BMC Biology

METHODOLOGY ARTICLE

Open Access

PNBACE: an ensemble algorithm to predict the effects of mutations on protein-nucleic acid binding affinity

Si-Rui Xiao[1], Yao-Kun Zhang[1], Kai-Yu Liu[1], Yu-Xiang Huang[1] and Rong Liu[1*]

Abstract

Background Mutations occurring in nucleic acids or proteins may affect the binding affinities of protein-nucleic acid interactions. Although many efforts have been devoted to the impact of protein mutations, few computational studies have addressed the effect of nucleic acid mutations and explored whether the identical methodology could be applied to the prediction of binding affinity changes caused by these two mutation types.

Results Here, we developed a generalized algorithm named PNBACE for both DNA and protein mutations. We first demonstrated that DNA mutations could induce varying degrees of changes in binding affinity from multiple perspectives. We then designed a group of energy-based topological features based on different energy networks, which were combined with our previous partition-based energy features to construct individual prediction models through feature selections. Furthermore, we created an ensemble model by integrating the outputs of individual models using a differential evolution algorithm. In addition to predicting the impact of single-point mutations, PNBACE could predict the influence of multiple-point mutations and identify mutations significantly reducing binding affinities. Extensive comparisons indicated that PNBACE largely performed better than existing methods on both regression and classification tasks.

Conclusions PNBACE is an effective method for estimating the binding affinity changes of protein-nucleic acid complexes induced by DNA or protein mutations, therefore improving our understanding of the interactions between proteins and DNA/RNA.

Keywords DNA mutation, Protein mutation, Binding affinity, Energy network, Differential evolution

Background

Protein-nucleic acid interactions (PNIs) play fundamental roles in transcription and translation processes [1]. Interactions between proteins and DNA/RNA molecules (PDIs and PRIs, respectively) are mediated and affected by various intermolecular forces, including hydrogen bonding, van der Waals attractions, and electrostatic interactions. Mutations appearing in nucleic acids or proteins could alter these factors, therefore leading to changes in their binding affinities. Investigating and quantifying the effects of residue and base mutations is beneficial to our understanding of the underlying mechanisms of PNIs. Although experimental techniques, such as surface plasmon resonance [2], isothermal titration calorimetry [3], and fluorescence resonance energy transfer [4], have been used to study the effects of mutations on PNIs, the process is laborious and time-consuming. With the exponential increase in genomic data, these traditional experimental methods may be unsuitable for

*Correspondence:
Rong Liu
liurong116@mail.hzau.edu.cn
[1] Hubei Key Laboratory of Agricultural Bioinformatics, College of Informatics, Huazhong Agricultural University, Wuhan 430070, People's Republic of China

high-throughput studies. Thus, there is a pressing need to develop computational tools for predicting protein-nucleic acid binding affinity changes.

Over the past decade, machine learning techniques and free energy calculations have been jointly or individually adopted to study the effects induced by different types of mutations in biomacromolecules [5, 6]. Especially, a series of algorithms have been developed to predict the impact of mutations on PNIs using energy- and/or knowledge-based features. Peng et al. established SAMPDI, which integrated an enhanced MM/PBSA approach with knowledge-based descriptors to estimate the binding affinity change of PDIs in response to protein mutations [7]. Developers of PremPDI utilized molecular force fields and fast side-chain optimization algorithms to assess the effect of protein mutations on PDIs and then created a tool called PremPRI that was an extension of PremPDI for PRIs [8, 9]. Our group invented PEMPNI, which employed an ensemble strategy incorporating novel energy- and nonenergy-based characteristics to estimate the binding affinity changes of PDIs and PRIs [10]. Generally, algorithms involving energy calculations possess higher computational costs than purely knowledge-based methods. Regarding the latter, Pires et al. developed mCSM-NA, which predicted the effect of protein mutations on PNIs using graph-based signatures in conjunction with pharmacophore modeling [11]. Moreover, they proposed an updated method (mmCSM-NA) capable of predicting the impact of multiple-point protein mutations [12]. PrabHot and PrPDH combined machine learning methods with structural and sequence properties to identify hotspot residues in PNIs [13, 14]. Recently, Li et al. proposed the SAMPDI-3D method, an improved version of SAMPDI, which uses gradient-boosted decision trees and a series of knowledge-based terms to assess the energy changes of PDIs caused by either DNA or protein mutations [15]. Altogether, the aforementioned works significantly advanced the development of protein-nucleic acid binding affinity change prediction.

Despite remarkable progress gained by the existing studies, several problems could be worthy of further exploration. First, previous computational studies mainly focused on the impact of residue mutations, but less attention was given to the effect of base mutations, probably due to the scarcity of relevant experimental data. SAMPDI-3D is the only method that can predict the effect of single-point DNA mutations. Second, despite the pioneering contribution of SAMPDI-3D, this method used physiochemical and structural descriptors to build models for DNA mutations and neglected the energy features that had been applied to protein mutations. In particular, the energy network of residues and bases for PNIs remains to be fully explored. Third, the developers of SAMPDI-3D started with different (specific) features for predicting the effects of DNA versus protein mutations. It would be interesting to investigate whether the same features and even the same computational framework could be used for these two types of mutations. Finally, our previous study's results implied that the prediction results could be enhanced by the ensemble strategy, but only two component models were used in PEMPNI through a weighted combination. If a greater number of multifaceted component models are integrated using advanced optimization techniques, their interplay would be beneficial for improving prediction accuracy.

Motivated by these problems, we first investigated the binding affinity changes caused by DNA mutations from different viewpoints. Then, we proposed a generalized algorithm called PNBACE (*p*rotein-*n*ucleic acid *b*inding *a*ffinity *c*hange *e*stimator) for predicting the energy influence on PNIs triggered by both DNA and protein mutations (Fig. 1). To this end, we designed energy-based topological features based on different energy networks and combined these novel terms with our previous partition-based energy features to build individual prediction models through feature selection. Furthermore, we created an ensemble model by integrating individual models using a differential evolution (DE) algorithm. In addition to predicting the impact of single-point mutations, PNBACE could predict the influence of multiple-point mutations and identify mutations significantly reducing binding affinities. Finally, we implemented our algorithm as a user-friendly webserver.

Results

Impacts of DNA mutations on PNIs from different aspects

Because the influence of residue mutations was investigated in our previous study, we mainly focused on the impact of base mutations on protein–DNA complexes in the current study. As shown in Additional file 1: Fig. S1, there was no significant difference between the binding affinity changes induced by DNA versus protein mutations. Among the 504 collected DNA mutations, 15 and 489 corresponded to single bases and base pairs, respectively. Figure 2A shows that substitutions between GC and CG (GC/CG-std) triggered greater changes than those between AT and TA (AT/TA-std), probably due to the higher number of proton donors in GC/CG base pairs, which may result in more stable binding with protein receptors [16]. Meanwhile, the smaller impact of GC/CG-ns (the original pair was GC or CG, and the mutant pair was any base pair other than GC and CG) compared to GC/CG-std may be attributed to the fact that substitutions between GC and CG involve the breakage and generation of three hydrogen bonds, leading to a greater

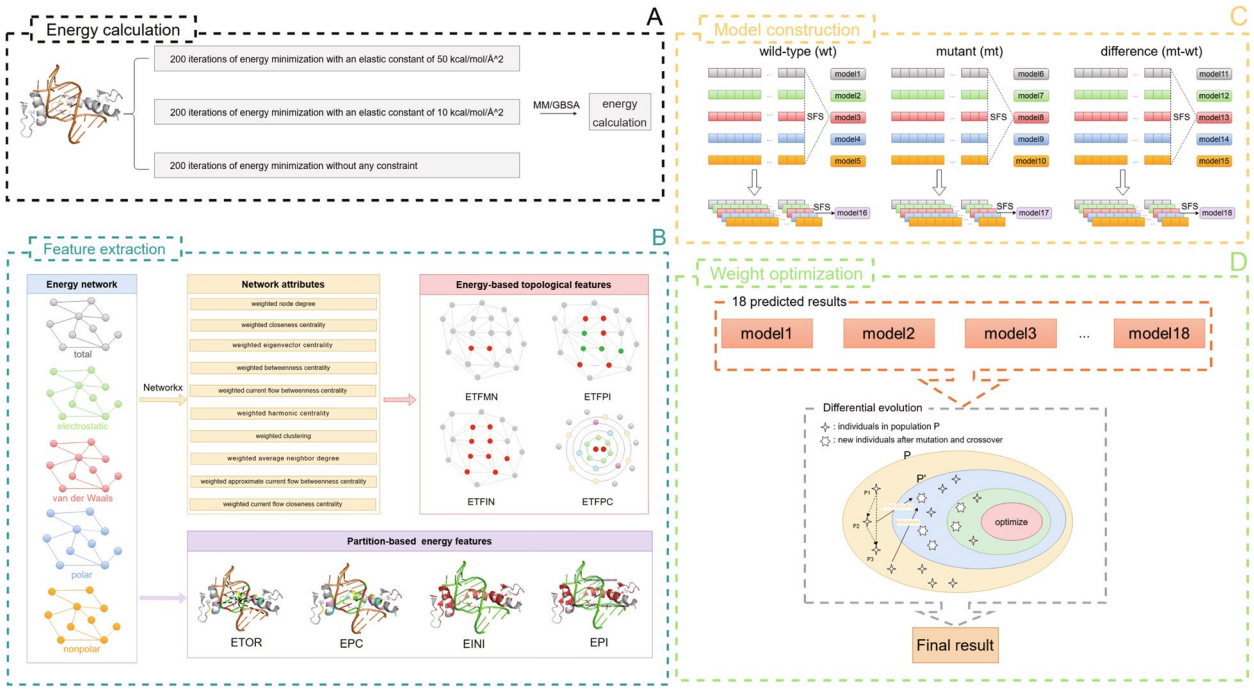

Fig. 1 PNBACE comprising four basic steps. **A** Energy calculation. The binding free energy and related energy terms are generated and decomposed in this step. **B** Feature extraction. The newly defined energy-based topological features combined with our previous partition-based energy features are extracted in this step. **C** Model construction. Eighteen individual models are constructed by using XGBoost methods and feature selections. **D** Weight optimization. An ensemble model is built by merging the results of individual models with the DE algorithm

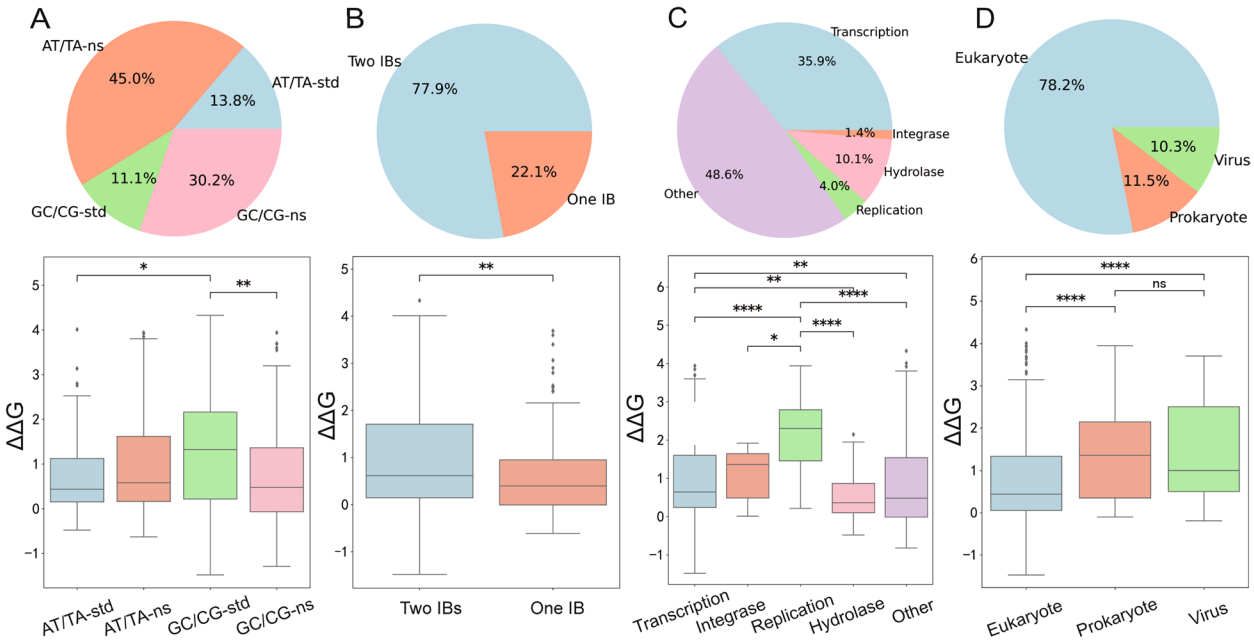

Fig. 2 DNA mutations leading to different degrees of protein-nucleic acid binding affinity changes. **A** Comparison of mutations according to base pair types. AT/TA-std: the substitution between AT and TA pairs, GC/CG-std: the substitution between GC and CG pairs, AT/TA-ns: the substitution between AT/TA (wide-type) and any pair other than AT/TA (mutant), and GC/CG-ns: the substitution between GC/CG (wide-type) and any pair other than GC/CG (mutant). **B** Comparison of mutations according to geometric locations. IB: interfacial base. **C** Comparison of mutations according to biological functions. **D** Comparison of mutations according to species. ****: $p < 0.0001$, ***: $0.0001 \le p < 0.001$, **: $0.001 \le p < 0.01$, *: $0.01 \le p < 0.05$, and ns: $p \ge 0.05$

impact on DNA structures and related interactions than other substitutions. In Fig. 2B, most base pairs appeared at the binding interface. Moreover, mutations having two interfacial bases had stronger effects than those having only one interfacial base. This result suggests that interfacial bases play a crucial role in maintaining PNIs.

According to the annotations provided by the PDB, we categorized the complexes into five functional groups: transcription, integrase, replication, hydrolase, and others. Figure 2C shows that replication-related complexes experienced the greatest changes in binding affinity in response to mutations, probably because the highly conserved structural basis of DNA binding for these complexes was changed [17]. In Fig. 2D, the mutation data were divided into three types based on species: eukaryotes, prokaryotes, and viruses. Base mutations in eukaryotes yielded smaller effects, possibly due to the sophisticated mechanisms used by eukaryotes for genetic information processing. These mechanisms could facilitate the timely repair of mutated regions, thereby reducing the impact on binding affinity [18]. Furthermore, we found that the differences in functional groups may be mainly determined by mutant pairs having two interfacial bases, while the differences in species may not be influenced by the positional effects of mutations (Additional file 1: Fig. S2).

Feature correlation and performance analysis

A total of 44 feature groups were generated for each combination of state and energy type. The correlation coefficients of these feature groups were calculated based on the training sets. As displayed in Additional file 1: Fig. S3, higher correlations were observed among the relevant features, such as the degree-related features (Groups 1, 8, 11, 18, 21, 28, 31, and 38 in Additional file 1: Table S1). However, lower correlations existed among the different categories of topological features, such as degree- and closeness-related groups. This suggests that the redundancy and complementarity of feature groups should be considered when we develop prediction models.

Accordingly, we implemented the SFS for base models of each mutation type. As shown in Fig. 3A, 70, 93 and 76 groups were reserved for the T298, MPD276, and MPR233 datasets, respectively. Moreover, the three datasets showed common and specific preferences for energy-based topological features. Regarding T298, wild-type state-based models favored degree-related features, whereas both the mutant and difference state-based models preferred betweenness-related features (Fig. 3B). For MPD276, the closeness-, betweenness-, and harmonic-based groups were most frequently used in the wild-type, mutant, and difference states, respectively (Fig. 3C), while betweenness-based features were most repeatedly selected across the three states of the MPR233 dataset (Fig. 3D). Generally, all three datasets commonly preferred betweenness-related attributes, suggesting the global importance of nodes in different energy networks for the binding affinity of PNIs. From the partition viewpoint, the ETFMN groups constituted the highest proportion among the reserved features for the three datasets, implying that the energy contributions of mutant residues or bases are good indicators of the changes in binding affinity. In terms of energy type, most features reserved for T298 and MPR233 were related to the electrostatic term, while those for MPD276 were associated with the nonpolar term. This may be because the electrostatic interactions between the positively charged residues of proteins and the negatively charged phosphates of nucleic acids play critical roles in determining the interaction strength between protein and DNA/RNA, and the nonpolar solvation energy could model the hydrophobic effect and is a key factor driving the binding between protein and DNA [19–21]. Additionally, the previous partition-based energy features accounted for a greater fraction for MPD276 than T298 and MPR233 (15%, 9%, and 6%, respectively). This was in line with our earlier work in which partition-based energy features were essential for MPD276.

As shown in Fig. 4 and Additional file 1: Fig. S4, the performances of all models on the three datasets were improved after performing the feature selection process. Particularly, the most significant improvements were observed for the polar term in the mutant state (polar_mt) of T298, the electrostatic term in the difference state (electro_mt-wt) of MPD276, and the polar term in the difference state (polar_mt-wt) of MPR233. The PCCs were increased by 0.438, 0.261, and 0.404, respectively. Meanwhile, the electrostatic term in the difference state

(See figure on next page.)

Fig. 3 Reserved feature groups of individual models after feature selections (each model corresponds to a combination of state and energy term). **A** Total number of selected feature groups of individual models under the same state based on training sets. **B** Selected feature groups of each model for DNA mutations. **C** Selected feature groups of each model for protein mutations involved in PDIs. **D** Selected feature groups of each model for protein mutations involved in PRIs. The circle denotes 44 feature groups (Additional file 1: Table S1), circle colors represent different categories of features, nodes represent selected features, and node colors represent different energy terms. In the circle, nodes with the same color are the selected features of an individual model

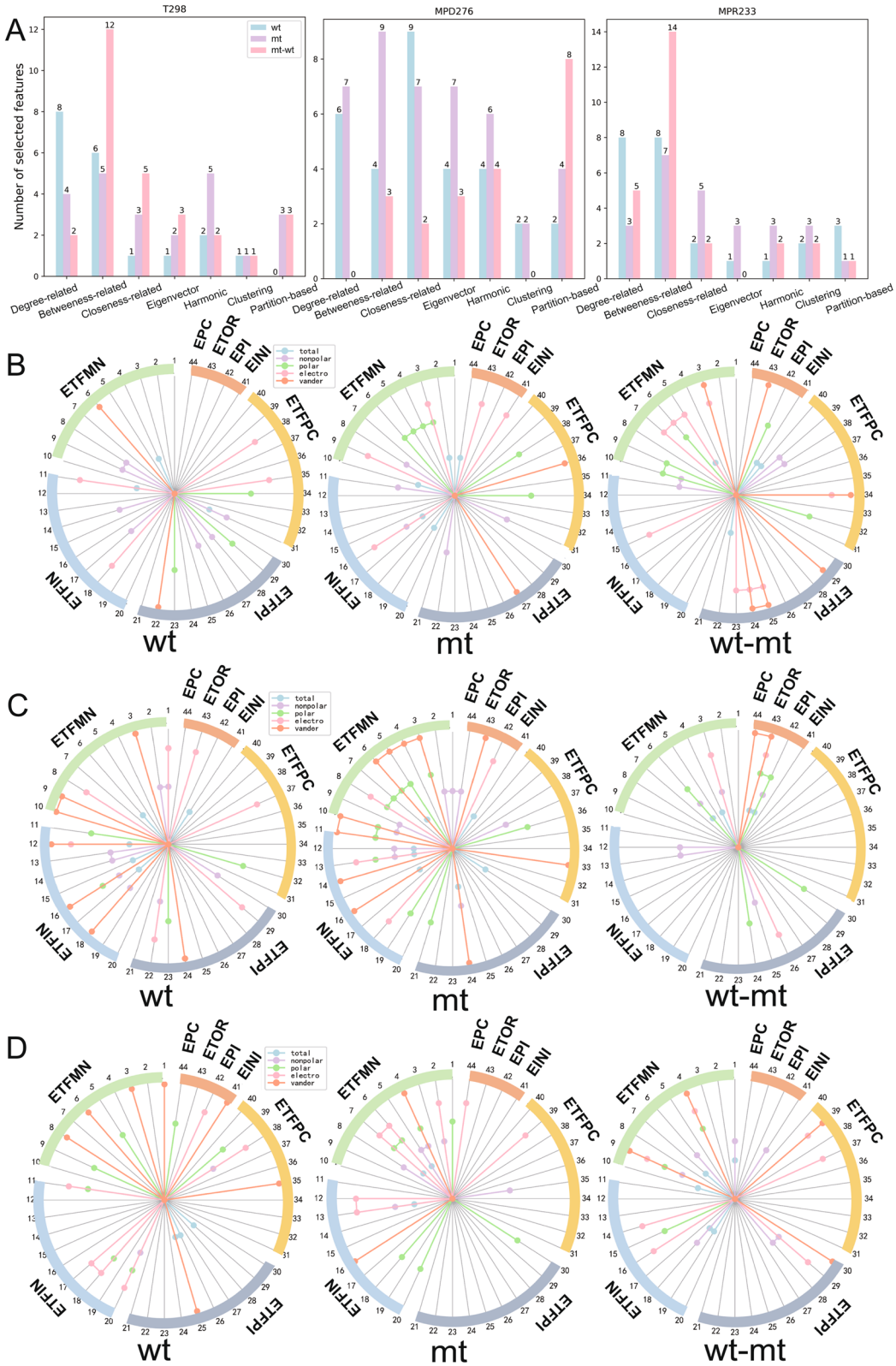

Fig. 3 (See legend on previous page.)

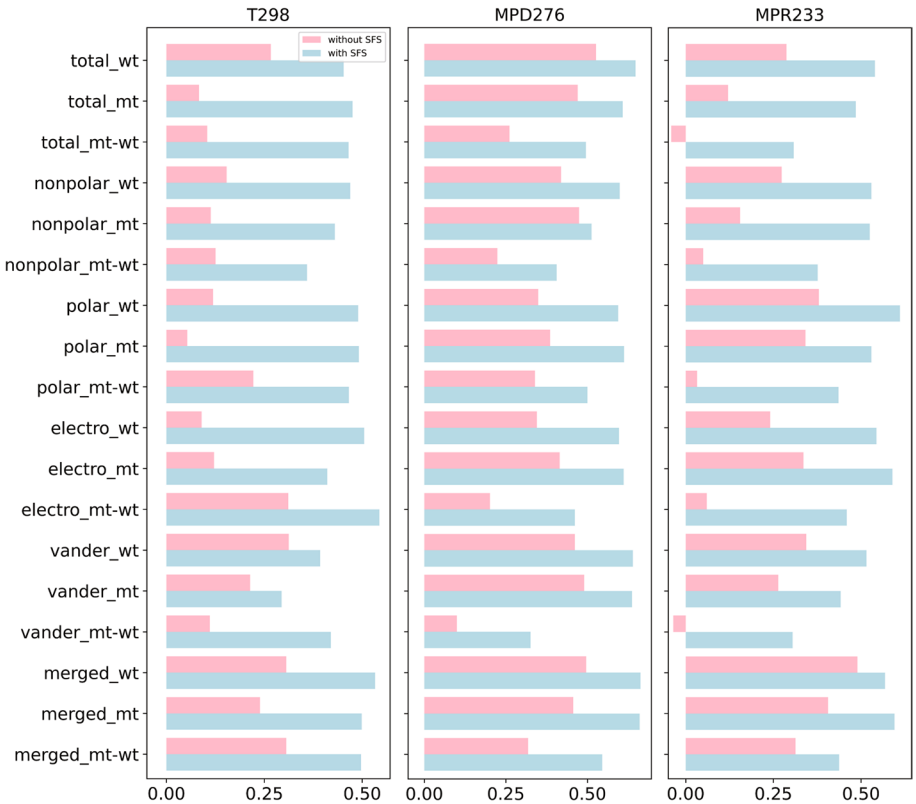

Fig. 4 PCC values of individual models without and with feature selections on training sets

(electro_mt-wt), the merged energy term in the wild-type state (merged_wt), and the polar term in the wild-type state (polar_wt) achieved optimal performance on T298, MPD276, and MPR233, respectively, with PCCs of 0.544, 0.663, and 0.611.

Furthermore, we adopted an integrative strategy to improve the performance using the interplay among the 18 models. Initially, we assigned the same weight to each model, and the ensemble model yielded higher PCCs of 0.655, 0.751, and 0.673. However, it should be noted that this approach did not fully consider the distinct contributions of the energy terms. We thus employed the DE algorithm to identify different weights for individual models and achieved improved PCCs of 0.704, 0.763, and 0.713. In addition to LOCOV, we used 80% of all complexes (or mutations) as the training set and the remaining 20% as the validation set. This procedure was performed 100 times. These results suggest that the performance of our models is robust (Additional file 1: Table S2).

Performance of PNBACE on different subsets

After obtaining the best performance, we separated the T298 dataset into several subsets as described in the first section of the "Results". As shown in Fig. 5A, the ensemble algorithm yielded comparable results for different base pair types and geometric locations. In terms of biological functions, PNBACE obtained excellent measures for integrase-related complexes but poor performance for hydrolase-related complexes. Moreover, we observed better metrics on the eukaryotic subset than the prokaryotic subset. Based on our previous work, we divided MPD276 and MPR233 into multiple subgroups (Fig. 5B and C). Compared with PEMPNI, the current method showed superior performance on most subsets. Regarding wild-type residues, our method was most effective for N in MPD276 and for E in MPR223. Considering the geometric locations, we obtained comparable performance on interface cores, interface rims, and surfaces in MPD276 but a significant decrease in correlations for the interface cores in MPR233. Moreover, PNBACE yielded optimal PCCs for mutations involved in minor grooves for both datasets. Regarding the major binding modes (the right 11 units), PNBACE generally performed well on the subsets of MPD276 but obtained relatively worse performance on partial subsets of MPR233 (e.g., Contact-C). These analyses revealed the levels of prediction difficulty of different subsets from the same dataset as well as those of corresponding subsets from different datasets. Subsets with high difficulty should receive more attention in the future.

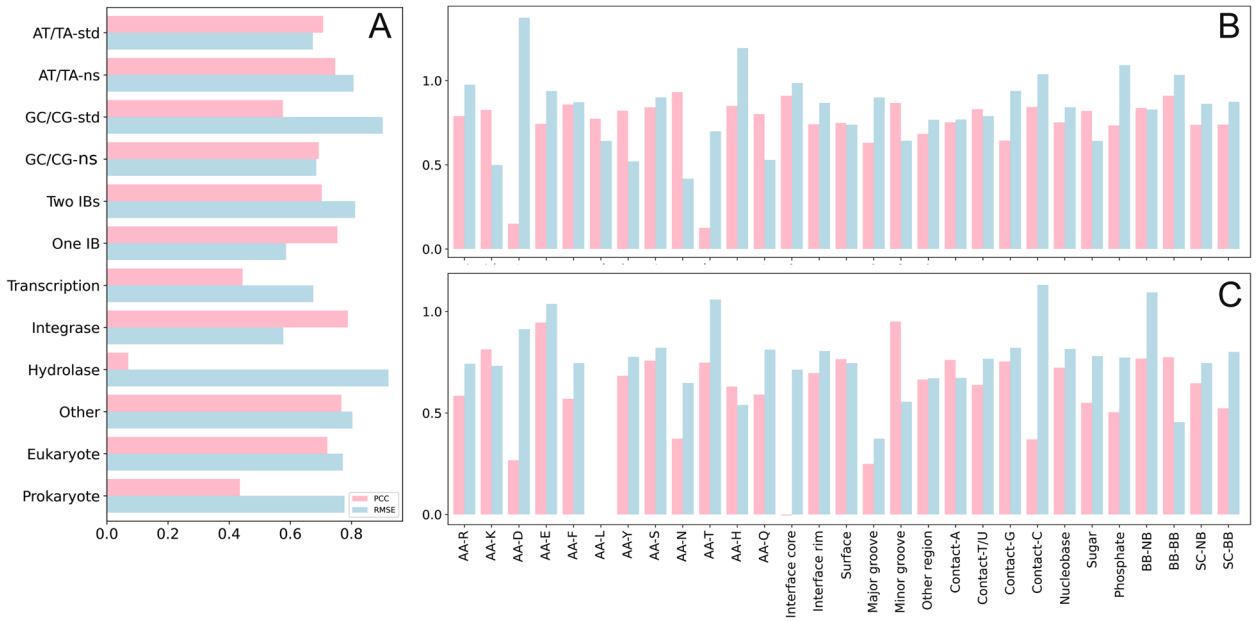

Fig. 5 PCC values of PNBACE on different mutation subsets. **A** Results on T298. **B** Results on MPD276. **C** Results on MPR233. IB: interfacial base, AA: amino acid, BB: backbone, NB: nucleobase, and SC: sidechain

Blind tests on nucleic acid and protein mutation data

Furthermore, we evaluated the performance of individual and ensemble models using test sets. Specifically, the DNA mutation datasets included T206 and T227, and the latter was constructed based on the binding score from the HT-SELEX experiments (Additional file 1: Table S3). Regarding T206, the PCC of the simple average method was slightly inferior to the result of the best individual model, namely, the van der Waals term in the wild-type state (0.487 and 0.494). When using the DE-based weights, however, the performance was obviously improved, with a PCC of 0.535. Regarding T227, the two integrative methods achieved the same PCC value of 0.473 and surpassed all component models. The protein mutation datasets comprised the four sets used by PEMPNI (MPD48, PDM, PDSI, and MPR79) and one set collected by SAMPDI-3D (S200). Note that if the training and test sets had overlapping complexes, the relevant samples were excluded from training. As presented in Additional file 1: Table S4, the best individual models achieved PCCs of 0.528, 0.643, 0.545, 0.322, and 0.386 on the five test sets. Furthermore, the results were improved by the DE-based ensemble model, with PCCs of 0.660, 0.711, 0.611, 0.395, and 0.417, respectively. Compared with the simple average method, the DE approach achieved a clearly better performance on mutations involved in PRIs and comparable results on mutations involved in PDIs. Based on the newly collected MPD248 and MPR134 datasets, our ensemble model yielded PCCs of 0.452 and 0.351, respectively (Additional file 1: Fig. S5). Overall, the proposed framework could enhance the prediction of binding free energy changes induced by both DNA and protein mutations.

In addition, multiple-point mutation datasets, namely, MD37 and MP125, were used to assess our method (Fig. 6 and Additional file 1: Table S5). Despite being trained on single-point data, most individual models effectively predicted the energy changes in response to multiple-point mutations. Meanwhile, the effect of DE-based optimization was more remarkable on DNA mutations than protein mutations. The PCCs for MD37 and MP125 were 0.550 and 0.495, respectively. This indicated that PNBACE could transfer the knowledge learned from single-point mutation data to the prediction task for multiple-point mutations. Notably, our method generated a narrow prediction range, especially for MD37. This is because PNBACE had deficiencies in capturing extreme changes in binding affinity. Nevertheless, the connections among the experimental values of these mutations could be effectively predicted. Details are presented in the first limitation of PNBACE (the last section of results).

Comparison of PNBACE and other methods

Based on the test sets, we conducted a thorough comparison between PNBACE and the other state-of-the-art algorithms (Table 1, Additional file 1: Table S6, and S7). Currently, SAMPDI-3D is the only available method suitable for DNA mutations and it adopts traditional

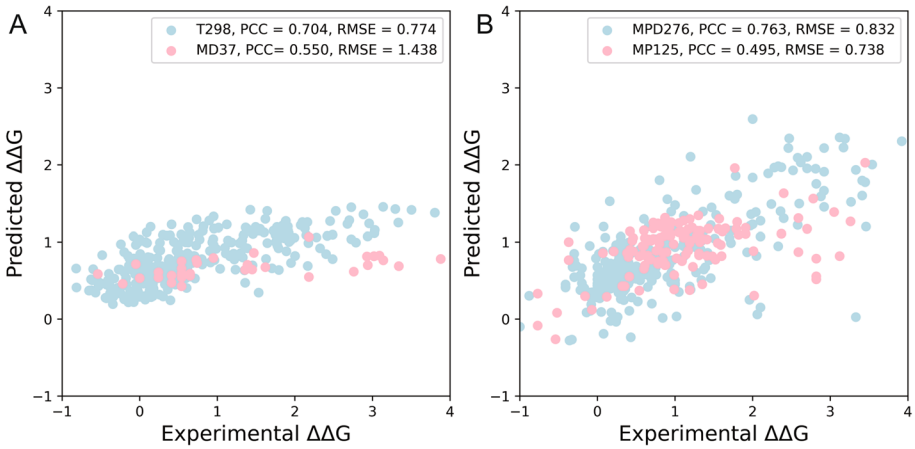

Fig. 6 Performance of PNBACE on multiple-point mutation datasets. **A** Results on MD37. **B** Results on MP125. The results on training sets are also provided and could be considered as the reference

Table 1 Comparison with other methods for regression tasks based on PCC

Dataset	mCSM-NA	SAMPDI	PremPDI	PremPRI	SAMPDI-3D	PEMPNI	PNBACE
T206	–	–	–	–	−0.129	0.193	0.535
T227	–	–	–	–	0.420	0.361	0.473
MPD48	−0.016	0.424	0.509	–	0.382	0.550	0.660
PDM	0.540	–	0.510	–	0.053	0.584	0.711
PDSI	–	0.530	0.740	–	0.251	0.478	0.611
S200	0.280	0.150	0.300	–	0.430	–	0.395
MPR79	0.564	–	–	0.279	–	0.407	0.417

"–" indicates that the result is unavailable

sequence and structural features as the input. We compared our method with this model using T227 and T206. PNBACE outperformed SAMPDI-3D in terms of both PCC and RMSE, suggesting that energy-related features may play a key role in improving model performance. Regarding single-point protein mutations in PDIs, PNBACE was compared with five existing approaches, including mCSM-NA, SAMPDI, PremPDI, SAMPDI-3D, and PEMPNI. mCSM-NA employs graph-based features to estimate the energy change, while the other methods (except SAMPDI-3D) combine energy and structural features to predict such changes. For single-point protein mutations in PRIs, the performance of PNBACE and three competing methods (mCSM-NA, PremPRI, and PEMPNI) was evaluated based on MPR79, and PremPRI was an extension of PremPDI for PRIs. As shown in Table 1, PNBACE surpassed the five methods on MPD48 and PDM (PCCs: 0.660 and 0.711) but obtained a relatively worse performance compared to PremPDI on PDSI and SAMPDI-3D on S200. Regarding MPR79, PNBACE performed better than both PEMPNI and PremPRI. Although the PCC of our method was inferior to the measure of mCSM-NA, our RMSE measure was markedly lower (0.728 versus 2.772). Additionally, we submitted the MP125 dataset to the mmCSM-NA server, which could predict changes upon multiple-point mutations. mmCSM-NA yielded a lower performance than our method, with a PCC and RMSE of 0.477 and 1.191, respectively. Finally, we compared PNBACE with the alanine scanning method, which is a purely physics-based model [22]. The competing method achieved PCCs of −0.290 and −0.003 on MPD48 and MPR79 (44 and 79 samples, respectively), compared to 0.662 and 0.417 for our method.

From the aforementioned results, we observed that our previous method (PEMPNI) achieved promising results on partial datasets, but the newly proposed PNBACE algorithm extended the application range and further boosted the prediction accuracy. The advantages of PNBACE over PEMPNI are summarized as follows: (1) In addition to the effect of protein mutations, PNBACE could be used to predict the impact of DNA mutations; (2) In addition to the partition-based energy features used by PEMPNI, the newly designed energy-based topological features based on different energy networks are used by PNBACE; and (3) In comparison to the relatively

simple ensemble module of PEMPNI, the current ensemble method is established on a greater number of component models along with a more advanced optimization algorithm (DE).

Prediction of mutations significantly decreasing binding affinities

To evaluate the classification performance of the model, we defined samples with $\Delta\Delta G \geq 1$ kcal/mol as mutations significantly reducing binding affinities. We adopted the XGBoost classification method instead of the regression method in the computational framework. In terms of the integrative strategy, overall, the DE method outperformed the simple average method on DNA mutations, and the two methods obtained comparable results on protein mutations (Additional file 1: Table S8). In addition, some individual models, such as the total energy term in the difference state of T227 and the wild-type state of MPR79, even achieved higher AUCs than the integrative models. This may be because the ensemble framework, together with its parameters, was directly transferred from the regression task and might need to be further optimized. Even so, PNBACE generated AUCs of 0.783, 0.803, and 0.694 for T298, MPD276, and MPR233, respectively. As shown in Additional file 1: Fig. S6, the subsets that were challenging for the regression task were also difficult for the classification task. Furthermore, we compared PNBACE and previous methods on test sets. As shown in Table 2 and Additional file 1: Table S7, SAMPDI-3D achieved worse results on both T227 and T206 than PNBACE. For protein mutations, PNBACE showed advantages over other methods on PDM and obtained comparable performance to PEMPNI on MPD48 and PremPDI on PDSI. For MPR79, our algorithm surpassed mCSM-NA and PremPRI but yielded lower measures than PEMPNI. Based on the newly collected MPD248 and MPR134 datasets, our method yielded AUCs of 0.713 and 0.566, respectively (Additional file 1: Fig. S5). When applied to multiple-point mutations, PNBACE demonstrated AUCs of 0.662 and 0.677 for MD37 and MP125, respectively, suggesting a certain transferability of our classification model. Thus, the proposed algorithm could identify different types of mutations that significantly reduce the binding affinity.

Major limitations of PNBACE

Despite the progress achieved here, there are three major limitations of this work. First, the binding affinity change values predicted by our model were concentrated within a narrow range (e.g., 0 ~ 2 kcal/mol). To investigate this issue, we divided the mutations of each dataset into two groups based on the actual changes in binding affinity: 0 ~ 2 kcal/mol and the remainder. The RMSE and PCC values of the second group were obviously higher than the corresponding measures of the first group (Additional file 1: Table S9). This suggests that the predicted values may not be accurate enough for mutations leading to extreme changes, but the relationships of the experimental values for these mutations can be effectively captured by our approach. Second, PNBACE could underestimate the experimental measures. As shown in Additional file 1: Table S9, the slopes of the fitting lines were between 0.1 and 0.4 for most datasets, implying that the actual energy changes were underestimated by a factor of 2.5 to 10. Moreover, the slopes for DNA mutation datasets were generally lower than those for protein mutation datasets, probably because the prediction of effects triggered by DNA mutations was more challenging. Third, our method had relatively higher computational costs (1 ~ 20 h per complex). For each mutation type, we chose a group of mutations from complexes with different lengths and recorded the running time of each sample (Additional file 1: Fig. S7). The higher computational cost of our algorithm was caused by pairwise residue energy decomposition and network feature extraction. Accordingly, the current model cannot be applied to genome-scale investigations. These limitations are worthy of further investigation in future work.

Discussion

Mutations occurring in either nucleic acids or proteins could impact the binding affinities of PNIs. Comprehensive studies have been devoted to understanding and predicting the effects induced by protein mutations.

Table 2 Comparison with other methods for classification tasks based on AUC

Dataset	mCSM-NA	SAMPDI	PremPDI	PremPRI	SAMPDI-3D	PEMPNI	PNBACE
T206	–	–	–	–	0.404	0.634	0.700
T227	–	–	–	–	0.618	0.700	0.718
MPD48	0.477	0.598	0.761	–	0.484	0.841	0.834
PDM	0.690	–	0.770	–	0.531	0.785	0.797
PDSI	–	0.790	0.850	–	0.681	0.778	0.849
MPR79	0.567	–	–	0.528	–	0.645	0.584

"–" indicates that the result is unavailable

However, limited attention has been directed toward predicting the effects of nucleic acid mutations as well as conducting a comparison between the methodologies for predicting the changes caused by the two types of mutations. In this work, we first showed that DNA mutations could cause varying degrees of binding affinity changes in terms of their mutation types, geometrical locations, biological functions, and species. Based on different energy networks of residues and bases, we developed a series of energy-based topological features, which were then combined with our previous partition-based energy features to develop prediction models. Through feature selections on single-point mutation datasets, we demonstrated that the features associated with betweenness and ETFMN played critical roles in predicting energy changes. Moreover, we compared the results of 18 individual models and two integrative models and observed that the DE method generally outperformed the simple average method and the best individual models for different types of mutations. Moreover, we showed that our method could not only estimate the energy influence induced by multiple-point mutations but also identify mutations significantly decreasing binding affinities. Extensive comparisons revealed that PNBACE largely performed better than existing methods on both regression and classification tasks. However, the current algorithm has some limitations, such as a concentrated prediction range, underestimating experimental values, and higher computational costs.

In addition to the three major limitations mentioned above, we could further improve our method from the following aspects. First, the settings of energy calculations were selected based on our experiences and may not be the optimal protocol. We also tried the new settings (i.e., ff19SB force fields, OPC water model and IGB66 implicit solvent) in this work [23, 24]. New settings performed worse than original settings for regression tasks but achieved better measures on partial datasets for classification tasks (Additional file 1: Table S10). The optimal protocol for MM/GBSA calculations is thus worthy of further exploration. Second, this method is a purely energy-based prediction algorithm that does not consider knowledge-based features. In the future, we can incorporate complementary structural features or models to improve its performance. Third, although there are many machine learning methods available, our individual models were solely dependent on the XGBoost method. It is possible to construct more effective ensemble models by combining the advantages of different machine learning algorithms. Fourth, here, we only adopted a DE algorithm and a simple average method to optimize the ensemble model. Other nature-inspired optimization algorithms could be used to tune the weights of individual models. Fifth, the datasets include limited experimental measurements for both DNA and protein mutations. Advanced data augmentation techniques may be considered to generate synthetic data for model construction. In summary, PNBACE could be a useful tool for estimating the binding affinity changes induced by both DNA and protein mutations, thus improving our understanding of PDIs and PRIs.

Conclusions

Existing computational studies have paid less attention to the impact of nucleic acid mutations on PNIs as well as a possible unified framework for predicting the changes caused by protein mutations and nucleic acid mutations. Herein, we developed a generalized algorithm called PNBACE to address the above problem. Leveraging different energy networks, we designed novel energy-based topological features and combined them with previous partition-based energy features to develop component prediction models using feature selection techniques. Furthermore, we utilized the DE algorithm to optimize the weights of component models to achieve more accurate ensemble models. PNBACE could not only predict the energy changes triggered by both DNA and protein mutations but also identify mutations significantly reducing binding affinities. Thus, this tool may be helpful in studying the influence of different types of mutations on PNIs.

Methods

Data collection

In this study, we collected experimentally measured binding affinity change data induced by residue and base mutations, both of which contained single-point and multiple-point entries. In total, we prepared 14 datasets for this work, among which six datasets were newly coined and the remaining datasets were derived from previous works. More details about the datasets are provided in Additional file 1: Text S1 and Table S11. Notably, the sequence and structural redundancy between the training set and the main test set was relatively low (Additional file 1: Text S1 and Fig. S8).

Single-point DNA mutation data

The single-point DNA mutation data were extracted from the D463 dataset collected by SAMPDI-3D and the records in the ProNAB database together with related literature [25]. From D463, we eliminated the entries containing nonstandard atoms, entries missing the coordinate information in the PDB file, and entries involving large proteins (over 1000 amino acids). This procedure resulted in 426 single-point mutations from 27 complexes. From the ProNAB database, we retrieved a total

of 20,090 records that were filtered out using the following criteria. We deleted the entries involved in PDIs without structural information, entries with multiple-point mutations, RNA mutations or protein mutations, entries missing energy measures for wild-type or mutant sites, entries involving large proteins, and entries appearing for the above 426 mutations. We obtained 69 mutations and manually checked the relevant literature, which led to an additional 9 mutations. By combining the filtered samples from different resources, we obtained 37 protein–DNA complexes, including 504 single-point DNA mutations. Subsequently, we built the training set (T298) based on 30 (80%) complexes and the test set (T206) based on 7 (20%) complexes. In addition, the T227 dataset prepared by SAMPDI-3D was also used for independent testing in this work.

Single-point protein mutation data

To evaluate whether our method could be applied to the binding affinity changes induced by protein mutations, we adopted the single-point mutation datasets used by our previous method (PEMPNI) [10]. For the PDIs, MPD276 was used as the training set, while MPD48, PDM, and PDSI were used as the test sets. Additionally, the other two datasets for the PRIs, namely, MPR233 and MPR79, were used for training and testing, respectively. We also built another two test sets based on the nonredundant datasets in the Nabe database [26]. By comparing against the training sets (MPD276/MPR233), we removed the complexes with sequence identities greater than 40% (for protein chains) and those with TM-scores generated by US-align [27] greater than 0.5 (for whole complexes). Moreover, we deleted the complexes without nucleic acid chains and the entries without structural information by checking the PDB files. Finally, we obtained 248 mutations (MPD248) from 80 protein–DNA complexes and 134 mutations (MPR134) from 36 protein–RNA complexes.

Multiple-point mutation data

The multiple-point DNA mutation dataset was constructed by manually checking entries in ProNAB. Based on the criteria mentioned above, 37 multiple-point mutation entries (MD37) from 7 protein–DNA complexes could be acquired as a test set. The multiple-point protein mutation data were selected from the dataset prepared by mmCSM-NA, which included 125 entries from 21 protein–DNA complexes and 16 entries from 7 protein–RNA complexes. Because of the very limited samples for PRIs, we only focused on multiple-point mutations in PDIs. Accordingly, 125 entries (MP125) were used as a test set in this work.

Overview of PNBACE

As shown in Fig. 1, the PNBACE algorithm is divided into four steps. First, the binding free energy and associated energy terms of protein-nucleic acid complexes were calculated and decomposed into pairwise energies between residues/bases using the MM/GBSA approach. Second, networks with pairwise energies as the weight were built for each complex according to different energy terms, and a series of features indicating the energy-based properties of different regions within the complex were designed based on the topological attributes of nodes in the energy network. Third, the extreme gradient boosting (XGBoost) algorithm combined with the feature selection procedure was used to construct specific energy term-based prediction models under three states (wild-type, mutant, and differences between them). Finally, an ensemble model that integrated the results of the component models was developed using a differential evolution algorithm.

Structure optimization and energy calculation

To generate mutant structures, we replaced the original residues and bases with mutant counterparts in wild-type structures using Modeller and 3DNA, respectively [28, 29]. Structure optimization and energy calculation were conducted with the ff14SB and parmbsc1 force fields in Amber18 [30, 31]. Each complex was embedded into a TIP3P water box 10 Å from the solute using the *tleap* module. Counter ions (Na + and Cl −) were then added to neutralize the total charge. Following our previous work, a three-step energy minimization strategy was applied to each complex (Fig. 1A). This process involved 200 iterations with an elastic constant of 50 kcal $mol^{-1}Å^{-2}$, another 200 iterations with an elastic constant of 10 kcal $mol^{-1}Å^{-2}$, and a final 200 iterations without any constraints. In each phase, there were 100 iterations of steepest descent minimization and 100 iterations of conjugate gradient minimization. Subsequently, the binding free energies of wild-type and mutant complexes could be calculated using the MM/GBSA approach implemented in the MMPBSA.py program based on the minimized structure as follows [32]:

$$\Delta G \approx \Delta E_{\mathrm{ele}} + \Delta E_{\mathrm{vdw}} + \Delta G_{\mathrm{GB}} + \Delta G_{\mathrm{SA}} \quad (1)$$

where ΔG denotes the total energy and includes the gas-phase interaction energy and the desolvation energy. The former contains electrostatic (ΔE_{ele}) and van der Waals (ΔE_{vdw}) interactions, which were calculated using the *sander* program. The latter contains polar (ΔG_{GB}) and nonpolar (ΔG_{SA}) components. The GB model developed by Hawkins et al. (GBH^{CT}) was selected to estimate the polar part based on our previous experience [33]. The

nonpolar energy was computed using the LCPO algorithm [34]. The total energy and each energy term were decomposed into residue/base pairwise energies by the MMPBSA program.

Novel features derived from energy networks

As illustrated in Fig. 1B, we built five energy networks for each complex based on the binding free energy and four energy terms. For each energy type, a complex was represented by a weighted network, where nodes denote residues and bases, and the weight of an edge denotes the decomposed energy value at the residue/base pair level. For each node, we then calculated 10 topological features, including weighted node degree, weighted closeness centrality, weighted eigenvector centrality, weighted betweenness centrality, weighted current flow betweenness centrality, weighted harmonic centrality, weighted clustering, weighted average neighbor degree, weighted approximate current flow betweenness centrality, and weighted current flow closeness centrality. The definitions of these attributes are given in Additional file 1:Text S2.

Based on each attribute, we designed four groups of novel energy descriptors to reflect the energy contributions of different partitions within each complex. First, we focused on the mutant nodes in the network. Regarding a specific topological feature, the average and sum of attribute values of mutant residues/bases in the complex were calculated. If the mutant section included only one residue/base, the average was equal to the sum; otherwise (e.g., a base pair or multiple points), these two measures were different. These descriptors were termed energy-based topological features of mutant nodes (ETFMN). Second, we separated each complex into two partitions, namely, the interface and noninterface regions. A residue-nucleotide contact was generated if at least one pair of nonhydrogen atoms was within a distance of 5 Å. We computed the average and sum of the topological features of nodes in the two regions. They were termed energy-based topological features of the interface and noninterface (ETFIN). Third, nodes involved in the interface regions could be further classified into interfacial residues and bases. Thus, the average and sum of the topological features were computed for both partitions. These descriptors were termed energy-based topological features of the partitioned interface (ETFPI). Fourth, we decomposed a complex into different regions in terms of the distances between the mutant nodes and other nodes, including 0 ~ 3 Å, 3 ~ 4 Å, 4 ~ 5 Å, 5 ~ 6 Å, and > 6 Å, and yielded the average and sum of topological features of each partition. They were termed energy-based topological features of the partitioned complex (ETFPC). Generally, these feature groups can be defined as follows:

$$S_{ij} = \sum_{n \in PT_i} topo\_feature_{nj} \tag{2}$$

$$A_{ij} = \frac{S_{ij}}{num\ (n)} \tag{3}$$

where i represents the index of a partition (PT) and j represents the index of a given topological feature ($topo\_feature$). n denotes nodes in a PT, and $num(n)$ denotes the number of nodes.

Our previous study applied partition-based energy features to predict the binding affinity changes that arise from protein mutations [10]. Herein, we also utilized these energy descriptors, including the energies between the target and other residues (ETOR), the energies of the partitioned complex (EPC), the energies of the interface and noninterface (EINI), and the energies of the partitioned interface (EPI), to complement the new features. Collectively, 44 energy-based feature groups were generated for each complex in this work (Additional file 1: Table S1).

Feature selection and individual model construction

We not only computed the aforementioned features of each energy type under the wild-type and mutant states but also calculated the differences in the measures of the two states. Due to the three states and five energy types, we adopted the XGBoost method to build 15 regression models, each of which corresponds to a combination of states and energy types (Fig. 1C). Default parameters were used for the XGBoost algorithm. From the 44 feature groups, the sequential forward selection (SFS) algorithm was utilized to choose the most effective features according to the Pearson correlation coefficient (PCC) measure. We started with the feature group that displayed the best performance in the first round and iteratively selected a new group from the remaining ones so that the combination of this group and the reserved groups in the preceding round could improve its performance to the greatest extent. This process was halted when the PCC started to decrease. In addition, to explore the complementarity between different energy terms, we built a novel feature pool by combining the feature groups of five energy types under each state. After the initial screening, the feature groups with a PCC higher than 0.15 were retained for the SFS process. We therefore generated another three regression models tailored to different states. For the classification task, the regression models were replaced by XGBoost classification algorithms, and the selected feature groups remained unchanged.

Ensemble model construction using a differential evolution algorithm

To make use of the complementarity between the outputs of the above 18 models, we constructed an ensemble model using a differential evolution (DE) algorithm [35], which iteratively used the mutation, crossover, and selection operations to achieve the weight for individual models (Additional file 1: Fig. S9). Specifically, we randomly initialized a population $P = \{P_1, P_2, \cdots, P_n\}$ consisting of 18-dimensional vectors, where each vector is an individual and denotes the weights of 18 models. Here, n was empirically set to 50. Subsequently, we mutated the individuals to keep the population evolving. In detail, three individuals were randomly selected from the initial population P, and a mutant individual P_m was generated by calculating the difference vector between any two individuals and adding it to the third individual (Formula 4). Additionally, a crossover operation was performed on the mutant individual P_m and the parent individual P_i to generate an offspring individual $P_n\prime$. As shown in Additional file 1: Fig. S10, a random number was assigned to the elements in the paired vectors. If the random number was less than or equal to the crossover probability, the element of the mutant individual was selected; otherwise, the element of the original individual was adopted (Formula 5). To select high-quality individuals, we compared offspring individuals with their parents in terms of fitness, which was the PCC measure. During the evolutionary process, individuals with a high fitness score were used as candidates for the next generation (Formula 6). The DE procedure was halted if the maximum number of iterations was reached (10 iterations assigned empirically), and the individual with the highest fitness score was chosen as the optimal weight. To enhance the robustness, we performed the DE procedure 5 times, and the averages of the optimal weights were finally used:

$$P_m = P_{r1} + F(P_{r2} - P_{r3}) \tag{4}$$

$$P'_{n,j} = \begin{cases} P_{m,j}, \, if \;\; r \; and \;\; (0,1) \leq CR \\ P_{i,j}, \, else \end{cases} \tag{5}$$

$$P_c = \begin{cases} P'_n, \, if \; fit(P'_n) \geq fit(P_i) \\ P_i, else \end{cases} \tag{6}$$

where P_m is a mutant individual, and P_{r1}, P_{r2}, and P_{r3} are individuals in population P. F, the weight of the difference vector, is set to 0.5. $P'_{n,j}$ is the element of an offspring individual, $P_{m,j}$ and $P_{i,j}$ are the elements of mutant and parent individuals, respectively, and j is the index of an element. $rand(0,1)$ is a random number, and CR, the crossover rate, is set to 0.5. P_c is a candidate individual with a high fitness score, and $fit(x)$ is the fitness function. Finally, the ensemble score generated by PNBACE can be presented as follows:

$$PNBACE_{score} = w_1 * Score_1 + w_2 * Score_2 + \cdots + w_{18} * Score_{18} \tag{7}$$

where $Score$ is the prediction value of each component model, and w is the weight assigned by the DE method. More details can be found in Additional file 1: Table S12.

Performance evaluation

We used leave-one-complex-out validation (LOCOV) to assess the model on the training sets. The dataset was divided into n folds (n is the number of complexes). Then, mutations from one complex were used for testing, whereas the other (n-1) folds were merged for training. This process was iterated n times to ensure that all complexes were tested. Independent test sets were also adopted to validate the model. The PCC and root mean squared error (RMSE) were used for the regression task. The Matthews correlation coefficient (MCC) and area under the curve (AUC) were used for classification. Statistical tests were used to evaluate significant differences in performance between different methods. For a given dataset, we randomly chose 70% of complexes 10 times and calculated the PCC (or AUC) value in each iteration. The Anderson–Darling test was then used to assess whether these values obey a normal distribution. Based on the normality assumption, the paired t-test or Wilcoxon rank-sum test was selected for statistical testing.

Abbreviations

AUC	Area under the curve
DE	Differential evolution
PCC	Pearson correlation coefficient
MCC	Matthews correlation coefficient
RMSE	Root mean squared error
ROC	Receiver operating characteristic
SFS	Sequential forward selection

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12915-024-02006-9.

Additional file 1. Texts S1-S2, Figs S1-S10, Tables S1-S12. Text S1 Details and comparison of datasets used in this work. Text S2 Definition of network attributes used in this work. Fig. S1 Comparison of the effects of DNA and protein mutations on PDIs. Fig. S2 Binding affinity changes caused by the positional effect of mutations. Fig. S3 Correlation coefficients between 44 types of features for T298 dataset. Fig. S4 RMSE values of individual models without and with feature selections on training sets. Fig. S5 Performance of PNBACE on newly collected test sets. Fig. S6 AUC values of PNBACE on different mutation subsets. Fig. S7 Running time for mutations from complexes with different lengths. Fig. S8 Sequence and structural redundancy between training sets and main test sets. Fig. S9 Flowchart of differential evolution algorithm. Fig. S10 Mutation and crossover operations in differential evolution algorithms. Table S1 44 types of features based on each

energy network. Table S2 Performance of PNBACE using alterative evaluation strategies. Table S3 PCC and RMSE values of T206 and T227 datasets. Table S4 PCC and RMSE values of different protein mutation datasets. Table S5 PCC and RMSE values of multiple-point mutations. Table S6 Comparison with other methods for regression tasks based on RMSE. Table S7 P-value of difference in PCC and AUC between PNBACE and other models. Table S8 AUC and MCC values of independent and ensemble models for classification tasks. Table S9 Evaluation measures on different datasets and their subsets. Table S10 Comparison of PNBACE with different settings for MM/GBSA calculations. Table S11 Mutation datasets used in this study. Table S12 Selected feature groups and DE-based weights of each component model.

Acknowledgements
Not applicable.

Authors' contributions
S.-R.X.: Methodology, Investigation, Data curation, Software, Visualization, Writing-original draft. Y.-K.Z.: Investigation, Data curation. K.-Y.L.: Investigation, Validation. Y.-X.H.: Investigation, Writing-original draft. R.L.: Conceptualization, Methodology, Supervision, Writing-review & editing, Funding acquisition. All authors read and approved the final manuscript.

Funding
This work was supported by the National Natural Science Foundation of China (32071249).

Availability of data and materials
The webserver, datasets, source codes of PNBACE can be available at http://liulab.hzau.edu.cn/PNBACE/. All data generated or analyzed during this study are included in this published article, its supplementary information files, and publicly available repositories. The PDM and PDSI datasets were obtained from [8]. The MPD276, MPD48, MPR233, and MPR79 datasets were obtained from [10]. The T277 and S200 datasets were obtained from [15]. The other datasets were newly collected in this work. More details are shown in Additional file 1.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 24 December 2023 Accepted: 3 September 2024
Published online: 11 September 2024

References

1. Ollis DL, White SW. Structural basis of protein-nucleic acid interactions. Chem Rev. 1987;87:981–95.
2. Teh HF, Peh WY, Su X, Thomsen JS. Characterization of protein—DNA interactions using surface plasmon resonance spectroscopy with various assay schemes. Biochemistry. 2007;46:2127–35.
3. Doyle ML. Characterization of binding interactions by isothermal titration calorimetry. Curr Opin Biotechnol. 1997;8:31–5.
4. Hillisch A, Lorenz M, Diekmann S. Recent advances in FRET: distance determination in protein–DNA complexes. Curr Opin Struc Biol. 2001;11:201–7.
5. Cain S, Risheh A, Forouzesh N. A physics-guided neural network for predicting protein–ligand binding free energy: from host–guest systems to the PDBbind database. Biomolecules. 2022;12:919.
6. Mobley DL, Gilson MK. Predicting binding free energies: frontiers and benchmarks. Annu Rev Biophys. 2017;46:531–58.
7. Peng Y, Sun L, Jia Z, Li L, Alexov E. Predicting protein–DNA binding free energy change upon missense mutations using modified MM/PBSA approach: SAMPDI webserver. Bioinformatics. 2018;34:779–86.
8. 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.
9. Zhang N, Lu H, Chen Y, Zhu Z, Yang Q, Wang S, et al. PremPRI: Predicting the Effects of Missense Mutations on Protein-RNA Interactions. Int J Mol Sci. 2020;21:5660.
10. Jiang Y, Liu H-F, Liu R. Systematic comparison and prediction of the effects of missense mutations on protein-DNA and protein-RNA interactions. PLoS Comput Biol. 2021;17:e1008951.
11. Pires DE, Ascher DB. mCSM–NA: predicting the effects of mutations on protein–nucleic acids interactions. Nucleic Acids Res. 2017;45:W241–6.
12. Nguyen TB, Myung Y, de Sá AG, Pires DE, Ascher DB. mmCSM-NA: accurately predicting effects of single and multiple mutations on protein–nucleic acid binding affinity. NAR Genom Bioinform. 2021;3:lqab109.
13. Zhang S, Zhao L, Zheng C-H, Xia J. A feature-based approach to predict hot spots in protein–DNA binding interfaces. Brief Bioinform. 2020;21:1038–46.
14. Pan Y, Wang Z, Zhan W, Deng L. Computational identification of binding energy hot spots in protein-RNA complexes using an ensemble approach. Bioinformatics. 2018;34:1473–80.
15. Li G, Panday SK, Peng Y, Alexov E. SAMPDI-3D: predicting the effects of protein and DNA mutations on protein–DNA interactions. Bioinformatics. 2021;37:3760–5.
16. Seeman NC, Rosenberg JM, Rich A. Sequence-specific recognition of double helical nucleic acids by proteins. Proc Natl Acad Sci. 1976;73:804–8.
17. Yang W, Gao Y. Translesion and repair DNA polymerases: diverse structure and mechanism. Annu Rev Biochem. 2018;87:239–61.
18. Xue B, Dunker AK, Uversky VN. Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life. J Biomol Struct Dyn. 2012;30:137–49.
19. Genheden S, Ryde U. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. Expert Opin Drug Dis. 2015;10:449–61.
20. Norberg J. Association of protein–DNA recognition complexes: electrostatic and nonelectrostatic effects. Arch Biochem Biophys. 2003;410:48–68.
21. Yu B, Pettitt BM, Iwahara J. Dynamics of ionic interactions at protein–nucleic acid interfaces. Acc Chem Res. 2020;53:1802–10.
22. Ramos RM, Moreira IS. Computational alanine scanning mutagenesis-an improved methodological approach for protein–DNA complexes. J Chem Theory Comput. 2013;9:4243–56.
23. Forouzesh N, Mishra N. An effective MM/GBSA protocol for absolute binding free energy calculations: a case study on SARS-CoV-2 spike protein and the human ACE2 receptor. Molecules. 2021;26:2383.
24. Tian C, Kasavajhala K, Belfon KA, Raguette L, Huang H, Migues AN, et al. ff19SB: amino-acid-specific protein backbone parameters trained against quantum mechanics energy surfaces in solution. J Chem Theory Comput. 2019;16:528–52.
25. 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:D1528–34.
26. Liu J, Liu S, Liu C, Zhang Y, Pan Y, Wang Z, et al. Nabe: an energetic database of amino acid mutations in protein–nucleic acid binding interfaces. Database. 2021; 2021: baab050.
27. Zhang C, Shine M, Pyle AM, Zhang Y. US-align: universal structure alignments of proteins, nucleic acids, and macromolecular complexes. Nat Methods. 2022;19:1109–15.
28. Eswar N, Eramian D, Webb B, Shen M-Y, Sali A. Protein structure modeling with MODELLER. Methods Mol Biol. 2008;426:145–59.
29. Lu XJ, Olson WK. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. Nucleic Acids Res. 2003;31:5108–21.
30. Maier JA, Martinez C, Kasavajhala K, Wickstrom L, Hauser KE, Simmerling C. ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. J Chem Theory Comput. 2015;11:3696–713.
31. Ivani I, Dans PD, Noy A, Pérez A, Faustino I, Hospital A, et al. Parmbsc1: a refined force field for DNA simulations. Nat Methods. 2016;13:55–8.

32. Miller III BR, McGee Jr TD, Swails JM, Homeyer N, Gohlke H, Roitberg AE. MMPBSA. py: an efficient program for end-state free energy calculations. J Chem Theory Comput. 2012; 8: 3314–3321.
33. Hawkins GD, Cramer CJ, Truhlar DG. Parametrized models of aqueous free energies of solvation based on pairwise descreening of solute atomic charges from a dielectric medium. J Phys Chem C. 1996;100:19824–39.
34. Weiser J, Shenkin PS, Still WC. Approximate atomic surfaces from linear combinations of pairwise overlaps (LCPO). J Theor Comput Chem. 1999;20:217–30.
35. Storn R, Price K. Differential evolution-a simple and efficient heuristic for global optimization over continuous spaces. J Global Optim. 1997;11:341.

Publisher's Note