuable insights for future research in protein design and therapeutic target exploration.

MATERIALS AND METHODS

In this section, we first introduce the details of the construction of our structure-based protein−protein binding affinity data set. Then, we elucidate the proposed ProAffinity-GNN (see Figure 1), including data processing, graph construction, and network building.

Data set Creation: Structure-Based Protein−Protein Complexes with Binding Affinity. In this study, we

https://doi.org/10.1021/acs.jcim.4c01850
J. Chem. Inf. Model. 2024, 64, 8796−8808

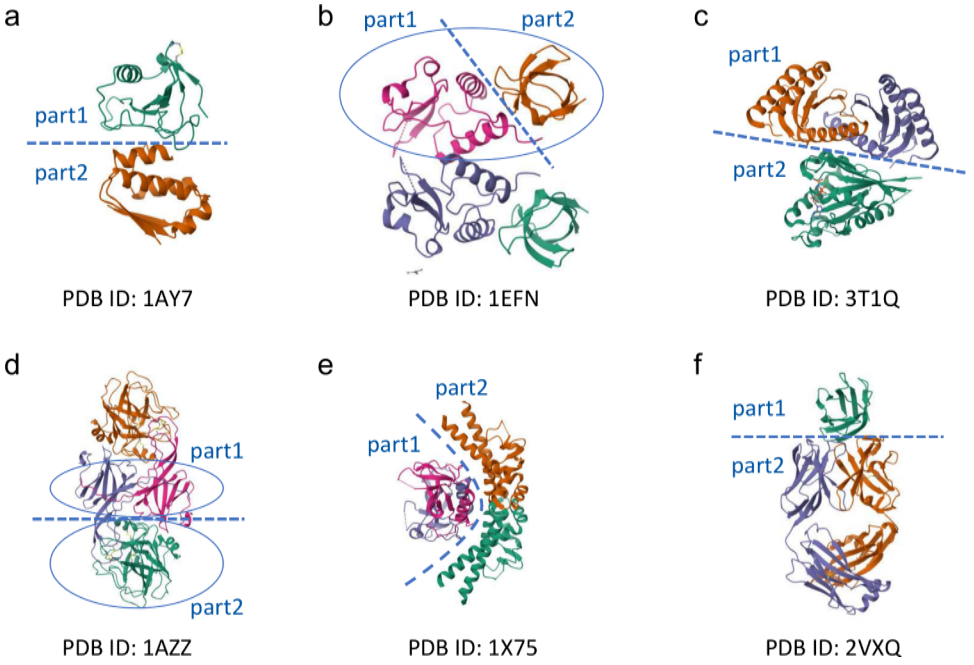


Figure 2. Common scenarios when dealing with protein−protein complexes. (a)−(e) are for complexes composed of 2 types of chains, and (f) is for complexes composed of 3 types of chains. (a) There are only two chains from different proteins. These chains constitute the components of the pairwise interaction. (b) The complex consists of identical units, such as symmetrical oligomeric complex. One of the smallest units is extracted, and subsequent processing is performed on this unit. (c) The complex comprises one chain and one dimer (or multimer) from different proteins, with both the chain and the entire dimer forming the interaction surface. The chain and the dimer, serve as the components of the pairwise interaction. (d) A dimer (or multimer) interacts with chains from different proteins, forming multiple identical interaction surfaces. One interface is selected, with the dimer and the chain that compose this interface serving as the components of the pairwise interaction. (e) A dimer (or multimer) interacts with another dimer (or multimer), forming an interaction surface. These two dimers are considered as the components of the pairwise interaction. (f) In most cases, within a complex composed of three types of chains, two chains that belong to the same protein form one component of the interaction, while the third chain, belonging to a different protein, forms the other component.

constructed a data set for structure-based protein−protein binding affinity prediction, utilizing the PDBbind[30] database as the foundation. The original PDBbind data set contains 2,852 entries, each comprising a 3D structure in PDB format, detailing the atomic-level structures of protein−protein complexes, and experimentally determined binding affinity values, represented as K_d (Dissociation Constant), K_i (Inhibition Constant), or IC_{50} (Half Maximal Inhibitory Concentration). To clearly label the interacting partners among chains in PDB protein complexes, we identified the chain components involved in the direct pairwise interactions of a minimum interaction unit. This approach aligns with 'structure-based benchmark database for protein−protein binding affinity',[31] which is an existing benchmark recording the chain components of pairwise binding in protein−protein complexes. Our focus was narrowed to complexes containing either two or three types of proteins, excluding more complex assemblies to facilitate the conversion of intricate protein−protein interactions into direct pairwise relationships. The labeling process is concluded as following specific patterns and corresponding rules: (i) In cases with only two chains belonging to separate proteins, the simplest scenario, we designated these two chains as the interaction components, such as Figure 2(a). (ii) If a complex comprised identical units, such as a symmetrical oligomeric complex, we extracted the smallest unit for further analysis, such as Figure 2(b). (iii) If one interacting entity of the complex is a homomultimer, with all chains contributing to the interaction surface, we consider the entire homomultimer as one component within the pairwise interaction, such as Figure 2(c). (iv) If one component within the complex interacts separately with identical chains from different proteins, the interfaces are typically the same. We extracted one such interaction surface and analyzed the pairwise component based on this surface, such as Figure 2(d). (v) If the two dimers compose the interacting interface, then they are the components of the pairwise interaction, such as Figure 2(e).

These guidelines were adhered to in our manual labeling process, with the exception of a few exceptionally challenging cases. For complicated complexes, the related publications were consulted for clarification, for instance, checking the precise interaction composition measured in the experimental setup. Here we concluded some common scenarios in the processing (shown in Figure 2). We also provided several complicated instances in Supporting Information. Ultimately, this resulted in a collection of 2,283 labeled pairwise PDB structures.

Data Preparation. In preparing the data set for training our ProAffinity-GNN model aimed at predicting binding affinities, we undertook a rigorous selection process. Initially, we filtered out any entries containing DNA/RNA structures due to their incompatibility with our model's focus, resulting in a data set comprising 2,269 entries. The data set is characterized by binding affinity values primarily denoted by K_d, alongside a smaller proportion labeled with K_i or IC_{50}. We removed the data with K_i and IC_{50}, only keeping the data with K_d. This filtering left 2,071 data. To standardize these values and address their typically low magnitudes, we adopted the negative logarithm to base 10 (pK_a) as the uniform measure for binding affinity, formulated as

$$pK_a = -\log_{10} K \tag{1}$$

https://doi.org/10.1021/acs.jcim.4c01850
J. Chem. Inf. Model. 2024, 64, 8796−8808

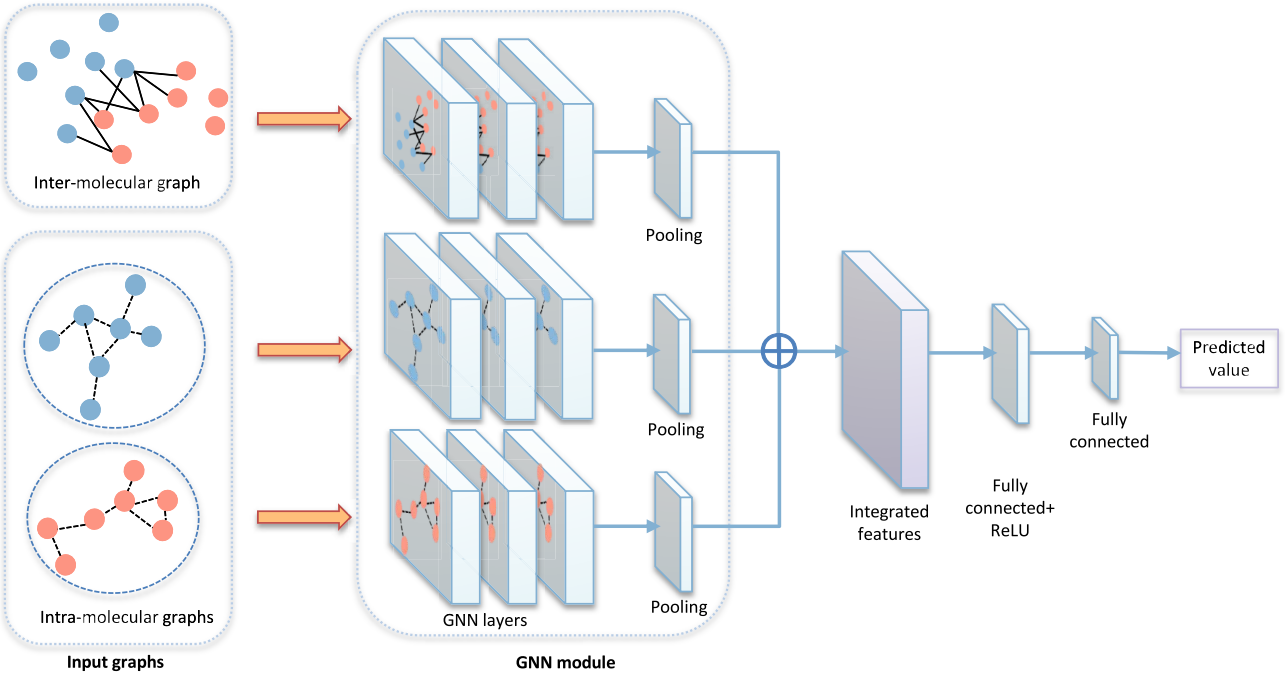


Figure 3. ProAffinity-GNN Architecture: Processes three distinct graphs (one intermolecular and two intramolecular) using GNN layers with an attention mechanism, followed by pooling for graph-level feature aggregation, concatenation for comprehensive interaction insights, and fully connected layers with ReLU activation for regression output.

where K represents the affinity values K_d with mol/L as the unit of measurement.

To preserve the integrity of our model evaluation, we first removed any entries from our data set that were also present in the test benchmarks, including 2 test sets and a subset of SKEMPI data set described subsequently, ensuring there was no overlap with the data sets used for testing. Moreover, we excluded any entries that contained sequences similar to those in the test benchmarks. Specifically, we defined a similarity score for complex pairs to assist with similarity filtering, as described below.

Given two protein complexes C_1 and C_2, each containing n chains (where $n = 2$ or $n = 3$, n should be only one number), denote the chains in C_1 as $A_1, ..., A_n$ and the chains in C_2 as $B_1, ..., B_n$. We define the sequence identity between chains A_i and B_j as identity$(A_i, B_j) \in [0, 1]$, where 1 denotes completely identical (calculated with Needleman–Wunsch algorithm[35]). There are $n!$ possible permutations σ that describe the matching combinations between the chains of C_1 and C_2. Each permutation σ corresponds to a specific matching of chains: $(A_1, B_{\sigma(1)}), (A_2, B_{\sigma(2)}), ..., (A_n, B_{\sigma(n)})$. For each permutation σ, we compute the minimum sequence identity value across all chain pairs in that matching:

$$m_\sigma = \min(\text{identity}(A_1, B_{\sigma(1)}),\ \text{identity}(A_2, B_{\sigma(2)}),\ ...,\ \text{identity}(A_n, B_{\sigma(n)})) \quad (2)$$

This minimum value m_σ ensures that we consider the least similar chain pair in each matching combination. Finally, the similarity score $S(C_1, C_2)$ between the two complexes is defined as the maximum of these minimum values across all permutations σ:

$$S(C_1, C_2) = \max_{\sigma \in \text{Perm}(n)} m_\sigma \quad (3)$$

where Perm(n) represents the set of all possible permutations of n elements.

This approach ensures that the similarity assessment considers all possible chain matching combinations and focuses on the overall similarity by maximizing the least similar link in each combination. By selecting the maximum of the minimum identity values, the method identifies the matching combination where the least similar chain pair still exhibits the highest possible similarity, thus providing a robust and comprehensive similarity score between the two complexes. In our work, we adopted the similarity score as 0.25, meaning that we considered complexes to be similar if the identity of each sequence in a complex was greater than 25% compared to each sequence in the test set complexes.

Following this similarity assessment, 169 data points identified as highly similar, which contain 148 data with 2 chains and 21 data with 3 chains, were consequently eliminated from the training set. The identity matrices comparing the training set with each test set are provided in the Supporting Information.

To enhance the data set's utility for our analyses, we converted these PDB files into PDBQT format using AutoDockFR.[36] This process also involved adding polar hydrogens for more complete molecular representation. Additionally, we extracted sequence information for each protein complex from Protein Data Bank[37] by retrieving the FASTA sequence, ensuring a comprehensive representation of each entry's molecular structure and properties.

The process resulted in a final count of 1,741 entries. The distribution of the target values is detailed in Figure S8. We subsequently divided this data set into a training set and a validation set using an 8:2 ratio, randomly. Moreover, we organized the data into five distinct groups to enable a 5-fold cross-validation approach.

https://doi.org/10.1021/acs.jcim.4c01850
J. Chem. Inf. Model. 2024, 64, 8796−8808

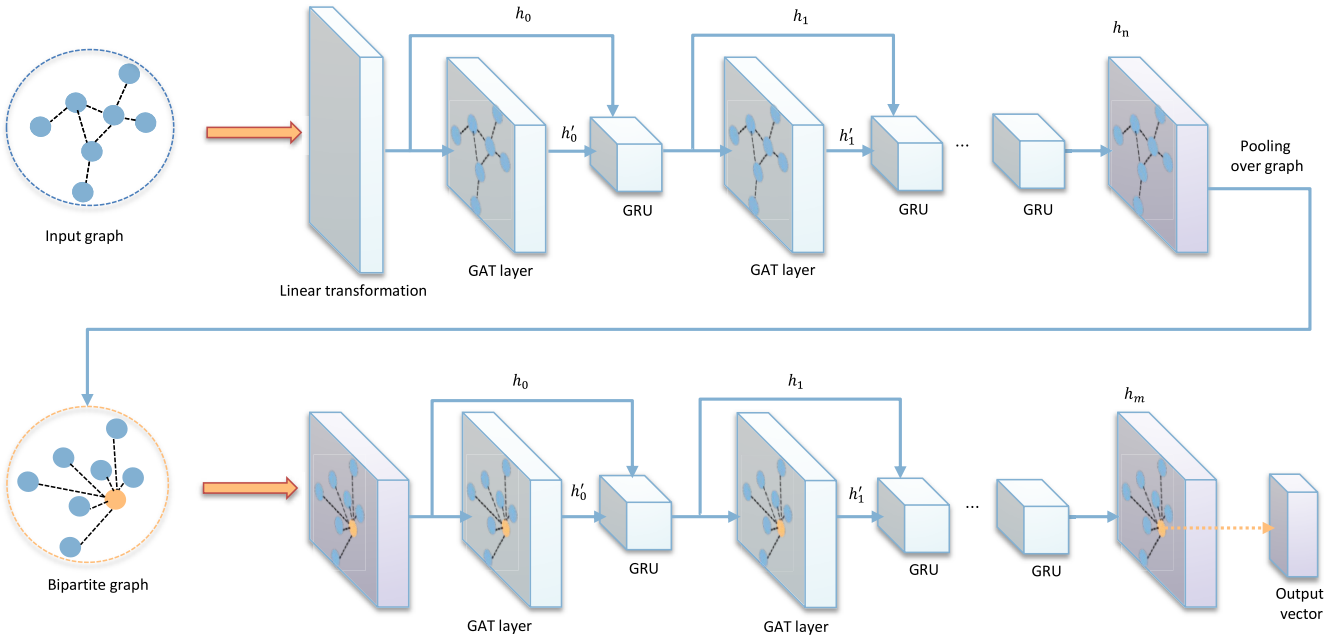


Figure 4. AttentiveFP Module: Begins with dimensionality adjustment via a fully connected layer, followed by feature refinement through cycles of GAT and GRU. Incorporates a super virtual node for comprehensive graph representation, leading to a bipartite structure that enhances information aggregation. Finalizes with ReLU-activated fully connected layers to distill graph-level features into a precise predictive value, merging local and global graph data for improved accuracy.

Graph Construction. Identifying the role of both inter-residue contacts and individual protein structures in protein−protein binding affinity,[21] we crafted two distinct types of graphs: the intramolecular graph highlights interactions within each protein, while the intermolecular graph captures interactions between protein pairs. These residue-based graphs depict individual amino acids as nodes. The adjacency matrix for intramolecular graphs, A_{intra}, is defined such that $A_{i,j} = 1$ if the distance between nodes i and j is within the intramolecular cutoff, and 0 otherwise. Similarly, for intermolecular graphs, A_{inter}, edges are established between the two components of the pairwise complex based on an intermolecular distance threshold. By doing parameter grid search (see Table S1), we set the intramolecular cutoff at 3.5 Å, and the intermolecular cutoff at 6 Å.

For both intra- and intermolecular graphs ($\mathcal{G}_{intra}$ and $\mathcal{G}_{inter}$), we utilized the Evolutionary Scale Modeling-2 (ESM-2),[33] a cutting-edge transformer-based protein language model, for node embedding. The model's efficiency and scale make it particularly suitable for contemporary protein research.[38,39] We used FASTA sequences of protein as the input to ESM-2, generating embeddings for each residue that were used as node features in the graphs. If one component in the pairwise interaction contains multiple chains, we generate the residue embeddings for each chain individually. The details of the node embedding can be seen in the Supporting Information.

Drawing on the principles of OnionNet,[40] which was developed for predicting protein−ligand binding affinities, we implemented a similar onion-like model to methodically categorize the spatial relationships of atom pairs within protein complexes (see Supporting Information).

Network. ProAffinity-GNN's architecture (Figure 3) is structured to process three distinct graphs: one intermolecular ($\mathcal{G}_{inter}$) and two intramolecular ($\mathcal{G}_{intra}$) graphs. The network employs multiple Graph Neural Network (GNN) layers featuring an attention mechanism to distill and learn intricate features from each graph. After these GNN layers update the node features within each graph, a pooling operation aggregates these updated features across all nodes in each graph to capture comprehensive graph-level representations.

These aggregated features from each graph are then concatenated, ensuring that the combined feature vector encapsulates comprehensive information from both the protein−protein interaction surface and the individual protein components. The network proceeds with several fully connected layers activated by ReLU functions, with the exception of the final layer, which employs a linear transformation to produce the regression output, predicting the binding affinity.

In our model, we integrate the AttentiveFP[41] framework (Figure 4) as the core of our GNN to process the input graphs efficiently. This module enhances the model by first individually updating the nodes within a graph. It then introduces a novel approach by creating a virtual super node that connects to all other nodes, encapsulating the entire graph's information. This methodology allows AttentiveFP to harness both detailed local node features and broader graph-level insights, facilitating more accurate predictions for graph-based tasks. The AttentiveFP architecture is detailed in the Supporting Information.

Evaluation Metrics. For assessing our model's performance, we adopted Mean Absolute Error (MAE) and Pearson's Correlation Coefficient (R) as primary metrics, aligning with standards set by existing benchmarks. To facilitate comparison with prior work,[28] we converted our model's output from pK_a to ΔG (change in Gibbs free energy upon binding), measured in kcal/mol, using the relationship between ΔG and K (K_d here):

$$\Delta G = RT \ln K \tag{4}$$

The definition of these metrics and the details of the conversion can be seen in the Supporting Information.

https://doi.org/10.1021/acs.jcim.4c01850
J. Chem. Inf. Model. 2024, 64, 8796−8808

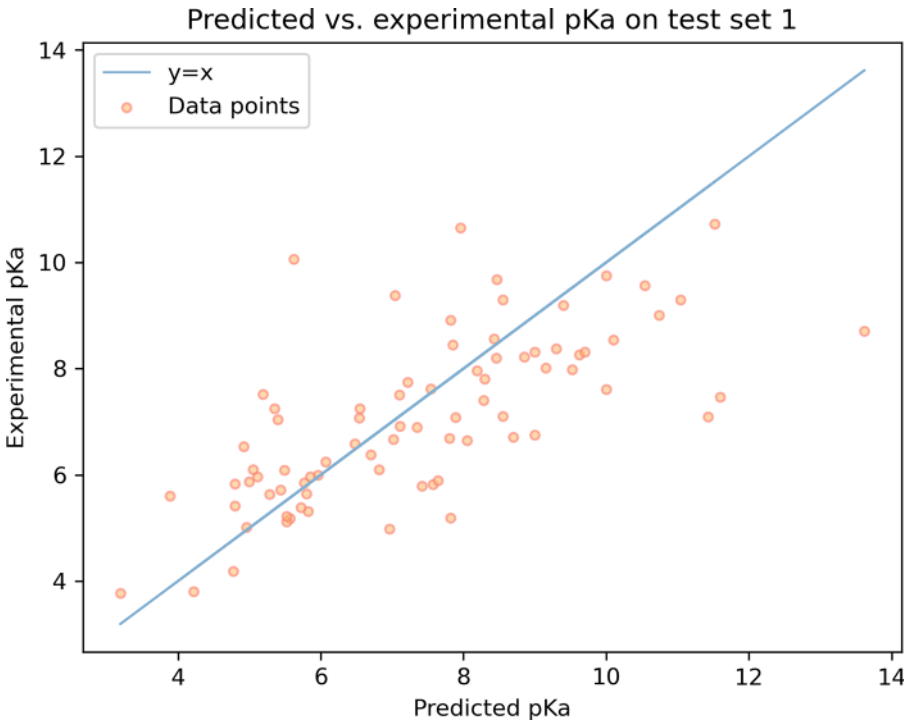


Figure 5. Correlation between predicted and experimental pK_a values for test set 1, with the identity line (y = x) indicating perfect prediction accuracy.

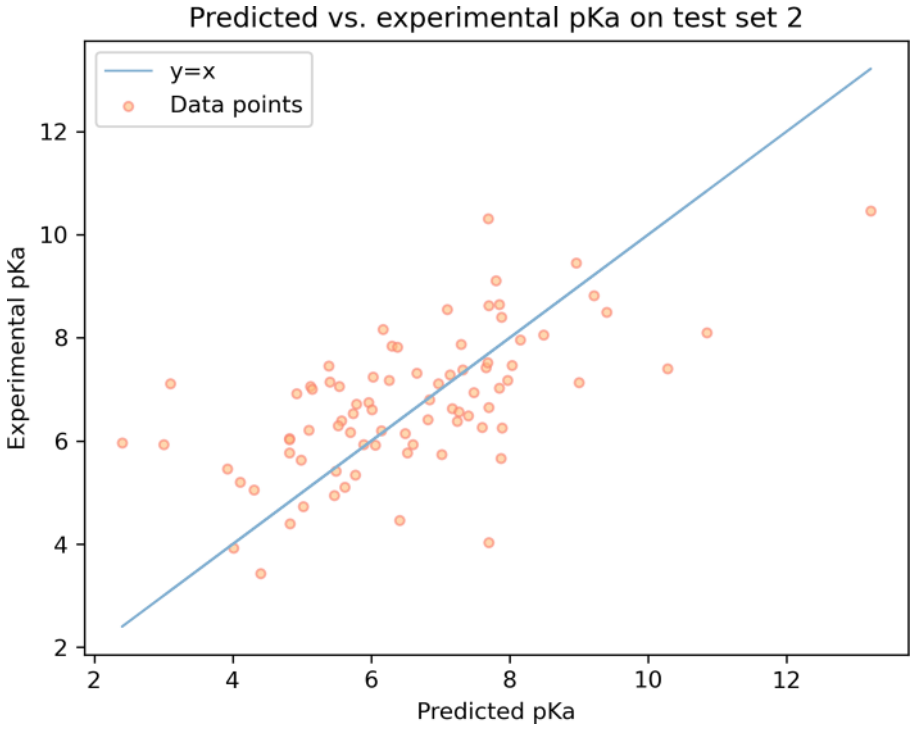


Figure 6. Correlation between predicted and experimental pK_a values for test set 2, with the identity line (y = x) indicating perfect prediction accuracy.

RESULTS AND DISCUSSION

We conducted a detailed assessment of the ProAffinity-GNN model, starting with 5-fold cross-validation on our data set. Following this, we compared the model's performance against established methods using specialized test sets for protein−protein binding affinity prediction. Additionally, to assess our data set's capacity, we conducted cross-tests by training several methods on our data set and a previously established data set, and then comparing their performance on benchmark tests. To further validate our framework's design, we also carried out an ablation study, which helped identify the impact of specific components on the overall framework efficacy.

Performance through 5-Fold Cross-Validation. To ascertain the generalizability and robustness of our model, we first employed 5-fold cross-validation on the data set comprising 1,741 entries. This method involved segmenting the data set into

https://doi.org/10.1021/acs.jcim.4c01850
J. Chem. Inf. Model. 2024, 64, 8796−8808

Table 1. Performance Compared with Existing Methods on Benchmarks

	Test set 1 (79)		Test set 2 (82)		Combined set (161)	
Method	R ↑	MAE (kcal/mol) ↓	R ↑	MAE (kcal/mol) ↓	R ↑	MAE (kcal/mol) ↓
PRODIGY[24]	**0.735**	**1.43**	0.334	2.52	0.456	1.98
DFIRE[42]	0.602	4.64	0.145	26.02	−0.005	15.53
CP_PIE[43]	0.517	8.80	0.111	7.26	0.167	8.02
ISLAND[44]	0.378	2.10	0.217	2.15	0.314	2.13
PPI-Affinity[28]	0.616	1.82	0.436	1.78	0.545	1.80
ProAffinity-GNN	0.697	1.52	**0.620**	**1.49**	**0.669**	**1.50**

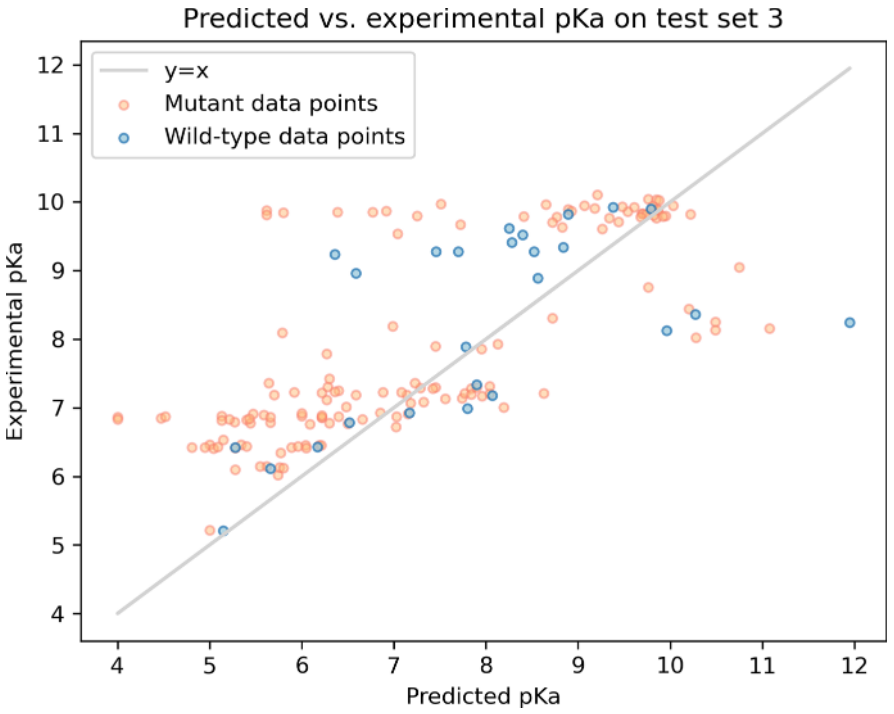


Figure 7. Correlation between predicted and experimental pK_a values for test set 3, with the identity line (y = x) indicating perfect prediction accuracy. The red data points indicate mutant data and the blue data points indicate wild-type data.

five equal parts, cyclically using one as the testing set while the remainder served for training. This cycle was repeated five times, ensuring each subset was used for testing once. The model's performance across these folds was recorded and averaged, with the standard deviation provided. Our model demonstrated a mean R = 0.63 ± 0.04 and a mean MAE = 1.54 ± 0.07 kcal/mol, showing a notable improvement over the previous PPI-Affinity model,[28] which reported an R value of 0.53 using an ensemble model on their development set.

Comparative Analysis with Existing Methods on Benchmark Sets. To rigorously assess the performance of our ProAffinity-GNN model, we compared it against existing methods using a series of established benchmark data sets for protein−protein binding affinity prediction. The first test set, comprising 79 data points with protein−protein complexes that may consist of two or more protein chains, is extracted from the structure-based benchmark database for protein−protein binding affinity.[31] This test set has been extensively utilized in related research.[23] The second test set, developed in the context of the PPI-Affinity,[28] is drawn from the PDBbind database and features 90 data points representing complexes solely composed of two protein chains. Similarly, we retained only those data entries with K_d representing the binding affinity, resulting in a total of 82 entries. While this set includes one homodimer complex, we opted to retain it for completeness, despite homodimers not being considered in our training data. The distributions of binding affinity for these two test sets are shown in Figure S9 and Figure S10. Table 5 and Table 6 show the scatter plot comparing predicted to experimental pK_a values from test set 1 and 2 respectively.

To provide a holistic view, we also combined these data from both test sets, yielding a comprehensive data set for evaluation. The evaluation results are presented in Table 1, highlighting the comparative performance of ProAffinity-GNN against the existing methods.

Our ProAffinity-GNN model exhibits performance on par with PRODIGY on test set 1 and demonstrates superior stability and accuracy across the remaining data sets, particularly test set 2 and the combined set. This suggests that our model offers notably enhanced generalizability compared to prior methods, which often show significant performance degradation on more challenging data sets.

Evaluation on Another External Data Set: SKEMPI. To substantiate the effectiveness of our approach, we conducted an assessment of our model using a segment of the SKEMPI (Structural database of Kinetics and Energetics of Mutant Protein Interactions) 2.0 database.[45] This database is dedicated to exploring the dynamics of mutant protein interactions, with a

https://doi.org/10.1021/acs.jcim.4c01850
J. Chem. Inf. Model. 2024, 64, 8796−8808

particular emphasis on how mutations influence protein−protein binding affinity. It is a valuable resource for benchmarking in the realm of binding affinity prediction.[27,46,47] The raw database encompasses 345 wild-type protein structures and an expansive collection of 7,085 mutants, each annotated with binding affinity in K_d. We took a subset previously established, consisting of 26 wild-type and 151 mutant entries.[28] We eliminated duplicate structures, resulting in a final compilation of 26 wild-type and 140 mutant protein−protein complexes. The distribution of binding affinity for this test set is shown in Figure S11. Utilizing the EvoEF2 toolkit,[48] we generated complete structures for the mutant complexes. Figure 7 shows the scatter plot comparing predicted to experimental pK_a values from test set 3. Then we evaluated ProAffinity-GNN's performance, as detailed in Table 2. This table also includes the performance of other protein−protein binding affinity prediction methods on SKEMPI/SKEMPI 2.0 or their subsets, as reported in previous studies.[29]

Table 2. Accuracy of Methods on SKEMPI/SKEMPI 2.0 or Their Subsets

Method	Evaluation on SKEMPI	Remark
mmCSM-PPI[46]	R = 0.75 on SKEMPI 2.0[a]	specific for mutation
GeoPPI[26]	R = 0.58 on SKEMPI, R = 0.52 on SKEMPI 2.0[a]	specific for mutation
TopNetTree[25]	R = 0.85 on SKEMPI, R = 0.79 on SKEMPI 2.0[a]	specific for mutation
MT-TopLap[49]	R = 0.88 on SKEMPI 2.0[b]	specific for mutation
TopLapNetGBT[50]	R = 0.87 on SKEMPI 2.0[b]	specific for mutation
PerSpect-EL[27]	R = 0.853 on SKEMPI[a]	
PPI-Affinity[28]	R = 0.77 on SKEMPI 2.0 subset[a]	
PIPR[51]	R = 0.873 on SKEMPI[a]	
PIPR + S2F[47]	R = 0.264 on SKEMPI 2.0 subset[a]	
ProAffinity-GNN	R = 0.722 and MAE = 1.36 kcal/mol on SKEMPI 2.0 subset	
	R = 0.742 and MAE = 1.34 kcal/mol on mutants	

[a]Data sourced from previous study.[29] [b]Data sourced from previous study.[52]

Notably, our model was not specifically trained on protein−protein complexes involving mutations, yet it still produced desirable results on data sets with mutant data. It is also important to highlight that while some other methods may appear to perform better, their results are likely overestimated due to a lack of sufficient consideration of data similarity in the test set. This accomplishment underscores ProAffinity-GNN's robustness and its capacity to generalize across a diverse range of protein interactions, including those not represented in the training data set.

Furthermore, we attempted to predict relative change in binding free energy ($\Delta\Delta G$) for these 140 mutant data points. Considering the complexity of the task and that our model was not specifically designed for it, the MAE between the experimental and predicted $\Delta\Delta G$ was 1.964 kcal/mol, indicating much room for improvement. We recognize that further refinement is necessary to enhance performance in this area.

Cross-testing with Existing Training Set. To demonstrate the capacity of our curated data set, we trained several methods on both our data set and an established structure-based protein−protein binding affinity data set, then compared their performances on test benchmarks. We adopted Sandra's training set used in PPI-Affinity,[28] which comprises 650 entries. As with previous procedures, we retained data entries with K_d as the binding affinity and filtered out those with similar sequences, resulting in a total of 516 entries. The methods tested include ProAffinity-GNN, two other types of GNN backbones: GAT and GCN, and PIPR,[51] a sequence-based method originally designed for binary protein−protein interactions. The performances are shown in Table 3. It is evident that in 10 out of 12 testing cases, training performances on our data set surpass those on Sandra's data set. The exception is PIPR when tested on test set 2, which may be attributed to PIPR's design that utilizes pairs of protein sequences as input; notably, all complexes in test set 2 consist of only two sequences.

Discussion on Sequence-Based Methods. Due to data set constraints and implementation differences, it is challenging to directly compare our method with certain sequence-based approaches for protein−protein binding affinity prediction on reliable benchmarks. We reviewed several sequence-based methods along with their performance metrics (see Table 4) and analyzed the key difference between the development of sequence-based and structure-based approaches.

Table 4. Performances of Sequence-Based Methods

Method	Performance
SVSBI[53]	R = 0.743 and the RMSE = 1.219 kcal/mol via 10-fold cross-validation
PIPR[51]	R = 0.873 on SKEMPI
TcellMatch[54]	R = 0.63 on 10x data set (TCR sequence)[55]

Sequence-based methods are often reported to achieve high performance, which can be primarily attributed to the availability of abundant data and well-developed pretrained methods, especially when compared to structure-based methods. However, when trained on small data sets, their

Table 3. Comparison of Prediction Method Performances Across Test Sets Using Different Training Data Sets[a]

	Training set 1 (ours)						Training set 2 (Sandra's)					
	Test set 1		Test set 2		Combined set		Test set 1		Test set 2		Combined set	
Method	MAE ↓	R ↑	MAE ↓	R ↑	MAE ↓	R ↑	MAE ↓	R ↑	MAE ↓	R ↑	MAE ↓	R ↑
ProAffinity-GNN	**1.52**	**0.697**	**1.49**	**0.62**	**1.50**	**0.669**	2.34	0.501	1.79	0.369	2.06	0.422
GAT	**1.69**	**0.604**	**1.66**	**0.584**	**1.67**	**0.606**	1.88	0.432	1.93	0.393	1.91	0.432
GCN	**1.70**	**0.619**	**1.70**	**0.555**	**1.70**	**0.598**	2.47	0.441	2.11	0.405	2.29	0.441
PIPR[51]	**1.98**	**0.400**	1.85	0.308	1.98	0.318	2.12	0.379	1.78	0.385	1.92	0.406
Mean of above	**1.72**	**0.580**	**1.67**	**0.517**	**1.71**	**0.548**	2.20	0.438	1.90	0.388	2.04	0.425

[a]The bold number indicates that the test performance from training on our data set is better than that from training on Sandra's data set.

https://doi.org/10.1021/acs.jcim.4c01850
J. Chem. Inf. Model. 2024, 64, 8796−8808

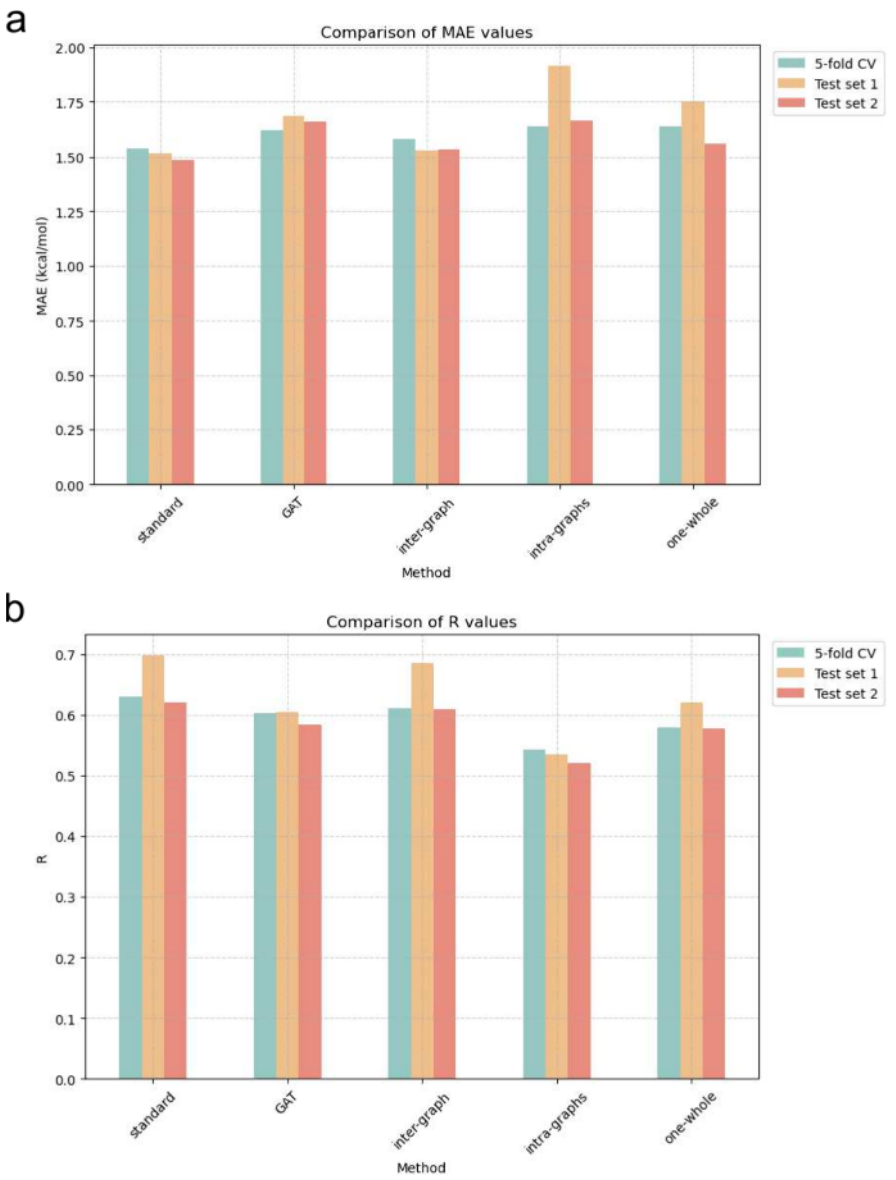


Figure 8. Results of the ablation study. (a) Test results of different methods on various sets, with respect to MAE. (b) Test results of different methods on various sets, with respect to R.

performance may not be as strong (see Table 3). Notably, as the amount of training data increased, the performance of our structure-based method improved dramatically, highlighting its potential.

With the advancement of powerful structure prediction tools such as AlphaFold,[56] access to a vast number of estimated protein structures has become readily available. Recent studies indicate that incorporating structural information can enhance the accuracy of PPI predictions,[57] and combining both sequence and structural data has also shown promise in binding affinity prediction.[58] Consequently, we believe that incorporating sequence and structural information will be highly beneficial and will likely play a pivotal role in future advancements.

Ablation Study. To ascertain the efficacy of the ProAffinity-GNN framework's design, we conducted an ablation study. This study aimed to dissect the impact of the important components of the structure on the model's performance, thereby clarifying their contributions to the framework's success in predicting protein−protein binding affinity.

Our ablation study comprised several GNN model variants, each with a specific modification—either through the removal or alteration of a component. The baseline model integrates an AttentiveFP approach with a combination of three graphs: one intermolecular graph and two intramolecular graphs. The variants tested include:

- **GAT**: Utilizes a Graph Attention Network (GAT) baseline with three layers without edge features, accompanied by three graphs.
- **Intergraph**: Features an AttentiveFP baseline with an intermolecular graph only.
- **Intragraphs**: Incorporates an AttentiveFP baseline with two intramolecular graphs only.
- **One whole graph**: Incorporates an AttentiveFP baseline with a single graph representing the entire structure, including both inter- and intraedge connections.

To maintain experimental consistency, each variant underwent evaluation using the same training procedure, data set, and metrics as the standard model. The outcomes of our ablation

https://doi.org/10.1021/acs.jcim.4c01850
J. Chem. Inf. Model. 2024, 64, 8796−8808

study are shown in Figure 8 and Table S2. From the results, we observe the following:

1. Solely employing GAT layers within the GNN model results in a performance decline. This indicates that GATs alone may not be optimally suited for this task. Conversely, the AttentiveFP structure more effectively aggregates information across the comprehensive graph.
2. A model that encompasses the entire complex information by combining both intermolecular information and individual protein details typically achieves superior results compared to models that focus on either aspect in isolation. Moreover, solely using a single graph to represent the entire structure, instead of learning from the combination of inter- and intragraphs, may lead to a decline in performance.
3. Respectable results can be achieved using intermolecular structures, without explicit individual protein structure features. This suggests that the model can capture the essential information of the interaction between the components, showing that the intermolecular information is the predominant contributor, while the individual protein structures play a supplementary role.

Case Study: Ranking Protein−Peptide Interactions. In this case study, we expanded the application of our protein−protein binding affinity prediction method to specifically address protein−peptide interactions within the PDZ domains of High-temperature requirement serine proteases (HtrAs), which are critical in the development of neurological disorders such as Alzheimer's and CARASIL.[59] Using our method, we evaluated and ranked the binding strengths of complexes formed between HtrAs and optimized peptides, including their variants.[60] The results of these computational predictions were then compared with experimentally obtained IC50 values[60] and evaluated against the performance of other methods.

We selected the PDZ domains of two human HtrAs, HtrA1 and HtrA3, and analyzed a collection of peptides interacting with them. The optimized complexes were based on PDB IDs 2JOA and 2P3W, respectively. Additional complex structures featuring various peptides were constructed through mutations using EvoEF2.[48] The ranking results for HtrA1 and HtrA3 are shown in Tables 5 and 6. Notably, for HtrA1, our method effectively identified the optimized peptide, ranking it as top 1.

Table 5. Comparison of Ranking by ProAffinity-GNN and Experimental IC50 Values with Protein−Peptide Interactions in HtrA1[a]

	ProAffinity-GNN		Experimental	
Peptide	pK_a	Ranking	IC50 (μM)	Ranking
DSRIWWV	6.450	1	0.9 ± 0.1	1
ASRIWWV	6.337	2	2.8 ± 0.3	4
DSAIWWV	6.262	3	2.5 ± 0.4	3
DSRAWWV	6.203	4	13 ± 1	9
DARIWWV	6.176	5	1.3 ± 0.1	2
DSRIWWA	6.169	6	3.5 ± 0.9	6
WDKIWHV	5.913	7	2.8 ± 0.3	4
DSRIAWV	5.824	8	40 ± 5	11
DIETWLL	5.554	9	23 ± 3	10
GWKTWIL	5.550	10	7.7 ± 0.6	8
DSRIWAV	5.437	11	6 ± 1	7

[a]The peptide in bold is the optimized one.

Table 6. Comparison of Ranking by ProAffinity-GNN and Experimental IC50 Values with Protein−Peptide Interactions in HtrA3[a]

	ProAffinity-GNN		Experimental	
Peptide	pK_a	Ranking	IC50 (μM)	Ranking
FGRWI	5.951	1	1.0 ± 0.1	4
FGRWF	5.924	2	7.7 ± 0.8	8
RSWWV	5.709	3	0.6 ± 0.1	1
FARWV	5.373	4	1.1 ± 0.2	5
FGRWV	5.339	5	0.6 ± 0.1	1
FGRWL	5.189	6	2.9 ± 0.3	6
FGRWA	5.168	7	3.5 ± 0.3	7
FGRAV	5.031	8	270 ± 110	9
FGAWV	4.988	9	0.9 ± 0.1	3

[a]The peptide in bold is the optimized one.

In contrast, for HtrA3, where the shorter peptide length offers less information, the task was more challenging, and our method ranked the optimized peptide fifth. However, it is observed that among the top 5 experimental entries, the differences are minor, within a range of 0.5 μM.

To further analyze performance, we enriched our evaluation by incorporating additional metrics to gauge ranking power and compare our approach with previous methods. We calculated the Pearson correlation coefficient, R, between the predicted values and $-\log_{10}IC_{50}$. Given the minimal variation in the top 5 experimental binding affinities (1.9 and 0.5 μM for HtrA1 and HtrA3, respectively), we adopted Top-5 Accuracy as a metric for ranking power. This metric quantifies the accuracy of the predictive model in identifying peptides within the experimentally determined top five based on binding affinity. The Top-5 Accuracy is defined as follows:

$$\text{Top-5 Accuracy} = \frac{|\{\text{Top 5 predicted}\} \cap \{\text{Top 5 actual}\}|}{5} \quad (5)$$

The comparative results of our method and other established methods are presented in Table 7. Our performance in correlation metrics is consistently high for HtrA1 relative to other methods. It is noteworthy that our method shows better performance for HtrA1 than for HtrA3. This discrepancy is likely due to the shorter peptide sequence associated with HtrA3, which presents a greater challenge for our predictive model. Nonetheless, Top-5 Accuracy appears to be a more relevant and intuitive metric for evaluating ranking power in this context. Notably, our method achieves the highest Top-5

Table 7. Comparison of Method Performances for HTRA1 and HTRA3

	R		Top-5 Accuracy		
Method	HtrA1	HtrA3	HtrA1	HtrA3	Average
Kdeep[61]	−0.077	−0.447	0.6	0.8	0.7
DFIRE[42]	−0.455	**-0.682**	0.6	0.6	0.6
CP_PIE[43]	0.109	0.152	0.6	0.6	0.6
RF-Score[62]	0.482	0.603	0.8	0.6	0.7
Prodigy[24]	0.102	−0.43	0.6	0.4	0.5
EvoEF2[48]	−0.032	−0.255	0.6	0.8	0.7
PPI-Affinity (peptide version)[28]	−0.548	−0.326	0.8	0.8	**0.8**
ProAffinity-GNN	**0.615**	0.293	0.8	0.8	**0.8**

https://doi.org/10.1021/acs.jcim.4c01850
J. Chem. Inf. Model. 2024, 64, 8796−8808

Accuracy for both HtrA1 and HtrA3. Furthermore, our method attains performance comparable to that of PPI-Affinity's peptide version, which is specifically tailored for protein−peptide binding affinity prediction.

This case study underscores the versatility of ProAffinity-GNN, demonstrating its ability to handle both protein−protein and protein−peptide interactions, as well as effectively predicting the impact of mutations, thereby highlighting the broader applicability of our methodologies in diverse biological contexts.

CONCLUSION

The prediction of protein−protein binding affinity occupies a pivotal role in the study of protein−protein interactions, heralding advancements in protein engineering and drug discovery. Despite its significance, the field is hampered by a notable scarcity of effective and efficient methodologies. A critical challenge is the significant lack of comprehensive data. Moreover, the processing procedures of existing methods are often overly complex and tailored to only a subset of the relevant concerns. In response to these limitations, we have developed an enriched data set based on PDBbind, focusing on pairwise interacting protein−protein complexes, and have manually added labels for identifying interacting two groups, each of which may contain multiple chains based on experimental setups. Additionally, we introduce a novel deep learning approach, ProAffinity-GNN, which leverages a protein language model and graph neural networks, integrating the intuitive spatial structure with the protein sequence that holds a lot of potential messages. This method is also distinguished by its simplicity in processing complexes without necessitating the cumbersome computation of physicochemical properties. Extensive evaluations demonstrate that ProAffinity-GNN not only achieves superior performance compared to existing techniques but also exhibits remarkable generalization capabilities across diverse external data sets, yielding more consistent prediction outcomes. Furthermore, the inclusion of an extended case study on protein−peptide interaction ranking underscores the application versatility and adaptability of ProAffinity-GNN.

The journey toward refining protein−protein binding affinity prediction is far from over, with significant strides still to be made in data set development and model innovation. A pressing issue remains the absence of a universally accepted evaluation data set, which is crucial for the equitable validation of model performance. It is our hope that the field will witness increased research efforts aimed at overcoming these challenges, further propelling the advancement of this critical area of study.

ASSOCIATED CONTENT

Data Availability Statement

The data set created for this study, along with the training and test sets, is available in the Supporting Information. The source code of implementation is available at https://github.com/legendzzy/ProAffinity-GNN.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.4c01850.

(Additional materials of the methods and experiments PDF)

PPI_data set_2283: The PPI data set we built (XLSX)

test_set_1: Detail of the test set 1 (XLS)

test_set_2: Detail of the test set 2 (XLS)

test_set_3_SKEMPI: Detail of the test set 3 (SKEMPI) (XLS)

training_list: PDB list for training (TXT)

AUTHOR INFORMATION

Corresponding Authors

Adams Wai-Kin Kong − *College of Computing and Data Science, Nanyang Technological University, 639798, Singapore*; Email: adamskong@ntu.edu.sg

Yuguang Mu − *School of Biological Sciences, Nanyang Technological University, 637551, Singapore*; orcid.org/0000-0002-2499-026X; Email: ygmu@ntu.edu.sg

Authors

Zhiyuan Zhou − *School of Biological Sciences, Nanyang Technological University, 637551, Singapore*

Yueming Yin − *Institute for Digital Molecular Analytics and Science (IDMxS), Nanyang Technological University, 636921, Singapore*

Hao Han − *School of Biological Sciences, Nanyang Technological University, 637551, Singapore*

Yiping Jia − *School of Pharmacy, Shanghai Jiao Tong University, 200240 Shanghai, China*

Jun Hong Koh − *School of Biological Sciences, Nanyang Technological University, 637551, Singapore*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.4c01850

Author Contributions

Zhiyuan Zhou conducted the main study, including developing the research concept, designing the study, analyzing the data, and writing the manuscript. Yueming Yin provided revisions to the study design. Hao Han, Yiping Jia, and Jun Hong Hoh contributed to data collection and data set development. Adams Wai-Kin Kong guided GNN model development and Yuguang Mu provided the overall research direction.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by Singapore Ministry of Education (MOE) Tier 1 RG97/22. Computations were mainly performed using the resources of the National Supercomputing Centre, Singapore (https://www.nscc.sg) and the HADLEY high-performance computing cluster of SCELSE. SCELSE is funded by Singapore's National Research Foundation, the Ministry of Education, NTU, and the National University of Singapore (NUS), and is hosted by NTU in partnership with NUS.

REFERENCES

(1) Peng, X.; Wang, J.; Peng, W.; Wu, F.-X.; Pan, Y. Protein−protein interactions: detection, reliability assessment and applications. *Briefings in bioinformatics* **2016**, *18*, 798−819.

(2) Ryan, D. P.; Matthews, J. M. Protein−protein interactions in human disease. *Curr. Opin. Struct. Biol.* **2005**, *15*, 441−446.

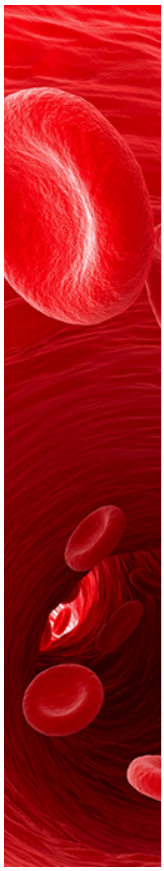
(3) Fry, D. C. Protein−protein interactions as targets for small molecule drug discovery. *Peptide Science: Original Research on Biomolecules* **2006**, *84*, 535−552.

(4) Hu, X.; Feng, C.; Ling, T.; Chen, M. Deep learning frameworks for protein−protein interaction prediction. *Computational and Structural Biotechnology Journal* **2022**, *20*, 3223−3233.

(5) Li, S.; Wu, S.; Wang, L.; Li, F.; Jiang, H.; Bai, F. Recent advances in predicting protein−protein interactions with the aid of artificial intelligence algorithms. *Curr. Opin. Struct. Biol.* **2022**, *73*, 102344.
(6) Casadio, R.; Martelli, P. L.; Savojardo, C. Machine learning solutions for predicting protein−protein interactions. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2022**, *12*, No. e1618.
(7) Tsuchiya, Y.; Yamamori, Y.; Tomii, K. Protein−protein interaction prediction methods: from docking-based to AI-based approaches. *Biophysical Reviews* **2022**, *14*, 1341−1348.
(8) Rogers, J. R.; Nikolényi, G.; AlQuraishi, M. Growing ecosystem of deep learning methods for modeling protein−protein interactions. *Protein Engineering, Design and Selection* **2023**, *36*, gzad023.
(9) Wang, X.; Flannery, S. T.; Kihara, D. Protein docking model evaluation by graph neural networks. *Frontiers in Molecular Biosciences* **2021**, *8*, 647915.
(10) Réau, M.; Renaud, N.; Xue, L. C.; Bonvin, A. M. DeepRank-GNN: a graph neural network framework to learn patterns in protein−protein interfaces. *Bioinformatics* **2023**, *39*, btac759.
(11) Das, S.; Chakrabarti, S. Classification and prediction of protein−protein interaction interface using machine learning algorithm. *Sci. Rep.* **2021**, *11*, 1761.
(12) Li, X.; Han, P.; Chen, W.; Gao, C.; Wang, S.; Song, T.; Niu, M.; Rodriguez-Patón, A. MARPPI: boosting prediction of protein−protein interactions with multi-scale architecture residual network. *Briefings in Bioinformatics* **2023**, *24*, bbac524.
(13) Dell'Orco, D. Fast predictions of thermodynamics and kinetics of protein−protein recognition from structures: from molecular design to systems biology. *Molecular BioSystems* **2009**, *5*, 323−334.
(14) Kaczor, A. A.; Bartuzi, D.; Stępniewski, T. M.; Matosiuk, D.; Selent, J. Protein−protein docking in drug design and discovery. *Computational Drug Discovery and Design* **2018**, *1762*, 285−305.
(15) Wang, L.; Jiang, J.; Zhang, L.; Zhang, Q.; Zhou, J.; Li, L.; Xu, X.; You, Q. Discovery and optimization of small molecules targeting the protein−protein interaction of heat shock protein 90 (Hsp90) and cell division cycle 37 as orally active inhibitors for the treatment of colorectal cancer. *Journal of medicinal chemistry* **2020**, *63*, 1281−1297.
(16) Zhou, M.; Li, Q.; Wang, R. Current experimental methods for characterizing protein−protein interactions. *ChemMedChem.* **2016**, *11*, 738−756.
(17) Kortemme, T.; Baker, D. *A simple physical model for binding energy hot spots in protein−protein complexes*. Proceedings of the National Academy of Sciences 2002, *99*, 14116−14121.
(18) Ma, X. H.; Wang, C. X.; Li, C. H.; Chen, W. Z. A fast empirical approach to binding free energy calculations based on protein interface information. *Protein engineering* **2002**, *15*, 677−681.
(19) Zhang, C.; Liu, S.; Zhu, Q.; Zhou, Y. A knowledge-based energy function for protein- ligand, protein- protein, and protein- DNA complexes. *Journal of medicinal chemistry* **2005**, *48*, 2325−2335.
(20) Flower, D. R.; Phadwal, K.; Macdonald, I. K.; Coveney, P. V.; Davies, M. N.; Wan, S. T-cell epitope prediction and immune complex simulation using molecular dynamics: state of the art and persisting challenges. *Immunome Research* **2010**, *6*, 1−18.
(21) Kastritis, P. L.; Rodrigues, J. P.; Folkers, G. E.; Boelens, R.; Bonvin, A. M. Proteins feel more than they see: fine-tuning of binding affinity by properties of the non-interacting surface. *Journal of molecular biology* **2014**, *426*, 2632−2652.
(22) Xiong, P.; Zhang, C.; Zheng, W.; Zhang, Y. BindProfX: assessing mutation-induced binding affinity change by protein interface profiles with pseudo-counts. *Journal of molecular biology* **2017**, *429*, 426−434.
(23) Vangone, A.; Bonvin, A. M. Contacts-based prediction of binding affinity in protein−protein complexes. *elife* **2015**, *4*, No. e07454.
(24) Xue, L. C.; Rodrigues, J. P.; Kastritis, P. L.; Bonvin, A. M.; Vangone, A. PRODIGY: a web server for predicting the binding affinity of protein−protein complexes. *Bioinformatics* **2016**, *32*, 3676−3678.
(25) Wang, M.; Cang, Z.; Wei, G.-W. A topology-based network tree for the prediction of protein−protein binding affinity changes following mutation. *Nature Machine Intelligence* **2020**, *2*, 116−123.
(26) Liu, X.; Luo, Y.; Li, P.; Song, S.; Peng, J. Deep geometric representations for modeling effects of mutations on protein-protein binding affinity. *PLoS computational biology* **2021**, *17*, No. e1009284.
(27) Wee, J.; Xia, K. Persistent spectral based ensemble learning (PerSpect-EL) for protein−protein binding affinity prediction. *Briefings in Bioinformatics* **2022**, *23*, bbac024.
(28) Romero-Molina, S.; Ruiz-Blanco, Y. B.; Mieres-Perez, J.; Harms, M.; Munch, J.; Ehrmann, M.; Sanchez-Garcia, E. PPI-affinity: A web tool for the prediction and optimization of protein−peptide and protein−protein binding affinity. *J. Proteome Res.* **2022**, *21*, 1829−1841.
(29) Guo, Z.; Yamaguchi, R. Machine learning methods for protein-protein binding affinity prediction in protein design. *Frontiers in Bioinformatics* **2022**, *2*, 1065703.
(30) Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind database: Collection of binding affinities for protein- ligand complexes with known three-dimensional structures. *Journal of medicinal chemistry* **2004**, *47*, 2977−2980.
(31) Kastritis, P. L.; Moal, I. H.; Hwang, H.; Weng, Z.; Bates, P. A.; Bonvin, A. M.; Janin, J. A structure-based benchmark for protein−protein binding affinity. *Protein Sci.* **2011**, *20*, 482−491.
(32) Bepler, T.; Berger, B. Learning the protein language: Evolution, structure, and function. *Cell systems* **2021**, *12*, 654−669.
(33) Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; dos Santos Costa, A.; Fazel-Zarandi, M.; Sercu, T.; Candido, S.; Rives, A. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **2023**, *379*, 1123−1130.
(34) Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; Yu, P. S. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems* **2021**, *32*, 4−24.
(35) Needleman, S. B.; Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology* **1970**, *48*, 443−453.
(36) Ravindranath, P. A.; Forli, S.; Goodsell, D. S.; Olson, A. J.; Sanner, M. F. AutoDockFR: advances in protein-ligand docking with explicitly specified binding site flexibility. *PLoS computational biology* **2015**, *11*, No. e1004586.
(37) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The protein data bank. *Nucleic acids research* **2000**, *28*, 235−242.
(38) Qiu, Y.; Wei, G.-W. Artificial intelligence-aided protein engineering: from topological data analysis to deep protein language models. *Briefings in Bioinformatics* **2023**, *24*, bbad289.
(39) Xu, X.; Bonvin, A. M. DeepRank-GNN-esm: a graph neural network for scoring protein−protein models using protein language model. *Bioinformatics Advances* **2024**, *4*, vbad191.
(40) Zheng, L.; Fan, J.; Mu, Y. Onionnet: a multiple-layer intermolecular-contact-based convolutional neural network for protein−ligand binding affinity prediction. *ACS omega* **2019**, *4*, 15956−15965.
(41) Xiong, Z.; Wang, D.; Liu, X.; Zhong, F.; Wan, X.; Li, X.; Li, Z.; Luo, X.; Chen, K.; Jiang, H.; Zheng, M. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *Journal of medicinal chemistry* **2020**, *63*, 8749−8760.
(42) Liu, S.; Zhang, C.; Zhou, H.; Zhou, Y. A physical reference state unifies the structure-derived potential of mean force for protein folding and binding. *Proteins: Struct., Funct., Bioinf.* **2004**, *56*, 93−101.
(43) Ravikant, D.; Elber, R. Pie—efficient filters and coarse grained potentials for unbound protein−protein docking. *Proteins: Struct., Funct., Bioinf.* **2010**, *78*, 400−419.
(44) Abbasi, W. A.; Yaseen, A.; Hassan, F. U.; Andleeb, S.; Minhas, F. U. A. A. ISLAND: in-silico proteins binding affinity prediction using sequence information. *BioData Mining* **2020**, *13*, 1−13.
(45) Jankauskaitė, J.; Jiménez-García, B.; Dapkunas, J.; Fernández-Recio, J.; Moal, I. H. SKEMPI 2.0: an updated benchmark of changes in protein−protein binding energy, kinetics and thermodynamics upon mutation. *Bioinformatics* **2019**, *35*, 462−469.

https://doi.org/10.1021/acs.jcim.4c01850
J. Chem. Inf. Model. 2024, 64, 8796−8808

(46) Rodrigues, C. H.; Pires, D. E.; Ascher, D. B. mmCSM-PPI: predicting the effects of multiple point mutations on protein−protein interactions. *Nucleic Acids Res.* **2021**, *49*, W417−W424.
(47) Xue, Y.; Liu, Z.; Fang, X.; Wang, F. Multimodal pre-training model for sequence-based prediction of protein-protein interaction. *Proceedings of the 16th Machine Learning in Computational Biology meeting* **2022**, 34−46.
(48) Huang, X.; Pearce, R.; Zhang, Y. EvoEF2: accurate and fast energy function for computational protein design. *Bioinformatics* **2020**, *36*, 1135−1142.
(49) Wee, J.; Chen, J.; Xia, K.; Wei, G.-W. Integration of persistent Laplacian and pre-trained transformer for protein solubility changes upon mutation. *Computers in Biology and Medicine* **2024**, *169*, 107918.
(50) Chen, J.; Qiu, Y.; Wang, R.; Wei, G.-W. Persistent Laplacian projected Omicron BA. 4 and BA. 5 to become new dominating variants. *Computers in Biology and Medicine* **2022**, *151*, 106262.
(51) Chen, M.; Ju, C. J.-T.; Zhou, G.; Chen, X.; Zhang, T.; Chang, K.-W.; Zaniolo, C.; Wang, W. Multifaceted protein−protein interaction prediction based on Siamese residual RCNN. *Bioinformatics* **2019**, *35*, i305−i314.
(52) Wee, J.; Wei, G.-W. Evaluation of AlphaFold 3's Protein−Protein Complexes for Predicting Binding Free Energy Changes upon Mutation. *J. Chem. Inf. Model.* **2024**, *64*, 6676−6683.
(53) Shen, L.; Feng, H.; Qiu, Y.; Wei, G.-W. SVSBI: sequence-based virtual screening of biomolecular interactions. *Communications biology* **2023**, *6*, 536.
(54) Fischer, D. S.; Wu, Y.; Schubert, B.; Theis, F. J. Predicting antigen specificity of single T cells based on TCR CDR 3 regions. *Molecular systems biology* **2020**, *16*, No. e9416.
(55) 10x Genomics, A new way of exploring immunity−linking highly multiplexed antigen recognition to immune repertoire and phenotype. *Technol. Rep.* **2019**.
(56) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Zidek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly accurate protein structure prediction with AlphaFold. *nature* **2021**, *596*, 583−589.
(57) Song, B.; Luo, X.; Luo, X.; Liu, Y.; Niu, Z.; Zeng, X. Learning spatial structures of proteins improves protein−protein interaction prediction. *Briefings in bioinformatics* **2022**, *23*, bbab558.
(58) Nikam, R.; Yugandhar, K.; Gromiha, M. M. Deep learning-based method for predicting and classifying the binding affinity of protein-protein complexes. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics* **2023**, *1871*, 140948.
(59) Clausen, T.; Southan, C.; Ehrmann, M. The HtrA family of proteases: implications for protein composition and cell fate. *Molecular cell* **2002**, *10*, 443−455.
(60) Runyon, S. T.; Zhang, Y.; Appleton, B. A.; Sazinsky, S. L.; Wu, P.; Pan, B.; Wiesmann, C.; Skelton, N. J.; Sidhu, S. S. Structural and functional analysis of the PDZ domains of human HtrA1 and HtrA3. *Protein Sci.* **2007**, *16*, 2454−2471.
(61) Jiménez, J.; Skalic, M.; Martinez-Rosell, G.; De Fabritiis, G. K. deep: protein−ligand absolute binding affinity prediction via 3d-convolutional neural networks. *J. Chem. Inf. Model.* **2018**, *58*, 287−296.
(62) Ballester, P. J.; Mitchell, J. B. A machine learning approach to predicting protein−ligand binding affinity with applications to molecular docking. *Bioinformatics* **2010**, *26*, 1169−1175.


CAS BIOFINDER DISCOVERY PLATFORM™
CAS BIOFINDER HELPS YOU FIND YOUR NEXT BREAKTHROUGH FASTER
Navigate pathways, targets, and diseases with precision
Explore CAS BioFinder
CAS
A division of the American Chemical Society

https://doi.org/10.1021/acs.jcim.4c01850
J. Chem. Inf. Model. 2024, 64, 8796−8808