Received: 30 June 2022 | Revised: 8 March 2023 | Accepted: 12 April 2023

DOI: 10.1002/prot.26503

RESEARCH ARTICLE

PROTEINS WILEY

An updated dataset and a structure-based prediction model for protein–RNA binding affinity

Xu Hong[1] | Xiaoxue Tong[1] | Juan Xie[1] | Pinyu Liu[1] | Xudong Liu[1] | Qi Song[2] | Sen Liu[2] | Shiyong Liu[1]

[1]School of Physics, Huazhong University of Science and Technology, Wuhan, Hubei, China

[2]Key Laboratory of Fermentation Engineering (Ministry of Education), Hubei University of Technology, Wuhan, China

Correspondence
Shiyong Liu, School of Physics, Huazhong University of Science and Technology, Wuhan, Hubei 430074, China.
Email: liushiyong@gmail.com

Funding information
National Natural Science Foundation of China, Grant/Award Numbers: 31100522, 32271267; Fundamental Research Funds for the Central Universities; National High Technology Research and Development Program of China, Grant/Award Number: 2012AA020402

Abstract

Understanding the process of protein–RNA interaction is essential for structural biology. The thermodynamic process is an important part to uncover the protein–RNA interaction mechanism. The regulatory networks between protein and RNA in organisms are dominated by the binding or dissociation in the cells. Therefore, determining the binding affinity for protein–RNA complexes can help us to understand the regulation mechanism of protein–RNA interaction. Since it is time-consuming and labor-intensive to determine the binding affinity for protein–RNA complexes by experimental methods, it is necessary and urgent to develop computational methods to predict that. To develop a binding affinity prediction model, first we update the dataset of protein–RNA binding affinity benchmark (PRBAB), which includes 145 complexes now. Second, we extract the structural features based on complex structure, and then we analyze and select the representative structural features to train the regression model. Third, we random select the subset from the PRBAB2.0 to fit the protein–RNA binding affinity determined by experiment. In the end, we tested our model on the nonredundant PDBbind dataset, and the results showed that Pearson correlation coefficient $r = .57$ and RMSE = 2.51 kcal/mol. The Pearson correlation coefficient achieves 0.7 while removing 5 complex structures with modified residues/nucleotides and metal ions. While testing on ProNAB, the results showed that 71.60% of the prediction achieves Pearson correlation coefficient $r = .61$ and RMSE = 1.56 kcal/mol with experiment values.

KEYWORDS
binding affinity, feature selection, protein–RNA interaction, regression model, structural features

1 | INTRODUCTION

Protein–RNA interactions participate in many biological functions in organisms.[1,2] The driving force of protein-RNA interaction is the kinetics, which ultimately regulate gene expression.[3–5] The mutation that occurred in the complex can increase or decrease the binding intensity, which can further lead to disorders of metabolism and cause diseases. Knowledge about protein-RNA binding affinity is essential for understanding protein–RNA recognition mechanism in post-transcriptional gene regulation. The protein–RNA binding affinity, described as the equilibrium dissociation constant, can be measured by isothermal titration calorimetry (ITC), surface plasmon resonance (SPR), electrophoretic mobility shift assay (EMSA), filter binding assay (FBA), and dynamic light scattering (DLS).[6–10] Since experimentally

Xu Hong and Xiaoxue Tong contributed equally to this work.

determining the binding affinity between protein and RNA is time consuming and laborious process, it is needed to develop the computational method instead.[11]

In the prediction of protein–RNA complex structures, the developed scoring functions based on the hypothesis that the near-native structure with the lowest energy are also used to calculate the binding affinity of the decoys, such as DRNA.[12] Besides, a series of docking scoring functions have been developed to rank the protein–RNA docking decoy structures.[13–16] Although these docking scoring functions can distinguish near native structures from decoys, our previous work shows that the predicted binding free energy from three available scoring functions has a low correlation with the experimental binding affinity.[17] After that, several methods have been developed to predict protein–RNA binding affinity during the last few years. Dias et al.[18] developed a generalized linear modeling (GLM)-score to predict protein–RNA binding affinity, which uses structural features including hydrogen bonds, hydrophobic contacts, van der Waals, and the deformation effect. Nithin et al.[19] predicted the binding affinity by using the molecules conformation change and other structural features with bound and unbound structure. In most instances, it is difficult to find the unbound structure of the complex. Therefore, it maybe not easy to apply it on large scales for prediction. In PredPRBA, Deng et al. combined the sequential and structural features of proteins and RNA and used a gradient boosted regression tree algorithms to predict the binding affinity of protein–RNA complexes.[20] They trained the prediction model based on complexes with different types. Therefore, these models cannot be used to predict binding affinity in other cases. Chen et al. predicted the protein–RNA binding affinity by the molecular mechanics/Poisson Boltzmann surface area (MM) and MM/generalized Born surface area approaches based on molecular dynamic simulations.[21] Though the method is time-consuming to predict the protein-RNA binding affinity, it can achieve a Pearson correlation 0.57 while testing.

The prediction of protein–RNA binding affinity lags far behind the study of protein–protein binding affinity prediction.[22–28] The main reason may be that there is not enough data, which makes it difficult to develop a new method to predict protein–RNA binding affinity. At present, there are three protein–RNA binding affinity datasets, which include PDBbind,[29] ProNAB,[30] and our own protein–RNA binding affinity benchmark (PRBAB).[17] In the current version of PDBbind (version 2020), it includes 142 protein–RNA binding affinity data. The authors do not collect the experimental conditions for these data, which makes it difficult to apply these data to train the binding affinity prediction model. For example, we used experimental temperature to calculate delta free energy in Equation (1). The ProNAB includes 264 sequence-based protein–RNA binding affinity data, but it does not remove sequence highly similarity structures which make the trained model easier and tend to be overfitting. In 2013, our group developed a dataset collected 73 nonredundant protein–RNA binding affinity data named PRBAB v1.0.[17] Based on this, in this article, we update the protein–RNA binding affinity dataset into a larger dataset with 145 nonredundant protein–RNA complexes at present. Compared to the other two binding affinity datasets, PRBAB v2.0 includes 47 new binding affinity data that are not included in the other datasets. Besides, it also provides a nonredundant binding affinity data to evaluate the scoring function in docking. Then, we fit a model PRdeltaGPred (protein–RNA delta G prediction) to predict protein–RNA binding affinity based on structural features, which are proved to be important in protein–protein binding affinity prediction. Since these features are highly correlated with each other, we cluster these features and select the representative features. In order to reduce the impact of experiment errors in the training process, we train the model by randomly selecting a subset of the dataset. Finally, to test the accuracy of our method, we predicted the binding affinity of the protein–RNA complex structures extracted from PDBbind[31] and compared it with the experiment values. The results show that our model achieves a Pearson correlation coefficient 0.57 while testing on 41 nonredundant protein–RNA complexes in PDBbind.

2 | MATERIALS AND METHODS

2.1 | The dataset of protein–RNA binding affinity

In PRBAB v1.0,[17] we had collected 73 protein–RNA complex structures with the binding affinity from PDB. But the number of protein–RNA complex structures in PDB has doubled since 2013. At the same time, the number of protein–RNA complex structures with binding affinity data are gradually increasing, which enables us to update the binding affinity dataset. More and more binding data can be used to build an effective protein–RNA binding model.

Compared with PRBAB v1.0, in PRBAB v2.0, we collect data in a semiautomatic way, which makes data collection more efficient. As of July 20, 2018, there are 1199 new added protein–RNA complex structures. As the same as before,[17] these structures whose protein or RNA is too shorter are filtered out. Finally, 876 protein–RNA complex structures are extracted. And then, these complex structures are grouped into 456 clusters by BLASTClust[32] with a threshold of protein sequences identity at 70%. After that, 73 clusters contain structures that we have collected before 2013, so the complexes in these clusters are removed, and 386 clusters are kept finally. Since the value of the binding affinity exists in the literature in the form of equilibrium dissociation constants (K_d, $1/K_a$, K_{off}/K_{on}), we search the literature through these key words. The data construction flowchart is shown in Figure 1. Finally, 145 protein–RNA complex structures with binding affinity in the dataset in total. So far, the updated dataset PRBAB v2.0 contains a total of 145 protein–RNA complex structures. The details, such as the pH value, temperature, and the reference, about PRBAB v2.0 are shown in Table S1. Unlike previous method,[21] which presents the free energy with pK_d representation, in order to be consistent with the data of PRBAB v1.0, we calculated it according to the Gibbs free energy formula[33] as follows.

$$\Delta G = RT \ln K_d \tag{1}$$

The value of the R is 8.314 J K^{-1} mol^{-1}, and the T represents the temperature (K).

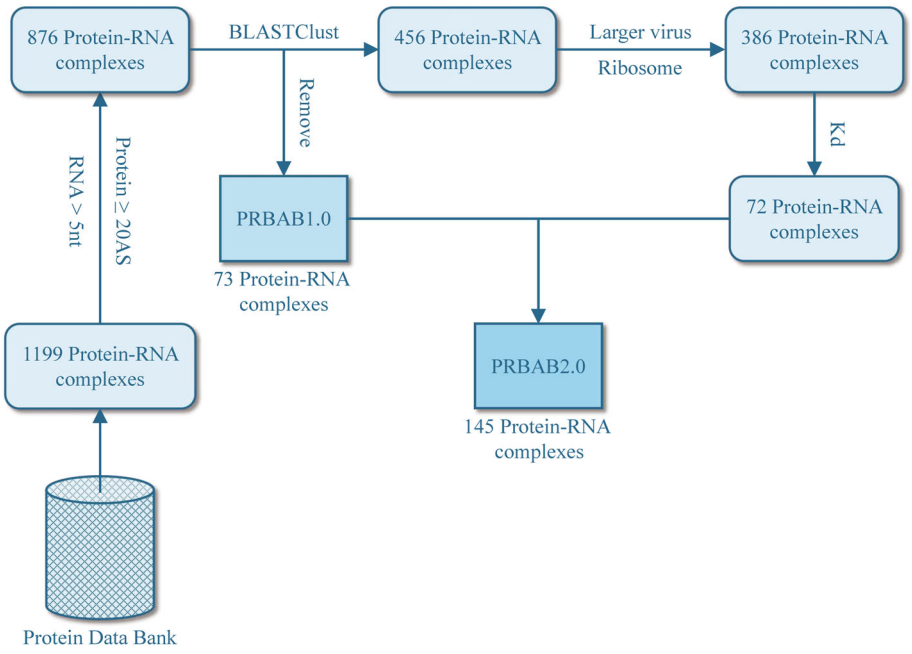

FIGURE 1 The flowchart of building the binding affinity dataset of protein-RNA complexes. We download 1199 protein–RNA complexes from PDB, 876 complexes are retained after deleting less than five nucleotides of RNA and less than 20 amino acids of protein. These complexes were clustered according to 70% sequence identity using BLASTClust, and the 73 protein–RNAs of protein–RNA binding affinity benchmark (PRBAB) v1.0 were removed, and 456 protein–RNA complexes are kept. We remove ribosomes and large viruses. After finding binding affinity data from the remaining complexes, and the K_d data of 72 complexes are finally obtained. Combined with PRBAB v1.0, PRBAB v2.0 contains 145 binding affinity data.

There are another two datasets that include protein–RNA binding affinity, which make it possible to evaluate our method with more data. The PDBbind database includes 1052 protein–nucleic acid binding affinity data. From the database, we selected protein–RNA complexes and removed the complex structures in which the protein sequence identity cutoff larger than 0.7 with the complex structures in PRBAB v2.0. Finally, there are 41 protein–RNA complexes in our first validation dataset (the ID of the complex, the experimental value, and the predicted values of our model are all in Table S2).

In addition to mentioned above, the ProNAB[30] also includes protein–RNA binding affinity data. We compare our dataset with PDBbind and ProNAB. As shown in Figure 3, there are 47 protein–RNA complexes in our datasets are unique (the comparison is discussed in Section 3).

2.2 | Structure-based features

Since the traditional physical interaction energy terms (such as hydrogen bonds, electrostatic potentials, desolation energy, and van der Waal interaction potentials)[34,35] and structural features (including noninteracting interaction and interface contact [IC]) have been proved useful to predict the protein–protein binding affinity.[36] In order to model the binding affinity of protein–RNA, some characteristics that have been proven correlated with protein–protein binding affinity are used in our binding affinity model. In the process of calculating structural features, as the same as the previous work,[26,36,37] we classified the residues into polarity, nonpolarity, and charged according to the different physicochemical characteristics. Polar residues include (C, H, N, Q, S, T, Y, W), nonpolar residues include (A, F, G, I, L, V, M, P), and charged residues (E, D, K, R). Details of these features are encoded in the following sections.

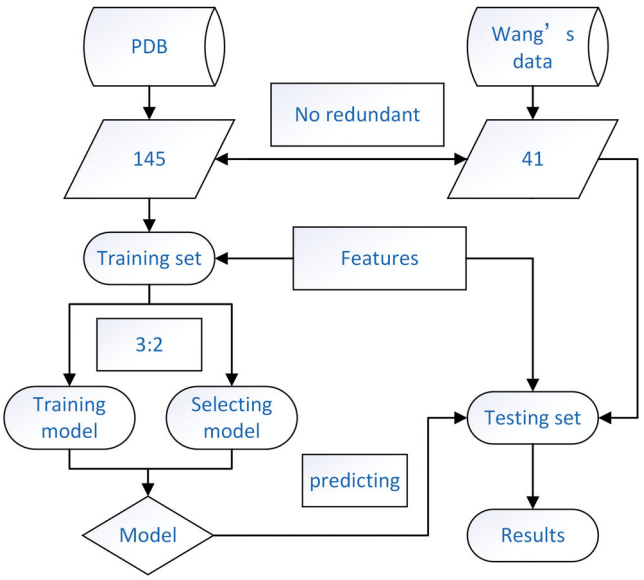

FIGURE 2 The pipeline of the protein–RNA binding affinity prediction model PRdeltaGPred. The multiple features are derived from the binding affinity prediction tools and scoring functions of protein–protein interaction. We calculated the features of protein-RNA on protein-RNA binding affinity benchmark (PRBAB) v2.0 and used the clustering method to get the better feature subsets. Three-fifths of the data in the database is chosen randomly as the training set, and the remaining part is used as a testing set. When the $R \geq 0.55$ on both the training set and the testing set, the model is retained.

2.2.1 | Hydrogen bond energy

Hydrogen bonds play a key role in molecular interactions[38] because hydrogen bonds can increase the equilibrium dissociation constants of two molecules.[39] Here, HBPLUS[35] is used to calculate hydrogen

bonds in the protein–RNA complex structures. It is shown that a single hydrogen bond can contribute to binding energy from 0.5 to 1.8 kcal/mol.[39] We simply count the number of hydrogen bonds and calculate the contribution of the entire hydrogen bond to free energy according to the energy with 0.5 kcal/mol per hydrogen bond.

2.2.2 | Solvation energy

The solvation energy play a key role in the prediction of protein–protein binding affinity in the previous study since the exposed areas of side chains of polar residues can increase the free energy of the system because these areas can reduce the entropy of water binding, indicating that water molecules change the energy of the system during the interaction.[34] In 1992, Horton et al. first pioneered the use of solvent-accessible surface areas and atomic solvent access parameters to calculate the binding free energy of protein–protein complexes.[24] In 1997, Zhang et al. improved the desolvation energy model by grouping the atoms in 20 residues into 18 atomic types according to property of the atom. And then they combined desolvation energy with electrostatic interaction to predict the binding affinity, the results showed that it has a high correlation between the experiment value and the predicted value in nine protease-inhibitors.[40] In 2007, Audie et al. predicted the changes of the binding free energy between proteins and ligands through the changes of desolvation free energy, the changes of contact free energy and then combined these with the energy contributed by hydrogen bonds and salt bridges.[22] In 2014, Janin et al. predicted the binding affinity by taking the solvent accessible area changes and conformational changes into account.[41]

To consider the effect of solvation energy on protein–RNA binding affinity, similar to protein–protein binding affinity prediction, we classify the atoms in residues and nucleotides both into five types, wherein the atomic types of residues are C, S, N/O, O^-, N^+, and the atomic solvent energy parameters (ASP) are derived from Zhou et al.'s.[42] We classify the atomic types of nucleotides as C, P, O^-, N, O and determine the contribution of different atomic types to binding free energy by calculating the size of the changes of the solvent accessible area upon binding. The surface accessible areas of bound protein–RNA complexes, unbound protein, and unbound RNA are calculated by the NACCESS.[43] As the same as,[34] the solvation energy is calculated by formula (2).

$$\Delta E_{\text{delsolvation}} = \sum_i A_i \cdot \Delta S_i \tag{2}$$

where A_i represents the atomic solvent energy parameters and ΔS_i represents the change of the solvent accessible area with corresponding atom.

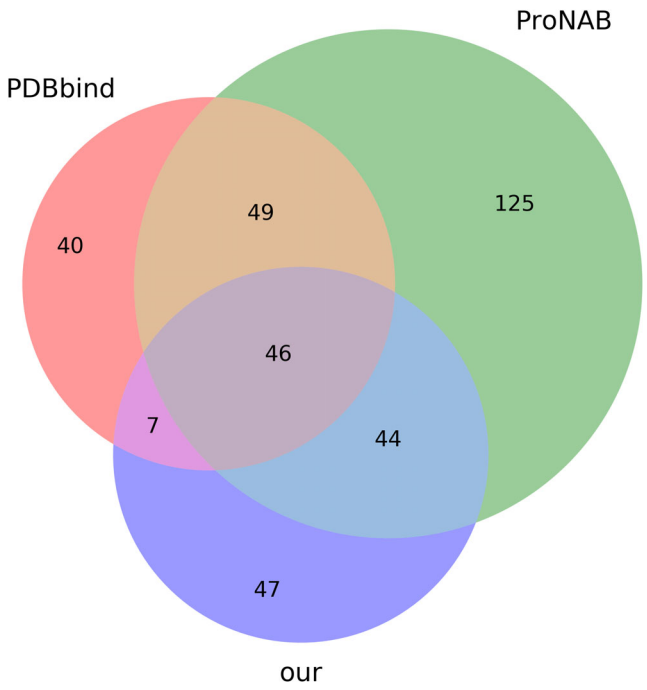

FIGURE 3 The comparison among our updated dataset, PDBbind and ProNAB. Our dataset includes 145 nonredundant protein–RNA binding affinity data. The PDBbind dataset includes 142 protein–RNA binding affinity data, and ProNAB includes 342 protein–RNA binding affinity data. Our dataset has 47 data that are not included in other two datasets.

2.2.3 | Salt bridge

As the same as the calculation of hydrogen bond energy, the number of salt bridges formed between NZ atoms or NE, CZ, NH1, NH2 atoms of lysine (LYS), and P/OP1/OP2 atoms of arginine (LYS) are used as the feature of salt bridges.[44]

2.2.4 | Noninteracting interface

Kastritis et al. found that the ICs and the noninteracting surfaces were highly correlated with protein–protein binding affinity, and then they considered these characteristics as related features in the prediction of protein–protein binding affinity.[26,36,45] In PRdeltaGPred, we also used these noninteracting interface (NIS) features.

2.2.5 | Interface contact

The key residues or bases on the protein-RNA interface play a key role in protein–RNA interactions. There are a series of scoring functions developed based on the protein–RNA interface.[14,16,46,47] Therefore, in order to characterize the role of these key residues, we encode the features of the interface by counting the interactions between different residue-nucleotide pairs. The amino acids in proteins are divided into three categories according to polarity, nonpolarity, and chargeability.

2.2.6 | Rpvescore

Due to the contribution of electrostatic and van der Waals interaction to the binding affinity. Here, we use RpveScore to calculate the van der Waals interaction energy, electrostatic potential energy, and the paired residue-base statistical potential. Rpvescore includes electrostatic and van der Waals interactions as well as residual-base pair interaction potentials, wherein electrostatic and van der Waals calculation formulas are from Zhang et al.[48]

2.2.7 | Noncovalent interaction

NCIPLOT can be used to calculate noncovalent interactions (NCI) between molecules based on electron density.[49] In LISA, NCIPLOT is used to calculate NCI terms including van der Waals interactions, hydrogen bonds, and stacking interactions to characterize key protein–protein characteristics. In the LISA model, the authors compute the characteristics of complexes by means of different residues in different solvent surface areas to count the number of promotions and exclusions of NCI to protein–protein interactions.[28] Wherein the residual base located in the solvent accessible surface is calculated according to the theory by Levy et al.[50] Therefore, similar to the way of LISA, we also take NCI into account in our prediction of protein-RNA binding affinity.

By defining statistically relevant features, wherein hydrogen bonds and salt bridges each contain a one-dimensional feature, the desolvation energy of complexes can be encoded to six-dimensional features, and the feature number of hydrogen bond and salt-bridge are one, respectively. According to the classified residues and four different nucleotides, we can obtain 7 noninteracting interface features, 12 interaction interface features, 7 Rpvescore features, and 28 noncovalent interaction features. Including a temperature, there are 63 features in total. These features are also listed in Table S5 accordingly.

2.3 | Feature selection

There are 63 features that can be used to predict the protein–RNA binding affinity. To avoid the redundancy between features which may bias the model trained by these features, we select features based on the correlation between features. We randomly selected 80% of these data from the training set to calculate the correlation between the features and repeat 10 times. Finally, we select the representative features by clustering with the threshold 0.8, which is higher than the criteria of similarity of 0.6 in the sequence-based prediction protein–protein binding affinity method.[51] The clustering algorithm groups similar features as far as possible through the threshold and obtains representative features according to the average distance among features. Feature clustering is done through the hierarchical clustering function hierarchy in the scipy library in Python.

2.4 | Model evaluation

In this article, our evaluation criteria is the same as other methods.[20] The model is evaluated by Pearson correlation coefficient R and root mean square error (RMSE). The calculations of Pearson correlation coefficient R and root mean variance RMSE are defined as following.

Given a pair of random variables (X, Y), R is calculated as

$$R = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} \tag{3}$$

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}\left(y_{\text{pred}}(t) - y_{\text{exp}}(t)\right)^2}{n}} \tag{4}$$

where E represents expectations, and μ_x and μ_y represents the averages of X and Y respectively, σ_x and σ_y are standard deviations. The Pearson correlation coefficient measures the correlation between two sets of variables. RMSE reflects the error between the predicted value and the experimental value.

3 | RESULTS

3.1 | The binding affinity dataset

For further analysis of the collected binding affinity data set, we plot a distribution based on the binding affinity data in PRBAB v2.0, which is shown in Figure S1. As can be seen from the statistical distribution, most of the complex with the binding affinity from −12 to −8 kcal/mol. There are only two complexes with binding affinity above −5 kcal/mol. The lower the value of binding affinity, the stronger the binding strength between protein and RNA.

3.2 | Compared with PDBbind and ProNAB

In PDBbind,[29] the author collected 23 496 biomolecular complexes binding affinity data, 142 of them are protein-RNA complexes. In ProNAB, the authors collected more than 20 000 experimental binding affinity data for protein–RNA/DNA, which also includes the binding free energy upon mutation and the dissociation constant. Here, we compared our data with these two datasets with protein–RNA complexes. The results shown in Figure 3. In our updated dataset, 47 out of 145 protein–RNA binding affinity data are not included in other two datasets. There are 46 complexes that exist in all three datasets. The ProNAB, in which 125 complexes do not appear in other two datasets, have the most complexes.

3.3 | The results of feature selection

Similar to the sequence-based method to predict protein–protein binding affinity, in order to avoid the redundancy among features in

the model, we calculated the Pearson correlation coefficient among different features by randomly selecting 60% of these data in PRBAB v2.0, and then we averaged the results calculated by repeating 10 times and clustered them by Pearson correlation coefficient 0.6. As shown in Figure 4, there are many features with a high correlation with each other ($R > 0.6$), which indicates that the structural features are highly relevant to each other quantified in different ways. Through hierarchical clustering to obtain representative features, only 39-dimensional features are kept from 63-dimensional features. Next, we use the remaining 39-dimensional features to train our binding affinity model.

3.4 | Regression model

Similar to protein–protein binding affinity prediction,[51] since there are difference when using different subsets to calculate the Pearson correlation coefficient between features, the fixed dataset used as a training set may bias the regression model, resulting in the model overfitting. In addition, because the low resolution of x-ray crystal structure also affects the prediction of protein–RNA binding affinity, and the experimental data obtained by experimental methods may have experimental errors. Therefore, taking these into account, we trained our model on the entire PRBAB v2.0 dataset and tested our model in PDBbind. In training, we randomly select a subset of PRBAB v2.0 to train the model, and another subset of PRBAB v2.0 is used to select the model. We randomly divided PRBAB v2.0 into two parts by a scale 3:2. Other words, 87 of 145 are randomly selected to train the model, and the left is used to select the model. And then, we calculated the Pearson correlation coefficient R between the predicted values and the experimental value, the model was kept if the Pearson correlation coefficient R large than 0.55 on both the training set and the test set. The process of model training, testing, and validation is shown in Figure 2.

We used the ordinary least square (OLS) for fitting the protein–RNA binding affinity. OLS is linear squares method for estimating the unknown parameters in a linear regression model. The training process is implemented by python.

3.5 | Model evaluation

During the training, we retained a model with a Pearson correlation coefficient greater than 0.55. According to the above conditions, we retained more than 200 models. As shown in Figure S2, the best RMSE on the training set is less than 1 kcal/mol, but the fluctuation of the RMSE is 2–18 kcal/mol. The Pearson correlation coefficient can be as high as 0.8, and by removing the model with a root mean

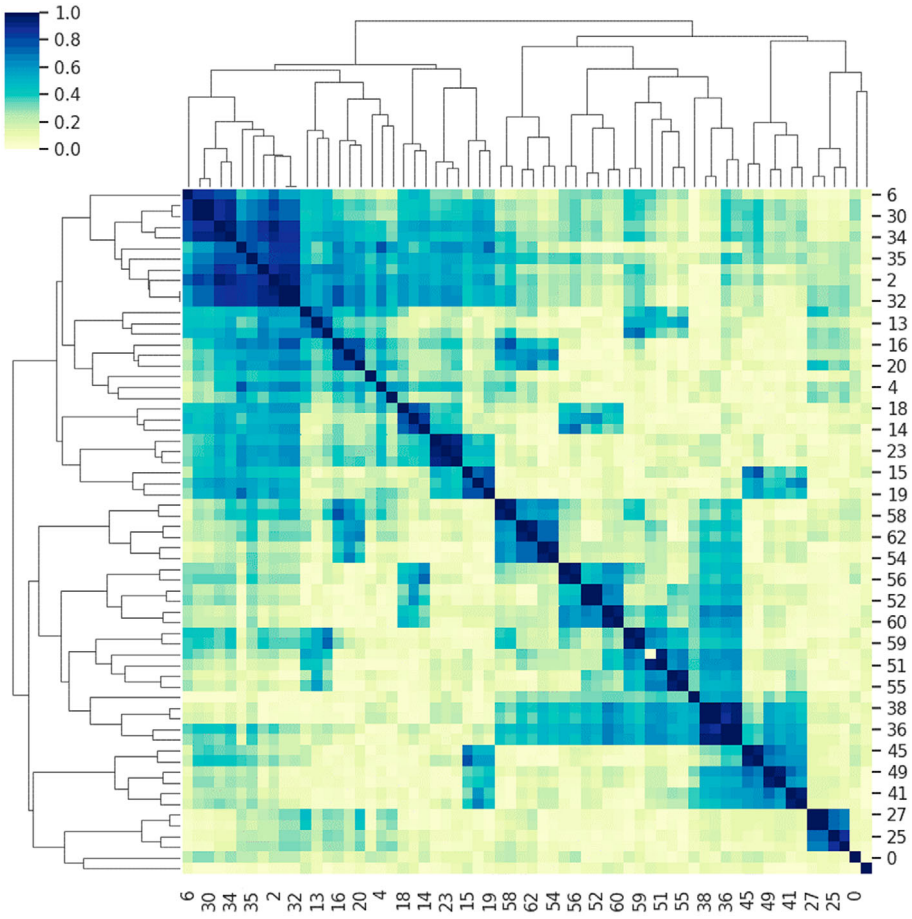

FIGURE 4 Clustering of features. We randomly selected a subset of data from the protein–RNA binding affinity benchmark (PRBAB) v2.0 10 times to calculate the average of the correlation between the two features. The bluer the color, the higher the correlation.

variance greater than 4 kcal/mol. The Pearson correlation coefficient and the root mean variance of the model are shown in Figure S3. By retaining a model with correlation greater than 0.55 and RMSE less than 4 kcal/mol on both the training set and the test set, one of the models is used to predict the training set and the test set. The results are shown in Figure 5.

Based on the Pearson correlation coefficients *R* and RMSE, six models are finally selected to predict protein–RNA binding affinity. And the six models are tested on PDBbind. As shown in Table S3, the best model is used to predict the protein–RNA binding affinity on the PDBbind, which achieves a Pearson correlation coefficient of 0.57 and a RMSE of 2.51 kcal/mol compared to experiment values. By comparing the experimental values one by one, we found that several complexes have a larger binding affinity deviation between the prediction value and the experiment value. In order to find out the reason why the model prediction error is too large, we analyzed the experimental value and the prediction value of more than 4 kcal/mol complexes and found that the IDs of these five complexes are 2L41, 5ID6, 4TUW, 4RCJ, and 4CSF. By manually examining the experimental data of the original article and analyzing the structure of the complexes, we found that the binding affinity of 2L41 in PDBbind was 763 μM, however, the experimental data was 48 μM (−5.88 kcal/mol) by examining the original article, the prediction value is −9.18 kcal/mol. In addition, complex 5ID6 contains metal ions, the amino acid in complex 4TUW and 4RCJ are phosphorylated and methylated, respectively. The RNA in 4CSF is highly flexible. Because there is currently too little binding affinity data known to contain metal ions and modified residues/base complexes, it is not possible to train and predict these complexes in our models. Taking these five complexes not into account, the Pearson correlation coefficient of the model can achieve 0.7 and the RMSE is 1.42 kcal/mol. The results predicted by our model on PDBbind's data are shown in Table S2.

We use the data on ProNAB to evaluate our best model. The comparison between the experiment and the prediction is shown in Figure 6. As shown in Figure 6, the prediction of 46 out of 81 (51.85%) complexes have high Pearson correlation ($r = .85$) with experiment values. The prediction of 61 out of 81 complexes have Pearson correlation ($r = .61$) with experiment values (The comparison between the experiment and prediction are listed in Table S4).

4 | CONCLUSION AND DISCUSSION

In this manuscript, we first update the protein–RNA binding affinity data, which adds 73 binding affinity data compared to the first protein–RNA binding affinity dataset. It has twice as much as the number of previous dataset PRBAB v1.0. In addition, by characterizing the structure of protein–RNA complexes, we developed a model to predict the protein-RNA binding affinity by using the least-squares method after removing redundancy of the structural features. By testing on 41 examples of protein-RNA complexes in the PDBbind dataset, the Pearson correlation coefficient R of the model can achieve 0.57 and RMSE is 2.51 kcal/mol. In the process of selecting features, we calculated the structural features of protein-RNA complexes in different ways. The results showed that there was a high correlation between structural features, indicating double counting in the characteristics of protein–RNA interactions derived from different patterns.

Though our methods achieved the Pearson correlation coefficient .57 with the experiment, the prediction of complex structures that contain ion or modified residues is not accurate at present for lack of

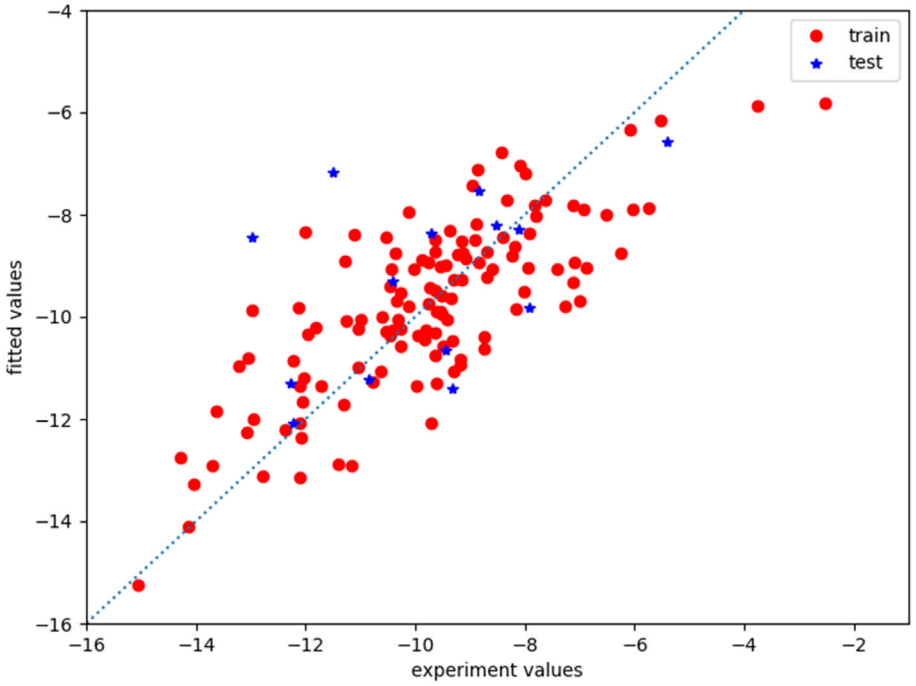

FIGURE 5 The comparison of the predicted and experimental values on the protein–RNA binding affinity benchmark (PRBAB) v2.0. Three fifths of the data are fitted on the protein–RNA binding affinity dataset and the other data are used as testing dataset. The Pearson correlation coefficient on the training set is .6 and the Pearson correlation coefficient on the test set is .55.

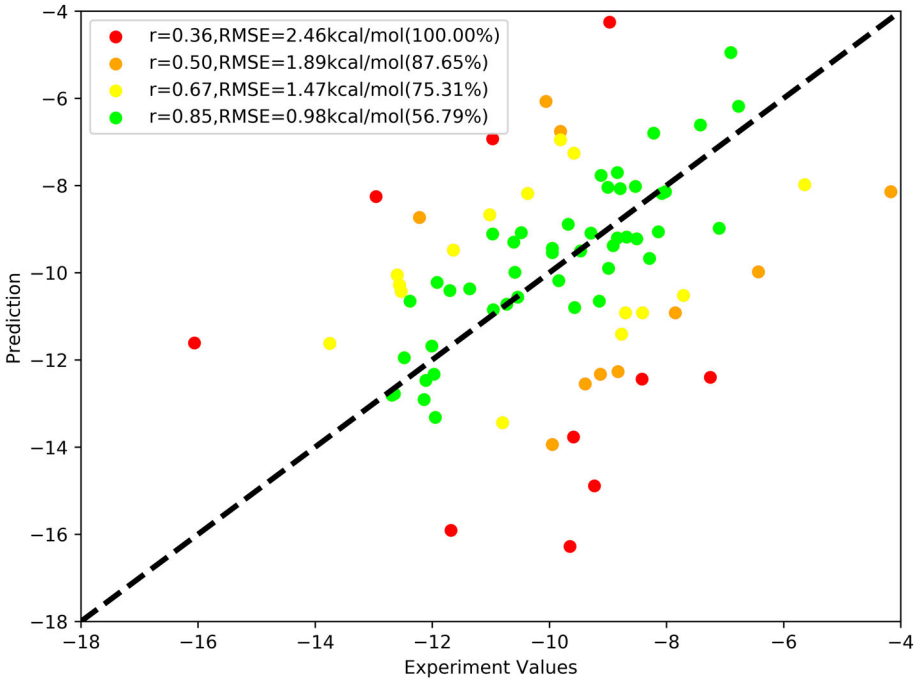

FIGURE 6 The comparison between the experiment values and the prediction on protein–RNA binding data of ProNAB. The different color represents the different error between the experiment and the prediction. The points colored green/yellow/orange/red represents that the error between the experiment value and the prediction are lower than 2/2–3/3–4/>4 kcal/mol. The prediction of 56.79% complexes has Pearson correlation .85 with experiment values.

amount of training data. Since the artificially defined features are coarse-grained, determining which structural features is important for protein–RNA interaction is undesirable. Although structure-based binding affinity prediction method limits its application in many cases in current stage. Because of AlphaFold, almost all protein structures in all kinds of proteome can be available. We believe that with more and more RNA 3D structure available, our structure-based binding affinity prediction method will be useful in future.

AUTHOR CONTRIBUTIONS

Conceptualization, Xu Hong, Xiaoxue Tong, and Shiyong Liu; Investigation Xu Hong, Xiaoxue Tong, and Shiyong Liu; PRdeltaGPred software development, Xu Hong and Shiyong Liu; Database and web server development, Xu Hong; Data Curation, Xiaoxue Tong, Pinyu Liu, Juan Xie, and Xu Hong; Writing–Original Draft, Xu Hong, Xiaoxue Tong, and Shiyong Liu; Writing–Review & Editing, Juan Xie, Xu Hong, Xiaoxue Tong, Qi Song, Sen Liu, Xudong Liu, and Shiyong Liu; Funding Acquisition, Shiyong Liu; Supervision, Shiyong Liu.

ACKNOWLEDGMENTS

This work has been supported by the National Natural Science Foundation of China [31100522] and [32271267]; National High Technology Research and Development Program of China [2012AA020402]; the Special Program for Applied Research on Super Computation of the NSFC-Guangdong Joint Fund (the second phase) [U1501501] and the Fundamental Research Funds for the Central Universities [2016YXMS017].

CONFLICT OF INTEREST STATEMENT

The authors declare no competing interests.

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1002/prot.26503.

DATA AVAILABILITY STATEMENT

Protein-RNA Binding Affinity Benchmark 2.0 (PRBAB2.0) and source code of protein-RNA binding affinity prediction model PRdeltaGPred was implemented in Python and is freely available at http://www.rnabinding.com/PRdeltaGPred.

ORCID

Shiyong Liu https://orcid.org/0000-0001-7986-5178

REFERENCES

1. Anantharaman V, Koonin EV, Aravind L. Comparative genomics and evolution of proteins involved in RNA metabolism. *Nucleic Acids Res*. 2002;30:1427-1464.
2. Glisovic T, Bachorik JL, Yong J, Dreyfuss G. RNA-binding proteins and post-transcriptional gene regulation. *FEBS Lett*. 2008;582:1977-1986.
3. Jankowsky E, Harris ME. Specificity and nonspecificity in RNA-protein interactions. *Nat Rev Mol Cell Biol*. 2015;16:533-544.
4. Muller-McNicoll M, Neugebauer KM. How cells get the message: dynamic assembly and function of mRNA-protein complexes. *Nat Rev Genet*. 2013;14:275-287.
5. Rissland OS. Dynamics of RNA-protein interactions studied in living cells. *Nature*. 2021;591:39-40.
6. Wong I, Lohman TM. A double-filter method for nitrocellulose-filter binding: application to protein-nucleic acid interactions. *Proc Natl Acad Sci USA*. 1993;90:5428-5432.
7. Myszka DG. Kinetic analysis of macromolecular interactions using surface plasmon resonance biosensors. *Curr Opin Biotechnol*. 1997;8:50-57.

8. Pierce MM, Raman CS, Nall BT. Isothermal titration calorimetry of protein-protein interactions. *Methods*. 1999;19:213-221.
9. Hellman LM, Fried MG. Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions. *Nat Protoc*. 2007;2:1849-1861.
10. Patel TR, Chojnowski G, Astha, Koul A, McKenna SA, Bujnicki JM. Structural studies of RNA-protein complexes: a hybrid approach involving hydrodynamics, scattering, and computational methods. *Methods*. 2017;118–119:146-162.
11. Jarmoskaite I, AlSadhan I, Vaidyanathan PP, Herschlag D. How to measure and evaluate binding affinities. *Elife*. 2020;9:1-34.
12. Yang Y, Zhao H, Wang J, Zhou Y. SPOT-Seq-RNA: predicting protein-RNA complex structure and RNA-binding function by fold recognition and binding affinity prediction. *Methods Mol Biol*. 2014;1137:119-130.
13. Huang SY, Zou X. A knowledge-based scoring function for protein-RNA interactions derived from a statistical mechanics-based iterative method. *Nucleic Acids Res*. 2014;42:e55.
14. Huang Y, Liu S, Guo D, Li L, Xiao Y. A novel protocol for three-dimensional structure prediction of RNA-protein complexes. *Sci Rep*. 2013;3:1887.
15. Setny P, Zacharias M. A coarse-grained force field for protein-RNA docking. *Nucleic Acids Res*. 2011;39:9118-9129.
16. Tuszynska I, Bujnicki JM. DARS-RNP and QUASI-RNP: new statistical potentials for protein-RNA docking. *BMC Bioinformatics*. 2011;12:348.
17. Yang X, Li H, Huang Y, Liu S. The dataset for protein-RNA binding affinity. *Protein Sci*. 2013;22:1808-1811.
18. Dias R, Kolazckowski B. Different combinations of atomic interactions predict protein-small molecule and protein-DNA/RNA affinities with similar accuracy. *Proteins*. 2015;83:2100-2114.
19. Nithin C, Mukherjee S, Bahadur RP. A structure-based model for the prediction of protein-RNA binding affinity. *RNA*. 2019;25:1628-1645.
20. Deng L, Yang W, Liu H. PredPRBA: prediction of protein-RNA binding affinity using gradient boosted regression trees. *Front Genet*. 2019;10:637.
21. Chen F, Sun H, Wang J, et al. Assessing the performance of MM/-PBSA and MM/GBSA methods. 8. Predicting binding free energies and poses of protein-RNA complexes. *RNA*. 2018;24:1183-1194.
22. Audie J, Scarlata S. A novel empirical free energy function that explains and predicts protein-protein binding affinities. *Biophys Chem*. 2007;129:198-211.
23. Bai H, Yang K, Yu D, Zhang C, Chen F, Lai L. Predicting kinetic constants of protein-protein interactions based on structural properties. *Proteins*. 2011;79:720-734.
24. Horton N, Lewis M. Calculation of the free energy of association for protein complexes. *Protein Sci*. 1992;1:169-181.
25. Moal IH, Agius R, Bates PA. Protein-protein binding affinity prediction on a diverse set of structures. *Bioinformatics*. 2011;27:3002-3009.
26. Vangone A, Bonvin AM. Contacts-based prediction of binding affinity in protein-protein complexes. *Elife*. 2015;4:e07454.
27. Vreven T, Hwang H, Pierce BG, Weng Z. Prediction of protein-protein binding free energies. *Protein Sci*. 2012;21:396-404.
28. Raucci R, Laine E, Carbone A. Local interaction signal analysis predicts protein-protein binding affinity. *Structure*. 2018;26:905-915.e4.
29. 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;47:2977-2980.
30. Harini K, Srivastava A, Kulandaisamy A, Gromiha MM. ProNAB: database for binding affinities of protein-nucleic acid complexes and their mutants. *Nucleic Acids Res*. 2022;50:D1528-D1534.
31. Liu Z, Li Y, Han L, et al. PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics*. 2015;31:405-412.
32. Wang Y, Xue ZD, Shi XH, Xu J. Prediction of pi-turns in proteins using PSI-BLAST profiles and secondary structure information. *Biochem Biophys Res Commun*. 2006;347:574-580.
33. Kastritis PL, Moal IH, Hwang H, et al. A structure-based benchmark for protein-protein binding affinity. *Protein Sci*. 2011;20:482-491.
34. Eisenberg D, McLachlan AD. Solvation energy in protein folding and binding. *Nature*. 1986;319:199-203.
35. McDonald IK, Thornton JM. Satisfying hydrogen bonding potential in proteins. *J Mol Biol*. 1994;238:777-793.
36. Kastritis PL, Rodrigues JP, Folkers GE, Boelens R, Bonvin AM. Proteins feel more than they see: fine-tuning of binding affinity by properties of the non-interacting surface. *J Mol Biol*. 2014;426:2632-2652.
37. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. *J Mol Biol*. 1982;157:105-132.
38. Meyer M, Wilson P, Schomburg D. Hydrogen bonding and molecular surface shape complementarity as a basis for protein docking. *J Mol Biol*. 1996;264:199-210.
39. Fersht AR. The hydrogen bond in molecular recognition. *Trends Biochem Sci*. 1987;12:3214-3219.
40. Zhang C, Vasmatzis G, Cornette JL, DeLisi C. Determination of atomic desolvation energies from the structures of crystallized proteins. *J Mol Biol*. 1997;267:707-726.
41. Janin J. A minimal model of protein-protein binding affinities. *Protein Sci*. 2014;23:1813-1817.
42. Zhou H, Zhou Y. Stability scale and atomic solvation parameters extracted from 1023 mutation experiments. *Proteins*. 2002;49:483-492.
43. Lee B, Richards FM. The interpretation of protein structures: estimation of static accessibility. *J Mol Biol*. 1971;55:379-400.
44. Sutch BT, Chambers EJ, Bayramyan MZ, Gallaher TK, Haworth IS. Similarity of protein-RNA interfaces based on motif analysis. *J Chem Inf Model*. 2009;49:2139-2146.
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;32:3676-3678.
46. Li CH, Cao LB, Su JG, Yang YX, Wang CX. A new residue-nucleotide propensity potential with structural information considered for discriminating protein-RNA docking decoys. *Proteins*. 2012;80:14-24.
47. Li H, Huang Y, Xiao Y. A pair-conformation-dependent scoring function for evaluating 3D RNA-protein complex structures. *PLoS One*. 2017;12:e0174662.
48. Zhang Z, Lu L, Zhang Y, et al. A combinatorial scoring function for protein-RNA docking. *Proteins*. 2017;85:741-752.
49. Contreras-Garcia J, Johnson ER, Keinan S, et al. NCIPLOT: a program for plotting non-covalent interaction regions. *J Chem Theory Comput*. 2011;7:625-632.
50. Levy ED. A simple definition of structural regions in proteins and its use in analyzing interface evolution. *J Mol Biol*. 2010;403:660-670.
51. Yugandhar K, Gromiha MM. Protein-protein binding affinity prediction from amino acid sequence. *Bioinformatics*. 2014;30:3583-3589.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Hong X, Tong X, Xie J, et al. An updated dataset and a structure-based prediction model for protein–RNA binding affinity. *Proteins*. 2023;1-9. doi:10.1002/prot.26503