# Accurately identifying nucleic-acid-binding sites through geometric graph learning on language model predicted structures

**Yidong Song[#], Qianmu Yuan[#], Huiying Zhao, Yuedong Yang**

Corresponding author: Yuedong Yang, School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510000, China, and Key Laboratory of Machine Intelligence and Advanced Computing of MOE, Sun Yat-sen University, Guangzhou 510000, China. Tel.: +86 020-37106046. E-mail: yangyd25@mail.sysu.edu.cn

[#]These two authors contributed equally.

Author biography:

**Yidong Song** is a Ph.D. student in the School of Computer Science and Engineering at Sun Yat-sen University. His research interests include protein function prediction, graph neural network, and protein disorder prediction.

**Qianmu Yuan** is a Ph.D. student in the School of Computer Science and Engineering at Sun Yat-sen University. His research interests include deep learning, protein function prediction, and graph neural network.

**Huiying Zhao** is an associate research fellow in the Sun Yat-sen Memorial Hospital at Sun Yat-sen University. Her research interests include protein function, pathogenic gene analysis, and RNA function prediction.

**Yuedong Yang** is a professor in the School of Computer Science and Engineering at Sun Yat-sen University. Currently he focuses on integrating HPC and AI algorithms for biomedical research.

## Abstract

The interactions between proteins and nucleic acids play crucial roles in various biological activities and the design of new drugs. How to identify the nucleic-acid-binding sites accurately remains a challenging task. Currently, the existing sequence-based methods have limited predictive performance due to only considering contextual features of the sequential neighbors, while structure-based methods are not suitable for proteins mostly without known tertiary structures. Though protein structures predicted by AlphaFold2 could be used, the extensive computing requirement of AlphaFold2 hinders its use for genome-wide applications. Based on the recent breakthrough of ESMFold for fast protein structure prediction, we propose GLMSite, an accurate predictor for identifying DNA and RNA-binding sites through geometric graph learning on ESMFold predicted structures. Here, the predicted protein structures are employed to construct protein structural graph with residues as nodes and spatially neighboring residue pairs for edges. The node representations are further enhanced through the pre-trained language model ProtTrans. The network was trained through the geometric vector perceptron, and the learned geometric embeddings are input into a common network to learn common binding characteristics, followed by two fully connected layers to learn specific binding patterns for DNA and RNA, respectively. Through comprehensive tests on DNA/RNA benchmark datasets, GLMSite was shown to surpass the latest sequence-based methods and be comparable with structure-based methods. Moreover, the prediction was shown useful for identifying nucleic-acid-binding proteins, demonstrating its potential for protein function discovery. The datasets, source codes, together with trained models are available at https://github.com/biomed-AI/nucleic-acid-binding.

**Keywords**: nucleic-acid-binding sites, geometric graph learning, pre-trained language model

## 1. Introduction

Interactions between proteins and nucleic acids play crucial roles in various biological processes and activities, such as gene replication and expression, repair, regulation, and signal transduction[1, 2]. The study of interactions between proteins and nucleic acids plays an important role in understanding protein function, analyzing genetic material, and designing new drugs. To address this issue, many experimental methods, such as nuclear magnetic resonance spectroscopy, X-ray[3], and fast ChIP[4] have been developed. However, they are usually costly and time-consuming. Therefore, it is necessary to develop computational methods for nucleic-acid-binding site prediction in a large-scale screening way[5].

Depending on the type of data used, the computational methods can be divided into two categories: sequence-based and structure-based methods. Sequence-based methods[5, 6] obtain local patterns of nucleic-acid-binding characteristics through sequence-derived features. For example, NCBRPred[7], by using the evolutionary information, solvent accessibility, and predicted secondary structures, learns local patterns for DNA and RNA-binding prediction through a sliding-window strategy. And DNAPred[8] identifies the DNA-binding sites with ensembled hyperplane-distance-based support vector machines. While these methods can make predictions for any protein from sequence, their performance may be limited since the potential patterns of binding sites can't be sufficiently derived from the sequences alone.

In contrast, structure-based methods are usually more accurate by inferring binding sites from known structures. The methods can be divided into template-based, machine learning based, and hybrid methods. Template-based methods, such as SPOT-Seq-RNA[9], DR_bind1[10] and TM-SITE[11], obtain dependable templates for specified query proteins using structural alignment, through which nucleic-acid-binding sites are identified. However, these methods don't apply to proteins without known templates. To address this issue, machine learning based methods[12] build classifiers using features from protein structures. For instance, GraphBind[13] uses graph neural networks (GNN) to obtain local tertiary patterns based on encoding protein structures as graphs. By comparison, hybrid methods combine both template-based methods and machine learning based methods, including COACH-D[14], DNABind[15], and NucBind[16]. Despite their good performance, structure-based methods are not suitable for proteins mostly without known tertiary structure due to the difficulty of determining protein structure experimentally.

With the recent breakthroughs in protein structure prediction by AlphaFold2, Yuan et al[17] proposed to identify the DNA-binding sites through the predicted structures. Unfortunately, the extensive computing requirement of AlphaFold2 hinders its use for genome-wide applications. To solve this problem, the pre-trained language model ESMFold[18] was constructed for fast and accurate structure prediction, achieving similar accuracy to AlphaFold2 but reducing the inference time by an order of magnitude faster than AlphaFold2. Such change enables better exploration of the structural space of metagenomic proteins. On the other hand, unsupervised language models (ProtTrans[19] and ESM-1b[20]) have been applied to extract features from protein sequences[21], which have shown fantastic results in ontology-based protein function[22], contact map prediction[20, 23], ion-binding site identification[24], and secondary and tertiary structural properties[25, 26]. However, these techniques haven't been fully utilized for nucleic-acid-binding site prediction.

In addition to the protein representations, effective learning of protein structure is essential for model performance. Protein structure could be learned through two types of techniques: graph neural networks (GNNs)[27, 28] and convolutional neural networks (CNNs)[29-31]. GNNs can perform

complex relational reasoning over structures[32], such as identifying key relationships among amino acids. By comparison, CNNs directly manipulate the geometry of the structure. Recently, the combination of two techniques is popular and showing better performance, a typical representative is geometric vector perceptron (GVP)[33]. GVP is a unifying architecture that bridges these two families of techniques by operating directly on both scalar and geometric features. This inspires us to consider using this method to achieve effective learning of protein structures.

In this study, we have proposed a novel method, GLMSite, which uses **G**eometric graph learning on **L**anguage **M**odel predicted structures for nucleic-acid-binding **site** identification. Specifically, the predicted protein structures are employed to construct protein structural graph with residues as nodes and spatially neighboring residue pairs for edges. The node representations are further enhanced through the pre-trained language model ProtTrans. The network was trained through the geometric vector perceptron, and the learned geometric embeddings are input into a common network to learn common binding characteristics, followed by two fully connected layers to learn specific binding patterns for DNA and RNA, respectively. Through comprehensive tests on DNA/RNA benchmark datasets, GLMSite was shown to surpass the latest sequence-based methods and be comparable with structure-based methods. Moreover, the prediction was shown useful for identifying nucleic-acid-binding proteins, demonstrating its potential for protein function discovery.

## 2. Materials and methods

### 2.1 Datasets

To evaluate the performance of GLMSite, we downloaded 14903 DNA and 13978 RNA-binding proteins from the BioLiP[34] database (released on March 30, 2022). The binding sites in this database were computed depending on the experimentally determined complex structures from the PDB database[35]. A target residue will be deemed to be a binding residue if the smallest atomic distance between it and the nucleic acid molecule is less than 0.5 Å plus the sum of Van der Waal's radius of the two nearest atoms.

One protein may bind different DNA/RNA, which are deposited in different PDB entries. To obtain the complete binding sites, we clustered the binding proteins with 95% sequence identity through MMseqs2[36], and chose the longest chain as the representative one. Following the previous studies[37, 38], the binding annotations were transferred from homology chains according to sequence alignment by blastp[39], leading to an increase of 28.6% and 32.4% in the number of DNA and RNA-binding sites, respectively. Then, the chains were removed through MMseqs2 at 30% sequence identity, resulting in DNA and RNA-binding data sets of size 915 proteins and 719 proteins, corresponding to 22866/261955 and 23045/219297 binding residues, respectively. For a strict evaluation, the proteins deposited before a specific date were used as the training and the afterwards as the test set. We set the deposition date as 18/12/2019 and 19/06/2019 for the DNA and RNA-binding data sets so that around 80% (735 and 577, respectively) proteins were used for the training. The details of training and test sets are summarized in **Table 1**.

To verify the ability of GLMSite to infer nucleic-acid-binding proteins from residue-level prediction, we constructed a new dataset PDB2770 from PDB[35] database (released after January 1, 2020). The homologous proteins were removed against the training set through MMseqs2 (30% sequence identity), resulting in 761 positive samples (nucleic-acid-binding proteins) and 2009 negative samples (non-nucleic-acid-binding proteins). The PDB2770 was later used to test the identification of DNA/RNA binding proteins.

### 2.2 Protein representations

According to the predicted structure from ESMfold, a protein is represented as a graph $G = (U, E)$, and each node $u_i \in U$ corresponding to an amino acid has node representations $h_u^{(i)}$. The set of edges is $E = \{e_{j\to i}\}_{i\neq j}$ for all $i$, $j$ where $u_j$ is among the k = 30 nearest neighbors of $u_i$ as measured by the distance between their Cα atoms. And each edge has an edge representations $h_e^{(j\to i)}$.

**Node representations**:

• Node vector features. Each node includes three vectors: the forward and reverse unit vectors in the directions of $\mathrm{C}_{\alpha_{i-1}} - \mathrm{C}_{\alpha_i}$ and $\mathrm{C}_{\alpha_{i+1}} - \mathrm{C}_{\alpha_i}$, and the unit vector in the direction of $\mathrm{C}_{\beta_i} - \mathrm{C}_{\alpha_i}$. Where the $C_\alpha$ is the central carbon atom in each amino acid residue and $C_\beta$ is the second carbon from the carboxyl carbon C.

• Structural properties. The DSSP[40] program was used to extract structure features, including (i) dihedral information $\{sin, cos\} \times (\Phi, \psi, \omega)$, where $\Phi, \psi, \omega$ are the dihedral angles computed from $\mathrm{C}_{i-1}, \mathrm{N}_i, \mathrm{C}_{\alpha_i}, \mathrm{C}_i$ and $\mathrm{N}_{i+1}$. (ii) solvent accessible surface area, standardized to relative solvent accessibility (RSA) by the maximal possible ASA of correlative amino acid type. (iii) nine one-hot secondary structure profile.

• Language model representations. To enhance the node representations, the protein language model ProtT5-XL-U50 (ProtTrans[19]) was employed to generate the protein embeddings. ProtTrans is a transformer-based auto-encoder called T5[41], performing self-supervised pretraining on UniRef50[42] to learn to predict masked amino acids. The hidden states from the last layer of the ProtTrans encoder were extracted as sequence features.

**Edge representations**:

• Edge vector features. The unit vector in the direction of $\mathrm{C}_{\alpha_j} - \mathrm{C}_{\alpha_i}$.

• Distance encoding. The distance encoding of $||\mathrm{C}_{\alpha_j} - \mathrm{C}_{\alpha_i}||_2$ according to the gaussian radial basis functions.

• Positional embedding. The positional embedding indicates the positioning of each neighbor $j$ by using the sinusoidal function of the gap $j - i$, where $i$ represents the current node.

### 2.3 The architecture of GLMSite

As shown in **Figure 1,** GLMSite uses ESMFold to predict protein structures while using ProtTrans to extract sequence embeddings, which are used to generate the node features and edge features. These are then fed into a geometric vector perceptron-based graph neural network (GVP-GNN). And the information is sent to two individual networks respective for DNA and RNA-binding site predictions.

#### 2.3.1 Geometric vector perceptron

The geometric vector perceptron (GVP) is a module for learning vector and scalar value functions based on geometric vectors and scalars. It is composed of two separate linear transformations acting on scalar and vector features, followed by two nonlinearities. Notably, the $L_2$ norm of the transformed vector features is concatenated before the scalar features are transformed; this guarantees that we can obtain rotation-invariant information from the input vectors. To control the output dimension independent of the number of norms, an additional linear transformation is added before the vector

nonlinearity. Specifically, given the scalar features s $\in R^n$ and vector features V $\in R^{v\times 3}$, the new features $(s', V') \in R^m \times R^{\mu\times 3}$ are computed as follows:

$$s' = \sigma(W_m s_{h+n} + b) \tag{1}$$

$$V' = \sigma^+\left(\left\|W_\mu W_h V\right\|_2\right) \odot \left(W_\mu W_h V\right) \tag{2}$$

Where $\sigma$ and $\sigma^+$ are nonlinearities, $W_m$, $W_h$ and $W_\mu$ are three separate linear transformations and $b$ is the bias term. The $s_{h+n} \in R^{h+n}$ represents the combination of s $\in R^n$ and $\|W_h V\|_2 \in R^h$, of which $h$ is the largest number of $v$ and $\mu$.

### 2.3.2 The GVP-based graph neural networks

The GVP-based graph neural networks (GVP-GNN) utilizes message passing[43] to updated node embeddings through the messages from neighboring nodes and edges. For each graph propagation, the protein graph defined above is fed into the architecture and the propagation steps are as follows:

$$h_m{}^{(j\to i)} := g\left(concat\left(h_u{}^{(j)}, h_e{}^{(j\to i)}\right)\right) \tag{3}$$

$$h_u{}^{(i)} \leftarrow LayerNorm\left(h_u{}^{(i)} + \frac{1}{k'} Dropout\left(\sum_{j:e_{j\to i}\in\epsilon} h_m{}^{(j\to i)}\right)\right) \tag{4}$$

where g is a sequence of GVPs, $h_u{}^{(i)}$ and $h_e{}^{(j\to i)}$ represent the embeddings of the node *i* and edge $(j \to i)$, respectively. And $h_m{}^{(j\to i)}$ is the message passed from node *j* to node *i*, while $k'$ is the number of incoming messages. Between graph propagation steps, a feed-forward point-wise layer is also used to update the node embeddings of all nodes *i*:

$$h_u{}^{(i)} \leftarrow LayerNorm\left(h_u{}^{(i)} + Dropout\left(g\left(h_u{}^{(i)}\right)\right)\right) \tag{5}$$

Both the scalar features and vector features at each node will be updated through these graph propagation and feed-forward steps.

### 2.3.3 Nucleic-acid-specific fully connected networks

The output of the GVP-GNN is transmitted to the nucleic-acid-specific fully connected networks to predict the DNA and RNA binding sites. Since different tasks have specific properties, we construct two independent fully connected networks for different tasks. For each specific task, only the corresponding specific network is updated, and the remaining specific network keeps unchanged.

### 2.3.4 Implementation details

On the training data, we performed 5-fold cross-validation (CV), in which the data was randomly divided into 5-folds. Each time, we trained a model on 4-folds and evaluated it on the rest of the data. We repeat this process five times and average the performances on 5-folds as the overall validation performance, which was used to optimize the hyperparameters and select the best feature combination. During the testing phase, all five models trained in CV were used to predict and averaged as the final predictions of our method.

Specifically, we utilized a 5-layer GVP encoder module with 128 hidden units. And we applied the Adam optimizer with a weight decay of 10−5, $\beta_1$ = 0.9, $\beta_2$ = 0.99 and a learning rate of $4 \times 10^{-4}$ for optimizing the model on the binary cross entropy loss. To avoid overfitting, the dropout rate is set to

0.1. The training process lasted up to 45 epochs and the early-stopping was performed with patience of 8 epochs according to the validation performance. Generally, in the training phase, it took ~45 min on an Nvidia GeForce RTX 3090 GPU. And only ~5s were taken to make predictions for a protein with pre-computed features.

### 2.3.5 Prediction of nucleic-acid-binding proteins

In this study, we found that the residue-level prediction can further infer the nucleic-acid-binding proteins. Referring to previous work[44], the following equation was employed to assign scores to identify nucleic-acid-binding proteins.

$$\alpha \frac{1}{n} \sum_{i=1}^{n} p_r(r_i) + (1-\alpha) p\left(\frac{n}{N}\right) \tag{6}$$

where the $p_r(r_i)$ is the $i$th highest binding probability of the residues, $N$ is the number of all residues in the protein, $p$ is a pdf of a gaussian fit to the distribution in the training set, $\alpha$ represents a weighting factor ($\alpha$ was set to 0.950 in this work), and $n$ is chosen to maximize this score.

### 2.3.6 Evaluation metrics

Referring to previous studies[45, 46], we employed precision (Pre), recall (Rec), F1-score (F1), Matthews correlation coefficient (MCC), the area under the precision-recall curve ($\mathrm{AUC_{PR}}$) and area under the receiver operating characteristic curve ($\mathrm{AUC_{ROC}}$) to measure the performance of the method developed:

$$Pre = \frac{TP}{TP + FP} \tag{7}$$

$$Rec = \frac{TP}{TP + FN} \tag{8}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{9}$$

$$MCC = \frac{TP \times TN - FN \times FP}{\sqrt{(TP + FN) \times (TP + FP) \times (TN + FN) \times (TN + FP)}} \tag{10}$$

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively.

## 3. Results

### 3.1 Performance on the 5-fold CV and independent tests

As shown in **Table 2,** We evaluated GLMSite through 5-fold CV and two independent tests. For DNA, GLMSite obtains $\mathrm{AUC_{ROC}}$ of 0.923, as well as $\mathrm{AUC_{PR}}$ of 0.634 on the 5-fold CV. Correspondingly, the $\mathrm{AUC_{ROC}}$ and $\mathrm{AUC_{PR}}$ are 0.929 and 0.571 on DNA-180-Test.

For RNA, the $\mathrm{AUC_{ROC}}$ and $\mathrm{AUC_{PR}}$ of the final model using 5-fold CV on the RNA-577-Train dataset are 0.882 and 0.545, which are 0.866 and 0.383 on RNA-142-Test, respectively. The consistency of $\mathrm{AUC_{ROC}}$ and $\mathrm{AUC_{PR}}$ on the CV and independent tests demonstrates the robustness of our model (see **Supplementary Tables S1 and S2** for more details).

To demonstrate the advantages of GLMSite and protein geometric knowledge, we compared GLMSite with a baseline method BiLSTM, which consists of a two-layer bidirectional long short-term memory network with 128 hidden units. We used this method as a geometric agnostic baseline to assess the influence of the geometric information for nucleic-acid-binding site prediction. As shown in **Table 3**, GLMSite surpasses BiLSTM on two independent test sets. The $AUC_{ROC}$, $AUC_{PR}$ and MCC of GLMSite is 1.4%, 6.7% and 6.3% higher than BiLSTM on DNA-180-Test, and the same is 2.9%, 8.2% and 11.0% higher than BiLSTM on RNA-142-Test. The results show that the protein geometric knowledge is crucial and GLMSite excels at extracting the geometric knowledge from the predicted structures.

To investigate why GLMSite achieved superior performance, we compared the performance of GLMSite and BiLSTM on amino acids containing different numbers of non-local contacts. The non-local contacts represent the contacts with residues that are >20 residues apart from sequence positions, but the atomic distance between Cα atoms ≤ 12 Å. As shown in **Figure 2**, the performance of GLMSite is consistently higher than that of BiLSTM on RNA-142-Test, and more importantly, as the non-local contact number increases, the performance gap between them gradually widens. This illustrates the effectiveness of GLMSite in capturing long-range contact information. Similarly, the same comparison was performed on DNA-180-Test (**Supplementary Figure S1**).

We further visualize the raw embeddings (size 1040) and the learned latent representations on DNA-180-Test. As shown in **Figure 3**, the binding and non-binding residues overlap each other and cannot be distinguished, while the learned ones tended to be clustered together. The results suggest that the latent representations learned by GLMSite is more discriminative for binding and non-binding residues. The visualization of the raw embeddings and learned latent representations on RNA-142-Test can be seen in **Supplementary Figure S2**.

### 3.2 Feature importance and ablation analysis

To demonstrate the importance of the features, we tested the model performance by using different features. As shown in **Table 4**, When using ProtTrans as node features, the highly processed protein representations from protein language model achieve satisfying performance with $AUC_{ROC}$ of 0.928 on DNA-180-Test and 0.862 on RNA-142-Test. When only the evolutionary profile (PSSM+HMM, denoted as Evo) is used, the $AUC_{ROC}$ of the model on DNA-180-Test and RNA-142-Test are 0.915 and 0.858, which are less than when using ProtTrans. This indicates that ProtTrans has a strong expressive ability while taking less time than Evo. ESM-1b was also tested, and its performance was lower than that of ProtTrans, which was not shown in this paper. When only the DSSP obtained from the predicted structure is used, the model still has considerable performance with $AUC_{ROC}$ of 0.895 on DNA-180-Test and 0.831 on RNA-142-Test, indicating that ESMFold can predict effective structures for downstream tasks. Besides, we tested the model performance with different feature combinations. We combined DSSP with traditional Evo and ProtTrans, respectively. As expected, when DSSP and ProtTrans are combined, the performance of the model is slightly higher than when DSSP and Evo are combined, with $AUC_{ROC}$ of 0.929 on DNA-180-Test and 0.866 on RNA-142-Test. This further proves the effectiveness of ProtTrans.

Since the predicted protein structures are used for geometric graph learning in GLMSite, the predicted structure quality of ESMFold is crucial for downstream prediction. For further demonstration, we tested the global distance test (GDT) between the native structures and the predicted structures calculated through SPalign[47]. As shown in **Figure 4**, the structure quality of ESMFold measured by

GDT is positively correlated with GLMSite performance measured by $AUC_{PR}$ on independent test DNA-180-Test. Closer inspection reveals that the top-20% proteins with the highest GDT (mean GDT = 0.938) have a mean $AUC_{PR}$ of 0.733. Whereas, the bottom-20% proteins with the lowest GDT (mean GDT = 0.167) have a mean $AUC_{PR}$ of 0.406, which is much lower than that of the top-20% proteins. These results demonstrate the importance of the quality of predicted protein structures for nucleic-acid-binding site prediction. For RNA, the same trend can be seen in **Supplementary Figure S3**.

### 3.3 Comparison with state-of-the-art methods

We compared GLMSite with six methods on DNA-180-Test while comparing it with four methods on RNA-142-Test. As shown in **Table 5**, GLMSite significantly surpassed the latest sequence-based methods and was comparable with structure-based methods. **Figure 5** compares the receiver operating characteristic curves on DNA-180-Test and RNA-142-Test.

On DNA-180-Test, we compared GLMSite with COACH-D, NucBind, SVMnuc, DNAPred, GraphBind, and GraphSite[17]. As shown in **Table 5**, the $AUC_{ROC}$, $AUC_{PR}$ and MCC of GLMSite are 0.929, 0.571, and 0.509, respectively, which are 1.3%, 9.6%, and 5.4% higher than the second-best method. Meanwhile, GLMSite outperforms all other methods with recall, precision, and F1 of 0.606, 0.490, and 0.542, respectively. Although GLMSite is a sequence-based predictor requiring only protein sequences as input, it outperforms other structure-based methods. In particular, it outperforms the state-of-the-art structure-based method GraphBind by 2.7% and 23.3% in $AUC_{ROC}$ and $AUC_{PR}$, respectively. When using the predicted structure, the performance of the structure-based methods will be further reduced. For example, the $AUC_{ROC}$ and $AUC_{PR}$ of GraphBind are reduced by 2.1% and 8.6% respectively, and the superiority of our method is more prominently reflected. Additionally, we also compared GLMSite with the latest AlphaFold2-based method GraphSite, and the results show that the $AUC_{ROC}$ and $AUC_{PR}$ of our method are 1.3% and 9.6% higher than GraphSite, respectively. This further demonstrates the superiority of our method and also reflects the high quality of the structures predicted by ESMFold.

Similarly, we compared GLMSite with COACH-D, NucBind, SVMnuc, and GraphBind on RNA-142-Test. GLMSite surpasses all other methods using predicted structures, with $AUC_{ROC}$, $AUC_{PR}$ and MCC of 0.866, 0.383, and 0.394, respectively, 9.8%, 39.3%, and 38.2% higher than the second-best method. When using native structures, the structure-based methods improve significantly, resulting in the $AUC_{PR}$ of GraphBind slightly outperforms our method, but the $AUC_{ROC}$ and MCC is still 1.4% and 2.3% lower than our method. This indicates that the information extracted from the ProtTrans embeddings and predicted structures are crucial, and comparable to the information contained in native structures. **Supplementary Figure S4** details the precision-recall curves of all methods on these two datasets.

### 3.4 Residue-level prediction allows the identification of nucleic-acid-binding proteins

To test the ability of GLMSite to infer nucleic-acid-binding proteins from residue-level prediction, a score was generated through the binding probabilities of residues assigned by model and the percentage of residues predicted by GLMSite to be binding residues with high confidence[44]. To calculate this score, the binding-residue percentage distribution of each protein in the training set was fit by a gaussian distribution (**Supplementary Figure S5**). The resulting score was the average of two metrics by weight. As shown in **Equation 6**, the first is the average probability of the top-n residues and the second is the possibility that a binding site should be composed of these n residues according to the

above distribution (gaussian term).

We compared the distribution of scores calculated for two types of proteins on PDB2770. As shown in **Figure 6**, the scores of nucleic-acid-binding proteins are higher than those of other proteins (non-nucleic-acid-binding proteins) greatly, which demonstrates the ability of our method to identify nucleic-acid-binding proteins. In addition, we also compared two methods for calculating scores: (i) using the average probability of all residues of a protein. (ii) using the average probability of top-n residues and gaussian term. The results show that when the gaussian term is used, the ability to identify nucleic-acid-binding proteins has been improved (**Supplementary Table S3**). We then compared the receiver operating characteristic and precision-recall curves of two different methods on PDB2770 (**Supplementary Figures S6 and S7**), which indicates the superiority of the proposed method. The above results suggest that the residue-level prediction is meaningful for inferring protein-level function.

### 3.5 Case study

As an example, we visualize one case (PDB ID: 7KX9, chain A) from RNA-142-Test predicted by GLMSite and the geometric-agnostic baseline method BiLSTM. **Figure 7** shows the nucleic-acid-binding site prediction results of GLMSite (A) and the geometric-agnostic baseline method BiLSTM (B). For this example, there are 59 RNA-binding residues over a total of 734 residues. GLMSite predicted 60 binding residues in which 40 are TP, leading to an F1 of 0.672, $\mathrm{AUC_{ROC}}$ of 0.967 and $\mathrm{AUC_{PR}}$ of 0.694 (**Supplementary Table S4**). As contrasted, BiLSTM predicted 70 binding residues in which only 36 are TP, leading to a lower F1 of 0.558, $\mathrm{AUC_{ROC}}$ of 0.931 and $\mathrm{AUC_{PR}}$ of 0.524.

## 4. Discussion and conclusion

Identifying nucleic-acid-binding sites is crucial for understanding protein function, analyzing genetic material, and designing new drugs. Existing sequence-based methods have limited predictive performance due to only considering contextual features of the sequential neighbors, while structure-based methods are not suitable for proteins mostly without known tertiary structures. Trained with the protein structures predicted by ESMFold and ProtTrans-based embeddings, GLMSite achieves excellent performance solely from protein sequences, solving the limitations of sequence-based and structure-based methods simultaneously. Here, the network was trained through the geometric vector perceptron, and the learned geometric embeddings are input into a common network to learn common binding characteristics, followed by two fully connected layers to learn specific binding patterns for DNA and RNA, respectively. In general, the superiority of GLMSite benefits from the following aspects: (I) the high quality of the predicted structures by ESMFold. (II) abundant information extracted from pre-trained language model ProtTrans. (III) crucial geometric embeddings obtained through the GVP module. (IV) the binding characteristics of different nucleic acids learned from a common network. Through comprehensive tests on the two independent test sets, GLMSite was shown to surpass the state-of-the-art methods.

In this study, we have an interesting observation that the residue-level prediction is meaningful for inferring protein-level function. Based on the residue-level prediction, a score was computed using the average probability of top-n residues and gaussian term. The results have shown that the scores of nucleic-acid-binding proteins were significantly higher than those of non-binding proteins. This inspires us that the residue-level prediction can be further extended to the protein-level function prediction. In the following work, we will also conduct more in-depth research on the interaction and promotion of the information between residue level and protein level.

Despite the advantages of GLMSite, the method can be improved in several ways. First, the performance of the proposed model is affected by the quality of the structures predicted by ESMFold. This might be relieved by containing other instructive sequence-derived features to enhance the robustness of the model. Second, the significant efficacy of common networks inspires us to employ more types of molecules to promote mutual learning. These challenges will be explored in our future work. In summary, we have demonstrated that GLMSite provides a deep-learning model for fast and accurate prediction of nucleic-acid-binding sites. GLMSite is also freely available at https://github.com/biomed-AI/nucleic-acid-binding.

**REFERENCES**


1. Charoensawan V, Wilson D, Teichmann SA. Genomic repertoires of DNA-binding transcription factors across the tree of life, Nucleic acids research 2010;38:7364-7377.
2. Hirota K, Miyoshi T, Kugou K et al. Stepwise chromatin remodelling by a cascade of transcription initiation of non-coding RNAs, Nature 2008;456:130-134.
3. Orengo CA, Michie AD, Jones S et al. CATH–a hierarchic classification of protein domain structures, Structure 1997;5:1093-1109.
4. Mandel-Gutfreund Y, Margalit H. Quantitative parameters for amino acid-base interaction: implications for prediction of protein-DNA binding sites, Nucleic acids research 1998;26:2306-2312.
5. Yan J, Kurgan L. DRNApred, fast sequence-based method that accurately predicts and discriminates DNA-and RNA-binding residues, Nucleic acids research 2017;45:e84-e84.
6. Hu J, Li Y, Zhang M et al. Predicting protein-DNA binding residues by weightedly combining sequence-based features and boosting multiple SVMs, IEEE/ACM transactions on computational biology and bioinformatics 2016;14:1389-1398.
7. Zhang J, Chen Q, Liu B. NCBRPred: predicting nucleic acid binding residues in proteins based on multilabel learning, Briefings in bioinformatics 2021;22:bbaa397.
8. Zhu Y-H, Hu J, Song X-N et al. DNAPred: accurate identification of DNA-binding sites from protein sequence by ensembled hyperplane-distance-based support vector machines, Journal of chemical information and modeling 2019;59:3057-3071.
9. Yang Y, Zhao H, Wang J et al. SPOT-Seq-RNA: predicting protein–RNA complex structure and RNA-binding function by fold recognition and binding affinity prediction, Protein structure prediction 2014:119-130.
10. Chen YC, Sargsyan K, Wright JD et al. Identifying RNA-binding residues based on evolutionary conserved structural and energetic features, Nucleic acids research 2014;42:e15-e15.
11. Yang J, Roy A, Zhang Y. Protein–ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment, Bioinformatics 2013;29:2588-2595.
12. Jiménez J, Doerr S, Martínez-Rosell G et al. DeepSite: protein-binding site predictor using 3D-convolutional neural networks, Bioinformatics 2017;33:3036-3042.
13. Xia Y, Xia C-Q, Pan X et al. GraphBind: protein structural context embedded rules learned by hierarchical graph neural networks for recognizing nucleic-acid-binding residues, Nucleic acids research 2021;49:e51-e51.
14. Wu Q, Peng Z, Zhang Y et al. COACH-D: improved protein–ligand binding sites prediction with refined ligand-binding poses through molecular docking, Nucleic acids research 2018;46:W438-W442.
15. Liu R, Hu J. DNABind: A hybrid algorithm for structure-based prediction of DNA-binding residues by combining machine learning-and template-based approaches, Proteins: Structure, Function, and Bioinformatics 2013;81:1885-1899.
16. Su H, Liu M, Sun S et al. Improving the prediction of protein–nucleic acids binding residues via multiple sequence profiles and the consensus of complementary methods, Bioinformatics 2019;35:930-936.
17. Yuan Q, Chen S, Rao J et al. AlphaFold2-aware protein–DNA binding site prediction using

graph transformer, Briefings in bioinformatics 2022;23:bbab564.
18. Lin Z, Akin H, Rao R et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction, bioRxiv 2022.
19. Elnaggar A, Heinzinger M, Dallago C et al. Prottrans: Toward understanding the language of life through self-supervised learning, IEEE transactions on pattern analysis and machine intelligence 2021;44:7112-7127.
20. Rives A, Meier J, Sercu T et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences, Proceedings of the National Academy of Sciences 2021;118:e2016239118.
21. Unsal S, Atas H, Albayrak M et al. Learning functional properties of proteins with language models, Nature Machine Intelligence 2022;4:227-245.
22. Yuan Q, Xie J, Xie J et al. Fast and accurate protein function prediction from sequence through pretrained language model and homology-based label diffusion, Briefings in bioinformatics 2023;24:bbad117.
23. Singh J, Litfin T, Singh J et al. SPOT-Contact-LM: improving single-sequence-based prediction of protein contact map using a transformer language model, Bioinformatics 2022;38:1888-1894.
24. Yuan Q, Chen S, Wang Y et al. Alignment-free metal ion-binding site prediction from protein sequence through pretrained language model and multi-task learning, Briefings in bioinformatics 2022;23:bbac444.
25. Singh J, Paliwal K, Litfin T et al. Reaching alignment-profile-based accuracy in predicting protein secondary and tertiary structural properties without alignment, Scientific reports 2022;12:7607.
26. Song Y, Yuan Q, Chen S et al. Fast and accurate protein intrinsic disorder prediction by using a pretrained language model, Briefings in bioinformatics 2023:bbad173.
27. Yuan Q, Chen J, Zhao H et al. Structure-aware protein–protein interaction site prediction using deep graph convolutional network, Bioinformatics 2022;38:125-132.
28. Chen J, Zheng S, Zhao H et al. Structure-aware protein solubility prediction from sequence through graph convolutional network and predicted contact map, Journal of cheminformatics 2021;13:1-10.
29. Lam JH, Li Y, Zhu L et al. A deep learning framework to predict binding preference of RNA constituents on protein surface, Nature communications 2019;10:4941.
30. Zheng S, Li Y, Chen S et al. Predicting drug–protein interaction using quasi-visual question answering system, Nature Machine Intelligence 2020;2:134-140.
31. Kozlovskii I, Popov P. Protein–peptide binding site detection using 3D convolutional neural networks, Journal of chemical information and modeling 2021;61:3814-3823.
32. Battaglia PW, Hamrick JB, Bapst V et al. Relational inductive biases, deep learning, and graph networks, arXiv preprint arXiv:1806.01261 2018.
33. Jing B, Eismann S, Suriana P et al. Learning from protein structure with geometric vector perceptrons, arXiv preprint arXiv:2009.01411 2020.
34. Yang J, Roy A, Zhang Y. BioLiP: a semi-manually curated database for biologically relevant ligand–protein interactions, Nucleic acids research 2012;41:D1096-D1103.
35. Berman HM, Westbrook J, Feng Z et al. The protein data bank, Nucleic acids research 2000;28:235-242.

36. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets, Nature biotechnology 2017;35:1026-1028.
37. Chen K, Mizianty MJ, Gao J et al. A critical comparative assessment of predictions of protein-binding sites for biologically relevant organic compounds, Structure 2011;19:613-621.
38. Zhang J, Kurgan L. Review and comparative assessment of sequence-based predictors of protein-binding residues, Briefings in bioinformatics 2018;19:821-837.
39. Johnson M, Zaretskaya I, Raytselis Y et al. NCBI BLAST: a better web interface, Nucleic acids research 2008;36:W5-W9.
40. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, Biopolymers: Original Research on Biomolecules 1983;22:2577-2637.
41. Raffel C, Shazeer N, Roberts A et al. Exploring the limits of transfer learning with a unified text-to-text transformer, The Journal of Machine Learning Research 2020;21:5485-5551.
42. Suzek BE, Huang H, McGarvey P et al. UniRef: comprehensive and non-redundant UniProt reference clusters, Bioinformatics 2007;23:1282-1288.
43. Gilmer J, Schoenholz SS, Riley PF et al. Neural message passing for quantum chemistry. In: International conference on machine learning. 2017, p. 1263-1272. PMLR.
44. Abdin O, Nim S, Wen H et al. PepNN: a deep attention model for the identification of peptide binding sites, Communications Biology 2022;5:503.
45. Do DT, Le TQT, Le NQK. Using deep neural networks and biological subwords to detect protein S-sulfenylation sites, Briefings in bioinformatics 2021;22:bbaa128.
46. Le NQK, Ho Q-T, Nguyen T-T-D et al. A transformer architecture based on BERT and 2D convolutional neural network to identify DNA enhancers from sequence information, Briefings in bioinformatics 2021;22:bbab005.
47. Yang Y, Zhan J, Zhao H et al. A new size-independent score for pairwise protein structure alignment and its application to structure classification and nucleic-acid binding prediction, Proteins: Structure, Function, and Bioinformatics 2012;80:2080-2088.

## Figure

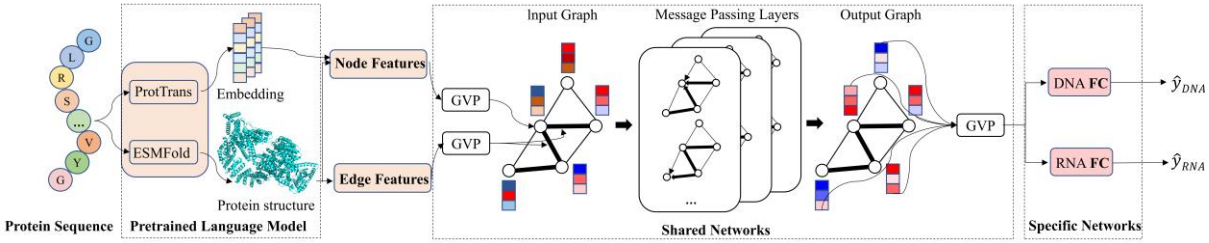


**Figure 1.** The protein sequence is input to ESMFold to predict protein structures while using ProtTrans to extract sequence embeddings, which are used to generate the node features and edge features. These are then fed into a geometric vector perceptron-based graph neural network (GVP-GNN). And the information is sent to two individual networks respective for DNA and RNA binding site predictions.

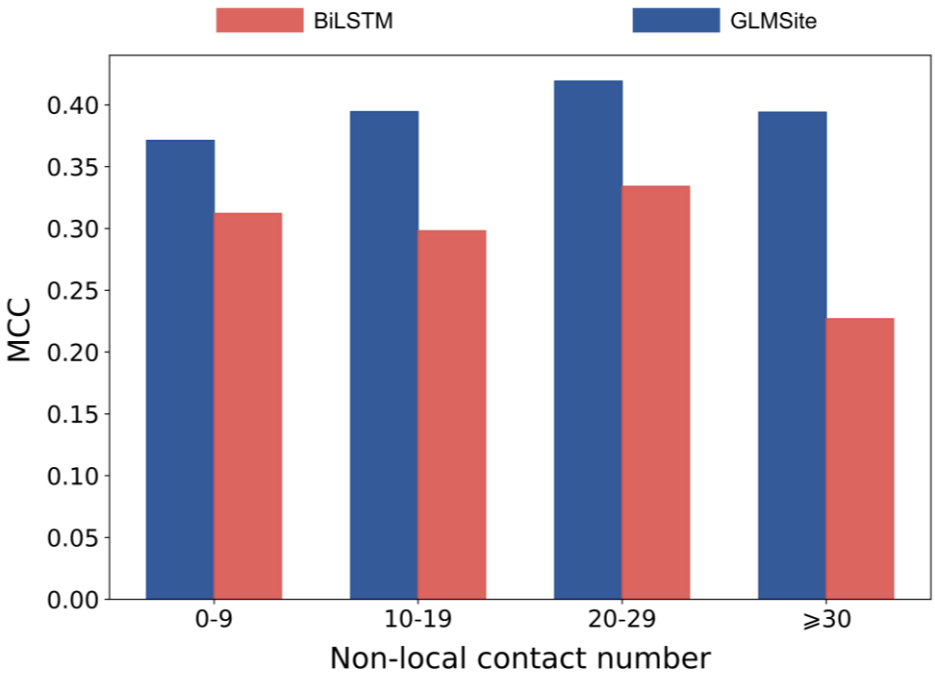


**Figure 2.** The MCC of GLMSite and BiLSTM on amino acids containing different numbers of non-local contacts in RNA-142-Test.

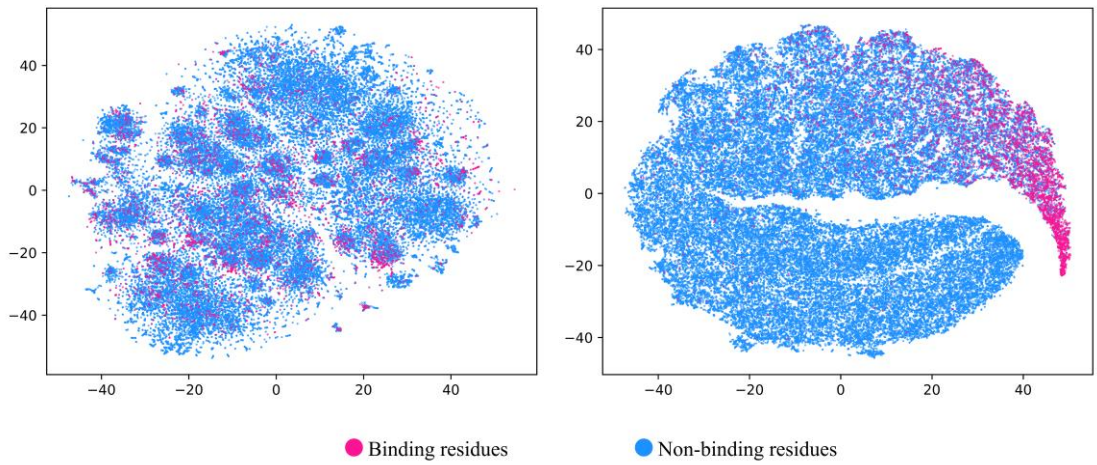


**Figure 3.** Visualization of the distributions of samples encoded by raw feature vectors **(A)** and latent feature vectors learned by GLMSite **(B)** on DNA-180-Test using t-SNE.

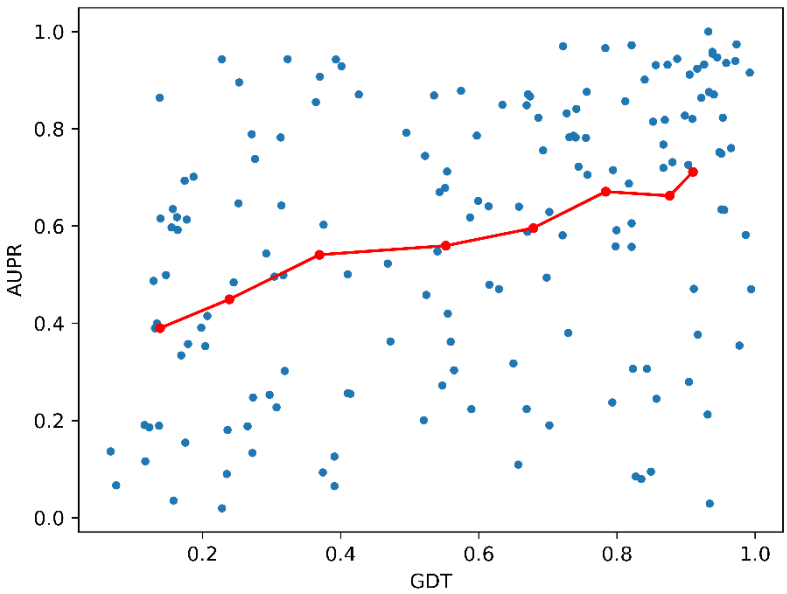


**Figure 4.** Model performance (measured by $AUC_{PR}$) varies with structural quality (measured by GDT) on DNA-180-Test. The blue scatters represent the GDT and $AUC_{PR}$ of each protein, and the red scatters represent the average GDT and $AUC_{PR}$ per bin after sorting the proteins by GDT and dividing them into eight bins.

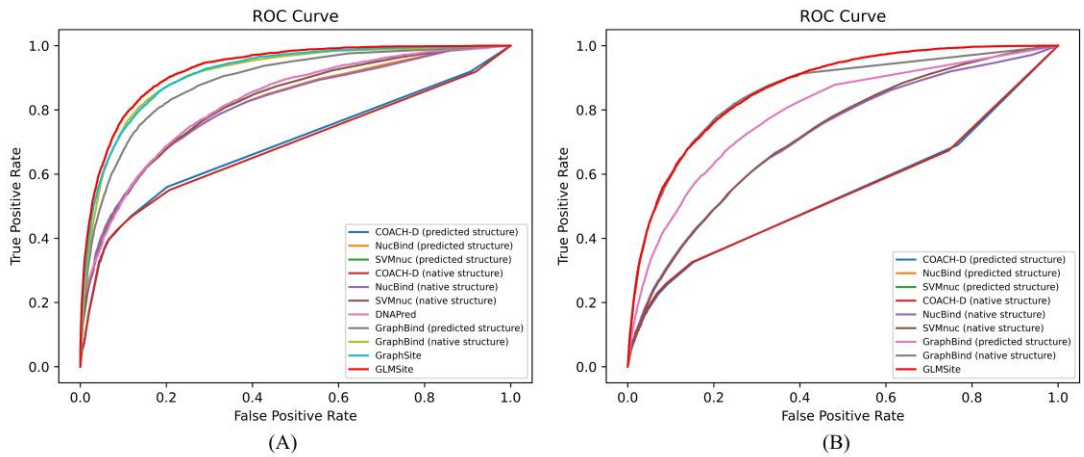


**Figure 5.** The receiver operating characteristic curves given by GLMSite and other methods on DNA-180-Test (A) and RNA-142-Test (B)

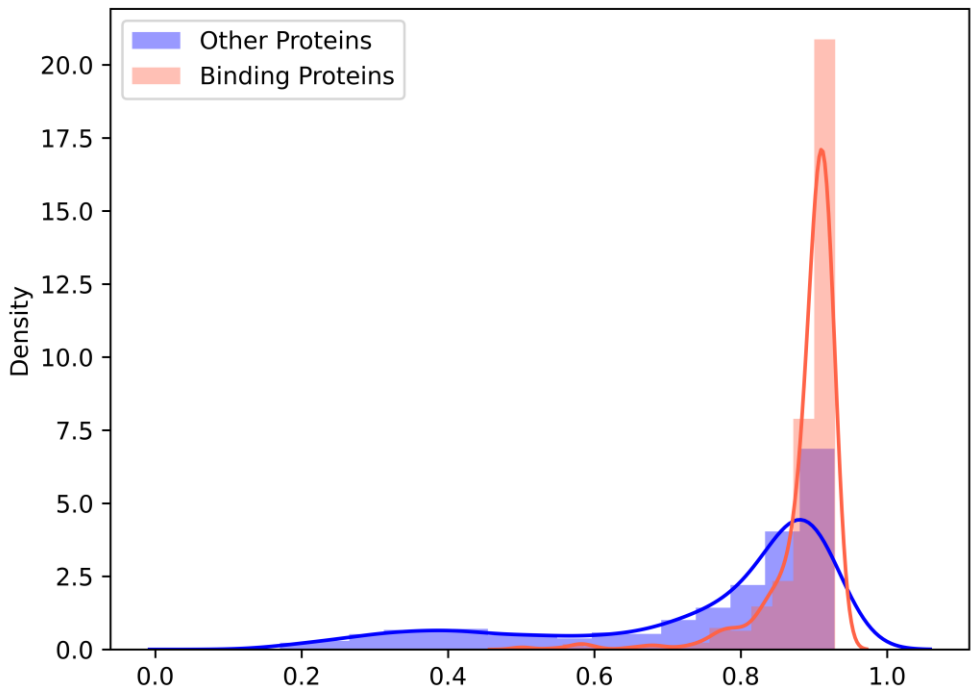


**Figure 6**. The score distribution of nucleic-acid-binding proteins and non-nucleic-acid-binding proteins (other proteins) on PDB2770.

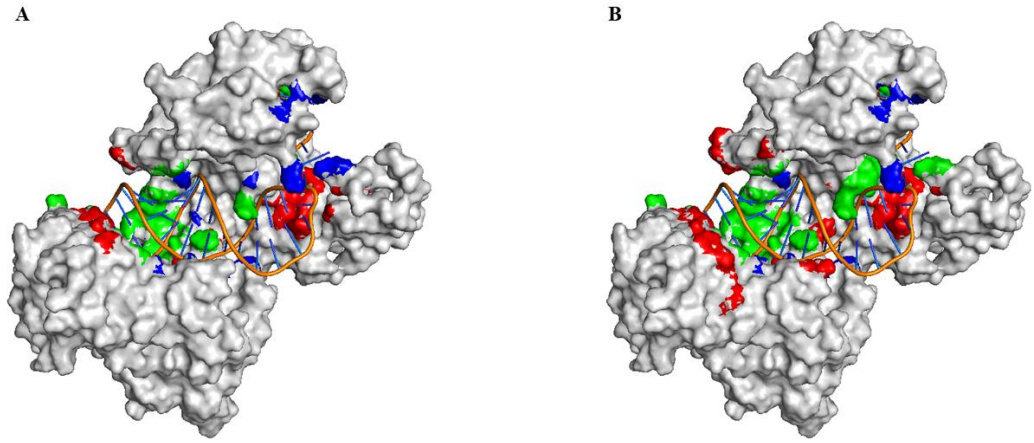


**Figure 7.** Visualization of one example (PDB ID: 7KX9, chain A) from RNA-142-Test predicted by GLMSite (**A**) and the geometric-agnostic baseline method BiLSTM (**B**). TP, FP, and FN are colored in green, red, and blue, respectively.

## Table

Table 1. Summary of training and test sets

| Type | Dataset | $N_{protein}$ [a] | $N_{pos}$ [b] | $N_{neg}$ [c] | PNratio [d] |
|---|---|---|---|---|---|
| DNA | DNA-735-Train | 735 | 18611 | 178125 | 0.104 |
| | DNA-180-Test | 180 | 4255 | 60964 | 0.070 |
| RNA | RNA-577-Train | 577 | 18564 | 143019 | 0.130 |
| | RNA-142-Test | 142 | 4481 | 53233 | 0.084 |

[a] Number of proteins; [b] Number of binding residues; [c] Number of non-binding residues; [d] PNratio = $N_{pos}/N_{neg}$.

Table 2. The performance of GLMSite on 5-fold CV and two independent test sets (DNA-180-Test and RNA-142-Test).

| Type | 5-fold CV | | Independent tests | |
|---|---|---|---|---|
| | $AUC_{ROC}$ | $AUC_{PR}$ | $AUC_{ROC}$ | $AUC_{PR}$ |
| DNA | 0.923 | 0.634 | 0.929 | 0.571 |
| RNA | 0.882 | 0.545 | 0.866 | 0.383 |

Table 3. The performance comparison of GLMSite and BiLSTM on two independent test sets DNA-180-Test and RNA-142-Test according to $AUC_{ROC}$, $AUC_{PR}$, MCC, Rec, Pre, and F1.

| Dataset | Method | $AUC_{ROC}$ | $AUC_{PR}$ | MCC | Rec | Pre | F1 |
|---|---|---|---|---|---|---|---|
| DNA-180-Test | BiLSTM | 0.916 | 0.535 | 0.479 | 0.546 | 0.485 | 0.514 |
| | GLMSite | 0.929 | 0.571 | 0.509 | 0.606 | 0.490 | 0.542 |
| RNA-142-Test | BiLSTM | 0.842 | 0.354 | 0.355 | 0.465 | 0.364 | 0.409 |
| | GLMSite | 0.866 | 0.383 | 0.394 | 0.565 | 0.359 | 0.439 |

Table 4. Ablation studies of GLMSite on DNA-180-Test and RNA-142-Test

| Feature | DNA-180-Test | | | RNA-142-Test | | |
|---|---|---|---|---|---|---|
| | $AUC_{ROC}$ | $AUC_{PR}$ | MCC | $AUC_{ROC}$ | $AUC_{PR}$ | MCC |
| Dssp | 0.895 | 0.455 | 0.436 | 0.831 | 0.333 | 0.328 |
| Evo | 0.915 | 0.512 | 0.479 | 0.858 | 0.365 | 0.366 |
| ProtTrans | 0.928 | 0.559 | 0.503 | 0.862 | 0.377 | 0.375 |
| Evo+Dssp | 0.921 | 0.532 | 0.491 | 0.865 | 0.380 | 0.378 |
| ProtTrans+Dssp (GLMSite) | **0.929** | **0.571** | **0.516** | **0.866** | **0.383** | **0.394** |

Table 5. Performance comparison of GLMSite with state-of-the-art methods on DNA-180-Test and RNA-142-Test

| Dataset | Method | $AUC_{ROC}$ | $AUC_{PR}$ | MCC | Rec | Pre | F1 |
|---|---|---|---|---|---|---|---|
| DNA-180-Test | COACH-D (predicted structure) | 0.691 | 0.263 | 0.312 | 0.307 | 0.403 | 0.349 |
| | NucBind (predicted structure) | 0.813 | 0.339 | 0.332 | 0.337 | 0.411 | 0.370 |
| | SVMnuc (predicted structure) | 0.820 | 0.333 | 0.321 | 0.324 | 0.403 | 0.359 |
| | COACH-D (native structure) | 0.685 | 0.266 | 0.318 | 0.311 | 0.410 | 0.354 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | NucBind (native structure) | 0.812 | 0.342 | 0.338 | 0.342 | 0.418 | 0.376 |
| | SVMnuc (native structure) | 0.820 | 0.332 | 0.320 | 0.323 | 0.402 | 0.358 |
| | DNAPred | 0.824 | 0.399 | 0.334 | 0.357 | 0.411 | 0.382 |
| | GraphBind (predicted structure) | 0.886 | 0.423 | 0.430 | 0.522 | 0.425 | 0.468 |
| | GraphBind (native structure) | 0.905 | 0.463 | 0.466 | 0.598 | 0.429 | 0.500 |
| | GraphSite | 0.917 | 0.521 | 0.483 | 0.557 | 0.484 | 0.518 |
| | GLMSite | **0.929** | **0.571** | **0.509** | **0.606** | **0.490** | **0.542** |
| RNA-142-Test | COACH-D (predicted structure) | 0.542 | 0.153 | 0.128 | 0.106 | 0.273 | 0.152 |
| | NucBind (predicted structure) | 0.714 | 0.201 | 0.168 | 0.166 | 0.284 | 0.210 |
| | SVMnuc (predicted structure) | 0.719 | 0.193 | 0.161 | 0.162 | 0.274 | 0.204 |
| | COACH-D (native structure) | 0.543 | 0.155 | 0.128 | 0.107 | 0.270 | 0.153 |
| | NucBind (native structure) | 0.714 | 0.200 | 0.168 | 0.167 | 0.285 | 0.210 |
| | SVMnuc (native structure) | 0.719 | 0.193 | 0.162 | 0.163 | 0.274 | 0.205 |
| | GraphBind (predicted structure) | 0.789 | 0.275 | 0.285 | 0.453 | 0.279 | 0.345 |
| | GraphBind (native structure) | 0.854 | **0.396** | 0.385 | 0.560 | 0.353 | 0.433 |
| | GLMSite | **0.866** | 0.383 | **0.394** | **0.565** | **0.359** | **0.439** |