Computational and Structural Biotechnology Journal 23 (2024) 460–472

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Research article

Systematic investigation of machine learning on limited data: A study on predicting protein-protein binding strength

Feifan Zheng, Xin Jiang, Yuhao Wen, Yan Yang, Minghui Li *

MOE Key Laboratory of Geriatric Diseases and Immunology, School of Biology and Basic Medical Sciences, Suzhou Medical College of Soochow University, Suzhou, Jiangsu Province 215123, China

ARTICLE INFO

Keywords:
Protein-protein binding affinity
Machine learning methods
Tools

ABSTRACT

The application of machine learning techniques in biological research, especially when dealing with limited data availability, poses significant challenges. In this study, we leveraged advancements in method development for predicting protein-protein binding strength to conduct a systematic investigation into the application of machine learning on limited data. The binding strength, quantitatively measured as binding affinity, is vital for understanding the processes of recognition, association, and dysfunction that occur within protein complexes. By incorporating transfer learning, integrating domain knowledge, and employing both deep learning and traditional machine learning algorithms, we mitigated the impact of data limitations and made significant advancements in predicting protein-protein binding affinity. In particular, we developed over 20 models, ultimately selecting three representative best-performing ones that belong to distinct categories. The first model is structure-based, consisting of a random forest regression and thirteen handcrafted features. The second model is sequence-based, employing an architecture that combines transferred embedding features with a multilayer perceptron. Finally, we created an ensemble model by averaging the predictions of the two aforementioned models. The comparison with other predictors on three independent datasets confirms the significant improvements achieved by our models in predicting protein-protein binding affinity. The programs for running these three models are available at https://github.com/minghuilab/BindPPI.

1. Introduction

Currently, deep learning techniques have demonstrated impressive predictive capabilities when applied to large datasets [1,2]. However, many biomedical problems suffer from a scarcity of experimental data, making it imperative to explore how existing techniques can be utilized to achieve enhanced predictive accuracy [3]. In previous work, we developed a set of data-driven machine learning methods to evaluate the impact of missense mutations on protein stability [4] and their interactions with other molecules [5–10]. These methods employed traditional machine learning algorithms alongside handcrafted features, establishing them as widely recognized tools [11–13]. Additionally, traditional machine learning algorithms provide a direct measure of feature importance, enhancing the interpretability of the model outcomes. However, this approach faced a bottleneck, leading to a lack of substantial improvement in predictive performance in recent years [14]. Despite the introduction of deep learning approaches in this field, their performance improvement has been severely hampered [15–17], primarily due to the limited availability of training data. Continued efforts are required to overcome the challenges posed by limited data in machine learning. In this study, we leveraged the advancements in protein-protein binding affinity methods to conduct an in-depth investigation of the application of machine learning on datasets with limited data availability.

Protein-protein interactions (PPIs) play a fundamental role in various mechanisms of protein biological functions [18,19] and are attractive targets for therapeutic intervention [20,21]. Numerous interactions between intracellular proteins are in principle possible but only a fraction of putative complexes form and prove functionally relevant. Thus, the determination and characterization of structures and binding strengths of protein-protein associations can gain significant insight into mechanisms in biological processes for disease research, such as prominent disorders of cancer and degenerative diseases associated with aberrant PPIs [22]. In therapy, one goal is to design new synthetic protein-protein complexes with the desired function, such as optimized antibody-antigen interactions with strong binding [23]. Thus,

* Corresponding author.
E-mail address: minghui.li@suda.edu.cn (M. Li).

https://doi.org/10.1016/j.csbj.2023.12.018
Received 3 October 2023; Received in revised form 14 December 2023; Accepted 16 December 2023
Available online 20 December 2023

the characterization of PPIs in terms of their binding strength is highly relevant to the design of new and improved therapeutics [24,25]. Binding affinity is a quantitative measure of the strength of the interaction between two or more molecules that bind reversibly. Experimental techniques, such as surface plasmon resonance [26], isothermal titration calorimetry [27], and fluorescence resonance energy transfer [28], require expensive experimental setup and are time-consuming. For this reason, developing computational methods to predict binding affinity is increasingly important, which can help evaluate and understand the significance of putative protein-protein interactions [29], the discovery of protein therapeutics [25], de novo interface design [30], etc.

The computational prediction of binding affinity has a long history and various methods have been proposed throughout the years, varying dramatically in terms of accuracy, computational cost, and physical plausibility [31]. Sophisticated approaches, such as free energy perturbation (FEP) [32] and thermodynamic integration (TI) [33], and end-point methods, like molecular mechanics Poisson–Boltzmann surface area (MMPBSA) [34,35], possess a relatively high level of accuracy in principle. However, these methods, which employ extensive molecular dynamics or Monte Carlo conformational searches, are computationally intensive and have a limited scope of application. Alternative, simplified empirical energy functions have been proposed to significantly reduce computational costs. One such method is statistical potentials, which use the observed relative positions of atoms or residues in experimental structures to infer a potential of mean force [36,37]. Another approach that has gained increasing popularity over the past decade is machine learning, where energy functions are determined through regression against experimentally measured binding affinities [16,38–46]. However, the prediction accuracy of currently available methods remains limited. Hence, it is essential to continue putting effort into developing accurate and reliable methods to tackle the challenge of predicting protein-protein binding affinity.

In this research, by exploring transfer learning, integrating domain knowledge, and utilizing both deep learning and traditional machine learning algorithms, we mitigate the impact of data limitations and make significant advancements in predicting protein-protein binding affinity. Specifically, we compiled a dataset of 802 protein-protein complexes with reliable experimental measurements of binding affinities and complex structures. We developed more than 20 predictive models belonging to four categories: structure-based models with handcrafted features, sequence-based models with transferred embedding features, ensemble models composed of structure-based and sequence-based models, and structure-sequence models with a combination of handcrafted and embedding features. Among these models, three were selected with the best performance representing the three categories. The structure-based model is composed of a random forest regression and thirteen carefully selected handcrafted features derived from the complex structures. Our sequence-based model consists of a multilayer perceptron and average pooling of embedded features extracted from ESM-2 [47]. By combining these two models, we obtained an ensemble model that outperforms each individual model. To validate the advancements achieved by our approaches, we compared the performance of our methods with other previously published predictors using three independent datasets. The results demonstrate the significant improvements achieved by our models.

2. Methods

2.1. Experimental datasets used for parameterizing our methods

The training dataset was compiled from four databases/datasets: Protein-Protein Binding Affinity Benchmark version 2 (PPBABv2) [48], SKEMPI 2.0 [49], PROXiMATE [50], and PDBbind version 2020 [51]. These databases contain experimentally measured binding affinities and three-dimensional (3D) complex structures for protein-protein interactions (PPIs). The binding affinity was calculated using the equation $\Delta G_{exp} = RT\ln(K_D) = RT\ln(K_i) = RT\ln(IC_{50})$. Specifically, we collected 179, 348, 118, and 1306 protein-protein interactions from PPBABv2, SKEMPI 2.0, PROXiMATE, and PDBbind, respectively. To ensure data quality, we removed entries with ambiguous affinity values. Then, we merged all these four datasets into a combined dataset. In cases where the same PPI entry had multiple affinity values, we first filtered out entries with standard deviations of multiple affinity values larger than 1.0 kcal mol^{-1}. Then, we selected only one affinity value based on the following criteria: (i) The priority was given to values that had been measured by more than one experimental technique or study, indicating a higher frequency of occurrence; (ii) We further prioritized values measured using surface plasmon resonance or isothermal titration calorimetry, as these methods are considered more reliable and accurate. If the above criteria still resulted in multiple values for a PPI entry, we calculated the average value. As a result, we obtained a total of 1562 unique protein-protein interactions with a single experimentally determined binding affinity value and their corresponding 3D structure.

One of the primary objectives of this study is to establish a structure-based model. To ensure the highest possible resemblance between the 3D complex structures employed in constructing the theoretical model and the proteins utilized for measuring binding affinity, we applied the following criteria to exclude certain complexes: (a) Protein-peptide and peptide-peptide complexes were removed if a chain has fewer than 50 amino acids, defining it as a peptide (427 complexes removed); (b) Complexes with metal coordination sites or containing modified/unknown/missing residues at the protein-protein binding interface were removed (225 complexes removed). The interface residues were defined as those with inter-atomic distances less than 6 Å between any heavy atoms of the interacting protein partners. The removed 652 complexes were used to compile the independent test sets (see details in the next section). In the final step, a total of 108 multimers were removed and set aside to serve as one of the test sets. Our training set, referred to as S802, comprised a meticulously selected 802 heterodimers. For more detailed information about the dataset, please refer to Fig. 1 and Supplementary Table 1.

2.2. Experimental datasets used for testing

Initially, we compiled two test sets using the removed 652 complexes during the construction of the training set. We excluded specific types of complexes, such as peptide-peptide complexes, complexes with unknown residues at the binding interface, and protein-protein complexes with peptides at the binding interface (A peptide is defined at the binding interface if any of its residues belong to the interface residue set). Additionally, complexes with any individual missing interval at the binding interface greater than or equal to five amino acids were omitted. As a result, we retained 192 protein-protein heterodimer complexes (referred to as S192) and 365 protein-peptide complexes (referred to as S365) as our independent test sets. Secondly, we utilized the 108 multimers, which were previously removed during the construction of the training set (referred to as S108), to evaluate the performance of our method on multimers. Then, we took the following procedures to repair the complex structures. We converted modified residues to their corresponding standard residues. For missing segments at the binding interface, we employed the Modeller software [52] to model them. Regarding complexes with metal coordination sites, we did not add the corresponding metals. For further information regarding the test sets, please refer to Fig. 1 and Table S1.

2.3. Structure optimization

The complex 3D structures were obtained from the Protein Data Bank (PDB) [53]. Only assigned interaction partners were retained in the calculation, and missing heavy side-chain and hydrogen atoms were added by VMD program [54] with the CHARMM36 force field parameters [55]. To optimize the structures, we tested several minimization

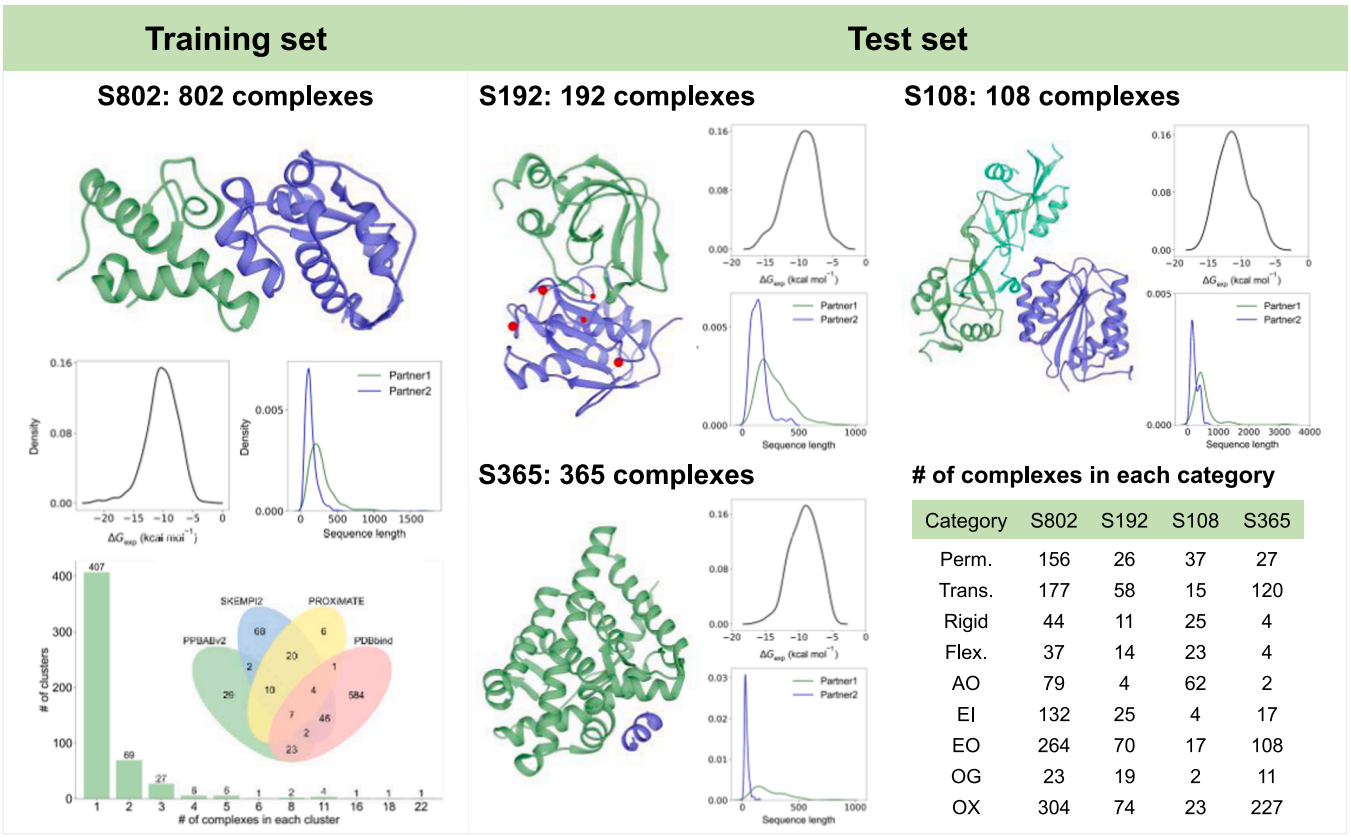

Fig. 1. Overview of the data sets used. S802: visualization of protein-protein heterodimer complex structure, the distribution of experimental binding affinity and sequence length for each interaction partner, the four sources compiled from, and the number of clusters based on sequence similarity analysis; S192: visualization of heterodimer having metal coordination sites or modified/missing residues at the binding interface; S108: visualization of multimer. S365: visualization of protein-peptide complex structure, a chain with less than 50 amino acids were defined as peptide. Statistics of different classification of complexes. See Table S1 for more information.

procedures in the gas phase for all complexes to remove steric clashes or repair possible distorted geometries. These procedures included: (a) A 100-step energy minimization with restraints on the backbone atoms (the force constant is 5 kcal mol^{-1} Å^{-2}); (b) A 2000-step energy minimization applying the same harmonic restraints as in (a); (c) A 2000-step minimization with restraints, followed by an unconstrained 5000-step minimization. The energy minimization was performed using the NAMD program (v 2.13) [56] based on the topology file of CHARMM36 force field.

2.4. Analysis of similarity among datasets

To assess the similarity between datasets, we employed three methods: sequence similarity analysis, structure similarity analysis, and their combined use. For sequence similarity analysis, we used the MMseqs2 software [57] and set the sequence identity threshold to 50%, with the alignment covering at least 50% of both query and target sequences. To be considered similar, two complexes must have both protein chains with similar sequences. For structure similarity analysis, we used TM-align [58] to perform structural alignments, focusing on the interface regions to compare the structure similarity of all protein-protein complexes. We generated a TM-score distance matrix encompassing all interface regions and conducted hierarchical clustering of the complexes using AgglomerativeClustering from the scikit-learn library [59], employing a distance threshold of 0.3 (TM-score $>$ 0.7) for clustering. The results of our analysis, as presented in Table S1b, reveal a substantial diversity among the complexes in the training set. Furthermore, the similarity between the test set and the training set is relatively low, indicating distinct characteristics between them.

2.5. Construction of structure-based models with handcrafted features

Deep learning has received considerable attention in recent times due to its impressive performance when trained on large datasets. However, in scenarios with limited training data, the combination of traditional machine learning algorithms and features identified by domain experts continues to demonstrate comparable or even superior predictive capabilities to deep learning methods [11]. In our study, we initially identified and calculated numerous handcrafted features associated with protein-protein binding. An overview of all these features is provided in Table S2. It is worth noting that while certain features can be computed solely using the protein sequences, in our study, we classified them as structure-based features because we specifically utilized those features derived from interface or surface amino acids. Overall, we generated a list of more than a thousand descriptors.

In our previous studies [4–7], the Random Forest (RF) algorithm has undergone rigorous validation and exhibited superior performance compared to other traditional machine learning algorithms. Moreover, RF offers a clear measure of feature importance and exhibits remarkable computational efficiency. Therefore, we first employed the RF algorithm to construct the structure-based models (Fig. 2a). Feature selection is an important step in traditional machine learning that enables us to identify and select the most relevant and informative features that can improve the model's performance on unseen data while reducing the dimensionality of the input data and computational time. It can also help to increase the interpretability of the model's results by focusing on the most important input features.

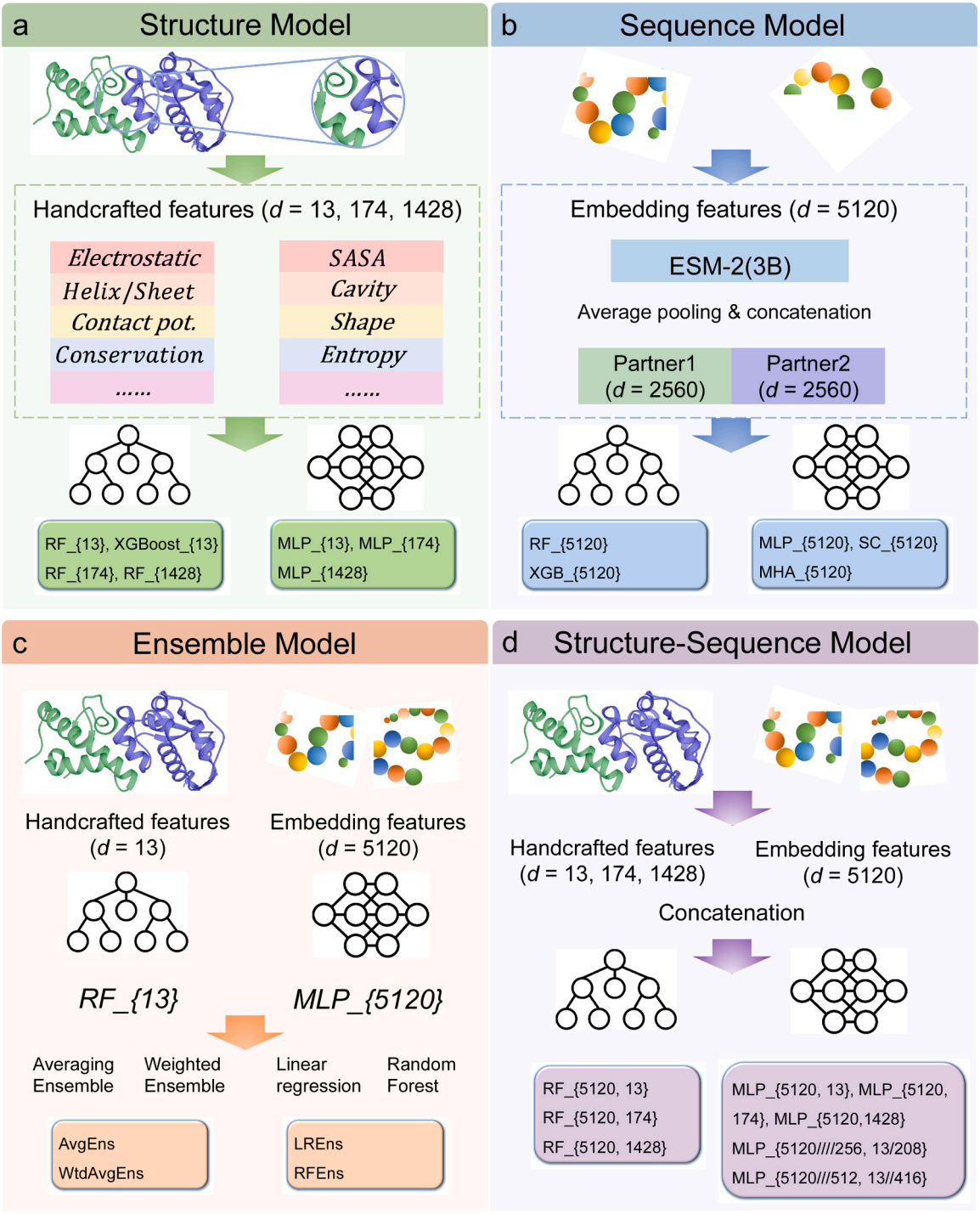

Fig. 2. Overview of the framework, consisting of four types of models. (a) Structure-based models with handcrafted features, which include a comprehensive set of physicochemical, evolutionary, sequence, and structural features for in-depth feature representation. (b) Sequence-based models with embedding features, which use a fixed-length concatenated vector representation for each complex (d = 5120) obtained from a large-scale pretrained language model, ESM-2(3B). (c) Ensemble models, combining the best-performing structure and sequence models obtained from (a) and (b). (d) Structure-sequence models, integrating a combination of raw sequence embedding features and handcrafted structure features. See Table S4 for the definition of models.

Here, we initially considered a large list of features provided in Table S2 and employed feature selection to reduce this list. The average Pearson correlation coefficient from five-fold cross-validation on the training data serves as our performance metric for feature selection. Initially, we sorted all features based on the average PCC of each feature and then selected the top 30 ranked features as the initial set to establish 30 initial models. Proceeding, we iterated over the remaining features to select subsequent features, employing the following criteria: prioritizing features not belonging to the same category as the previously selected ones, then choosing features that would maximize the increase in PCC, and finally, favoring features with shorter computation times. This iterative process persisted until the model's PCC value ceased to improve, deemed non-improving if the PCC improvement was less than 0.005. We employed a two-step scheme for feature selection: the first step involved selecting features from the 174 structure-based features, and the second step included selecting features from 1254 amino acid

descriptors primarily sourced from the AAindex database. Finally, the ultimate model selection adhered to the following principles: maximizing the Pearson correlation coefficient, minimizing the number of features, maximizing their contribution to the model, ensuring low correlation among selected variables, and enhancing feature interpretability. When these principles are satisfied, priority is determined by computational cost. As a result, a total of thirteen features from ten categories were selected. The selected features are described below, and their respective contributions to the model (referred to as RF_{13}) are presented in Table S3.

- SA^{p1} and SA^{p2} are the solvent accessible surface areas (SASA) of unbound partners (p1: partner 1 and p2: partner 2), which are calculated using the CORMAN module of CHARMM [60].
- $P^{p1}_{L/SA}$ and $P^{p2}_{L/SA}$ are the ratios of sequence lengths and solvent accessible surface areas for two unbound partners, which measure how tightly the protein structure is packed.
- P^{IF}_{Helix} and P^{IF}_{Sheet} represent the percentages of helices and sheets at the binding interface, respectively, which are calculated by dividing the number of residues assigned in helix/sheet conformation by the total number of interface residues. The secondary structure elements are assigned using the DSSP program [61,62].
- N^{IF}_{HAP} is the number of interactive heavy atom pairs between two partners. Two heavy atoms are considered interactive if their distance is within 6 Å.
- P^{IF}_{CS} stands for the ratio of the number of conserved residues at the binding interface to the total number of interface residues. A residue is considered conserved if the score for a residue mutated to alanine, as calculated by PROVEAN [63], is no more than -2.5.
- ΔE_{elec} is the electrostatic interaction between two interacting partners, which is calculated as the difference in electrostatic energies between a bound complex and each interacting partner. The calculation is performed using the ENERGY module of CHARMM [60].
- P^{Surf}_{Charge} represents the percentage of charged amino acids on the surface of complex. It is calculated by dividing the number of surface charged residues by the total number of surface residues. A surface residue is defined by a SASA ratio greater than 0.2 between the residue in the complex and the extended tripeptide [64]. The SASA values for the residue in the extended tripeptide and complex are obtained from [65] and calculated using DSSP program [61], respectively.
- S^{IF}_{AI} is the sum of the amphiphilicity index of amino acids located at the binding interface. The amphiphilicity index of amino acids is obtained from Amino Acid Index Database [66] with identifier MITS020101 [67].
- E^{IF}_{CE} is the inter-protein contact energy calculated using atom-atom (AACE167) statistical contact potentials derived from the Potts model implemented in iPot program [68]. The contact distance cutoff is $d_{\max} = 10$ Å, and the sequence separation is $k_{\min} = 5$.
- V_{Cavity} represents the number of water molecules that can be accommodated in the cavities of the complexes, calculated using the McVol program [69]. Cavities are defined as empty spaces with sufficient volume to accommodate a water molecule.

In order to compare with the RF_{13} model, we employed a different traditional machine learning algorithm, eXtreme Gradient Boosting (XGBoost), and experimented with a deep learning multilayer perceptron (MLP) neural network that incorporated the handcrafted features. Additionally, we conducted a thorough evaluation to assess the importance of feature selection, analyzing not only the 13 selected features but also all available features (Table S4 provides a detailed description of all models).

2.6. Construction of sequence-based models with transferred embedding features

For each protein sequence with a length of L, we extracted amino acid-level embeddings from the ESM-2(3B) model [47], a large-scale pretrained language model based on the BERT transformer architecture. The ESM (Evolutionary Scale Modeling) models were trained to predict masked amino acids using the surrounding amino acids in the sequence. The ESM-2(3B) model consists of 36 transformer layers, containing 3 billion parameters, and was trained on over 60 million protein sequences. For each amino acid, the ESM-2(3B) model outputs a feature vector of dimension $d = 2560$. To represent a protein, we computed the average of all amino acid embeddings over the L amino acids using average pooling. This process resulted in a fixed-length vector representation of $d = 2560$ for each protein. For cases where one interaction partner contains multiple chains, we first averaged each individual chain and then calculated the average for all chains within the partner. Subsequently, we concatenated the two protein-level embeddings to obtain a 5120-dimensional feature vector for a protein-protein complex. This concatenated feature vector was used as the input for the sequence-based models in our study (Fig. 2b and Fig. S1a).

Initially, we constructed the sequence-based models using the Random Forest and eXtreme Gradient Boosting algorithms (Table S4). Subsequently, we employed a multilayer perceptron (MLP) neural network to develop the predictive model (refer to Fig. S1a). The MLP architecture consists of multiple fully connected hidden layers followed by an output layer, with a rectified linear unit (ReLU) activation function applied after each hidden layer. The mean squared error (MSE) is used as the loss function to assess the model's performance. To determine the optimal number of epochs, we incorporated an early stopping technique. We set a patience value of 20, considering an epoch improved only if its validation MSE loss surpassed the previous loss by a tolerance of 0. To ensure the robustness of our model, we conducted 10 iterations of 5-fold cross-validation. The selection of the optimal hyperparameter combination for each cross-validation was based on the Pearson correlation coefficient calculated on the validation set. Instead of refitting the model with the entire training set, we used the mean of 50 models for prediction. The results of our experiments confirmed that this architecture achieved the highest performance among the evaluated models and was consequently selected as our final sequence-based model (named as MLP_{5120}).

In addition to the aforementioned architectures, we enhanced the feature processing procedure by incorporating a multi-head attention network and skip connections. One architecture, shown in Fig. S1b, utilizes two fully connected layers to capture patterns within individual proteins. Another architecture, illustrated in Fig. S1c, involves using a multi-head attention (MHA) layer to obtain fused embeddings for each protein, which could capture the interactions between proteins. The transformed embeddings from these layers were then concatenated. To further facilitate information flow, skip connections were introduced in these two architectures. Finally, both models incorporate an MLP architecture to further process and utilize the embeddings obtained from the skip connections. By incorporating these techniques, including multi-head attention, skip connections, and MLPs, we aim to enhance the representation of individual proteins and protein complexes, capture their complex relationships, and improve the overall model performance. However, It's worth noting that these architectures do not outperform MLP_{5120}.

2.7. Construction of ensemble models combining structure and sequence models

Based on the correlation analysis of the prediction results obtained from the sequence models and the structural models (Fig. S2), we proposed that combining these two types of models can further improve

prediction performance (refer to Fig. 2c and Table S4). To implement the ensemble, we explored three distinct approaches: averaging, weighted averaging, and stack-based methods. In the weighted average approach, weights are determined based on the Pearson correlation coefficient values obtained from 5-fold cross-validation for each model. To ensure that the scaling of the predicted values is not affected, we normalized the weights of each model in each weighted ensemble combination. For the stacking method, the predicted results from the individual models were used as input features for training the meta-model. The meta-regressor was constructed using two algorithms: linear regression and random forest. The results confirmed that the average and weighted ensemble approaches achieved the highest performance, and consequently, the simplest average ensemble was selected as our final ensemble model.

2.8. Construction of structure-sequence models combining handcrafted and embedding features

Finally, we integrated a combination of sequence-based embedding features and structure-based handcrafted features to build the predictive models using the MLP and RF algorithms, respectively (Fig. 2d and Table S4). First, we directly concatenated the 5120-dimensional embedding features with the handcrafted features, which have dimensions of 13, 174, and 1428, as inputs to construct MLP and RF models. However, due to the substantial dimensionality gap between the sequence features ($d = 5120$) and the 13 selected one-dimensional structure features, we adopted a two-step method to address this issue, as depicted in Fig. S1d. In the first step, we independently adjusted the dimensionality of the structural and sequence features. For the structural features, we increased their dimensionality by employing one or two layers, resulting in either 16 or 32 dimensions for each feature. Consequently, the dimensionality of the 13 structural features increased to either 208 or 416 dimensions. Simultaneously, we reduced the dimensionality of the 5120-dimensional embedding to 512 or 256 dimensions using three or four layers, respectively, to match the dimensionality level of the structural features. In the second step, we concatenated the up-sampled and down-sampled features, which were now at compatible dimensional levels, and utilized them as inputs to perform the MLP architecture.

2.9. Hyperparameter tuning

The hyperparameters of the RF and XGBoost models were optimized through a grid search approach within a predefined hyperparameter search space. The complete list of hyperparameters can be found in Table S5. The optimal combination of hyperparameters was chosen based on the average Pearson correlation coefficient obtained from 5-fold cross-validation. The final RF and XGBoost models were trained on the entire training set using these selected hyperparameters.

To improve computational efficiency, a sequential search strategy was employed for hyperparameter selection in MLP models. The hyperparameters were determined sequentially, beginning with the number of hidden layers, followed by the learning rate, batch size, hidden layer dimension, and weight decay, as indicated in Table S5. The order of selection was based on the relative importance of each hyperparameter's impact on the model's performance. The Pearson correlation coefficient calculated on the validation set was used to choose the optimal combination of hyperparameters for each cross-validation. Instead of retraining the model with the entire training set, the mean of 50 models (10 repetitions × 5 folds) was used for prediction, enabling a more stable and reliable estimation of model performance.

2.10. Classification of protein-protein interactions

In this research, protein-protein interactions were categorized into three distinct classes. The first classification was based on the strength of the interaction: Permanent interactions were identified by an interaction strength with a ΔG_{exp} value of ≤ -12.27 kcal mol^{-1}, while transient interactions were defined by an interaction strength with a ΔG_{exp} value of ≥ -8.18 kcal mol^{-1}. [70,71] The second classification was based on the flexibility of the complexes, dividing them into rigid-body and flexible complexes. Rigid-body complexes were characterized by an interface C-alpha root-mean-square deviation (I-RMSD) value of ≤ 1.0 Å, whereas flexible complexes exhibited an I-RMSD value of > 1.0 Å. [44] To calculate the I-RMSD, the unbound components were superimposed onto their bound complexes, considering the C-alpha atoms of the interface residues. Only a subset of entries possesses 3D structures of unbound components. It should be emphasized that unbound structures were solely utilized for characterizing the flexibility of the complexes and were not employed in the predictions. Fig. 1 and Table S1c present the number of complexes falling into each category.

The third classification was conducted based on the functional categorization of protein-protein complexes. Out of the four available data resources, namely PPBABv2, SKEMPI 2.0, PROXiMATE, and PDBbind, only the Protein-Protein Binding Affinity Benchmark (PPBABv2) offered a functional classification for complexes, including six classes such as Enzyme-Inhibitor and Antibody-Antigen [70]. Given that our training sets, S802 and S665, contain only 149 complexes from PPBABv2, we opted to independently classify all complexes by referring to the categorization in PPBABv2. The following steps were taken to classify complexes into six functional classes: AN (Antigen-Nanobody): complexes where at least one chain can be found in the SAbDab-nano database [72]; AA (Antibody-Antigen): complexes where at least one chain can be found in the SAbDab database [73]; As a result of the limited number of instances in the AA and AN categories, we combined them into a single category named AO. EI (Enzyme-Inhibitor): one of the proteins in the complex has an EC number obtained from PDBe (Protein Data Bank in Europe) or PDB, and the other protein is annotated with the term "inhibit" in Pfam, InterPro, SCOP, or CATH databases integrated within PDBe [74]; EO (Enzyme-Others): complexes where any chain is annotated with an EC number, except for those falling under the EI classification; OG (G-protein-Others): complexes where any chain has annotations containing the term "G protein" in the molecule function of GO [74], as well as annotations containing the terms "G protein" or "GTPase activity" in Pfam, InterPro, SCOP or CATH databases; OX (Others-miscellaneous): all remaining complexes that do not fall into any of the aforementioned functional categories.

2.11. Comparison with other methods

We performed a comprehensive comparison between our models and eight other state-of-the-art methods used to calculate the binding energy, including PRODIGY [44,45], PPI-Affinity [38], PPA_Pred2 [16], Minpredictor [42], ISLAND [43], FoldX [75], Rosetta [76], and MMPBSA [8,77]. PPA_Pred2 and ISLAND are sequence-based approaches, while the rest are structure-based methods. Among these methods, PRODIGY, PPI-Affinity, PPA_Pred2, Minpredictor, and ISLAND were trained on protein-protein binding affinity data. The number of overlapping complexes between their training sets and our datasets is provided in Table S6. It is worth mentioning that except for PRODIGY, all the other four machine learning methods are limited to calculating the binding energy for dimeric complexes.

The three methods, FoldX, Rosetta, and MMPBSA, are commonly used for calculating absolute energy in protein systems. The AnalyseComplex module of FoldX was used to calculate the binding energy. For Rosetta, the binding energy was calculated using the ddG Mover in RosettaScripts with the beta_nov16 score function. The resulting energy values are reported in Rosetta Energy Unit (REU), which are correlated with kcal mol^{-1}. [78] MMPBSA combines molecular mechanical energies with the Poisson−Boltzmann continuum representation of the solvent, which was expressed as the sum of van der Waals interaction energy, polar solvation energy, and nonpolar solvation energy. For further information regarding the specific details of the MMPBSA

implementation, please refer to our previous study [8].

Additionally, we assessed the predictive performance of ten docking scores obtained from the CCharPPI webserver [79] in estimating binding affinity. Ten docking scores include: ZRANK [80], ZRANK2 [81], RosettaDock [82], pyDock [83], FireDock [84], FireDock (antibody-antigen energy function) [84], FireDock (enzyme-inhibitor energy function) [84], PISA [85], PIE [86], and SIPPER [87].

2.12. Performance evaluation and statistical analysis

Pearson correlation coefficient (PCC) and root-mean-square error (RMSE) were used to quantify the agreement between experimentally-determined and predicted values of binding affinities. A two-tailed *t*-test was used to assess whether the correlation coefficient is statistically significant from zero. RMSE (kcal mol^{-1}) is the standard deviation of the prediction errors, calculated by taking the square root of the average squared difference between predicted and experimental estimates. To evaluate the statistical significance in the difference of PCC between our models and other methods, we employed the Hittner2003 test [88], which is used for comparing two correlation coefficients based on dependent groups. Furthermore, we compared the receiver operating characteristics (ROC) curves using the DeLong test [89].

To assess the performance of the proposed approaches in distinguishing permanent or transient interactions from others, ROC analysis was conducted. The true positive rate (TPR) and false positive rate (FPR) were calculated as follows: TPR = TP/(TP+FN) and FPR = FP/(FP+TN), where TP represents true positives, TN denotes true negatives, FP signifies false positives, and FN stands for false negatives. To account for imbalances in the labeled dataset, the Matthews correlation coefficient (MCC) was also computed. The Hittner2003 test was performed using the *cocor* package [90] in R, while the remaining evaluation metrics were implemented using the Python SciPy [91] and Scikit-learn packages [59].

The assessment of uncertainties in PCC and RMSE metrics used a bootstrap approach. To gauge the reliability of our results, we performed 1000 resamplings with replacement on pairs of experimentally obtained and calculated ΔG values. The 95% confidence interval was derived from these bootstrap samples, presented as x_{lower}^{upper}. Here, x represents the mean statistic, while the lower and upper bounds are determined by the 2.5th and 97.5th percentiles of the sorted list of bootstrap samples.

3. Results and Discussion

3.1. Structure-based models with handcrafted features

In our study, a total of 1428 handcrafted features related to protein-protein binding were identified and computed (an overview of all these features is provided in Table S2). Subsequently, seven structure-based models were constructed using those features with the Random Forest, eXtreme Gradient Boosting, and multilayer perceptron neural network algorithms, as depicted in Fig. 2a and detailed in Table S4. The Pearson correlation coefficient (PCC) and root mean square error (RMSE) values between the predicted and experimentally-determined binding affinities across all datasets are shown in Table 1. Overall, the models built using the RF algorithm exhibit better performance compared to those built using MLP and XGBoost algorithms. Furthermore, for the three RF models, an increase in the number of features results in a decrease in performance on the training set, while there is no significant change in performance on the test sets. In contrast, the MLP model performs poorly when built using only 13 features; however, as the number of features increases, the model's performance improves significantly. This implies that the MLP model has better adaptability to more comprehensive feature representations.

Hence, considering model performance, interpretability, and computational efficiency, RF_{13}, comprising a random forest regression and thirteen carefully selected handcrafted features derived from

Table 1
Performance of all models. The best model for each subgroup is shown in bold.

Model	S802		S192		S108		S365		S665	
	PCC	RMSE	PCC	RMSE	PCC	RMSE	PCC	RMSE	PCC	RMSE
Structure-based models with handcrafted features										
RF_{13}	**0.63**	**2.13**	**0.52**	**2.01**	**0.33**	**2.20**	**0.37**	**2.05**	**0.46**	**2.06**
RF_{174}	0.61	2.17	0.54	1.98	0.35	2.17	0.40	1.95	0.49	2.00
RF_{1428}	0.58 * *	2.22	0.51	2.03	0.41	2.08	0.39	2.02	0.48	2.03
XGBoost_{13}	0.62	2.14	0.51	2.01	0.37	2.31	0.22 * *	2.18	0.41 *	2.16
MLP_{174}	**0.51**	**2.33**	**0.40**	**2.21**	**0.47**	**2.25**	**0.32**	**2.10**	**0.41**	**2.16**
MLP_{13}	0.34 * *	2.57	0.19 * *	2.40	0.28 *	2.58	0.22	2.09	0.23 * *	2.27
MLP_{1428}	0.51	2.34	0.44	2.13	0.41	2.21	0.32	2.14	0.44	2.15
Sequence-based models with embedding features										
RF_{5120}	**0.67**	**2.04**	**0.48**	**2.06**	**0.34**	**2.26**	**0.40**	**2.04**	**0.45**	**2.08**
XGBoost_{5120}	0.69 *	1.97	0.48	2.04	0.27	2.31	0.31 * *	2.02	0.43	2.08
MLP_{5120}	**0.65**	**2.06**	**0.47**	**2.10**	**0.41**	**2.21**	**0.38**	**2.00**	**0.47**	**2.06**
SC_{5120}	0.67 * *	2.01	0.47	2.10	0.38	2.30	0.36 *	2.06	0.46	2.11
MHA_{5120}	0.63 *	2.12	0.47	2.11	0.38	2.42	0.38	2.03	0.47	2.12
Ensemble models combining structure and sequence models										
AvgEns	**0.68**	**2.00**	**0.56**	**1.94**	**0.52**	**2.10**	**0.44**	**1.91**	**0.54**	**1.95**
WtdAvgEns	0.68	2.00	0.56	1.94	0.52	2.10	0.44	1.91	0.54	1.95
LREns	0.68	1.97	0.56	1.92	0.52	2.09	0.44	1.90	0.54	1.94
RFEns	0.63 * *	2.12	0.53	2.01	0.46	2.18	0.39	1.99	0.49 *	2.03
Structure-sequence models combining handcrafted and embedding features										
RF_{5120, 174}	**0.68**	**2.02**	**0.53**	**2.00**	**0.39**	**2.20**	**0.44**	**1.99**	**0.51**	**2.03**
RF_{5120, 13}	0.67 *	2.04	0.49	2.04	0.35	2.25	0.40	2.05	0.46 * *	2.08
RF_{5120, 1428}	0.68	2.02	0.55	1.97	0.31	2.27	0.42	2.01	0.49	2.04
MLP_{5120///512, 13//416}	**0.68**	**1.98**	**0.53**	**2.02**	**0.46**	**2.10**	**0.47**	**1.89**	**0.54**	**1.96**
MLP_{5120, 13}	0.62 * *	2.14	0.50	2.07	0.46	2.14	0.49	1.86	0.54	1.97
MLP_{5120, 174}	0.54 * *	2.39	0.48	2.24	0.48	2.09	0.36 *	2.25	0.47 * *	2.22
MLP_{5120, 1428}	0.49 * *	2.46	0.49	2.11	0.49	2.02	0.36 * *	2.28	0.45 * *	2.19
MLP_{5120////256, 13/208}	0.66 * *	2.03	0.52	2.02	0.46	2.11	0.46 *	1.90	0.54 * *	1.97

PCC: Pearson correlation coefficient between experimental and predicted binding affinities. RMSE (kcal mol^{-1}): root-mean-square error. All presented values of correlation coefficients are statistically significantly different from zero ($P < 0.05$, *t*-test). * $P < 0.05$/ * * $P < 0.005$ compared to the best model in each subgroup (Hittner2003 test).

the complex structures, was chosen as our final structure-based model. This model exhibits superior performance across the majority of datasets in comparison to other models, as demonstrated by higher PCC and lower RMSE values (Fig. 3 and Table 1). To optimize the hyperparameters of RF_{13}, we conducted 5-fold cross-validation to evaluate different combinations of decision trees and features considered when splitting a node. Through this process, the optimal settings for RF_{13} are 230 decision trees and 2 features for splitting a node. In addition, we employed 100-step energy minimized complex structures for feature calculation of RF_{13}, as it shows no statistically significant difference compared to the 2000-step protocol while enhancing the model's efficiency. Furthermore, it exhibits slightly better performance than the 7000-step protocol (Fig. S3). The reason behind these observations stems from the impact of different structure optimization protocols on the conformations of protein complexes, as shown in Fig. S4a. Since RF_{13} is a structure-based model, its features are intricately linked to the underlying structures. Figs. S4c and S4d present the correlation analysis of features and feature contribution values across different minimization protocols. This analysis offers further insights into the origins and reasons for the variations in predicted values observed under different minimization protocols (Fig. S4b).

Following this, we conducted an assessment of uncertainties to investigate the impact of different runs of the RF_{13} model on the results. For each complex in S192, we performed 50 repeated runs, generating 50 ΔG values. The distribution of the standard deviation of 50 ΔG values for 192 complexes is depicted in Fig. S5. The results indicate low uncertainties in ΔGs, with only a few complexes showing slight deviations. Upon scrutinizing the factors contributing to these discrepancies, we found that the conformations of the complexes remained unchanged across different runs. The source of this discrepancy can be traced to the V_{Cavity} feature, which represents the number of water molecules that can be accommodated in the cavities of the complexes. For certain complexes, the values computed by the McVol program varied in different runs, a phenomenon attributed to the intrinsic stochastic nature of the Monte Carlo method used for determining the volume of the molecule. This also explains the lowest correlation observed among different minimization procedures for this feature V_{Cavity} (Fig. S4). In conclusion, our assessment of uncertainties through 50 repeated runs of the RF_{13} model for each complex reveals minimal deviations in ΔG values, indicating overall robustness.

3.2. Sequence-based models with transferred embedding features

Although AlphaFold and similar methods have greatly advanced our ability to predict the structures of individual protein monomers, predicting the structures of protein complexes remains a complex and challenging task [92]. Due to the limited availability of experimentally-determined protein complex structures, it is necessary to develop sequence-based models for our purpose of predicting protein-protein binding affinity. Here, we developed five models, including two traditional machine learning models and three deep learning models (Fig. 2b, Fig. S1a-c and Table S4). These models utilize the embeddings extracted from ESM-2(3B) as their inputs. Pre-trained models, trained on large-scale data, have learned complex feature representations of protein sequences and structures, and are widely used as features for numerous downstream tasks. In this scenario, ESM-2(3B) has been pre-trained on a large-scale protein sequence dataset, and these embedded features are then utilized for predicting protein-protein binding affinity. This approach can be seen as the transfer of knowledge learned in one domain (discerning evolutionary patterns across protein sequences) to another related domain (protein-protein binding affinity prediction).

The performance evaluation based on PCC and RMSE values indicates that the RF model slightly outperforms the XGBoost model on the

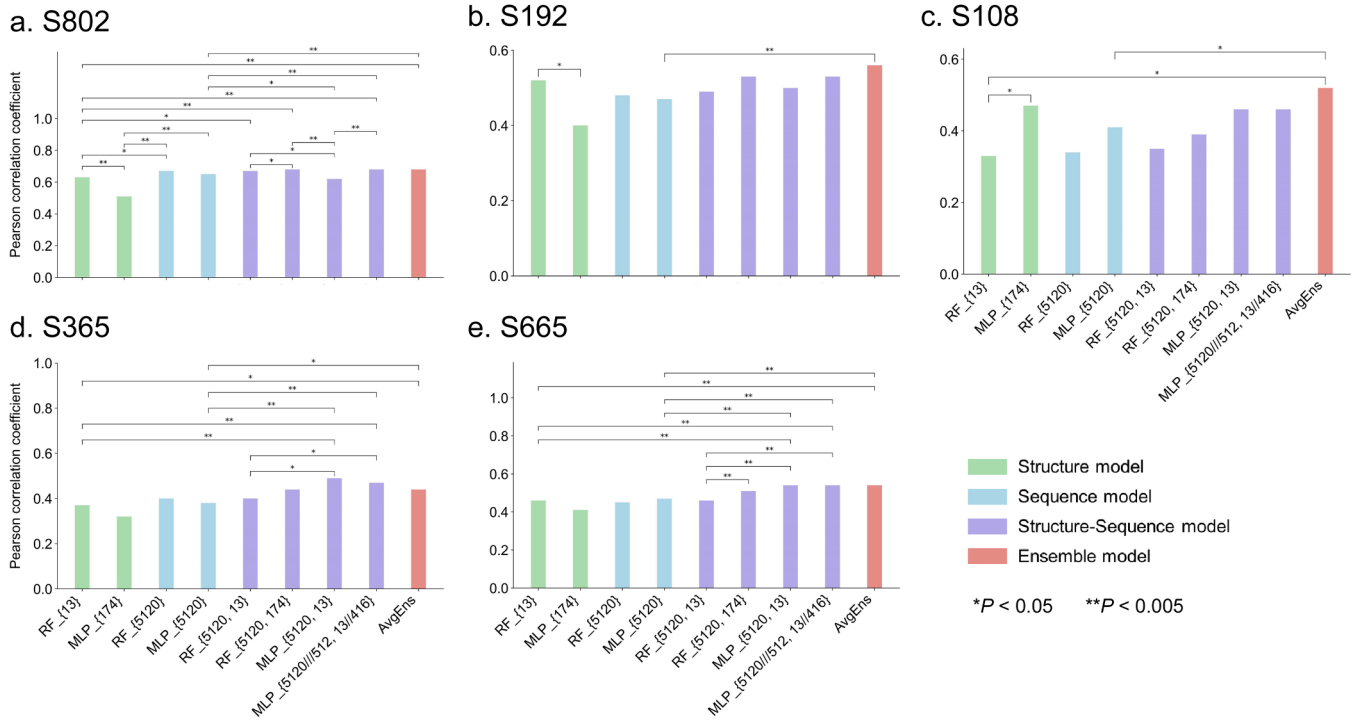

Fig. 3. Pearson correlation coefficient between experimentally-determined and predicted values of binding affinities for nine models tested on five datasets. (a) S802, 5-fold cross-validation results are shown for S802. (b-d) three independent test sets of S192, S108, and S365. (e) S665, a combination of the three test sets. All correlation coefficients presented are significantly different from zero ($P < 0.005$, *t*-test). Significant comparisons were performed within four individual models and four structure-sequence models, as well as between the four structure-sequence models and RF_{13} and MLP_{5120}, respectively. The averaging ensemble model was also compared to RF_{13}, MLP_{5120}, and MLP_{5120///512, 13//416}, respectively. P values were calculated using the Hittner2003 test for comparing two correlation coefficients.

S108 and S365 test sets (see Table 1). While exploring deep learning architectures, it was observed that integrating multi-head attention and skip connections did not lead to a performance improvement beyond that of the simple MLP model. This observation suggests that the MLP architecture is already proficient in extracting relevant information from the input features, rendering the additional complexity of multi-head attention and skip connections unnecessary for our specific prediction task with limited data.

In the comparative analysis between the traditional machine learning model, RF_{5120}, and the deep learning model, MLP_{5120}, it is observed that the PCC values cannot sufficiently differentiate between them (refer to Fig. 3). Consequently, we proceeded to investigate the distribution of predicted values and their performance across three distinct classification tasks. The results reveal that the RF model tends to generate more concentrated predictions within a limited range, while the MLP exhibits a broader distribution of predictions (Fig. S6). Evaluation across the three classification tasks, as illustrated in Fig. S7, indicates that the RF model demonstrates comparatively lower predictive capability than the MLP model in predicting complexes associated with flexibility and functional classifications. In the classification based on interaction strength, most models, like RF_{13} and MLP_{5120}, perform better in predicting permanent complexes in S802 than transient complexes. However, the trend is reversed in S665. Only a few models, such as RF_{5120}, consistently excel in predicting permanent complexes in both S802 and S665 datasets. Further analysis of each component of the MCC reveals that, on S665, the enhanced prediction of permanent complexes by RF_{5120} is primarily attributed to the contribution of TP*TN (TP*TN=63 *439 for permanent complexes and TP*TN=54 *446 for transient complexes). Similarly, MLP_{5120} stands out in predicting transient complexes over permanent ones for the same reason (TP*TN=51 *463 for permanent complexes and TP*TN=148 *285 for transient complexes).

In conclusion, our comparative analysis of RF_{5120} and MLP_{5120} emphasizes distinctions in their predictive patterns and underscores the importance of considering distribution characteristics and task-specific performance metrics. These observations suggest that deep learning architectures, exemplified by the MLP model, possess an enhanced capacity to capture intricate patterns and representations from high-dimensional feature vectors when compared to traditional machine learning techniques. Based on these findings, MLP_{5120} was selected as our final sequence model.

3.3. Ensemble models combining structure and sequence models

As shown in Fig. S2, there are relatively low correlations between the prediction results of sequence and structural models, exemplified by a PCC of 0.47 between RF_{13} and MLP_{5120} when evaluated on the S665 dataset. Consequently, we explored various combinations of the sequence and structural models and found that the averaging ensemble of the two top-performing models, RF_{13} and MLP_{5120}, yielded the top-level overall performance. This outcome is rationalized by the high correlation between the sequence models and the structural models themselves. The combined performance of RF_{13} and MLP_{5120 surpasses that of each individual model (Fig. 3). Furthermore, weighted averaging and stack-based ensemble methods were explored, but they did not exhibit higher performance compared to the simplest average ensemble (Table 1). Therefore, the simplest average ensemble, denoted as AvgEns, was chosen as our ultimate ensemble model. This selection allows us to effectively leverage the unique strengths of each model while alleviating the limitations of their standalone predictions, resulting in a more robust and accurate overall predictive framework.

3.4. Structure-sequence models combining handcrafted and embedding features

Finally, we employed a combination of raw sequence-based embedding features and structure-based handcrafted features to construct a total of eight predictive models, utilizing both MLP and RF algorithms, as illustrated in Fig. 2d and Table S4. The evaluation results, presented in Fig. 3 and Table 1, indicate that incorporating handcrafted features into MLP_{5120} does not significantly improve predictive performance on the test sets of S192 and S108. This observation holds true even after substantial efforts were made to narrow the dimensionality gap between structural and sequence features. Similarly, integrating embedding features into the traditional RF models does not notably enhance predictive performance across all test sets. Based on the outcomes from S802 and S192, we selected MLP_{5120///512, 13//416} as the representative model within this category.

In pursuit of a comprehensive understanding of the predictive reliability of our models, we initiated a statistical uncertainty assessment using the bootstrap approach, and the results are presented in Table S7 and Fig. S8. The Interquartile Ranges (IQR) of the PCCs and RMSEs demonstrate a relatively small dispersion, ranging from 0.09 to 0.14 for PCCs and 0.19 to 0.39 for RMSEs across all models applied to S802 and S665. Notably, as the size of the datasets (S192, S108, and S365) decreases, there is an increase in statistical uncertainty. This phenomenon is attributed to the inherent challenge that arises when working with smaller datasets. In such cases, the new samples obtained through bootstrap may be more susceptible to noise or local features present in the original dataset, leading to a higher variance in the estimation results. In conclusion, our bootstrap analysis affirms that our models consistently provide reliable and stable predictions, especially in larger datasets. This examination further fortifies our confidence in the robustness and generalizability of the predictive capabilities demonstrated by our models.

In summary, we have selected four models: RF_{13}, MLP_{5120}, AvgEns, and MLP_{5120///512, 13//416}, each representing distinct approach in predicting protein-protein binding affinity. RF_{13} stands as a traditional machine learning model incorporating 13 meticulously selected structure-based handcrafted features, affording interpretability to the predictions. On the other hand, MLP_{5120} adopts a deep learning approach utilizing sequence-based transferred embedding features, offering computational speed advantage and independence from the 3D structure of the complex. Both models exhibit comparable performance, as demonstrated in Fig. 3. The AvgEns model synergizes the predictions of RF_{13} and MLP_{5120}, harnessing the complementary strengths of both models and yielding superior predictive performance compared to individual models (Fig. 3 and Fig. 4). Finally, MLP_{5120///512, 13//416} represents an endeavor to integrate raw sequence and structure-based features. However, the integrated model's performance falls short of that achieved by AvgEns. Consequently, AvgEns stands as our ultimate combination model for predicting protein-protein binding affinity.

3.5. Performance on protein-protein interaction classification

In this investigation, we undertook a classification analysis of protein-protein interactions based on three distinctive characteristics: interaction strength, flexibility, and function of the complexes. The evaluation utilized the S802 and S665 datasets, and the outcomes are delineated in Fig. 5. In the prediction of interaction strength, although the classification performance is not high, there is still a significant differentiation from random. The ensemble model AvgEns significantly outperforms each individual model ($P < 0.05$, DeLong test), while RF_{13} and MLP_{5120} show comparable performance. The challenge of limited data volume at both extremes is a common obstacle faced by nearly all machine learning prediction models when dealing with extreme values. Concerning the prediction of flexibility, RF_{13} manifested higher PCC values for rigid-body complexes compared to MLP_{5120}, although the difference was not statistically significant ($P > 0.05$, Hittner2003 test). The AvgEns model shows the best performance on both rigid-body and flexible complexes. Addressing functional

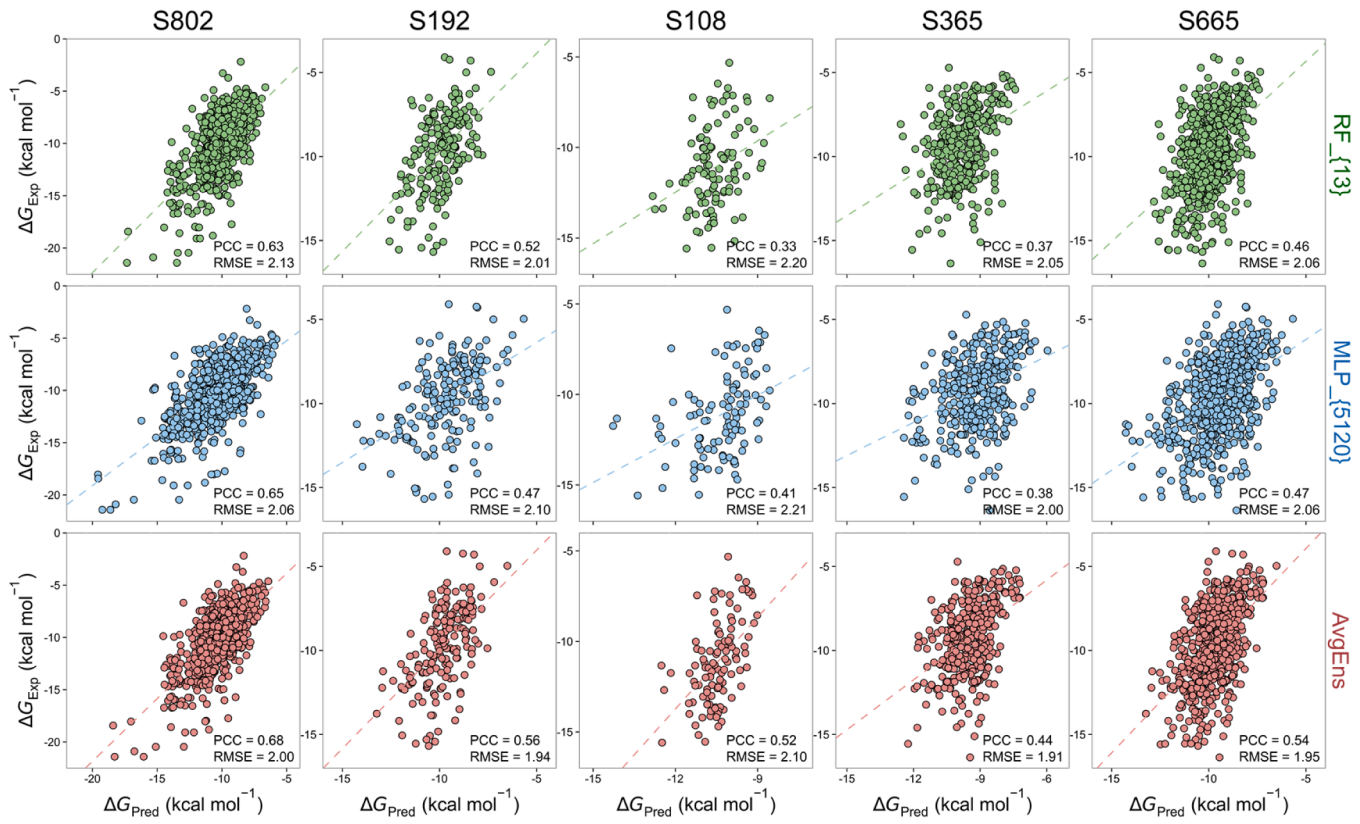

Fig. 4. Performance of three selected representative methods, RF_{13}, MLP_{5120}, and AvgEns, on five datasets. 5-fold cross-validation results are shown for S802. All correlation coefficients are statistically significantly different from zero ($P < 0.005$, *t*-test). PCC: Pearson correlation coefficient, RMSE (kcal mol^{-1}): root-mean-square error.

classes, previous studies have indicated disparate predictability for different types of functional complexes [41,46]. Our three models exhibit commendable performance across three types of complexes (EI, EO, and OX), with statistically significant PCC values. However, no statistically significant correlation is observed for OG complexes from S665. Predicting the binding affinity of complexes involving antibodies has perpetually presented challenges. Nevertheless, our sequence-based model achieved notable PCC values in this context, signifying promising performance in forecasting the binding strength of antibody complexes.

3.6. Comparative analysis with other approaches

Fig. 6 and Table S8 present a comprehensive performance comparison among our three representative methods (RF_{13}, MLP_{5120}, and AvgEns), five machine learning approaches trained on affinity data, three absolute energy calculation methods, and ten docking scores. Our methods consistently demonstrate superior performance across all datasets in comparison to the other approaches. Among the five machine learning approaches, PPI-Affinity shows relatively good performance with PCC values of 0.42 and 0.34 for S802 and S192, respectively. This performance may be attributed to the overlap between its training set and the S802 and S192 datasets (Table S6). The three absolute energy calculation methods (FoldX, MMPBSA, and Rosetta) are observed to be sensitive to different structure optimization approaches (Fig. S9). Hence, the highest PCC values for these methods on each dataset are reported. However, these methods demonstrate limited predictive power, with non-significant or very low PCC values for the S802 and S192 datasets. Additionally, the large RMSE values suggest that the predicted values from these methods cannot be directly interpreted as binding energy. As for the ten docking scores, they also exhibit very limited ability to predict affinity. This observation aligns with the understanding that the native conformation of the complex may not necessarily correspond to the one with the lowest binding energy, as the interaction is intended to generate a specific biological function rather than solely achieving high affinity [93,94].

Overall, in this study, we explored a range of approaches that incorporate transfer learning, integrate domain knowledge, and employ a combination of deep learning and traditional machine learning algorithms. These strategies collectively mitigate the impact of data limitations and lead to significant advancements in predicting protein-protein binding affinity. Our study yields the following insights:

- The integration of features extracted from extensive pre-trained models, when combined with deep learning techniques, yields promising predictive performance.
- Traditional machine learning methods, when coupled with carefully curated structural features grounded in prior knowledge, remain a valuable choice.
- The synergy between models leveraging transferred embedding features and those incorporating handcrafted features results in improved performance.

In light of these discoveries, we devised sequence-based and structure-based models capable of accurately estimating protein-protein binding affinity. These methods hold great potential for aiding protein engineering endeavors by offering valuable starting points, minimizing the risk of unsuccessful laboratory experiments, and facilitating the design and development of therapeutic proteins.

Despite the advancements made by our models in predicting affinity, persistent challenges related to prediction accuracy, robustness, generalizability, and interpretability necessitate ongoing attention. Our study introduces three models—RF_{13}, MLP_{5120}, and their ensemble AvgEns—each presenting specific limitations and potential constraints.

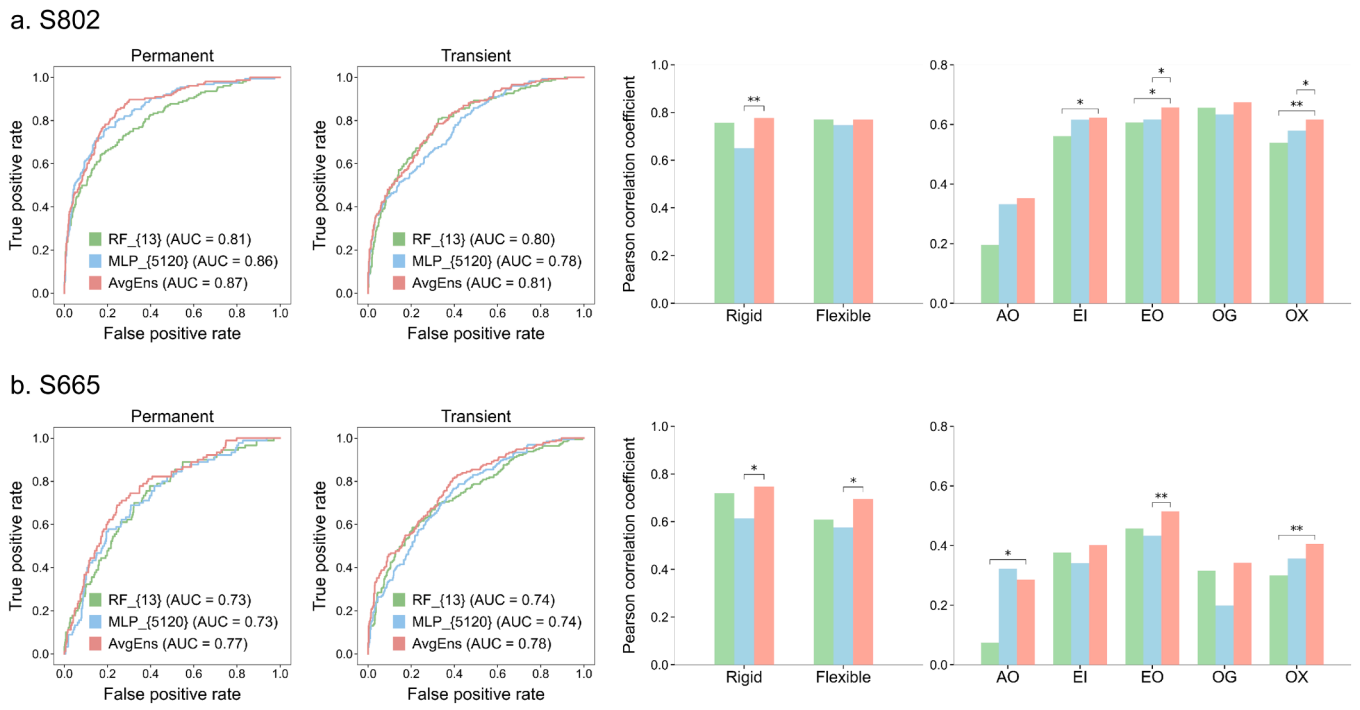

Fig. 5. Performance of three methods, RF_{13}, MLP_{5120}, and AvgEns, was evaluated on datasets S802 (a) and S665 (b) for three distinct interaction classifications. Receiver operating characteristics curves for three approaches to distinguish permanent and transient protein-protein interactions from others. Pearson correlation coefficients for rigid-body and flexible complexes, and for five functional categorizations of complexes. The number of complexes in each category is provided in Fig. 1. P values were calculated using the Hittner2003 test for comparing two correlation coefficients (* $P < 0.05$ and ** $P < 0.005$). The PCC values for RF_{13} applied on AO from both datasets do not have statistically significant difference from zero ($P > 0.05$, t-test). Additionally, the PCC values for all three methods applied on OG from S665 are not statistically significant either. The rest of PCC values are significantly different from zero ($P < 0.05$, t-test). See Supplementary Figure 7 for significant analysis of all PCC values.

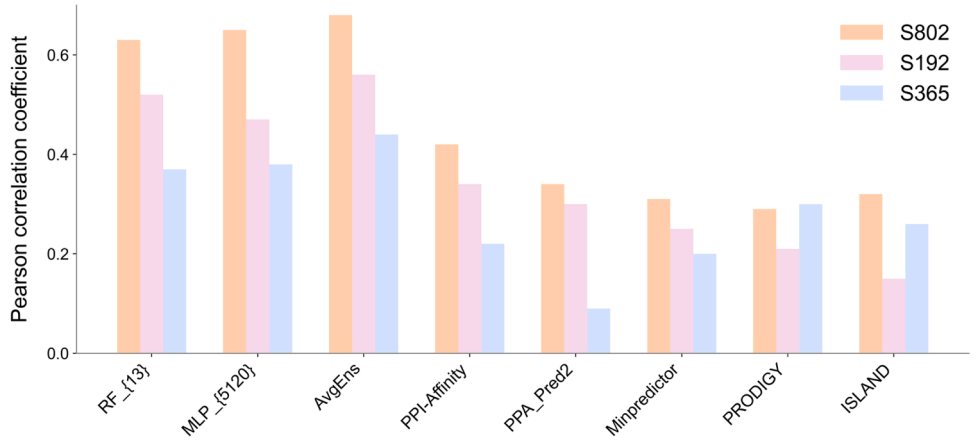

Fig. 6. Comparison of methods' performances. Pearson correlation coefficients are shown for our three representative methods and five machine learning approaches trained on binding affinity data. Among these five other approaches, only PRODIGY has the capability to predict multimers. Across all datasets, our method of AvgEns significantly outperform all the other approaches ($P < 0.05$, Hittner2003 test). The PCC value for PPA_Pred2 tested on S365 does not exhibit a statistically significant difference from zero ($P > 0.05$, t-test). For more detailed results, refer to Table S8.

It is imperative to acknowledge these factors when deploying these models across different scenarios or datasets.

- Dependency on data quality and size: The performance of all models relies on the quality and representativeness of the training data. Limitations in the diversity and size of the training dataset may impact the generalizability of the models to new and unseen data.
- Interpretability vs. applicability trade-off: RF_{13}, being a traditional machine learning model with 13 handcrafted features, offers interpretability to predictions. However, the scarcity of experimentally-determined protein complex structures limits its application. In contrast, MLP_{5120}, a deep learning model that only requires protein sequences as inputs, enhances its applicability but sacrifices interpretability.
- Challenges in predicting extreme values: The models encounter limitations in effectively predicting extreme values. This constraint is acknowledged as a common issue in various machine learning prediction models due to the limited data volume at both extremes in the training sets.

- Limited predictive power for certain complex types: Despite overall good performance, the models face challenges in accurately predicting the binding affinity for certain complex types, such as AO and OG complexes.

CRediT authorship contribution statement

Conceptualization, M.L.; Methodology, F.Z., X.J., Y.W., and M.L.; Software, F.Z. and X.J.; Validation, F.Z., X.J., and M.L.; Formal Analysis, F.Z.; Investigation, F.Z., X.J., and M.L.; Data Curation, F.Z. and X.J.; Writing – Original and Revised Drafts, M.L.; Writing – Review & Editing, M.L.; Visualization, F.Z., Y.Y., and M.L.; Supervision, M.L.; Project Administration, M.L.; Funding Acquisition, M.L.

Declaration of Competing Interest

The authors declare no competing interests.

Data Availability

The compiled experimental datasets and computational results that support our findings are publicly available on GitHub at https://github.com/minghuilab/BindPPI. The programs for running our three representative models, RF_{13}, MLP_{5120}, and AvgEns, are also available at https://github.com/minghuilab/BindPPI. Our methods are named as "BindPPI". Additional data files and codes that support the findings of this study are available from the corresponding authors upon request

Acknowledgments

This work was supported by the National Natural Science Foundation of China [32070665] and the Priority Academic Program Development of Jiangsu Higher Education Institutions. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2023.12.018.

References

[1] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–44.
[2] Jumper J, et al. Highly accurate protein structure prediction with AlphaFold. Nature 2021;596:583–9.
[3] Min S, Lee B, Yoon S. Deep learning in bioinformatics. Brief Bioinform 2017;18:851–69.
[4] Chen Y, Lu H, Zhang N, Zhu Z, Wang S, Li M. PremPS: Predicting the impact of missense mutations on protein stability. PLoS Comput Biol 2020;16:e1008543.
[5] Zhang N, et al. MutaBind2: predicting the impacts of single and multiple mutations on protein-protein interactions. iScience 2020;23:100939.
[6] Li M, Simonetti FL, Goncearenco A, Panchenko AR. MutaBind estimates and interprets the effects of sequence variants on protein-protein interactions. Nucleic Acids Res 2016;44:W494–501.
[7] Sun T, Chen Y, Wen Y, Zhu Z, Li M. PremPLI: a machine learning model for predicting the effects of missense mutations on protein-ligand interactions. Commun Biol 2021;4:1311.
[8] Li M, Petukh M, Alexov E, Panchenko AR. Predicting the impact of missense mutations on protein-protein binding affinity. J Chem Theory Comput 2014;10:1770–80.
[9] Zhang N, et al. PremPRI: predicting the effects of missense mutations on protein-RNA interactions. Int J Mol Sci 2020;(21, ().
[10] Zhang N, Chen Y, Zhao F, Yang Q, Simonetti FL, Li M. PremPDI estimates and interprets the effects of missense mutations on protein-DNA interactions. PLoS Comput Biol 2018;14:e1006615.
[11] Pancotti C, et al. Predicting protein stability changes upon single-point mutation: a thorough comparison of the available tools on a new dataset. Brief Bioinform 2022;23.
[12] Huang YQ, Sun P, Chen Y, Liu HX, Hao GF, Song BA. Bioinformatics toolbox for exploring target mutation-induced drug resistance. Brief Bioinform 2023;24.
[13] Sequeiros-Borja CE, Surpeta B, Brezovsky J. Recent advances in user-friendly computational tools to engineer protein function. Brief Bioinform 2021;22.
[14] Pucci F, Schwersensky M, Rooman M. Artificial intelligence challenges for predicting the impact of mutations on protein stability. Curr Opin Struct Biol 2022;72:161–8.
[15] Benevenuta S, Pancotti C, Fariselli P, Birolo G, Sanavia T. An antisymmetric neural network to predict free energy changes in protein variants. J Phys D: Appl Phys 2021;54:245403.
[16] Nikam R, Yugandhar K, Michael Gromiha M. Discrimination and prediction of protein-protein binding affinity using deep learning approach. Intell Comput Theor Appl 2018:809–15.
[17] Wang B, Mao J, Wei M, Qi Y, Zhang JZH. SeBPPI: a sequence-based protein–protein binding predictor. J Comput Biophys Chem 2022;21:729–37.
[18] Jones S, Thornton JM. Principles of protein-protein interactions. Proc Natl Acad Sci USA 1996;93:13–20.
[19] Wodak SJ, Vlasblom J, Turinsky AL, Pu S. Protein-protein interaction networks: the puzzling riches. Curr Opin Struct Biol 2013;23:941–53.
[20] Loregian A, Palu G. Disruption of protein-protein interactions: towards new targets for chemotherapy. J Cell Physiol 2005;204:750–62.
[21] Goncearenco A, Li M, Simonetti FL, Shoemaker BA, Panchenko AR. Exploring protein-protein interactions as drug targets for anti-cancer therapy with in silico workflows. Methods Mol Biol 2017;1647:221–36.
[22] Blazer LL, Neubig RR. Small molecule protein-protein interaction inhibitors as CNS therapeutic agents: current progress and future hurdles. Neuropsychopharmacology 2009;34:126–41.
[23] Wang B, Gallolu Kankanamalage S, Dong J, Liu Y. Optimization of therapeutic antibodies. Antib Ther 2021;4:45–54.
[24] Kastritis PL, Bonvin AM. On the binding affinity of macromolecular interactions: daring to ask why proteins interact. J R Soc Interface 2013;10:20120835.
[25] Dar KB, et al. Exploring proteomic drug targets, therapeutic strategies and protein - protein interactions in cancer: mechanistic view. Curr Cancer Drug Targets 2019;19:430–48.
[26] Willander M, Al-Hilli S. Analysis of biomolecules using surface plasmons. Methods Mol Biol 2009;544:201–29.
[27] Ladbury JE, Chowdhry BZ. Sensing the heat: the application of isothermal titration calorimetry to thermodynamic studies of biomolecular interactions. Chem Biol 1996;3:791–801.
[28] Phillip Y, Kiss V, Schreiber G. Protein-binding dynamics imaged in a living cell. Proc Natl Acad Sci 2012;109:1461–6.
[29] Aloy P, Russell RB. Structural systems biology: modelling protein interactions. Nat Rev Mol Cell Biol 2006;7:188–97.
[30] Fleishman SJ, et al. Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. Science 2011;332:816–21.
[31] Siebenmorgen T, Zacharias M. Computational prediction of protein–protein binding affinities. WIREs Comput Mol Sci 2019;10.
[32] Wang L, Berne BJ, Friesner RA. On achieving high accuracy and reliability in the calculation of relative protein-ligand binding affinities. Proc Natl Acad Sci USA 2012;109:1937–42.
[33] Bhati AP, Wan S, Wright DW, Coveney PV. Rapid, accurate, precise, and reliable relative free energy prediction using ensemble based thermodynamic integration. J Chem Theory Comput 2017;13:210–22.
[34] Rastelli G, Del Rio A, Degliesposti G, Sgobba M. Fast and accurate predictions of binding free energies using MM-PBSA and MM-GBSA. J Comput Chem 2010;31:797–810.
[35] Panday SK, Alexov E. Protein-Protein Binding Free Energy Predictions with the MM/PBSA Approach Complemented with the Gaussian-Based Method for Entropy Estimation. ACS Omega 2022;7:11057–67.
[36] Su Y, Zhou A, Xia X, Li W, Sun Z. Quantitative prediction of protein-protein binding affinity with a potential of mean force considering volume correction. Protein Sci 2009;18:2550–8.
[37] Zhang C, Liu S, Zhu Q, Zhou Y. A knowledge-based energy function for protein-ligand, protein-protein, and protein-DNA complexes. J Med Chem 2005;48:2325–35.
[38] Romero-Molina S, et al. PPI-Affinity: a web tool for the prediction and optimization of protein-peptide and protein-protein binding affinity. J Proteome Res 2022;21:1829–41.
[39] Gromiha MM, Yugandhar K, Jemimah S. Protein-protein interactions: scoring schemes and binding affinity. Curr Opin Struct Biol 2017;44:31–8.
[40] Moal IH, Agius R, Bates PA. Protein-protein binding affinity prediction on a diverse set of structures. Bioinformatics 2011;27:3002–9.
[41] Vreven T, Hwang H, Pierce BG, Weng Z. Prediction of protein-protein binding free energies. Protein Sci 2012;21:396–404.
[42] Choi JM, et al. Minimalistic predictor of protein binding energy: contribution of solvation factor to protein binding. Biophys J 2015;108:795–8.
[43] Abbasi WA, Yaseen A, Hassan FU, Andleeb S, Minhas F. ISLAND: in-silico proteins binding affinity prediction using sequence information. BioData Min 2020;13:20.
[44] Vangone A, Bonvin AM. Contacts-based prediction of binding affinity in protein-protein complexes. Elife 2015;4:e07454.
[45] Xue LC, Rodrigues JP, Kastritis PL, Bonvin AM, Vangone A. PRODIGY: a web server for predicting the binding affinity of protein-protein complexes. Bioinformatics 2016;32:3676–8.
[46] Yugandhar K, Gromiha MM. Protein-protein binding affinity prediction from amino acid sequence. Bioinformatics 2014;30:3583–9.
[47] Lin Z, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 2023;379:1123–30.
[48] Vreven T, et al. Updates to the integrated protein-protein interaction benchmarks: docking benchmark version 5 and affinity benchmark version 2. J Mol Biol 2015;427:3031–41.

[49] Jankauskaite J, Jimenez-Garcia B, Dapkunas J, Fernandez-Recio J, Moal IH. SKEMPI 2.0: an updated benchmark of changes in protein-protein binding energy, kinetics and thermodynamics upon mutation. Bioinformatics 2019;35:462–9.
[50] Jemimah S, Yugandhar K, Michael Gromiha M. PROXiMATE: a database of mutant protein-protein complex thermodynamics and kinetics. Bioinformatics 2017;33:2787–8.
[51] Wang R, Fang X, Lu Y, Wang S. The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. J Med Chem 2004;47:2977–80.
[52] Webb B, Sali A. Comparative protein structure modeling using MODELLER. 5 6 1-5 6 37 Curr Protoc Bioinforma 2016;54. 5 6 1-5 6 37.
[53] Berman HM, et al. The Protein Data Bank. Nucleic Acids Res 2000;28:235–42.
[54] Humphrey W, Dalke A, Schulten K. VMD: visual molecular dynamics. J Mol Graph 1996;14(33-38):27–38.
[55] MacKerell AD, et al. All-atom empirical potential for molecular modeling and dynamics studies of proteins. J Phys Chem B 1998;102:3586–616.
[56] Phillips JC, et al. Scalable molecular dynamics with NAMD. J Comput Chem 2005;26:1781–802.
[57] Steinegger M, Soding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol 2017;35:1026–8.
[58] Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 2005;33:2302–9.
[59] Pedregosa F., *et al.* Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12, 2825–2830.
[60] Brooks B.R., Bruccoleri R.E., Olafson B.D., States D.J., Swaminathan Sa, Karplus MJJocc. CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. 4, 187–217 (1983).
[61] Joosten RP, et al. A series of PDB related databases for everyday needs. Nucleic Acids Res 2011;39:D411–419.
[62] Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–637.
[63] Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and indels. PLoS One 2012;7:e46688.
[64] Hou Q, Kwasigroch JM, Rooman M, Pucci F. SOLart: a structure-based method to predict protein solubility and aggregation. Bioinformatics 2020;36:1445–52.
[65] Rose GD, Geselowitz AR, Lesser GJ, Lee RH, Zehfus MH. Hydrophobicity of amino acid residues in globular proteins. Science 1985;229:834–8.
[66] Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M. AAindex: amino acid index database, progress report 2008. Nucleic Acids Res 2008;36:D202–205.
[67] Mitaku S, Hirokawa T, Tsuji T. Amphiphilicity index of polar amino acids as an aid in the characterization of amino acid preference at membrane-water interfaces. Bioinformatics 2002;18:608–16.
[68] Anishchenko I, Kundrotas PJ, Vakser IA. Contact potential for structure prediction of proteins and protein complexes from potts model. Biophys J 2018;115:809–21.
[69] Till MS, Ullmann GM. McVol - a program for calculating protein volumes and identifying cavities by a Monte Carlo algorithm. J Mol Model 2010;16:419–29.
[70] Kastritis PL, et al. A structure-based benchmark for protein-protein binding affinity. Protein Sci 2011;20:482–91.
[71] La D, Kong M, Hoffman W, Choi YI, Kihara D. Predicting permanent and transient protein-protein interfaces. Proteins 2013;81:805–18.
[72] Schneider C, Raybould MIJ, Deane CM. SAbDab in the age of biotherapeutics: updates including SAbDab-nano, the nanobody structure tracker. D1368-D1372 Nucleic Acids Res 2022;50. D1368-D1372.
[73] Dunbar J, et al. SAbDab: the structural antibody database. D1140-1146 Nucleic Acids Res 2014;42. D1140-1146.
[74] Gutmanas A, et al. PDBe: Protein Data Bank in Europe. D285-D291 Nucleic Acids Res 2014;42. D285-D291.
[75] Guerois R, Nielsen JE, Serrano L. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. J Mol Biol 2002;320:369–87.
[76] Fleishman SJ, et al. RosettaScripts: a scripting language interface to the rosetta macromolecular modeling suite. PLOS ONE 2011;6:e20161.
[77] Kollman P.A., *et al.* Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models.
[78] Kellogg EH, Leaver-Fay A, Baker D. Role of conformational sampling in computing mutation-induced changes in protein structure and stability. Proteins 2011;79:830–8.
[79] Moal IH, Jimenez-Garcia B, Fernandez-Recio J. CCharPPI web server: computational characterization of protein-protein interactions from structure. Bioinformatics 2015;31:123–5.
[80] Pierce B, Weng Z. ZRANK: reranking protein docking predictions with an optimized energy function. Proteins 2007;67:1078–86.
[81] Pierce B, Weng Z. A combination of rescoring and refinement significantly improves protein docking performance. Proteins 2008;72:270–9.
[82] Chaudhury S, Lyskov S, Gray JJ. PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. Bioinformatics 2010;26:689–91.
[83] Cheng TM, Blundell TL, Fernandez-Recio J. pyDock: electrostatics and desolvation for effective scoring of rigid-body protein-protein docking. Proteins 2007;68:503–15.
[84] Andrusier N, Nussinov R, Wolfson HJ. FireDock: fast interaction refinement in molecular docking. Proteins 2007;69:139–59.
[85] Viswanath S, Ravikant DV, Elber R. Improving ranking of models for protein complexes with side chain modeling and atomic potentials. Proteins 2013;81:592–606.
[86] Ravikant DV, Elber R. PIE-efficient filters and coarse grained potentials for unbound protein-protein docking. Proteins 2010;78:400–19.
[87] Pons C, Talavera D, de la Cruz X, Orozco M, Fernandez-Recio J. Scoring by intermolecular pairwise propensities of exposed residues (SIPPER): a new efficient potential for protein-protein docking. J Chem Inf Model 2011;51:370–7.
[88] Hittner JB, May K, Silver NC. A Monte Carlo evaluation of tests for comparing dependent correlations. J Gen Psychol 2003;130:149–68.
[89] Delong ER, Delong DM, Clarkepearson DI. Comparing the areas under 2 or more correlated receiver operating characteristic curves - a nonparametric approach. Biometrics 1988;44:837–45.
[90] Diedenhofen B, Musch J. cocor: a comprehensive solution for the statistical comparison of correlations. PLOS ONE 2015;10:e0121945.
[91] Virtanen P., *et al.* SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods 17, 261–272.
[92] Zhu W, Shenoy A, Kundrotas P, Elofsson A. Evaluation of alphafold-multimer prediction on multi-chain protein complexes. Bioinformatics 2023;39.
[93] Kozakov D, et al. The ClusPro web server for protein-protein docking. Nat Protoc 2017;12:255–78.
[94] Pantsar T, Poso A. Binding affinity via docking: fact and fiction. Molecules 2018;23.