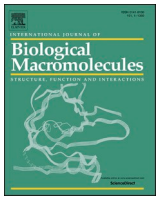
International Journal of Biological Macromolecules 281 (2024) 136147

Contents lists available at ScienceDirect

International Journal of Biological Macromolecules

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



# PDNAPred: Interpretable prediction of protein-DNA binding sites based on pre-trained protein language models

Lingrong Zhang, Taigang Liu *

*College of Information Technology, Shanghai Ocean University, Shanghai 201306, China*

ARTICLE INFO

*Keywords:*
Protein-DNA binding site
Pre-trained protein language model
Model interpretability

ABSTRACT

Protein-DNA interactions play critical roles in various biological processes and are essential for drug discovery. However, traditional experimental methods are labor-intensive and unable to keep pace with the increasing volume of protein sequences, leading to a substantial number of proteins lacking DNA-binding annotations. Therefore, developing an efficient computational method to identify protein-DNA binding sites is crucial. Unfortunately, most existing computational methods rely on manually selected features or protein structure information, making these methods inapplicable to large-scale prediction tasks. In this study, we introduced PDNAPred, a sequence-based method that combines two pre-trained protein language models with a designed CNN-GRU network to identify DNA-binding sites. Additionally, to tackle the issue of imbalanced dataset samples, we employed focal loss. Our comprehensive experiments demonstrated that PDNAPred significantly improved the accuracy of DNA-binding site prediction, outperforming existing state-of-the-art sequence-based methods. Remarkably, PDNAPred also achieved results comparable to advanced structure-based methods. The designed CNN-GRU network enhances its capability to detect DNA-binding sites accurately. Furthermore, we validated the versatility of PDNAPred by training it on RNA-binding site datasets, showing its potential as a general framework for amino acid binding site prediction. Finally, we conducted model interpretability analysis to elucidate the reasons behind PDNAPred's outstanding performance.

## 1. Introduction

Protein-DNA interactions play a pivotal role in a wide range of biological processes, including DNA replication, signal transduction, transcriptional regulation, and gene expression [1–3]. Gaining insight into protein-DNA interactions holds tremendous potential for predicting protein function [4–6], elucidating disease pathogenesis [7,8], and identifying novel drug targets [9,10]. Due to the critical role of protein-DNA binding sites, many experiment-based approaches have been developed to identify the protein-DNA binding sites, such as Fast ChIP [11], X-ray crystallography [12], and electrophoretic mobility shift analysis (EMSA) [13]. However, these experimental approaches, although providing high-quality protein annotations, are often costly, time-consuming, and unable to keep pace with the exponential growth of protein sequences in the post-genomic era. Therefore, there are a substantial number of sequenced proteins that still lack DNA binding annotations. With the rapid development of artificial intelligence, several advanced techniques have been extensively applied in the field of bioinformatics, leading to significant breakthroughs and underscoring the pivotal role of computer technology in this domain [14–16]. The development of an efficient and accurate computational method for the identification of protein-DNA binding residues is imperative [17].

Current computational methods for protein-DNA binding site prediction can be categorized into two types: sequence-based methods and structure-based methods [18,19]. To be specific, structure-based methods utilize either natural or predicted 3D structural information of proteins. Protein structures contain abundant information and play a crucial role in determining protein function. Hence, utilizing protein structure information often leads to better performance in predicting protein-DNA binding sites when compared to sequence-based methods. Existing structure-based methods include DeepSite [20], GraphBind [21], DNABind [22], COACH-D [23], NucBind [24], GraphSite [17], and so on. For instance, GraphBind [21] utilizes structural information and spatial neighborhoods relationship to construct graphs, which are then classified using hierarchical graph neural networks (HGNNs). DNABind [22] employs a hybrid approach that combines machine learning techniques with template-based strategies, relying on structural alignment,

* Corresponding author.
*E-mail address:* tgliu@shou.edu.cn (T. Liu).

https://doi.org/10.1016/j.ijbiomac.2024.136147
Received 28 May 2024; Received in revised form 11 September 2024; Accepted 27 September 2024
Available online 1 October 2024

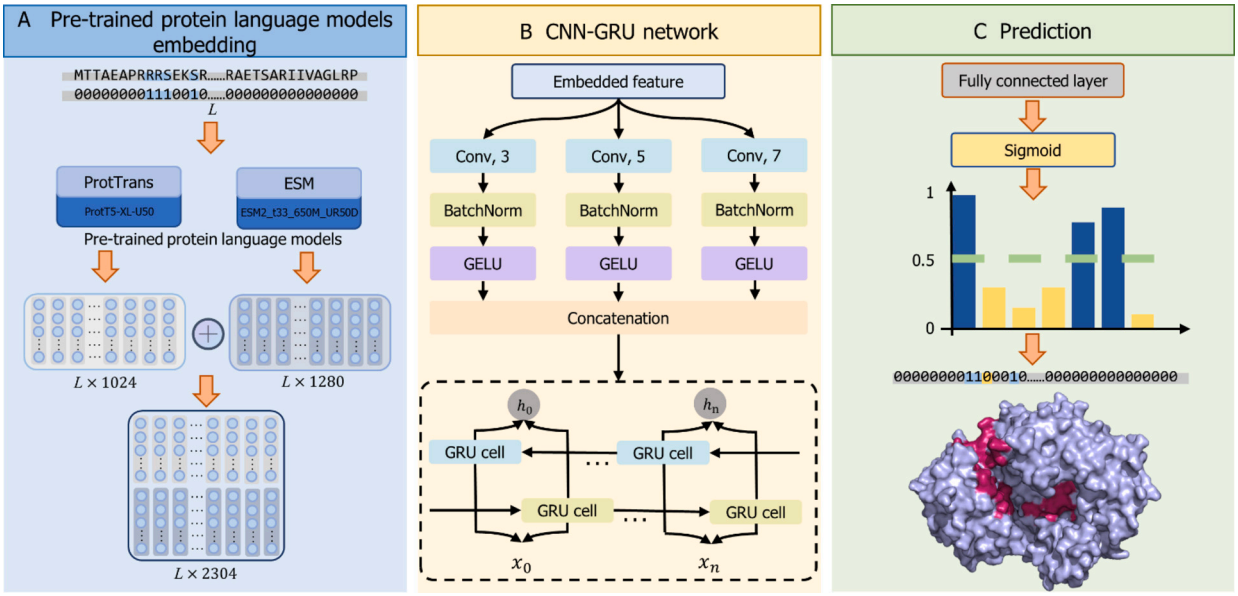



**Fig. 1.** The workflow of PDNAPred.

to accurately classify DNA-binding sites. However, existing structure-based methods often demand extensive biological structure information, leading to high computational resource requirements. Additionally, these methods may struggle to accurately predict binding sites when the protein or nucleic acid undergoes substantial conformational changes upon binding [25,26]. Moreover, many proteins still lack experimentally determined structural information, leading to existing structure-based methods unsuitable for large-scale prediction.

Sequence-based methods have received significant attention due to their data availability. These methods solely rely on sequence information to identify DNA-binding sites, making them more user-friendly and computationally efficient. These methods include DRNAPred [27], DNAPred [28], SVMnuc [24], NCBRPred [29], DBPred [30], CLAPE [31], and so on. However, most existing sequence-based methods still exhibit subpar performance in practical applications, primarily due to their reliance on manual feature engineering. For example, DRNAPred [27] utilizes 15 physicochemical features and 8 biochemical features to represent protein sequences. DNAPred [28] employs features including position-specific scoring matrix (PSSM), predicted secondary structure (PSS), predicted relative solvent accessibility (PRSA), and amino acid frequency difference between binding and nonbinding (AAFD-BN) to represent residues. It is apparent that the effectiveness of sequence-based models heavily relies on the careful selection and representation of features, which presents a significant obstacle in achieving comparable performance to structure-based methods. Furthermore, it must be acknowledged that traditional manual feature extraction methods are often time-consuming and labor-intensive. Moreover, protein sequence representations that rely on the physicochemical characteristics of proteins lack more efficient potential evolutionary information. Evolutionary information plays an important role in protein function prediction. Although PSSM-based features contain rich evolutionary information, their acquisition cost is often high due to their reliance on multiple sequence comparison methods. Therefore, it is important to adopt a more effective protein sequence feature extraction method.

Recently, the significant advancements made in large-scale language modelings have extended to various fields, including the study of amino acids in proteins [25]. The pre-trained protein language model (PPLM) is a result of applying natural language processing (NLP) techniques to bioinformatics [32,33]. By leveraging the vast amounts of protein sequence data available, PPLMs can learn complex biological information from billions of protein sequences, allowing for the extraction of comprehensive features that solely rely on sequences [34]. This is particularly advantageous in overcoming the limitations of small and potentially biased labeled datasets [35]. In the task of identifying DNA-binding sites, researchers have made progress by using PPLMs. For instance, Liu et al. developed CLAPE [31], which employed the ProtT5-XL-UniRef50 [36] (referred to as ProtT5) model to extract features from sequences and utilized a convolutional neural network to identify DNA-binding sites, achieving impressive performance. Similarly, Zhu et al. introduced ULDNA [37], which leveraged the ESM-2 [38], ProtT5 [36], and ESM-MAS [39] models for feature extraction. They combined these features with an LSTM-attention network to accurately identify DNA binding sites, yielding remarkable outcomes.

It is important to acknowledge that the features obtained from PPLMs capture general information across various proteins and may lack personalized features specific to particular proteins [40,41]. Consequently, there is a crucial demand to develop a more powerful neural network architecture that can forcefully utilize the embedded features extracted from PPLMs. Additionally, the imbalanced distribution of DNA-binding sites in datasets introduces several challenges during model training, including the potential for biased performance. Therefore, it is necessary to devise and implement strategies that can productively mitigate the impact of class imbalance on training high-quality models. Furthermore, despite some approaches demonstrating the effectiveness of leveraging PPLMs, there remains a lack of in-depth exploration of model interpretability. The specific reasons why the PPLMs model can achieve such remarkable results have not been analyzed in detail from the perspective of the model. Given these factors, it is essential to comprehensively address the challenges mentioned above and develop a robust and reliable method for accurately identifying DNA binding sites.

In this study, we proposed a novel sequence-based method called PDNAPred, which accurately identified protein-DNA binding sites by combining PPLMs with a specifically designed CNN-GRU network. Specifically, we adopted a transfer learning approach to extract residue-level embeddings learned by PPLMs, utilizing both ProtT5 and ESM-2 models. Subsequently, we employed the CNN-GRU network to capture the subtle features within the embedded representations crucial for identifying DNA-binding sites. Additionally, we decreased the impact of imbalanced data during model training by employing focal loss. PDNAPred was systematically evaluated on two independent DNA-binding sites test sets. The computational results demonstrated that

**Table 1**
Summary of benchmark protein-DNA binding datasets.

| Datasets | Dataset 1 | | Dataset 2 | | |
|---|---|---|---|---|---|
| | TR646 | TE46 | TR573 | TE181 | TE129 |
| Number of proteins | 646 | 46 | 573 | 181 | 129 |
| Number of sites | 314,139 | 10,876 | 159,883 | 75,258 | 37,515 |
| Number of binding sites | 15,636 | 965 | 14,479 | 3208 | 2240 |
| Number of non-binding sites | 298,503 | 9911 | 145,404 | 72,050 | 35,275 |
| % of binding sites | 4.98 | 8.87 | 9.06 | 4.26 | 5.97 |

PDNAPred outperformed existing state-of-the-art sequence-based methods while achieving performance competitive with structure-based approaches. Furthermore, PDNAPred also exhibited promising performance on RNA-binding site datasets, indicating its potential as a general framework for predicting protein-nucleic acid binding sites. Finally, to provide a more comprehensive understanding of the factors contributing to PDNAPred's impressive performance, an analysis of model interpretation was conducted. Fig. 1 depicts the flowchart of PDNAPred, providing a visual representation of its methodology and processes.

## 2. Methods and materials

### 2.1. Dataset description

In this study, to evaluate the proposed PDNAPred, we utilized two classical benchmark datasets. For the sake of clarity, we referred to these two datasets as Dataset 1 and Dataset 2, with the training sets labeled as TR and the testing sets as TE. In these datasets, a residue was classified as a DNA-binding site if the smallest atomic distance between the target residue and the DNA molecule was less than 0.5 Å plus the sum of the Van der Waals radius of the two nearest atom [17].

Dataset 1 was initially developed for the study of DBPred, which was proposed by Patiyal et al. for identifying protein DNA binding residues [30]. It contains a training set and a test set. The training set is denoted as TR646, which includes 646 proteins with 15,636 binding sites and 298,503 non-binding sites. The test set is denoted as TE46, which consists of 46 proteins with 965 binding sites and 9911 non-binding sites.

Dataset 2 was compiled from two studies. Among them, TR573 and TE181 are from the GraphBind, a method proposed by Xia et al. for identifying nucleic acid binding residues using structural information [21]. TR573 is a training set of 573 proteins with 14,479 binding sites and 145,404 non-binding sites.TE181 is a test set of 181 proteins, which has 3208 binding sites and 72,050 non-binding sites. Additionally, TE129 is derived from the GraphSite, a model proposed by Yuan et al. for predicting protein-DNA binding sites using AlphaFold2 [17]. It excludes redundant proteins with more than 30 % protein sequence identity to those in TE181. Specifically, TE129 contains 129 proteins with 2240 binding sites and 35,275 non-binding sites.

Furthermore, the CD-HIT [42] was employed with a threshold set at 0.3, ensuring that no proteins exhibited a similarity exceeding 30 % between the training and test sets within each dataset. It is important to note that data imbalances were identified in both Dataset 1 and Dataset 2, as depicted in Table 1. Specifically, in TR646, only 4.98 % of the residues are binding residues. Likewise, in TR573, 9.06 % of the residues are binding residues. This suggests that the prevalence of negative samples (non-binding residues) could introduce bias in model training and prediction, resulting in weak discrimination of binding sites. To gain deeper insights into the frequency distribution of the 20 amino acids across binding and non-binding sites within these datasets, we generated Fig. S1. Evidently, the distribution patterns of amino acids in both binding and non-binding residues remain largely consistent across the training and test sets. This observation underscores the challenge of accurately identifying DNA binding residues.

### 2.2. Feature extraction module

Self-supervised techniques in NLP leverage contextual information to predict missing words, which can provide insights into word meanings. Protein Language Models extend the application of diverse language models to the field of biochemistry. By processing protein sequences, they acquire knowledge about the underlying biochemical properties, secondary and tertiary structures, and functional patterns within these sequences [43]. The learned representations exhibit a hierarchical structure, capturing information from the amino acid properties to the distant relationships between proteins. These representations are subsequently employed as embeddings for downstream analysis tasks using transfer learning.

In this study, we employed the ESM-2 [38] and ProtT5 [36] to facilitate rapid mining and automatic extraction of potentially discriminative representations from protein sequences associated with DNA-binding sites. Specifically, for the ESM-2 model, we used the "esm2_t33_650M_UR50D" pre-trained model to extract feature embeddings, which was pre-trained on the UniRef50 database [44]. Each protein with $L$ residues will produce an embedded feature matrix of size $L \times 1280$ when using the ESM-2 model, resulting in a feature vector with 1280 dimensions for each residue. As for the ProtT5, we used "ProtT5-XL-U50", a 24-layer Transformer-based pre-trained model, was initially trained on the Big Fantastic Database (BFD) [45] and further fine-tuned with UniRef50 database [44]. Similar to ESM-2, ProtT5 will produce an embedded feature matrix of size $L \times 1024$, with each residue's feature vector having 1024 dimensions. Given that evolutionary knowledge from multiple database sources may be complementary [46], for each protein with a length of $L$, we concatenate the two feature embedding matrices to form a hybrid embedding matrix with dimensions of $L \times 2304$ as shown in Fig. 1A.

### 2.3. CNN-GRU network

The designed CNN-GRU network includes a CNN layer, a BiGRU layer, a fully connected layer, and an output layer, as shown in Fig. 1B.

The CNN layer was designed with a parallel architecture consisting of three convolutional layers. Each layer will reduce the dimension from 2304 to 256. Specifically, the three convolutional layers share the same input channels but have distinct receptive fields of $3 \times 3$, $5 \times 5$, and $7 \times 7$. This configuration enables the network to capture information at various spatial scales, enhancing its ability to detect relevant features. Following each convolutional operation, batch normalization is applied to normalize the activations and stabilize the learning process. Batch normalization accelerates training and improves the generalization ability of the network by reducing internal covariate shifts. To introduce non-linearity and increase the network's expressive power, the gaussian error linear units (GELU) activation function [47] is employed after each batch normalization steps. The outputs of the three convolutional layers are concatenated to form a feature with dimensions of 768. This output is subsequently fed into the BiGRU layer.

The BiGRU layer consists of a Bidirectional Gated Recurrent Unit (BiGRU) [48], which utilizes the Gated Recurrent Unit (GRU) cell as its primary component. GRU is a type of recurrent neural network (RNN) designed to address the gradient problem encountered during the processing of long-term memory and backward forward [49,50]. Unlike Long-Short Term Memory (LSTM) architecture, the GRU unit employs an update gate in place of input and forget gates, streamlining the computation of the hidden state within the network and enhancing efficiency, particularly when dealing with extensive training data. Its fundamental role lies in the storage and filtration of feature information through the update gate and reset gate mechanisms:

$$rt = \sigma(W_r \bullet [h(t-1), xt]) \quad (1)$$

$$zt = \sigma(W_Z \bullet [h(t-1), xt]) \quad (2)$$

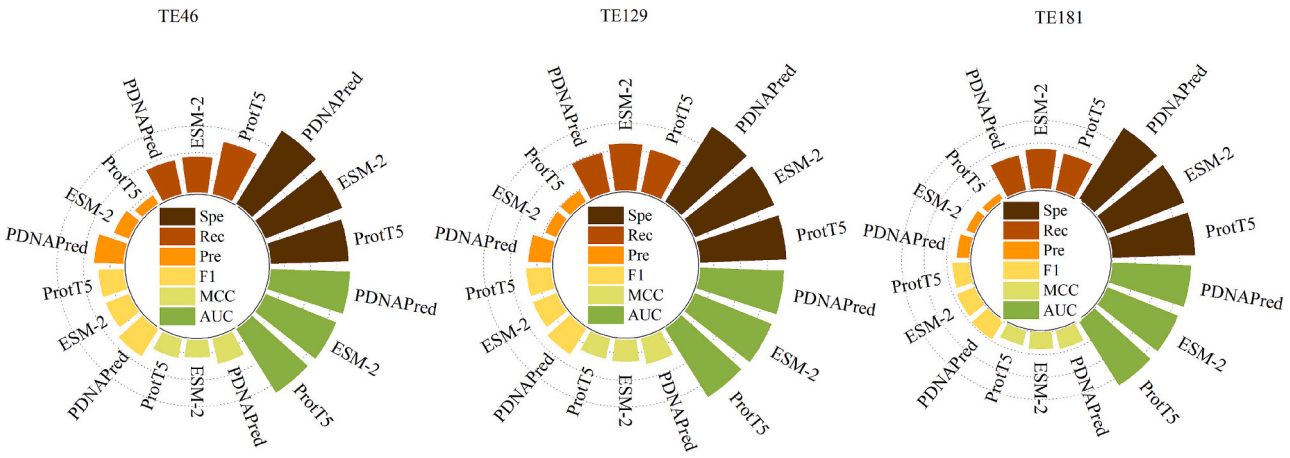



**Fig. 2.** The performance values of three feature embeddings on three test sets.

$$\widetilde{ht} = (W \bullet [rt^{*}h(t-1), xt]) \tag{3}$$

$$ht = (1-zt)^{*}h(t-1) + zt^{*}\widetilde{ht} \tag{4}$$

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{5}$$

where *rt* and zt represent the gate control states, with *rt* determining the integration of new input information with previous memory, and zt dictates the retention level of the previous memory for the current time step. The output value ht is obtained by combining the previous hidden state $h(t-1)$ with the current node input, involving the forgetting of certain information and the incorporation of new information. The output of the BiGRU layer has dimensions of 256 for each residue, obtained by concatenating the hidden states from all GRU cells across all time steps.

The output of the BiGRU layer is then passed to a fully connected layer, which consists of two linear layers followed by a Rectified Linear Unit (ReLU) activation function. This configuration reduces the feature dimension from 256 to 1. Finally, the output layer, employing a sigmoid function, produces a confidence score for each residue. A threshold of 0.5 is set, such that residues with an output confidence greater than 0.5 are considered DNA-binding sites, while those with a confidence lower than 0.5 are classified as non-DNA-binding sites.

### 2.4. Loss function

The protein-DNA binding site datasets are imbalanced dataset with fewer positive samples (binding sites) than negative samples (non-binding sites). Consequently, the deep learning model may assign greater weight to the negative samples in the loss function, thus paying less attention to the positive samples. However, accurate identification of positive sample is of greater significance in protein-DNA binding site prediction tasks. To solve this problem, we adopt focal loss [51] as the loss function of the deep learning model. It is an effective strategy to mitigate the negative impact of sample imbalance, which will assign greater weight to positive samples, thus improving their identification. The formula for the focal loss function is as follows:

$$\text{focal loss}\,(p_t) = -\alpha_t(1-p_t)^{\gamma} log(p_t) \tag{6}$$

where $p_t$ is the predicted probability of the positive class, $\alpha_t$ is a weighting factor that assigns different weights to positive and negative, and $\gamma$ controls the focusing parameter, which determines how much emphasis is placed on difficult negatives. In this study, the $\gamma$ is set to 2, while the $\alpha_t$ ranges from 0.2 to 0.8. The final values of $\alpha_t$ are determined using grid search across different training sets.

### 2.5. Evaluation metrics

In this study, the benchmark dataset utilized demonstrates an imbalance between positive and negative samples. To comprehensively evaluate the effectiveness of the proposed method, we employed six metrics commonly applied in imbalanced classification tasks: accuracy (ACC), specificity (Spe), recall (Rec), precision (Pre), F1-score (F1), and Mathews Correlation Coefficient (MCC). These metrics are calculated using the following formulas:

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \tag{7}$$

$$Spe = \frac{TN}{TN+FP} \tag{8}$$

$$Rec = \frac{TP}{TP+FN} \tag{9}$$

$$Pre = \frac{TP}{TP+FP} \tag{10}$$

$$F1 = \frac{2TP}{2TP+FP+FN} \tag{11}$$

$$MCC = \frac{TP \times TN - FN \times FP}{\sqrt{(TP+FP)\times(TP+FN)\times(TN+FP)\times(TN+FN)}} \tag{12}$$

where TP (true positive) and TN (true negative) represent the numbers of correctly predicted binding and non-binding sites, respectively; FP (false positive) and FN (false negative) denote the numbers of incorrectly predicted binding and non-binding sites. The ACC metric evaluates the overall correct prediction, capturing both TP and TN classifications. Additionally, the recall and precision metrics measure the predictive capability of a classifier for identifying binding sites, while the F1-score provides an assessment of the overall performance of a classifier. Moreover, MCC evaluates the predictive ability of both positive and negative classes of the model and is particularly suitable for imbalanced datasets. Additionally, we have incorporated two additional evaluation metrics: the receiver operation characteristic (ROC) curve and the precision-recall (PR) curve, and computed the area under the ROC curve (AUC) and the area under the PR curve (AUPRC) to evaluate the overall predictive performance. Higher values indicate better and more robust performance. In this study, due to dataset imbalance, our priority lies in comparing MCC values and AUC values with other existing methods.

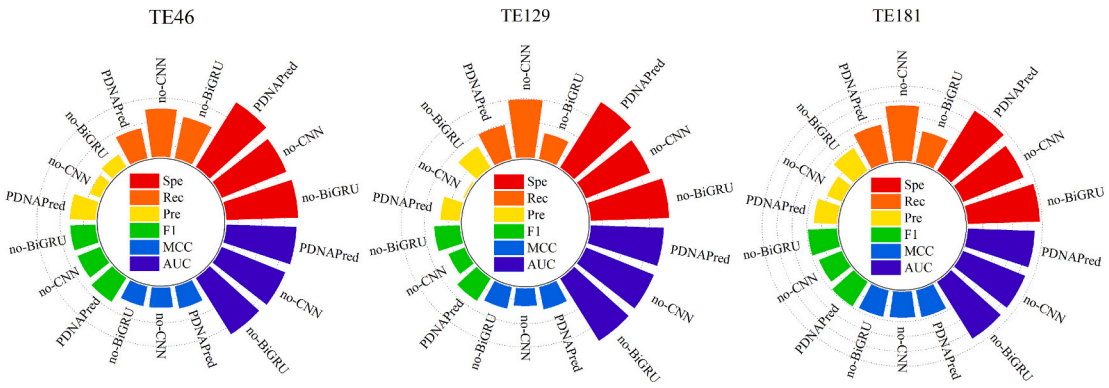



**Fig. 3.** The performance values of three models on two training sets and three test sets.

## 3. Results and discussion

### 3.1. Contribution analysis of different protein language models

To assess the contribution of two PPLMs, ESM-2 and ProtT5, we conducted comparative experiments on the benchmark datasets. The features comprised individual embeddings extracted from ESM-2 and ProtT5 models, along with a hybrid feature created by combining the two embeddings, denoted as ESM-2 + ProtT5. Fig. 2 presents the performance comparison of three embeddings across three test sets (TE46, TE129, and TE181) for independent validation, while Table S1 presents the detailed results. Additionally, we utilized the t-Distributed Stochastic Neighbor Embedding (t-SNE) program [52] to visualize these features, as shown in Fig. S2.

The experimental results show that effective complementarity can be achieved between ProtT5 and ESM-2. On the TE46 test set, ProtT5 is better at identifying positive samples, while ESM-2 has an advantage in identifying negative samples. Specifically, the specificity of ProtT5 is 0.885, and the specificity of ESM-2 is 0.923. However, at this time, the recall of ProtT5 is 0.701, and the recall of ESM-2 is only 0.570. In addition, similar conclusions can also be drawn on TE129. On TE181, ProtT5 and ESM-2 exhibit similar performance in identifying protein DNA binding and non-binding sites. MCC can comprehensively evaluate the performance of the model when the positive and negative samples are imbalanced. It can be found that when using fused features, the MCC value of the model further increases. Taking TE181 as an example, the MCC of the model using ProtT5 is 0.331, the MCC of the model using ESM-2 is 0.350, and the MCC using fused features reaches 0.364, achieving a performance improvement of 1.4 % -3.3 %. In addition, AUC is an indicator that is not affected by classification thresholds, and it can also be observed that models using fused features have higher AUC values than those using a single feature. These findings suggest that the two language models, pre-trained on different sequence databases, complement each other in improving the performance of protein-DNA binding site prediction. In addition, it can also be seen from Fig. S2 that ESM-2 + ProtT5 achieves better sample aggregation than ESM-2 or ProtT5 alone.

### 3.2. Exploration of the optimal architecture of the PDNAPred

To comprehensively assess the effectiveness of the proposed PDNAPred and compare the contribution of each module, we conducted a series of experiments on two benchmark datasets. We deleted the CNN layer or BiGRU layer from the original model to generate no-CNN model and no-BiGRU model, subsequently comparing their performance with the original model. The results, presented in Fig. 3, indicated that the original model achieved the best performance. Specifically, as seen in Figure3 and Table S2, on the TE46, PDNAPred achieved a precision of 0.519, an F1-score of 0.539, an MCC of 0.564 and an AUC of 0.897. Compared to the no-BiGRU model, PDNAPred showed improvements of 9.6 % in precision, 2.4 % in F1-score, 2.2 % in MCC and 3 % in AUC. Additionally, it outperformed the no-CNN model by 12.3 % in precision, 3.1 % in F1-score, 2.3 % in MCC and 1.3 % in AUC.

Investigating the individual contributions of the CNN and BiGRU model revealed their essential roles in identifying protein-DNA binding residues. Omitting either component led to a decline in performance metrics such as specificity, recall, precision, and MCC. Specifically, the no-BiGRU module shows increased specificity, implying that the CNN module plays a vital role in accurately identifying negative instances. Conversely, the no-CNN model demonstrates higher recall and F1 scores, suggesting that the BiGRU model significantly contributes to the accurate identification of positive cases. The integration of both CNN and BiGRU modules synergistically complemented their respective strengths, achieving a harmonious balance between specificity and recall, thereby yielding superior performance across diverse datasets.

### 3.3. Impact of the loss function and thresholds in the performance of the PDNAPred

In this section, to further explore the effectiveness of the focal loss employed by PDNAPred, we conduct a delicate experiment where we replaced the focal loss in the model with the Binary Cross Entropy loss (BCE loss). The experimental results, shown in Table S3, revealed an interesting phenomenon: the model with BEC loss exhibited a higher specificity but lower recall. This suggests that while the model effectively identified negative samples, its ability to predict positive samples was weak, leading to a suboptimal overall MCC value. For the task of DNA-binding site identification, accurately identifying positive samples is crucial. Our evaluation of the two different loss functions underscores the superiority of focal loss. Specifically, on the TR646, the model with focal loss achieved an MCC value of 0.564, surpassing the model with BCE loss that achieved an MCC value of 0.416. Similarly, on the TR573 dataset, the model with focal loss achieved an MCC value of 0.557, surpassing the model with BCE loss that achieved an MCC value of 0.443. These findings highlight the efficacy of focal loss in mitigating the imbalance between positive and negative samples, resulting in more robust model performance across all three test sets.

In our classification network, a sigmoid activation function was employed alongside a threshold set at 0.5. Instances surpassing this confidence level are classified as protein-DNA binding sites, while those below are classified as non-binding sites. It is widely acknowledged that threshold selection influences model predictions. Lower thresholds enhance correct positive sample predictions, whereas higher thresholds improve correct negative sample predictions. To explore this, we systematically adjusted the threshold from 0 to 1 in increments of 0.1, with the experimental outcomes detailed in Table S4. Analysis of these results reveals that threshold variations minimally impact the predictive performance of PDNAPred, underscoring its robustness as a method.

**Table 2**
Performance comparison between PDNAPred and other classical deep learning algorithms on the test sets.

| Dataset | Models | Spe | Rec | Pre | F1 | MCC | AUC |
|---|---|---|---|---|---|---|---|
| TE46 | CNN | 0.912 | 0.659 | 0.423 | 0.515 | 0.471 | 0.867 |
| | BiLSTM | 0.895 | 0.710 | 0.396 | 0.508 | 0.470 | 0.884 |
| | MLP | 0.852 | 0.776 | 0.339 | 0.471 | 0.444 | 0.892 |
| | PDNAPred | 0.949 | 0.561 | 0.519 | 0.539 | 0.493 | 0.897 |
| TE181 | CNN | 0.924 | 0.603 | 0.261 | 0.364 | 0.357 | 0.863 |
| | BiLSTM | 0.875 | 0.748 | 0.211 | 0.329 | 0.351 | 0.901 |
| | MLP | 0.851 | 0.780 | 0.189 | 0.304 | 0.334 | 0.897 |
| | PDNAPred | 0.949 | 0.512 | 0.309 | 0.386 | 0.364 | 0.896 |
| TE129 | CNN | 0.935 | 0.678 | 0.397 | 0.500 | 0.479 | 0.897 |
| | BiLSTM | 0.890 | 0.803 | 0.316 | 0.454 | 0.458 | 0.927 |
| | MLP | 0.865 | 0.837 | 0.283 | 0.423 | 0.436 | 0.925 |
| | PDNAPred | 0.957 | 0.595 | 0.466 | 0.523 | 0.493 | 0.923 |

**Table 3**
Performance comparison between PDNAPred and other sequence-based predictors on the test sets.

| Dataset | Models | Spe | Rec | Pre | F1 | MCC | AUC |
|---|---|---|---|---|---|---|---|
| TE46 | DRNAPred[a] | 0.692 | 0.677 | 0.185 | 0.291 | 0.226 | 0.775 |
| | DNAPred[a] | 0.655 | 0.671 | 0.157 | 0.254 | 0.194 | 0.730 |
| | SVMnuc[a] | 0.666 | 0.668 | 0.154 | 0.250 | 0.192 | 0.715 |
| | NCBRPed[a] | 0.674 | 0.677 | 0.165 | 0.265 | 0.207 | 0.713 |
| | DBPred[a] | 0.784 | 0.708 | 0.243 | 0.362 | 0.320 | 0.794 |
| | CLAPE-DB[a] | 0.835 | 0.747 | 0.306 | 0.434 | 0.401 | 0.871 |
| | ULDNA[b] | 0.696 | 0.800 | 0.204 | 0.325 | 0.296 | 0.831 |
| | PDNAPred | 0.949 | 0.561 | 0.519 | 0.539 | 0.493 | 0.897 |
| TE181 | DNAPred[a] | 0.948 | 0.334 | 0.223 | 0.267 | 0.233 | 0.802 |
| | SVMnuc[a] | 0.960 | 0.289 | 0.242 | 0.263 | 0.229 | 0.803 |
| | NCBRPred[a] | 0.964 | 0.259 | 0.241 | 0.250 | 0.215 | 0.771 |
| | CLAPE-DB[a] | 0.931 | 0.413 | 0.212 | 0.280 | 0.252 | 0.824 |
| | ULDNA[b] | 0.917 | 0.585 | 0.238 | 0.339 | 0.331 | 0.851 |
| | PDNAPred | 0.949 | 0.512 | 0.309 | 0.386 | 0.364 | 0.896 |
| TE129 | DRNAPred[a] | 0.937 | 0.233 | 0.190 | 0.210 | 0.155 | 0.693 |
| | DNAPred[a] | 0.954 | 0.396 | 0.353 | 0.373 | 0.332 | 0.845 |
| | SVMnuc[a] | 0.966 | 0.316 | 0.371 | 0.341 | 0.304 | 0.812 |
| | NCBRPed[a] | 0.969 | 0.312 | 0.392 | 0.347 | 0.313 | 0.823 |
| | CLAPE-DB[a] | 0.955 | 0.464 | 0.396 | 0.427 | 0.389 | 0.881 |
| | ULDNA[b] | 0.911 | 0.725 | 0.340 | 0.463 | 0.452 | 0.893 |
| | PDNAPred | 0.957 | 0.595 | 0.466 | 0.523 | 0.494 | 0.923 |

[a] Data excerpted from CLAPE-DB [31].
[b] Results computed using the standalone program of ULDNA downloaded at https://github.com/yiheng-zhu/ULDNA.

### 3.4. Performance comparison with classical deep learning algorithms

To examine whether classical deep learning methods trained on these two datasets could provide comparable performance to that of PDNAPred, we selected deep learning algorithms: CNN, bidirectional long short-term memory network (BiLSTM), and multi-layer perceptron (MLP). These models were trained on TR646 and TR573 and evaluated on their respective test sets: TE46, TE181, and TE129. Additionally, we employed 5-fold cross-validation to train these models on the training sets and set the classifier threshold at 0.5. The results are summarized in Table 2, and the ROC and PR curves on the independent test set are illustrated in Fig. S3. It is evident that the CNN network achieved the highest MCC values of 0.471 on TE46, 0.357 on TE181, and 0.479 on TE129 among the three algorithms compared. Although BiLSTM is tailored for sequence modeling tasks, our findings suggest that CNN outperformed it in predicting DNA-binding sites. This observation might stem from the fact that DNA-binding residues are primarily influenced by spatial structures rather than simple sequential order [31]. CNN models protein sequences using sliding windows, thereby inherently integrating relative positional information of amino acids. Conversely, RNN models treat amino acids as independent tokens [53].

### 3.5. Performance of the hybrid test set on dataset 2

In this section, we amalgamated TE181 and TE129 to establish a novel, autonomous test set for validating model performance. TE181 curated experimentally validated protein-DNA binding site data from the BioLiP database spanning January 6, 2016, to December 5, 2018. Complementarily, TE129 serves as an extension to TE181, encompassing data released from December 6, 2018, to August 19, 2021, and excluding sequences with over 30 % similarity to TE181. The amalgamation of TE129 and TE181 represents a comprehensive compilation of datasets from the BioLiP database spanning January 6, 2016, to August 19, 2021. By merging these two test sets and assessing them with the TR573-trained model, the outcomes are detailed in Table S5. The amalgamated test set exhibited a specificity of 0.951, a recall of 0.547, a precision of 0.361, an F1-score of 0.435, an MCC of 0.401, and an AUC value of 0.908. Overall, the performance of the combined TE129 and TE181 test set falls between that of TE129 and TE181 individually, thereby offering supplementary insights for comparing the fusion's performance with the two original test sets.

### 3.6. Comparison with existing sequence-based protein-DNA binding site predictors

In this section, we conducted a comprehensive evaluation of PDNAPred in comparison with other sequence-based protein-DNA binding site predictors, including DRNAPred [27], DNAPred [28], SVMnuc [24], NCBRPed [29], DBPred [30], CLAPE-DB [31], ULDNA [37]. Table 3 illustrates the performance of PDNAPred across benchmark datasets. Notably, on the TE46, PDNAPred demonstrated promising results, boasting a specificity of 0.949, a precision of 0.519, an F1-score of 0.539, and an MCC of 0.493. Furthermore, the robust AUC value of 0.897 from the test dataset further validated the effectiveness and robustness of PDNAPred. Compared to other predictors, PDNAPred surpassed CLAPE-DB and DBPred, enhancing the AUC indicator by 0.092 and 0.173, respectively.

In the case of the TE81, PDNAPred achieved a specificity of 0.949, recall of 0.512, precision of 0.309, F1 of 0.386, MCC of 0.364, and AUC of 0.896. When compared with deep learning-based methods such as PDNAPred, NCBPRed, CLAPE-DB, and ULDNA, alongside machine learning-based methods like SVMnuc, PDNAPred demonstrates improvements in MCC values of 0.131, 0.149, 0.112, 0.033, and 0.135, respectively. Furthermore, it exhibited substantial increases in the F1-score, with enhancements of 0.136 and 0.123 compared to NCBPred and SVMnuc, respectively.

For the TE129, PDNAPred exhibited remarkable performance across various metrics: recall at 0.595, precision at 0.466, F1-score at 0.523, MCC at 0.493, and AUC at 0.923. In comparison to DNAPred, NCBRPed, CLAPE-DB, and ULDNA, PDNAPred exhibits improvements in MCC indicators of 0.161, 0.189, and 0.104, as well as in recall indicators of 0.113, 0.54, 0.07, and 0.042, respectively.

CLAPE-DB and ULDNA notably adopted the feature encoding strategy akin to PDNAPred, utilizing a PPLM. This underscored the effectiveness of this methodology over manually crafted features, substantiated by their superior performance. Moreover, the utilization of deep learning models has been demonstrated to elevate model performance in contrast to machine learning-based approaches. It is noteworthy that while ULDNA has demonstrated commendable performance among existing methodologies, its reliance on a pre-trained protein language model grounded in multi-sequence comparison results in a relatively lower efficiency compared to PDNAPred.

### 3.7. Comparison with existing structure-based protein-DNA binding site predictors

In this section, we comprehensively evaluated PDNAPred alongside other structure-based protein-DNA binding site predictors, including

**Table 4**
Performance comparison between PDNAPred and other structure-based predictors on the test sets.

| Dataset | Models | Spe | Rec | Pre | F1 | MCC | AUC |
|---|---|---|---|---|---|---|---|
| TE181 | COACH-D[c] | 0.971 | 0.254 | 0.280 | 0.266 | 0.235 | 0.655 |
| | NucBind[c] | 0.960 | 0.293 | 0.248 | 0.269 | 0.234 | 0.796 |
| | DNABind[c] | 0.904 | 0.535 | 0.199 | 0.290 | 0.279 | 0.825 |
| | GraphBind[c] | 0.933 | 0.624 | 0.293 | 0.399 | 0.392 | 0.904 |
| | GraphSite[a,c] | 0.958 | 0.517 | 0.354 | 0.420 | 0.397 | 0.917 |
| | GLMSite[d] | 0.805 | 0.829 | 0.209 | 0.311 | 0.334 | 0.899 |
| | EquiPNAS[d] | 0.958 | 0.436 | 0.346 | 0.366 | 0.353 | 0.907 |
| | EGPDI[d] | 0.952 | 0.558 | 0.346 | 0.424 | 0.407 | 0.914 |
| | PDNAPred | 0.949 | 0.512 | 0.309 | 0.386 | 0.364 | 0.896 |
| TE129 | COACH-D[c] | 0.958 | 0.367 | 0.357 | 0.362 | 0.321 | 0.710 |
| | NucBind[c] | 0.966 | 0.330 | 0.381 | 0.354 | 0.317 | 0.811 |
| | DNABind[c] | 0.926 | 0.601 | 0.346 | 0.440 | 0.411 | 0.858 |
| | GraphBind[b, c] | – | 0.439 | 0.310 | 0.362 | 0.320 | 0.816 |
| | GraphBind[a, c] | 0.941 | 0.676 | 0.425 | 0.522 | 0.499 | 0.927 |
| | GLMSite[d] | 0.816 | 0.848 | 0.287 | 0.405 | 0.412 | 0.918 |
| | EquiPNAS[d] | 0.956 | 0.516 | 0.471 | 0.462 | 0.443 | 0.919 |
| | EGPDI[d] | 0.961 | 0.612 | 0.503 | 0.549 | 0.522 | 0.941 |
| | PDNAPred | 0.957 | 0.595 | 0.466 | 0.523 | 0.493 | 0.923 |

[a] Indicates the GraphBind using experimental protein structures.
[b] Indicates the GraphBind using predicted protein structures.
[c] Data excerpted from CLAPE-DB [31].
[d] Data excerpted from EGPDI [56].

**Table 5**
Summary of benchmark protein-RNA binding datasets.

| Datasets | Dataset 3 | | Dataset 4 | |
|---|---|---|---|---|
| | TR545 | TE161 | TR495 | TE117 |
| Number of proteins | 545 | 161 | 495 | 117 |
| Number of sites | 190,438 | 51,315 | 136,899 | 37,345 |
| Number of binding sites | 18,559 | 6966 | 14,609 | 2031 |
| Number of non-binding sites | 171,879 | 44,349 | 122,290 | 35,314 |
| % of binding sites | 9.75 | 13.58 | 10.76 | 5.44 |

COACH-D [23], NucBind [24], DNABind [22], GraphBind [21], GraphSite [17], GLMSite [54], EquiPNAS [55], and EGPDI [56]. The results in Table 4 demonstrated that PDNAPred achieved comparable or even superior performance to several state-of-the-art structure-based predictors.

Specifically, on the TE181, when compared with CPACH-D, NucBind, DNABind, GLMSite, and EquiPNAS, our proposed PDNAPred exhibited improvements of 0.129, 0.13, 0.085, 0.03, and 0.011 in MCC. For the TE129, in comparison to COACH-D, NucBind, DNABind, GLMSite, and EquiPNAS showed improvements in MCC indicators of 0.172, 0.176, 0.082, 0.081, and 0.05 as well as in AUC indicators of 0.213, 0.112, 0.065, 0.005, and 0.004, respectively. Based on these results, it is evident that PDNAPred, despite not incorporating any structural information, outperformed several structure-based models. Notably, the GraphBind model, which utilized predicted protein structure, exhibited poor performance with an AUC of 0.816, lower than that of PDNAPred. This result suggests that structure-based models rely heavily on accurate protein structure information to attain acceptable prediction results. In addition, compared with the latest proposed EGPDI based on multi view feature fusion, PDNAPred still has some room for improvement.

### 3.8. Comparison with general RNA binding sites predictors

To further validate the performance of PDNAPred in identifying RNA-binding sites, we collected two mainstream benchmark datasets of protein-RNA binding sites and trained PDNAPred on them. Specifically, we denoted these two datasets as Dataset 3 and Dataset 4. Table 5 and Text S1 provides a comprehensive description of the datasets. To fully evaluate the generalization ability of PDNAPred, we compared it with existing state-of-the-art RNA-binding site predictors, including

**Table 6**
Comparison of the proposed PDNAPred and other methods on the RNA117 test set.

| Models | Rec | Pre | F1 | MCC | AUC |
|---|---|---|---|---|---|
| RNABindPlus[c] | 0.273 | 0.227 | 0.248 | 0.202 | 0.717 |
| SVMnuc[c] | 0.231 | 0.240 | 0.235 | 0.192 | 0.729 |
| CLAPE-RB[c] | 0.467 | 0.201 | 0.281 | 0.240 | 0.800 |
| COACH-D*[,c] | 0.221 | 0.252 | 0.235 | 0.195 | 0.663 |
| NucBind*[,c] | 0.231 | 0.235 | 0.233 | 0.189 | 0.715 |
| aaRNA*[,c] | 0.484 | 0.166 | 0.237 | 0.214 | 0.771 |
| NucleicNet*[,c] | 0.371 | 0.201 | 0.261 | 0.216 | 0.788 |
| GraphBind*[,a, c] | 0.303 | 0.171 | 0.218 | 0.168 | 0.718 |
| GraphBind*[,b, c] | 0.463 | 0.294 | 0.358 | 0.322 | 0.854 |
| PDNAPred | 0.335 | 0.298 | 0.315 | 0.274 | 0.829 |

Note: *Indicates structure-based models.
[a] Indicates results using predicted protein structures.
[b] Indicates results using experimental protein structures.
[c] Data excerpted from CLAPE-DB [31].

RNABindPlus [57], SVMnuc [24], CLAPE-RB [31], COACH-D [23], NucBind [24], aaRNA [58], NucleicNet [59], and GraphBind [21] for Dataset 3. PDNAPred demonstrated good generalization performance and excelled in RNA-binding site prediction tasks, as shown in Table 6. Notably, PDNAPred outperformd GraphBind based on incorrectly predicted protein structures, highlighting its potential to overcome the limitations of structure-based models.

For Dataset 4, we compared PDNAPred with DRNAPred [27], NCBRPred [29], RNABindR-Plus [57], NucBind [24], iDRNA-ITF [60], and MucLiPred [61]. The results in Table S6 indicated that PDNAPred outperformed all other methods except MucLiPred, a newly proposed protein and molecular binding disability predictor equipped with a dual contrastive learning mechanism. PDNAPred achieves higher accuracy while maintaining the same AUC as MucLiPred.

Our results demonstrate that PDNAPred is a versatile framework for predicting nucleic acid binding sites for various protein sequences and nucleic acids. Furthermore, our experimental results demonstrate that PDNAPRed achieved high performance even without structural information.Fig. 4 illustrate the ROC and PR curves of PDNAPred on the RNA binding site prediction task.

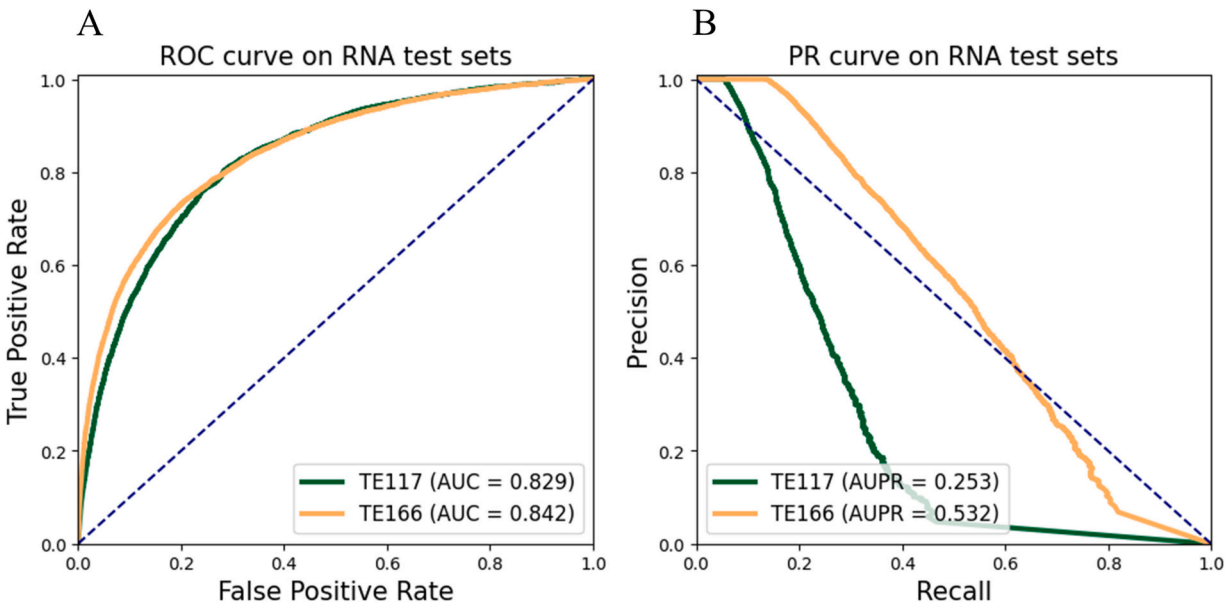



**Fig. 4.** PDNAPred's general ability to predict nucleic acid binding sites. (A- B) The ROC and PR curves of PDNAPred models in predicting RNA-binding sites.

**Table 7**
The modeling results of two DNA-binding site prediction methods on two representative examples.

| Method | 1GCC_A | | | | 1KQQ_A | | | |
|---|---|---|---|---|---|---|---|---|
| | TP | FP | TN | FN | TP | FP | TN | FN |
| CLAPE | 8 | 6 | 47 | 2 | 7 | 11 | 114 | 7 |
| PDNAPred | 13 | 1 | 46 | 3 | 7 | 11 | 119 | 2 |

### 3.9. Case study

To visually compare the prediction performance of DNA-binding residues in PDNAPred, we selected two experimentally validated proteins that were not included in datasets: 1GCC_A (PBD ID: 1GCC, chain A, denoted as 1GCC_A) [62] and 1KQQ_A (PBD ID: 1KQQ, chain A, denoted as 1KQQ_A) [63]. We compared PDNAPred with CLAPE, which is currently considered the leading algorithm for sequence-based DNA binding residue prediction. Table 7 presents the confusion matrices of the two algorithms tested on the selected samples, the prediction results are shown in Table S7, and the corresponding visualization results are displayed in Fig. 5. The results indicated that these sequence-based

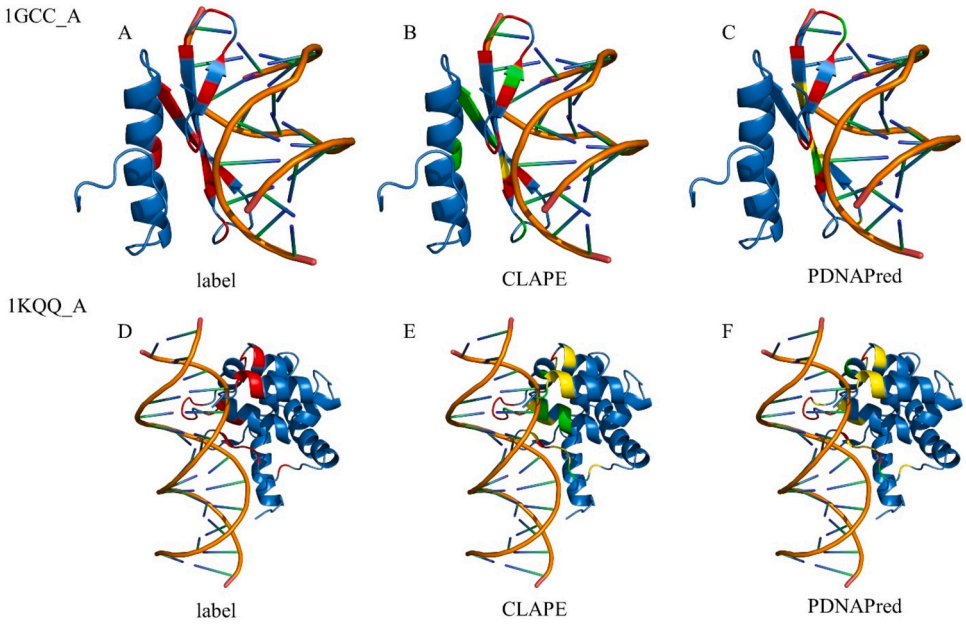



**Fig. 5.** Comparative and empirical case studies. (A-C) Analysis of the DNA-binding sites for protein 1GCC_A, where (A) represents the experimental results, (B) and (C) represent the results predicted by CLAPE and PDNAPred. (D—F) Analysis of the DNA-binding sites for protein 1KQQ_A, where (D) represents the experimental results, (E) and (F) represent the results predicted by CLAPE and PDNAPred. The atomic-level native structure of each protein is downloaded from the PDB database and then plotted as a cartoon picture using PyMOL software [64]. The colour scheme is used as follows: the DNA is orange, the DNA-binding site is red, the non-DNA-binding site is sky blue, the false negative is green, and the true negative is yellow.

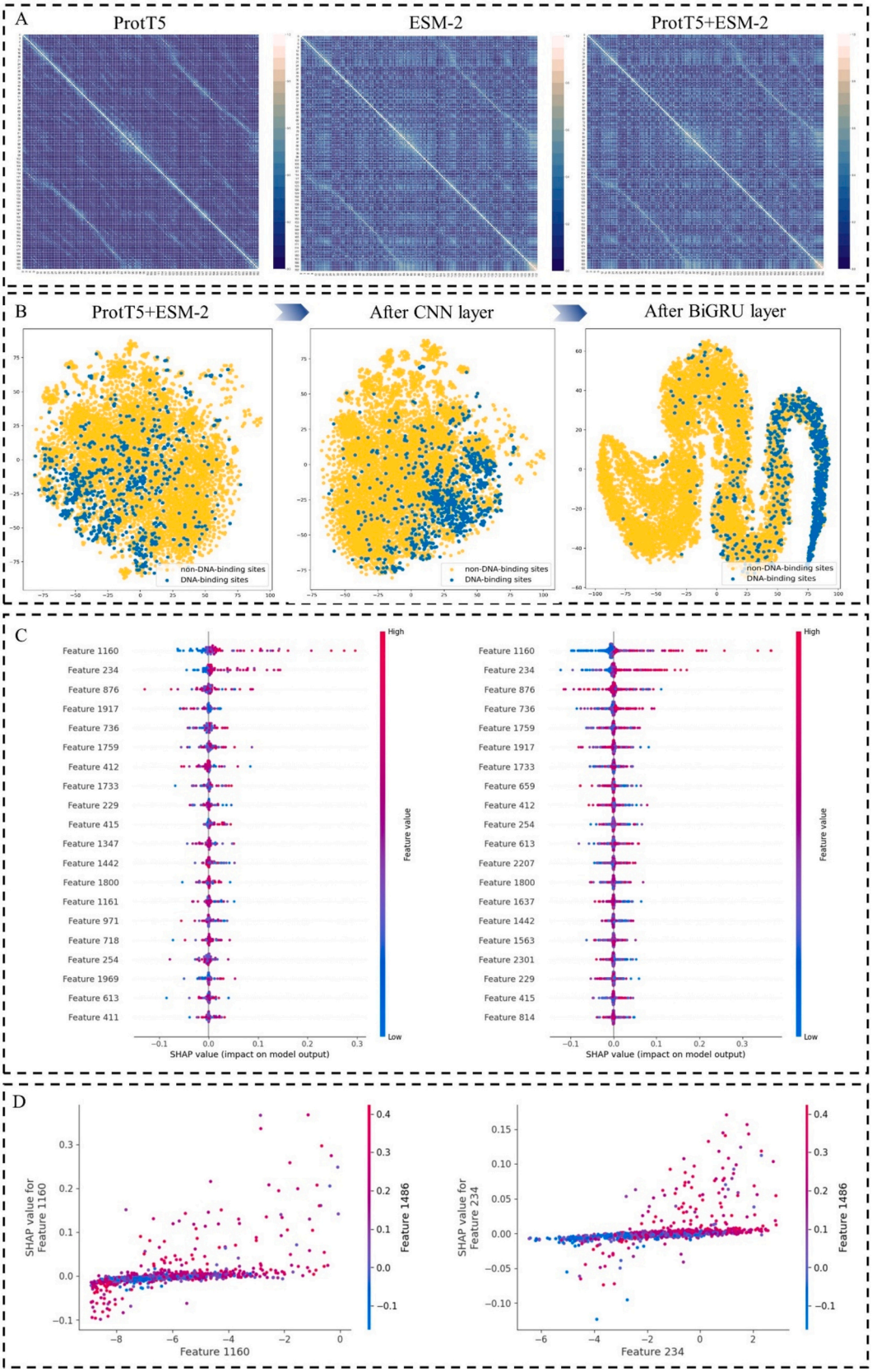



**Fig. 6.** Presentation of the interpretability experiment results. A. Correlation heat map of each residue under ProtT5 embeddings, ESM-2 embeddings, and ProtT5 + ESM-2 embeddings; B. The t-SNE results show the feature variation results of the output of different intermediate layers in the PDNAPred architecture; C. The top 20 features with the greatest impact on predicting DNA-binding sites and non-DNA-binding sites; D. The effect of interactions between Feature 1160 and Feature 234 with other features.

methods exhibited satisfactory performance, with PDNAPRed outperforming CLAPE in these two samples. In the 1GCC_A sample, CLAPE correctly identified 8 DNA binding residues, while PDNAPred correctly identified 13 DNA binding sited. In the 1KQQ_A sample, both CLAPE and PDNAPred accurately identified 7 DNA binding residues. However, PDNAPred showed a lower overall misclassification rate, leading to better performance.

Additionally, we observed interesting occurrences. For example, in the 1GCC_A sample, CLAPE correctly identified 10P as a non-DNA-binding residue, while PDNAPred misclassified it. In contrast, PDNAPred correctly identified 8Q as a non-DNA-binding residue, while CLAPE made an incorrect prediction. Therefore, for practical applications, we suggest considering a combination of different methods. Overall, PDNAPred demonstrates excellent proficiency in identifying DNA binding residues compared to existing tools.

### 3.10. Model interpretability analysis

Model interpretability has become increasingly important in the development of models [65]. It aims to clarify how features contribute to model predictions and offer insights into the model's effectiveness. In this study, we employ two commonly used model interpretability methods, Shapley Additive exPlanations (SHAP) [66] and t-SNE [52], to analyze the PDNAPred model. Using the TE46 dataset, we presented the explanatory results in Fig. 6.

Firstly, we extracted embedded features from the first protein sample (PDB ID:4JBM) in the TE 46 using the ProtT5 and ESM-2 models, respectively. These features were also concatenated to form a hybrid feature. Pearson correlation coefficient was employed to describe the correlation between amino acid residues in the embedding features of each amino acid, as illustrated in Fig. 6A. It was evident from the figure that each amino acid residue exhibited a strong correlation with its adjacent amino acids. Interestingly, the ProtT5 and ESM-2 captured specific relationships between amino acids even as the distance increased. This indicated that PPLMs can mitigate the decline in model performance resulting from missing structural information in protein sequences. Furthermore, combining these two features enhanced the association between amino acids.

Second, we utilized the t-SNE program for dimensionality reduction and visualization of the features, as shown in Fig. 6B. The results revealed that the combined embedding features contribute to the prediction of DNA binding sites. After feature embedding, the positive samples appeared clustered in a narrow range, albeit exhibiting a generally random distribution. Subsequently, the CNN layer preserved crucial classification features, causing positive samples to cluster over a broader range. Following the BiGRU layer, the aggregation of positive samples became more pronounced. This outcome underscored the utility of our designed CNN-GRU network in identifying DNA binding sites. Finally, we investigated the influence of various features on protein binding site prediction by employing SHAP for explanation. Fig. 6C showcased the 20 features that exerted the greatest impact on DNA binding site prediction. Here, red denotes higher eigenvalues, while blue indicates lower eigenvalues. It is noteworthy that Feature 0 to Feature 1023 represent the embedded features of the ProtT5 model, whereas Feature 1024 to Feature 2303 represent those of the ESM-2 model. Our findings demonstrated that both ProtT5 and ESM-2 collaboratively enhanced the accuracy of DNA binding site prediction. To illustrate, we examined the influence of the two features with the greatest impact, namely Feature 1160 and Feature 234, on PDNAPred's prediction outcomes. As depicted in Fig. 6D, both features exhibited the strongest interaction with Feature 1486, and these features can promote each other. This underscores the global context dependence inherent in the ProtT5 and ESM-2 models, as their feature representations are based on joint action with different features, thereby further validating the efficacy of the ProtT5 and ESM-2 models in DNA binding sites identificatory tasks [67].

## 4. Conclusion

Predicting protein-DNA binding residues holds significant importance in biomedical research, offering insights into biological functions, disease mechanisms, and drug discovery in biomedical research. In this study, we introduce PDNAPred, a novel deep learning method designed for accurately identifying DNA binding residues within protein sequences. Leveraging pre-trained models ESM-2 and ProtT5, PDNAPred autonomously generates enhanced representations of protein sequences. Our crafted CNN-GRU network effectively captures subtle features associated with DNA binding residues, resulting in improved identification accuracy. To address data imbalance, we employ the focal loss as the loss function during training.

Through comprehensive comparisons with existing advanced deep learning models on widely used benchmark datasets, PDNAPred demonstrates outstanding performance. It surpasses sequence-based methods and achieves commensurable results to structure-based methods. Furthermore, PDNAPred exhibits commendable performance on RNA binding site datasets, showcasing its versatility in identifying amino acid binding residues in proteins. The interpretability analysis provides insight into PDNAPred's internal workings, enhancing its credibility. With superior computational efficiency and speed, PDNAPred is suitable for large-scale binding site prediction tasks. The standalone PDNAPred package is freely available for academic use at the provided link: https://github.com/zlr-zmm/PDNAPred. In terms of method efficiency, the prediction time for 10 protein sequences with the length of 1000 is under 1 min when utilizing the Nvidia RTX 3090 GPU, rendering it suitable for large-scale predictions.

While PDNAPred has exhibited satisfactory results, it harbors untapped potential for further enhancement. Structural approaches excel in discriminating conformational variations and considering biological assembly capabilities, showcasing notable advantages, especially in intricate structure predictions [68]. These are advantages that sequence-based methods cannot achieve. In future endeavors, we intend to address the limitations inherent in sequence-centric methodologies, particularly concerning the complexities of protein structures. Furthermore, our subsequent research will focus on refining PDNAPred to become a robust predictor of protein-nucleic acid binding sites. Drawing from esteemed prior studies [29,69,70], cross-validation has garnered considerable attention, highlighting its significance in assessing the reliability of predictive models. Our aspiration is to enhance the efficiency of our methodology for predicting protein-ligand binding sites, thereby aiding the research community in their endeavors.

## CRediT authorship contribution statement

**Lingrong Zhang:** Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Taigang Liu:** Writing – review & editing, Supervision, Funding acquisition.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

The data and code for this study can be found in a GitHub repository accompanying this manuscript: https://github.com/zlr-zmm/PDNAPred.

## Acknowledgments

This research was funded by the National Natural Science Foundation of China (grant number 11601324). We would like to thank the reviewers for their professional review work, constructive comments,

and valuable suggestions on our manuscript.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijbiomac.2024.136147.

## References


[1] G.D. Stormo, Y. Zhao, Determining the specificity of protein–DNA interactions, Nat. Rev. Genet. 11 (11) (2010) 751–760.
[2] L.A. Gallagher, E. Velazquez, S.B. Peterson, J.C. Charity, M.C. Radey, M. J. Gebhardt, F. Hsu, L.M. Shull, K.J. Cutler, K. Macareno, Genome-wide protein–DNA interaction site mapping in bacteria using a double-stranded DNA-specific cytosine deaminase, Nat. Microbiol. 7 (6) (2022) 844–855.
[3] H. Zhao, Y. Yang, Y. Zhou, Structure-based prediction of DNA-binding proteins by structural alignment and a volume-fraction corrected DFIRE-based energy function, Bioinformatics 26 (15) (2010) 1857–1863.
[4] N. Bhardwaj, H. Lu, Residue-level prediction of DNA-binding sites and its application on DNA-binding protein predictions, FEBS Lett. 581 (5) (2007) 1058–1066.
[5] J. Konc, M. Hodošček, M. Ogrizek, J. Trykowska Konc, D. Janežič, Structure-based function prediction of uncharacterized protein using binding sites comparison, PLoS Comput. Biol. 9 (11) (2013) e1003341.
[6] K. Ponnuraj, K.M. Saravanan, Dihedral angle preferences of DNA and RNA binding amino acid residues in proteins, Int. J. Biol. Macromol. 97 (2017) 434–439.
[7] R. Kumar, M.A. Corbett, B.W. Van Bon, J.A. Woenig, L. Weir, E. Douglas, K. L. Friend, A. Gardner, M. Shaw, L.A. Jolly, THOC2 mutations implicate mRNA-export pathway in X-linked intellectual disability, Am. J. Hum. Genet. 97 (2) (2015) 302–310.
[8] S. Wang, K. Liang, Q. Hu, P. Li, J. Song, Y. Yang, J. Yao, L.S. Mangala, C. Li, W. Yang, JAK2-binding long noncoding RNA promotes breast cancer brain metastasis, J. Clin. Invest. 127 (12) (2017) 4498–4515.
[9] R. Esmaeeli, A. Bauzá, A. Perez, Structural predictions of protein–DNA binding: MELD-DNA, Nucleic Acids Res. 51 (4) (2023) 1625–1636.
[10] E. Kim, Y.-J. Kim, Z. Ji, J.M. Kang, M. Wirianto, K.R. Paudel, J.A. Smith, K. Ono, J.-A. Kim, K. Eckel-Mahan, ROR activation by Nobiletin enhances antitumor efficacy via suppression of IκB/NF-κB signaling in triple-negative breast cancer, Cell Death Dis. 13 (4) (2022) 374.
[11] J.D. Nelson, O. Denisenko, K. Bomsztyk, Protocol for the fast chromatin immunoprecipitation (ChIP) method, Nat. Protoc. 1 (1) (2006) 179–185.
[12] M. Smyth, J. Martin, X ray crystallography, Mol. Pathol. 53 (1) (2000) 8.
[13] M.A. Heffler, R.D. Walters, J.F. Kugel, Using electrophoretic mobility shift assays to measure equilibrium dissociation constants: GAL4-p53 binding DNA as a model system, Biochem. Mol. Biol. Educ. 40 (6) (2012) 383–387.
[14] J.M. Sagendorf, R. Mitra, J. Huang, X.S. Chen, R. Rohs, Structure-based prediction of protein-nucleic acid binding using graph neural networks, Biophys. Rev. (2024) 1–18.
[15] R. Mitra, J. Li, J.M. Sagendorf, Y. Jiang, A.S. Cohen, T.-P. Chiu, C.J. Glasscock, R. Rohs, Geometric deep learning of protein–DNA binding specificity, Nat. Methods (2024) 1–10.
[16] C.J. Glasscock, R. Pecoraro, R. McHugh, L.A. Doyle, W. Chen, O. Boivin, B. Lonnquist, E. Na, Y. Politanska, H.K. Haddox, Computational Design of Sequence-specific DNA-binding Proteins, bioRxiv, 2023.
[17] Q. Yuan, S. Chen, J. Rao, S. Zheng, H. Zhao, Y. Yang, AlphaFold2-aware protein–DNA binding site prediction using graph transformer, Brief. Bioinform. 23 (2) (2022) bbab564.
[18] K. Qu, L. Wei, Q. Zou, A review of DNA-binding proteins prediction methods, Curr. Bioinforma. 14 (3) (2019) 246–254.
[19] H. Zhang, Y. Wu, Y. Zhu, L. Ge, J. Huang, Z. Qin, Identification and functional analysis of a serine protease inhibitor using machine learning strategy, Int. J. Biol. Macromol. 265 (2024) 130852.
[20] J. Jiménez, S. Doerr, G. Martínez-Rosell, A.S. Rose, G. De Fabritiis, DeepSite: protein-binding site predictor using 3D-convolutional neural networks, Bioinformatics 33 (19) (2017) 3036–3042.
[21] Y. Xia, C.-Q. Xia, X. Pan, H.-B. Shen, GraphBind: protein structural context embedded rules learned by hierarchical graph neural networks for recognizing nucleic-acid-binding residues, Nucleic Acids Res. 49 (9) (2021) e51.
[22] R. Liu, J. Hu, DNABind: a hybrid algorithm for structure-based prediction of DNA-binding residues by combining machine learning-and template-based approaches, Proteins: Struct., Funct., Bioinf. 81 (11) (2013) 1885–1899.
[23] Q. Wu, Z. Peng, Y. Zhang, J. Yang, COACH-D: improved protein–ligand binding sites prediction with refined ligand-binding poses through molecular docking, Nucleic Acids Res. 46 (W1) (2018) W438–W442.
[24] H. Su, M. Liu, S. Sun, Z. Peng, J. Yang, Improving the prediction of protein–nucleic acids binding residues via multiple sequence profiles and the consensus of complementary methods, Bioinformatics 35 (6) (2019) 930–936.
[25] L. Jing, S. Xu, Y. Wang, Y. Zhou, T. Shen, Z. Ji, H. Fang, Z. Li, S. Sun, CrossBind: collaborative cross-modal identification of protein nucleic-acid-binding residues, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2024, pp. 2661–2669.
[26] M. Chen, S.J. Ludtke, Deep learning-based mixed-dimensional Gaussian mixture model for characterizing variability in cryo-EM, Nat. Methods 18 (8) (2021) 930–936.
[27] J. Yan, L. Kurgan, DRNApred, fast sequence-based method that accurately predicts and discriminates DNA-and RNA-binding residues, Nucleic Acids Res. 45 (10) (2017) e84.
[28] Y.-H. Zhu, J. Hu, X.-N. Song, D.-J. Yu, DNAPred: accurate identification of DNA-binding sites from protein sequence by ensembled hyperplane-distance-based support vector machines, J. Chem. Inf. Model. 59 (6) (2019) 3057–3071.
[29] J. Zhang, Q. Chen, B. Liu, NCBRPred: predicting nucleic acid binding residues in proteins based on multilabel learning, Brief. Bioinform. 22 (5) (2021) bbaa397.
[30] S. Patiyal, A. Dhall, Raghava GPS: a deep learning-based method for the prediction of DNA interacting residues in a protein, Brief. Bioinform. 23 (5) (2022).
[31] Y. Liu, B. Tian, Protein–DNA binding sites prediction based on pre-trained protein language model and contrastive learning, Brief. Bioinform. 25 (1) (2024) bbad488.
[32] Rao R, Meier J, Sercu T, Ovchinnikov S, Rives A: Transformer protein language models are unsupervised structure learners. Biorxiv 2020:2020.2012. 2015.422761.
[33] R. Rao, N. Bhattacharya, N. Thomas, Y. Duan, P. Chen, J. Canny, P. Abbeel, Y. Song, Evaluating protein transfer learning with TAPE, in: Advances in Neural Information Processing Systems, 2019.
[34] Z. Yan, F. Ge, Y. Liu, Y. Zhang, F. Li, J. Song, D.-J. Yu, TransEFVP: a two-stage approach for the prediction of human pathogenic variants based on protein sequence embedding fusion, J. Chem. Inf. Model. 64 (4) (2024) 1407–1418.
[35] J.-S. Wu, Y. Liu, F. Ge, D.-J. Yu, Prediction of protein-ATP binding residues using multi-view feature learning via contextual-based co-attention network, Comput. Biol. Med. 172 (2024) 108227.
[36] A. Elnaggar, M. Heinzinger, C. Dallago, G. Rehawi, Y. Wang, L. Jones, T. Gibbs, T. Feher, C. Angerer, M. Steinegger, Prottrans: toward understanding the language of life through self-supervised learning, IEEE Trans. Pattern Anal. Mach. Intell. 44 (10) (2021) 7112–7127.
[37] Y.-H. Zhu, Z. Liu, Y. Liu, Z. Ji, D.-J. Yu, ULDNA: integrating unsupervised multi-source language models with LSTM-attention network for high-accuracy protein–DNA binding site prediction, Brief. Bioinform. 25 (2) (2024) bbae040.
[38] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, Evolutionary-scale prediction of atomic-level protein structure with a language model, Science 379 (6637) (2023) 1123–1130.
[39] R.M. Rao, J. Liu, R. Verkuil, J. Meier, J. Canny, P. Abbeel, T. Sercu, A. Rives, MSA transformer, in: International Conference on Machine Learning, 2021. PMLR: 8844–8856.
[40] Y. Li, Y. Wei, S. Xu, Q. Tan, L. Zong, J. Wang, Y. Wang, J. Chen, L. Hong, Y. Li, AcrNET: predicting anti-CRISPR with deep learning, Bioinformatics 39 (5) (2023) btad259.
[41] Y. Fang, F. Xu, L. Wei, Y. Jiang, J. Chen, L. Wei, D.-Q. Wei, AFP-MFL: accurate identification of antifungal peptides using multi-view feature learning, Brief. Bioinform. 24 (1) (2023) bbac606.
[42] L. Fu, B. Niu, Z. Zhu, S. Wu, W. Li, CD-HIT: accelerated for clustering the next-generation sequencing data, Bioinformatics 28 (23) (2012) 3150–3152.
[43] A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C.L. Zitnick, J. Ma, Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences, Proc. Natl. Acad. Sci. 118 (15) (2021) e2016239118.
[44] B.E. Suzek, Y. Wang, H. Huang, P.B. McGarvey, C.H. Wu, C. UniProt, UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches, Bioinformatics 31 (6) (2015) 926–932.
[45] M. Steinegger, M. Mirdita, J. Soding, Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold, Nat. Methods 16 (7) (2019) 603–606.
[46] M. Manfredi, C. Savojardo, P.L. Martelli, R. Casadio, E-pRSA: Embeddings improve the prediction of residue relative solvent accessibility in protein sequence, J. Mol. Biol. 168494 (2024).
[47] Hendrycks D, Gimpel K: Gaussian Error Linear Units (GELUS). arXiv preprint 2016.
[48] Dey R, Salem FM: Gate-variants of gated recurrent unit (GRU) neural networks. In: 2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS): 2017. IEEE: 1597–1600.
[49] Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y: Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation. arXiv preprint arXiv:14061078 2014.
[50] Chung J, Gulcehre C, Cho K, Bengio Y: Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. arXiv preprint arXiv:14123555 2014.
[51] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, IEEE Trans. Pattern Anal. Mach. Intell. 42 (2) (2018) 318–327.
[52] L. Van der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (11) (2008) 2579–2605.
[53] K.K. Yang, N. Fusi, A.X. Lu, Convolutions are competitive with transformers for protein sequence pretraining, Cell Syst. 15 (3) (2024) 286–294.
[54] Y. Song, Q. Yuan, H. Zhao, Y. Yang, Accurately identifying nucleic-acid-binding sites through geometric graph learning on language model predicted structures, Brief. Bioinform. 24(6):bbad360 (2023).
[55] R. Roche, B. Moussad, M.H. Shuvo, S. Tarafder, D. Bhattacharya, EquiPNAS: improved protein–nucleic acid binding site prediction using protein-language-model-informed equivariant deep graph neural networks, Nucleic Acids Res. 52 (5) (2024) e27.
[56] M. Zheng, G. Sun, X. Li, Y. Fan, EGPDI: identifying protein–DNA binding sites based on multi-view graph embedding fusion, Brief. Bioinform. 25 (4) (2024).

[57] R.R. Walia, L.C. Xue, K. Wilkins, Y. El-Manzalawy, D. Dobbs, V. Honavar, RNABindRPlus: a predictor that combines machine learning and sequence homology-based methods to improve the reliability of predicted RNA-binding residues in proteins, PLoS One 9 (5) (2014) e97725.
[58] S. Li, K. Yamashita, K.M. Amada, D.M. Standley, Quantifying sequence and structural features of protein–RNA interactions, Nucleic Acids Res. 42 (15) (2014) 10086–10098.
[59] J.H. Lam, Y. Li, L. Zhu, R. Umarov, H. Jiang, A. Héliou, F.K. Sheong, T. Liu, Y. Long, Y. Li, A deep learning framework to predict binding preference of RNA constituents on protein surface, Nat. Commun. 10 (1) (2019) 4941.
[60] N. Wang, K. Yan, J. Zhang, Liu B: iDRNA-ITF: identifying DNA-and RNA-binding residues in proteins based on induction and transfer framework, Brief. Bioinform. 23 (4) (2022) bbac236.
[61] J. Zhang, R. Wang, L. Wei, MucLiPred: multi-level contrastive learning for predicting nucleic acid binding residues of proteins, J. Chem. Inf. Model. 64 (3) (2024) 1050–1065.
[62] M.D. Allen, K. Yamasaki, M. Ohme-Takagi, M. Tateno, M. Suzuki, A novel mode of DNA recognition by a β-sheet revealed by the solution structure of the GCC-box binding domain in complex with DNA, EMBO J. 17 (18) (1998) 5484–5496.
[63] J. Iwahara, M. Iwahara, G.W. Daughdrill, J. Ford, R.T. Clubb, The structure of the dead ringer–DNA complex reveals how AT-rich interaction domains (ARIDs) recognize DNA, EMBO J. 21 (5) (2002) 1197–1209.
[64] S. Yuan, H.S. Chan, Z. Hu, Using PyMOL as a platform for computational drug design, Wiley Interdiscip. Rev.: Comput. Mol. Sci. 7 (2) (2017) e1298.
[65] L. Lin, Y. Long, J. Liu, D. Deng, Y. Yuan, L. Liu, B. Tan, H. Qi, FRP-XGBoost: identification of ferroptosis-related proteins based on multi-view features, Int. J. Biol. Macromol. 130180 (2024).
[66] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Advances in Neural Information Processing Systems, 2017.
[67] Z. Hou, Y. Yang, Z. Ma, K.-c. Wong, X. Li, Learning the protein language of proteome-wide protein-protein binding sites via explainable ensemble deep learning, Communications Biology 6 (1) (2023) 73.
[68] J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A.J. Ballard, J. Bambrick, Accurate structure prediction of biomolecular interactions with AlphaFold 3, Nature (2024) 1–3.
[69] J. Zhang, S. Basu, L. Kurgan, HybridDBRpred: improved sequence-based prediction of DNA-binding amino acids using annotations from structured complexes and disordered proteins, Nucleic Acids Res. 52 (2) (2024) e10.
[70] F. Zhang, B. Zhao, W. Shi, M. Li, L. Kurgan, DeepDISOBind: accurate prediction of RNA-, DNA-and protein-binding intrinsically disordered residues with deep multi-task learning, Brief. Bioinform. 23 (1) (2022) bbab521.