Accurate Nucleic Acid-Binding Residue Identification based Domain-Adaptive Protein Language Model and Explainable Geometric Deep Learning

Wenwu Zeng, Liangrui Pan, Boya Ji, Liwen Xu, Shaoliang Peng*

College of Computer Science and Electronic Engineering, Hunan University, Changsha, China
{wwz_cs, panlr, byj, xuliwen, slpeng}@hnu.edu.cn

Abstract

Protein-nucleic acid interactions play a fundamental and critical role in a wide range of life activities. Accurate identification of nucleic acid-binding residues helps to understand the intrinsic mechanisms of the interactions. However, the accuracy and interpretability of existing computational methods for recognizing nucleic acid-binding residues need to be further improved. Here, we propose a novel method called GeSite based the domain-adaptive protein language model and E(3)-equivariant graph neural network. Prediction results across multiple benchmark test sets demonstrate that GeSite is superior or comparable to state-of-the-art prediction methods. The MCC values of GeSite are 0.522 and 0.326 for the one DNA-binding residue test set and one RNA-binding residue test set, which are 0.57 and 38.14% higher than that of the second-best method, respectively. Detailed experimental results suggest that the advanced performance of GeSite lies in the well-designed nucleic acid-binding protein adaptive language model. Additionally, interpretability analysis exposes the perception of the prediction model on various remote and close functional domains, which is the source of its discernment ability.

Code — https://github.com/pengsl-lab/GeSite
Datasets — https://huggingface.co/zengwenwu/GeSite
Extended version — https://www.biorxiv.org/content/10.1101/2024.12.11.628078v1

Introduction

Protein-nucleic acid interactions serve as a fundamental role in various biological processes like gene regulation and expression in organisms (Lambert, et al., 2018). Understanding these interactions is crucial for studying protein function and facilitating drug development. Accurate identification of nucleic acid-binding residue (NBS) is a key step in elucidating the underlying mechanisms of these interactions. Traditional biological wet-lab experimental methods, including chromatin immunoprecipitation on microarrays, nuclear magnetic resonance, and X-ray crystallography, have significantly advanced the study of protein-nucleic acid interactions. However, these methods are limited by high costs and long lead times, making it impractical to determine all NBSs across the vast number of protein sequences in the post-genomic era. As of 5 March 2024, there are 664,681,050 protein sequences recorded in the UniParc database (Leinonen, et al., 2004), while the number of resolved protein-DNA (or RNA) complex structures recorded in the Nucleic Acid Database (Narayanan, et al., 2014) is only 6,296 (or 2,846). Although recent methods like AlphaFold3 (Abramson, et al., 2024) and RoseTTAFold All-Atom (Krishna, et al., 2024) attempted to directly predict the 3D structure of protein-nucleic acid complexes, the accuracy remains suboptimal due to the limited number of known complex structures that they heavily depend on. Accurate binding sites can further assist in modeling protein-nucleic acid complexes. Consequently, the development of rapid and accurate method for predicting NBS remains essential. Numerous computational methods have been proposed to address this challenge. In the early days, researchers used statistical and machine learning-based methods to capture conserved information of NBS for prediction. Despite the progress made, these methods generally suffered from poor accuracy and generalizability. In the past decade, deep learning-based methods have gained significant attention. These methods capture the intricate non-linear relationships between the sequence, structure, and function of proteins, enabling highly accurate NBS prediction. Depending on the feature sources employed, these methods can be broadly classified into two categories: sequence-driven methods, e.g., DNApred (Y. H. Zhu, et al., 2019), Pprint2 (Patiyal, et al., 2022), DRNAPred (Yan and Kurgan, 2017), iDRNA-ITF (N. Wang, et al., 2022), ESM-NBR (Zeng, et al., 2023), ULDNA (Y.-H. Zhu, et al., 2024), CLAPE (Liu and Tian, 2024), hybridRNAbind (F. Zhang, et

*Corresponding author.

al., 2023), and HybridDBRpred (J. Zhang, Basu, and Kurgan, 2024); as well as structure-driven methods, e.g., GraphSite (Yuan, et al., 2022), GraphBind (Xia, et al., 2021), CrossBind (L. Jing, et al., 2024), and EquiPNAS (Roche, et al., 2024). Sequence-driven methods typically explore evolutionary information through primary sequences to identify nucleic acid-binding residues. While these methods are convenient and easily scalable, their accuracy is often limited due to the inherent difficulty of extracting useful discriminative information directly from primary sequences. On the other hand, structure-driven methods leverage the high conservatism and specificity of NBS on 3D structure, potentially achieving better performance. However, the longstanding lack of high-quality protein 3D structures has hindered the scalability of such methods. Promising to change this dilemma are recent major advances (Baek, et al., 2021; Humphreys, et al., 2021; Tunyasuvunakool, et al., 2021) in deep learning-based protein 3D structure prediction like AlphaFold2 (Jumper, et al., 2021). Using the predicted 3D structure as a substitute can complement the inadequacy of the resolved native structure. Moreover, the role of protein language model (PLM) in studying protein function and structure (Lin, et al., 2023; Rives, et al., 2021) provides new insights into the analysis of protein-nucleic acid interactions. Integrating protein structure features with knowledge from PLM is expected to significantly advance the study of protein-nucleic acid interactions.

Several methods have been developed to predict nucleic acid-binding residues based on protein structure (whether native or predicted) or universal PLM. For example, CrossBind (L. Jing, et al., 2024) predicts NBS by integrating amino acid-level PLM and atom-level feature representation. Liu *et al.* (Liu and Tian, 2024) proposed a DNA-binding residue (DBS) predictor called CLAPE that combines the PLM ProtTrans (Elnaggar, et al., 2021) and the contrastive learning strategy. In EquiPNAS (Roche, et al., 2024), ESM2 feature embedding, MSA representation extracted from AlphaFold2, and a variety of local structural features serve as inputs to train an E(3)-equivariant graph neural network (EGNN) (Satorras, Hoogeboom, and Welling, 2021).

Despite the good results achieved by the aforementioned methods, there is still room for improvement. Firstly, PLM-based NBS prediction methods typically extract sequence embedding as feature representations directly from the original universal PLM trained on massive general protein sequences. While these PLMs are excellent for characterizing properties common to all protein families, such as tertiary structure, they are under-explored for specificity when focusing on particular families like Nuclear Receptor and Forkhead, which are typical DNA-binding protein (DBP) families. The ability of proteins to bind to nucleic acids comes from a pocket folded by a small, highly conserved sequence motif. Further probing of these typical nucleic acids-binding motifs will undoubtedly further enhance the ability of PLM to characterize nucleic acids-binding protein (NBP), thereby improving NBS prediction accuracy. Second, to capture contextual information about spatial structures, previous methods (Xia, et al., 2021; Yuan, et al., 2022) use various GNN variants such as graph transformer and Gate Recurrent Unit (GRU)-based GNN (Dey and Salem, 2017), which, despite their good performance, do not provide visualization and interpretability, making it difficult to intuitively understand what the model has actually learned.

In this study, we propose a novel structure-based NBS prediction method named GeSite based on DNA- and RNA-binding protein domain-adaptive PLM and EGNN. In GeSite, we directly utilized the previous study, ESM-DBP (Zeng, et al., 2024), as the DBP adaptive PLM to extract sequence embedding as input feature for DBS prediction. For RNA-binding protein (RBP) adaptive PLM, similar to ESM-DBP, we collected 459,656 non-redundant RBP sequences to fine-tuning the parameters of the last five transformer blocks of ESM2, resulting in ESM-RBP for RNA-binding residue (RBS) prediction. Subsequently, the MSA file is generated using the HHblits (Remmert, et al., 2012) tool to search the Uniclust30 database (Mirdita, et al., 2017) and fed into the ESM-MSA model to obtain the embedding matrix as the representation of evolutionary information of protein sequence. The embedding matrix extracted from the well-trained ESM-D/RBP is concatenated with the output of ESM-MSA to serve as the feature representation of each protein sequence in the benchmark dataset. Finally, a protein graph based on residue distance map is constructed for input into EGNN to predict the binding probability of each residue. The prediction results on the three benchmark test sets show that the performance of GeSite on both the DBS and RBS are better than or comparable to state-of-the-art (SOTA) methods. The AUC values of GeSite on DNA-129_Test, DNA-181_Test, and RNA-117_Test are 0.941, 0.919, and 0.861, which are 0.75, 0.22, and 9.96% higher than that of the second-best method separately. The experimental results demonstrate that the superior performance of the proposed method is based on the domain-adaptive PLM that provides better sequence characterization of NBS than the universal PLM. Additionally, interpretability analysis and visualization of the GeSite reveal the sensitivity of GeSite to various nucleic acids-binding domains, which is the source of its recognition ability. Especially, the focus on remote domain is also shown on DNA ligase of African swine fever virus.

Type	Dataset	$N_{protein}$	N_{posi}	N_{nega}	*Ratio*
DBS	DNA-573_Train	573	14,479	145,404	0.100
	DNA-129_Test	129	2,240	35,275	0.064
	DNA-181_Test	181	3,208	72,050	0.044
RBS	RNA-495_Train	495	14,609	122,290	0.119
	RNA-117_Test	117	2,031	35,314	0.058

Table 1: Composition of the training and testing data sets.

Methods

Benchmark datasets

As in the ESM-DBP, to retrain an RBP domain-adaptive PLM, we first collected 9,743,473 redundant RBPs (up to February 6, 2024) from UniProtKB database (Bateman, et al., 2019); then, to prevent the model from overfitting the RBP family with high redundancy, CD-HIT tool (Fu, et al., 2012) was used to remove those high similarity sequence using a cluster threshold of 0.4 and remaining 459,656 non-redundant RBPs named UniRBP40 as the pretraining data set.

To facilitate the verification of the performance of GeSite, two nucleic acids-protein binding datasets used extensively in previous studies are employed. Specifically, for DBS prediction, one training data set named DNA-573_Train and two independent test sets, i.e., DNA-129_Test and DNA-181_Test, are employed; for RBS prediction, one training data set RNA-495_Train and one test set RNA-117_Test are employed. Of these benchmark datasets, DNA-573_Train, DNA-129_Test, RNA-495_Train, and RNA-117_Test are collected from BioLip database (Yang, Roy, and Zhang, 2012; C. Zhang, et al., 2024) by Xia *et al*. in GraphBind (Xia, et al., 2021); DNA-181_Test is the newly released protein (from 6 December 2018 to August 2021) in BioLip database and collected by Yuan *et al.* in GraphSite (Yuan, et al., 2022). To ensure a fair and objective performance evaluation, the CD-HIT program with a cluster threshold of 0.4 was used to remove protein sequences that hold high similarities between the test set and the training set. The detailed components of these datasets are listed in Table 1.

Domain-adaptive protein language model

Recent advances in protein function and structure prediction based on PLM demonstrate that PLM learns amino acid dependencies to efficiently characterize protein sequence. This learning paradigm is largely inspired by the BERT-based large language model (LLM) in the field of natural language processing (NLP) (Devlin, et al., 2019). A recent study (Gururangan, et al., 2020) about NLP showed that domain-adaptive pretraining can provide significant gains in downstream task performance. This idea can be naturally transferred to PLM. In ESM-1b (Rives, et al., 2021), Rives *et al.* mentioned that the PLM after self-supervision learning encodes the MSA knowledge into the sequence representation. It is easy to imagine that if PLM is ulteriorly trained on particular protein families, the sequence characterization of these particular families will be further improved. In the previous study ESM-DBP, through domain-adaptive pretraining on massive DBP sequence data, the proposed DBP domain-adaptive PLM improves prediction performance and outperforms SOTA methods on several DBP-related tasks. Here, similar to ESM-DBP, to construct ESM-RBP, we continue to train the ESM2 model consisting of 33 transformer blocks with 650 million parameters by randomly masking and then predicting 15% residues of each sequence in UniRBP40 (see Figure 1A). Slightly different, since the number of protein sequences in UniRBP40 (459,656) is greater than UniDBP40 (170,264), we increase the number of updatable transformer blocks to 5 about 100 million parameters. The first 28 transformer blocks hold the fundamental biological knowledge that ESM2 learned from about 65 million general sequences in UniRef50, and the last 5 transformer blocks possess the RBP-specific knowledge learned through continued training from massive RBPs. In the pretraining phase, the cross-entropy and Adam optimizer are used to calculate the loss and update the parameters, respectively. Each input sequence consists of 512 tokens, short sequences are filled with token of <pad>, and sequences longer than 512 are split. The batch size is set to 230 according to the available memory. The model was optimized for ~35,000 steps over three days using two Tesla A40 GPUs with 48 GB memory each.

Protein Graph Representation

We converted the protein structure into a graph representation to learn spatial feature of target residue for NBS prediction. Specifically, a protein of length s is represented by a graph $\boldsymbol{G} = (\boldsymbol{V}, \boldsymbol{E})$, where $\boldsymbol{V} = \{v^0, \dots, v^i, \dots, v^{s-1}\}$ represents the set of all residue nodes; $e^{ij} \in \boldsymbol{E}$ represents the set of edges of interacting residues. The detailed descriptions of generation process of node and edge features are as follows.

Node feature. We treat each residue as a node in the graph $\boldsymbol{G}$. For each protein sequence of length s, we first input it into the ESM-D/RBP to generate an embedding matrix of size $s \times 1280$; then the Multiple Sequence Alignment (MSA) file is generated by searching Uniclust30 database (Mirdita, et al., 2017) using HHblits tool (Remmert, et al., 2012) and fed to ESM-MSA (Rao, et al., 2021) to obtain an embedding matrix of size $s \times 768$ for portraying the evolutionary information of protein (see Supplementary Text S2 in extend version); finally, these two embedding matrixes are concatenated into the feature representation matrix $\boldsymbol{f} = \{f_0, \dots, f_i, \dots, f_{s-1}\}$ of size $s \times 2{,}048$ of the target protein. Each feature of residue node v^i is a vector f_i of length 2,048.

Edge feature. Edges portray associations between nodes and are an important source of spatial information about neighboring residues. Here, we define edge for residue pairs that are close in spatial distance. In particular, if the Euclidean distance between the Cα atoms of two residues is less than 14Å, then they are considered to be in contact. The edge feature e^{ij} of target residue pair (R^i, R^j) is $|R^i, R^j|/D_{max}$, where $|R^i, R^j|$ means Euclidean distance between the Cα atoms of residues R^i and R^j; D_{max} means the maximum residue distance in the target protein.

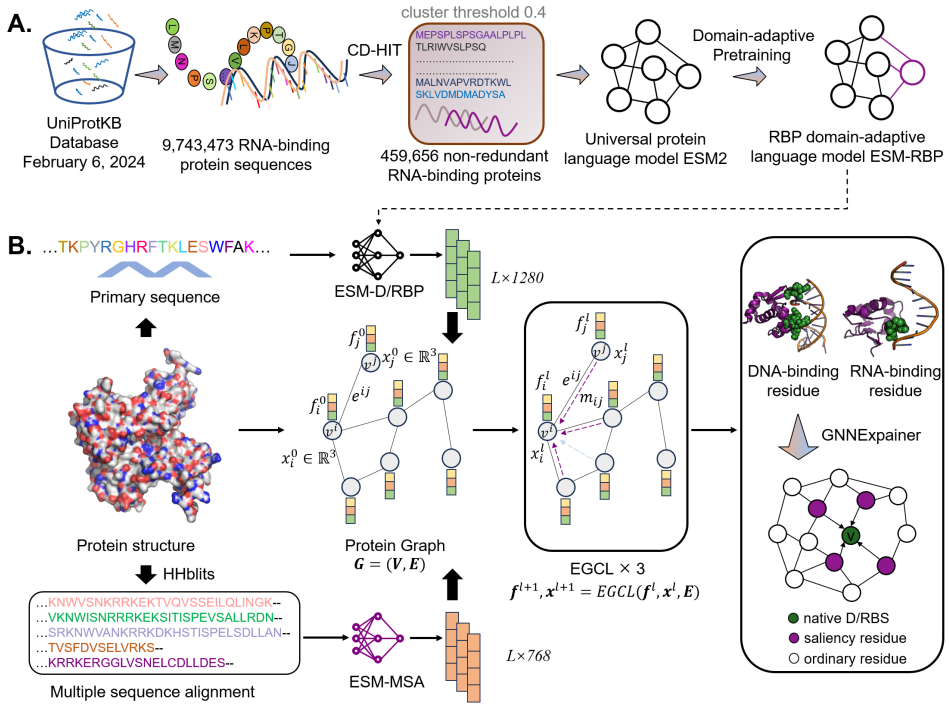

Figure 1: The overall architecture of GeSite. **(A).** flow chart of construction of domain-adaptive protein language model; **(B).** nucleic acid-binding residue prediction based on ESM-D/RBP and E(3) equivariant graph neural network. ESM-DBP (Zeng, et al., 2024) is the previous study, and ESM-RBP is trained de novo in this study.

EGNN also introduces coordinate feature x_i for each node v^i. Here, x_i is denoted by the 3D coordinates of Cα atom of residue v^i. EGNN retains equivariance to rotations and translations on coordinate set $\boldsymbol{x}$ and to permutations on node set $\boldsymbol{V}$ (Satorras, et al., 2021).

E(3) Equivariant Graph Neural Network

GNN and the variants are widely used to capture knowledge of protein 3D structure since they are adept at extracting spatial contextual embeddings of neighboring residues. The equivariant properties of EGNN in rotations, translations, reflections and alignments allow it to maintain the symmetry of the graph structure. In this study, EGNN is employed as prediction model to learn more abundant structure representation of protein than traditional graph convolutional network (GCN). See Figure 1B, EGNN is composed of three Equivariant Graph Convolutional Layers ($EGCL$) which takes the residue node feature set $\boldsymbol{f}^l$, edge set $\boldsymbol{E}$, and coordinate set $\boldsymbol{x}^l$ as input and performs a transformation $EGCL(\boldsymbol{f}^l, \boldsymbol{x}^l, \boldsymbol{E})$ as follows:

$$m_{ij} = \emptyset_e\left(f_i^l, f_j^l, \left\|x_i^l - x_j^l\right\|^2, e^{ij}\right) \quad (1)$$

$$x_i^{l+1} = x_i^l + C\sum_{j\neq i}\left(x_i^l - x_j^l\right)\emptyset_x(m_{ij}) \quad (2)$$

$$m_i = \sum_{j\neq i} m_{ij} \quad (3)$$

$$f_i^{l+1} = \emptyset_f\left(f_i^l, m_i\right) \quad (4)$$

where $\emptyset_e$ and $\emptyset_f$ mean edge and node operations respectively based on Multilayer Perceptron and $Swish()$ activation function which are similar to typical GCN; $C = 1/(s-1)$ means taking the average; $\emptyset_x = \{Linear() \rightarrow Swish() \rightarrow Linear()\}$ converts m_{ij} into a scalar value as the weight of relative difference $\left(x_i^l - x_j^l\right)$. The outputted node embedding $\boldsymbol{f}^{l+1}$ and coordinate set $\boldsymbol{x}^{l+1}$ are used as inputs of the next $EGCL()$. The involvement of the coordinate information in the updating of the node embedding is the main difference that distinguishes EGNN from traditional GCN, and is the source of its equivariance on rotations and translations.

The GeSite models are implemented using the Pytorch and the DGL frameworks (M. Wang, et al., 2019). Limited by available memory size, the batch size is set to 1, that is, each batch uses one protein graph for forward and back propagation. The cross-entropy and AdamW optimizer with a learning rate of 1e-4 are used to portray the loss and optimize parameters. Considering the category imbalance, the loss weights for the positive and negative samples are 0.7 and 0.3, respectively. To avoid overfitting, regularization with a coefficient of 1e-04 is employed to restrict the parameters. The entire training process lasted 50 epochs on a Tesla V100 GPU with the memory of 16G. Notably, our GeSite consists of two independent single-task models predicting DNA- and RNA-binding residues, respectively.

Test set	Feature	Spe	Rec	Pre	F_1	MCC	AUC	AP
DNA-129_Test	ESM2	0.948	0.618	0.431	0.508	0.480	0.927	0.518
	ESM-DBP	**0.956**	**0.638**	**0.481**	**0.549**	**0.522**	**0.941**	**0.563**
DNA-181_Test	ESM2	**0.931**	0.581	0.274	0.373	0.362	0.904	0.334
	ESM-DBP	0.925	**0.650**	**0.279**	**0.391**	**0.389**	**0.919**	**0.367**
RNA-117_Test	ESM2	0.838	**0.652**	0.188	0.293	0.285	0.837	0.245
	ESM-RBP	**0.909**	0.550	**0.258**	**0.352**	**0.326**	**0.861**	**0.271**

Table 2: Performance comparison of the sequence representation of original ESM2 and ESM-D/RBP on independent test sets.

Experiments

Role of Domain-adaptive Language Model

To demonstrate the advantages of the NBP domain-adaptive language models over ESM2 for the task of NBS prediction, we replace the ESM-D/RBP sequence embeddings in GeSite with the original ESM2 sequence embeddings and then retrain the GeSite models for comparison. From Table 2, the MCC values of GeSite using ESM-D/RBP feature embeddings on DNA-129_Test, DNA-181_Test, and RNA-117_Test are 0.522, 0.389, and 0.326, which are 8.75, 7.45, and 14.38% higher than those of GeSite using ESM2 sequence embeddings respectively. Considering other evaluation indexes, the model using ESM-D/RBP is also better than that using original ESM2 at prediction performance. For DNA-129_Test, the Spe, Rec, Pre, F_1, AUC, and AP values of the former are 0.948, 0.618, 0.431, 0.508, 8.07, 1.51, and 8.68% than those of the latter respectively.

In addition, for a more intuitionistic comparison at the protein-level, Figure 2 illustrates a head-to-head comparison of the MCC values in the three test sets. By looking at Figure 2, regardless of the test set, on most of the proteins, GeSite has better prediction results than the ESM2. For example, for DBS predictions, 77 of the 129 DBPs in DNA-129_Test have higher MCC values for GeSite than ESM2; for RBS prediction, 69 of the 117 RBPs in RNA-117_Test have higher MCC values for GeSite than ESM2. We also note that most of the RBPs (Figure 2C) are located closer to the bottom left than the DBPs (Figures 2A and 2B), implying that the overall predictive performance of the RBS is lower than that of the DBS. There are two main potential reasons for this: first, the number of DBPs (573) in the training set is higher than that of RBPs (495); and second, the binding patterns of proteins and RNAs are more complex to be captured well by the model. Nevertheless, in Figure 2C, it is still intuitively clear that GeSite using the ESM-RBP sequence embedding performs much better than that using the original ESM2 embedding representation. The above experimental results show that the domain-adaptive PLM pays attention to more in-depth identification knowledge of nucleic acid-binding patterns which provides a better sequence characterization of NBP after the domain-adaptive pretraining on a large number of nucleic acid-binding sequences, and thus improves the NBS prediction performance.

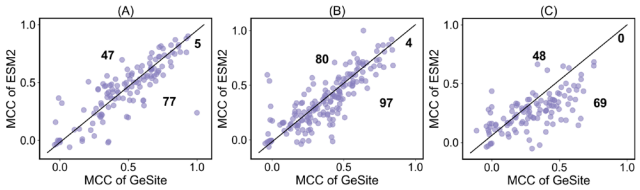

Figure 2: Head-to-head comparison of the MCC values of GeSite and ESM2 on the three test sets at the protein-level. Each dot represents a protein. **(A).** on DNA-129_Test; **(B).** on DNA-181_Test; **(C).** on RNA-117_Test.

Comparison with state-of-the-art methods

To future demonstrate the validity of the proposed GeSite for predicting NBS, 9 DBS predictors and 7 RBS predictors are employed as control.

Table 3 shows the detailed prediction results of these methods on three independent test sets, i.e., DNA-129_Test, DNA-181_Test, and RNA-117_Test. From Table 3, in both DBS and RBS prediction, GeSite demonstrates outstanding prediction performance that is superior to most of the other methods. Specifically, for DBS prediction, taking the DNA-129_Test as an example, the MCC value of GeSite is 0.522, which is 7.85, 0.58, 14.47, 34.19, 236.77, 19.72, 34.88, 51.3, and 198.29% higher than those of other methods, respectively. For RBS prediction, the MCC value of GeSite on RNA-117_Test is 0.326, achieving improvements of 35.83, 38.14, 94.05, 76.22, 1317.39, 77.17, and 171.66% over the control methods separately. The leading performance in other indicators also demonstrates the comprehensive performance of the proposed method.

To observe the difference between the proposed method and the existing methods, the PCC and *p*-value are calculated. Specifically, we calculate PCC using the probabilities that all residues in the test set are predicted to be NBSs. While for *p*-value, the probability of native NBS being predicted a positive sample is employed since the limited computational accuracy. The highest PCC of 7.31e-01 is given by hybridDBRpred on the DNA-129_Test, which is still quite a difference. The *p*-values against most methods are statistically significant, except that against ESM-NBR and CLAPE on DNA-129_Test are relatively high, which are 3.39e-02 and 7.85e-01 separately.

Test set	Predictor	Spe	Rec	Pre	F_1	MCC	AUC	AP	PCC	*p*-value
DNA-129_Test	GraphBind	0.948	0.625	0.434	0.512	0.484	0.916	0.497	6.91e-01	3.86e-42
	GraphSite	0.950	**0.665**	0.460	0.543	0.519	0.934	0.544	-	-
	ESM-NBR	0.971	0.463	0.511	0.486	0.456	0.893	0.483	3.34e-01	3.39e-02
	CLAPE	0.955	0.464	0.396	0.427	0.389	0.881	0.250	4.52e-01	7.85e-01
	DRNAPred	0.937	0.233	0.190	0.210	0.155	0.693	0.142	2.96e-01	6.02e-210
	ULDNA	0.925	0.647	0.355	0.459	0.436	0.907	0.462	6.96e-01	0.00e+00
	iDRNA-ITF	0.953	0.466	0.391	0.425	0.387	0.883	0.400	6.21e-01	3.28-e38
	DNApred	**0.988**	0.241	**0.564**	0.338	0.345	0.845	0.366	4.97e-01	1.81e-04
	hybridDBRpred	0.769	0.551	0.131	0.212	0.175	0.713	0.141	7.31e-01	4.33e-141
	GeSite	0.956	0.637	0.481	**0.549**	**0.522**	**0.941**	**0.563**	**-**	**-**
DNA-181_Test	GraphBind	0.949	0.505	0.304	0.380	0.357	0.893	0.317	-	-
	GraphSite	0.958	0.517	**0.354**	**0.420**	**0.397**	0.917	**0.369**	-	-
	ESM-NBR	0.952	0.472	0.305	0.371	0.345	0.857	0.324	6.41e-01	1.00e-12
	CLAPE	0.931	0.413	0.212	0.280	0.252	0.824	0.148	4.26e-01	2.80e-03
	DRNAPred	0.932	0.226	0.129	0.164	0.122	0.702	0.102	3.34e-01	1.89e-241
	ULDNA	0.906	0.611	0.225	0.329	0.327	0.877	0.289	6.93e-01	0.00e+00
	iDRNA-ITF	**0.968**	0.289	0.287	0.288	0.256	0.752	0.235	5.55e-01	5.38e-169
	DNApred	0.903	0.362	0.143	0.205	0.173	0.690	0.148	3.24e-01	8.47e-32
	hybridDBRpred	0.822	0.397	0.090	0.147	0.113	0.671	0.085	3.05e-01	5.22e-260
	GeSite	0.925	**0.651**	0.279	0.391	0.389	**0.919**	0.366	-	-
RNA-117_Test	CLAPE	0.642	**0.673**	0.097	0.171	0.148	0.718	0.134	3.96e-01	0.00e+00
	iDRNA-ITF	0.964	0.349	0.235	0.281	0.236	0.760	0.186	4.86e-01	5.39e-19
	GraphBind	0.936	0.303	0.171	0.218	0.168	0.718	0.268	6.06e-01	1.81e-07
	ESM-NBR	0.939	0.271	0.204	0.233	0.185	0.783	0.190	3.98e-01	2.92e-44
	DRNAPred	0.971	0.085	0.045	0.059	0.023	0.489	0.058	3.85e-01	0.00e+00
	hybridRNAbind	**0.973**	0.179	**0.266**	0.214	0.184	0.704	0.176	2.24e-01	9.54e-80
	Pprint2	0.922	0.225	0.143	0.175	0.120	0.578	0.107	1.72e-01	1.10e-03
	GeSite	0.909	0.550	0.258	**0.352**	**0.326**	**0.861**	**0.271**	**-**	**-**

Table 3: Performance comparison of GeSite and the SOTA NBS prediction methods on three independent test sets.

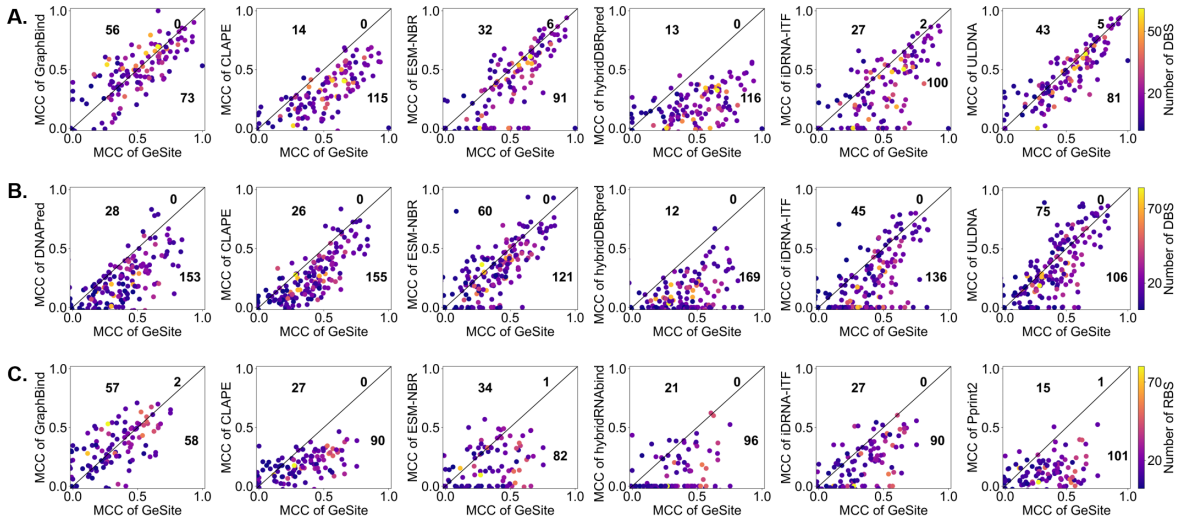

Figure 3: Comparison among GeSite and SOTA methods. (**A**). DNA-129_Test, (**B**). DNA-181_Test, (**C**). RNA-117_Test.

Additionally, GeSite performs better at protein-level on the vast majority of proteins regardless of the test set (see Figure 3). For DBS prediction on DNA-181_Test, out of 181 DBPs, there are 153, 155, 121, 169, 136, and 106 cases where GeSite possesses a higher MCC value than six control methods, respectively. For RBS prediction, out of 117 RBPs, there are 58, 90, 82, 96, 90, and 101 cases where GeSite outperforms other predictors separately. These results highlight the excellent performance of the proposed method at the single protein-level. A detailed description of the sources of the prediction results of control methods sees Supplementary Text S3 in extend version.

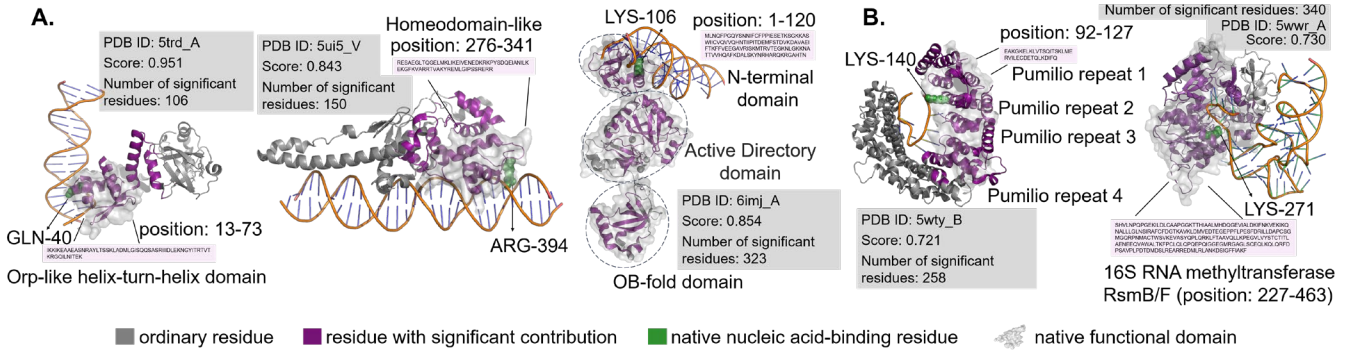

Figure 4: GeSite is enlightened by native nucleic acid-binding domains. Residues selected by GNNExplainer that contribute significantly to the prediction of target NBS (green spheres) are highlighted in purple. The region wrapped by the protein surface is the recorded nucleic acid-binding and related functional domain. (**A**). on three DBS cases; (**B**). on two RBS cases.

Interpretability and visualization

Generally, the ability of proteins to bind nucleic acids derives primarily from a small conserved nucleic acid-binding domain (NBD) capable of highly specific recognition of nucleic acid sequences. The GeSite backbone neural network, EGNN, can learn valid knowledge related to spatially neighboring nodes of a single node. Theoretically, the model can perceive the specific knowledge of the diverse NBDs from the protein graph, thus enabling accurate NBS recognition.

We selected three DBSs (Figure 4A) and two RBSs (Figure 4B) as representatives to find out the key residue nodes for the prediction of these five NBSs from test sets using GNNExplainer (Ying, et al., 2019). Apparently, those regions that are significant for identification cover the native NBD. For DBS prediction task, GeSite successfully identifies the DBS GLN-40 of 5trd_A as a positive with a score of 0.951. The GNNExplainer algorithm reports 106 salient residues (marked in purple) close to the DNA chain that cover a typical DNA-binding domain (DBD) named Orp-like helix-turn-helix (HTH) domain located at position 13-73 (coated with protein surface). For the residue ARG-394 of 5ui5_V, GeSite successfully identified it with a probability of 0.843. Similarly, the saliency region of this site, consisting of 150 residues, is highly overlapping with a Homeodomain-like region located at position 376-341. The HTH domain is widely present in a variety of prokaryotic and eukaryotic organisms and plays a fundamental regulatory role (Aravind, et al., 2005). The Homeodomain is also known as a classic DBD in eukaryotes. These two cases exemplify the effective utilization of GeSite of near single structure domain discriminative information. In the chain A of 6imj (DNA ligase of African swine fever virus), GeSite presents the attention for multi-domains, both remote and close. Concretely, in the third subfigure of Figure 4A, 6imj_A contains three functional domains, namely N-terminal domain, Active Directory domain, and OB-fold domain, whose positive effects on DNA-protein interactions have been demonstrated in previous study (Chen, et al., 2019). Clearly, the significant residues for prediction of LYS-106 are distributed in all three domains even though the OB-fold domain is far from LYS-106. The ability comes primarily from the ability of the message-passing mechanism of EGNN to allow the network to capture features from remote nodes.

For RBS prediction task, similar phenomena are observed. For instance, a prediction score of 0.721 on residue LYS-140 of 5wty_B indicates that GeSite correctly predicted it as an RBS. This chain contains multiple pumilio repeat regions that regulate gene expression by specifically recognizing and binding to RNA sequences (Edwards, et al., 2001). The 258 residues with significant contribution for the recognition of LYS-140 cover the pumilio repeat 1 to 4. Likewise, for LYS-271 of chain A of 5wwr, the purple prominent structure overlaps with 16S RNA methyltransferase RsmB/F region. These findings suggest that, like DBS prediction, GeSite performs identification through the discriminatory information of the RNA-binding domain in the spatial context of the target site. Overall, the above study conveys the idea that the proposed GeSite is inspired by the NBD in the periphery of the target NBS to be predicted to acquire discriminative knowledge and thus perceive the pattern of nucleic acid-binding. This capability can be extended to multiple functional domains in remote locations.

Conclusions

In this paper, we propose GeSite based on nucleic acid-binding protein domain-adaptive protein language model and E(3)-equivariant graph neural network for accurately predicting nucleic acid-binding residue. Predicted results on multiple test sets demonstrate the excellent performance of GeSite. Meanwhile, interpretability analysis on graph neural networks uncovers that the prediction model captures key information about nucleic acid-binding domains thereby helping to identify native nucleic acid-binding residue.

Acknowledgments

This work was supported by NSFC-FDCT Grants 62361166662; NSFC Grants U19A2067; National Key R&D Program of China 2022YFC3400400; Key R&D Program of Hunan Province 2023GK2004, 2023SK2059, 2023SK2060; Top 10 Technical Key Project in Hunan Province 2023GK1010; Key Technologies R&D Program of Guangdong Province (2023B1111030004 to FFH). The Funds of State Key Laboratory of Chemo/Biosensing and Chemometrics, the National Supercomputing Center in Changsha (http://nscc.hnu.edu.cn/), and Peng Cheng Lab.

References

Abramson, J.; Adler, J.; Dunger, J.; Evans, R.; Green, T.; Pritzel, A.; Ronneberger, O.; Willmore, L.; Ballard, A. J.; Bambrick, J., et al. 2024. Accurate structure prediction of biomolecular interactions with AlphaFold3. *Nature,* 630(8016), 493-500.

Aravind, L.; Anantharaman, V.; Balaji, S.; Babu, M. M., and Iyer, L. M. 2005. The many faces of the helix-turn-helix domain: transcription regulation and beyond. *FEMS microbiology reviews,* 29(2), 231-262.

Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G. R.; Wang, J.; Cong, Q.; Kinch, L. N.; Schaeffer, R. D., et al. 2021. Accurate prediction of protein structures and interactions using a three-track neural network. *Science,* 373(6557), 871-876.

Bateman, A.; Martin, M. J.; Orchard, S.; Magrane, M.; Alpi, E.; Bely, B.; Bingley, M.; Britto, R.; Bursteinas, B.; Busiello, G., et al. 2019. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research,* 47(D1), D506-D515.

Chen, Y.; Liu, H.; Yang, C.; Gao, Y.; Yu, X.; Chen, X.; Cui, R.; Zheng, L.; Li, S.; Li, X., et al. 2019. Structure of the error-prone DNA ligase of African swine fever virus identifies critical active site residues. *Nature Communications,* 10(1), 387.

Devlin, J.; Chang, M. W.; Lee, K., and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Conference of the North-American-Chapter of the Association-for-Computational-Linguistics-Human Language Technologies (NAACL-HLT)*, 4171-4186.

Dey, R., and Salem, F. M. 2017. Gate-variants of gated recurrent unit (GRU) neural networks. In *IEEE 60th international midwest symposium on circuits and systems (MWSCAS)*, 1597-1600.

Edwards, T. A.; Pyle, S. E.; Wharton, R. P., and Aggarwal, A. K. 2001. Structure of Pumilio reveals similarity between RNA and peptide binding motifs. *Cell,* 105(2), 281-289.

Elnaggar, A.; Heinzinger, M.; Dallago, C.; Rehawi, G.; Wang, Y.; Jones, L.; Gibbs, T.; Feher, T.; Angerer, C., and Steinegger, M. 2021. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence,* 44(10), 7112-7127.

Fu, L.; Niu, B.; Zhu, Z.; Wu, S., and Li, W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics,* 28(23), 3150-3152.

Gururangan, S.; Marasović, A.; Swayamdipta, S.; Lo, K.; Beltagy, I.; Downey, D., and Smith, N. A. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Humphreys, I. R.; Pei, J.; Baek, M.; Krishnakumar, A.; Anishchenko, I.; Ovchinnikov, S.; Zhang, J.; Ness, T. J.; Banjade, S., and Bagde, S. R. 2021. Computed structures of core eukaryotic protein complexes. *Science,* 374(6573), eabm4805.

Jing, L.; Xu, S.; Wang, Y.; Zhou, Y.; Shen, T.; Ji, Z.; Fang, H.; Li, Z., and Sun, S. 2024. CrossBind: Collaborative Cross-Modal Identification of Protein Nucleic-Acid-Binding Residues. In *Proceedings of the AAAI Conference on Artificial Intelligence,* 38(3), 2661-2669.

Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A., and Potapenko, A. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature,* 596(7873), 583-589.

Krishna, R.; Wang, J.; Ahern, W.; Sturmfels, P.; Venkatesh, P.; Kalvet, I.; Lee, G. R.; Morey-Burrows, F. S.; Anishchenko, I.; Humphreys, I. R., et al. 2024. Generalized biomolecular modeling and design with RoseTTAFold All-Atom. *Science,* 384(6693), eadl2528.

Lambert, S. A.; Jolma, A.; Campitelli, L. F.; Das, P. K.; Yin, Y.; Albu, M.; Chen, X.; Taipale, J.; Hughes, T. R., and Weirauch, M. T. 2018. The human transcription factors. *Cell,* 172(4), 650-665.

Leinonen, R.; Diez, F. G.; Binns, D.; Fleischmann, W.; Lopez, R., and Apweiler, R. 2004. UniProt archive. *Bioinformatics,* 20(17), 3236-3237.

Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y., et al. 2023. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science,* 379(6637), 1123-1130.

Liu, Y., and Tian, B. 2024. Protein-DNA binding sites prediction based on pre-trained protein language model and contrastive learning. *Briefings in Bioinformatics,* 25(1), bbad488.

Mirdita, M.; Von Den Driesch, L.; Galiez, C.; Martin, M. J.; Söding, J., and Steinegger, M. 2017. Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Research,* 45(D1), D170-D176.

Narayanan, B. C.; Westbrook, J.; Ghosh, S.; Petrov, A. I.; Sweeney, B.; Zirbel, C. L.; Leontis, N. B., and Berman, H.

M. 2014. The Nucleic Acid Database: new features and capabilities. *Nucleic Acids Research,* 42(D1), D114-D122.

Patiyal, S.; Dhall, A.; Bajaj, K.; Sahu, H., and Raghava, G. P. S. 2022. Prediction of RNA-interacting residues in a protein using CNN and evolutionary profile. *Briefings in Bioinformatics,* 24(1), bbac538.

Rao, R. M.; Liu, J.; Verkuil, R.; Meier, J.; Canny, J.; Abbeel, P.; Sercu, T., and Rives, A. 2021. MSA Transformer. In *Proceedings of the 38th International Conference on Machine Learning*, 139, 8844-8856.

Remmert, M.; Biegert, A.; Hauser, A., and Söding, J. 2012. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature methods,* 9(2), 173-175.

Rives, A.; Meier, J.; Sercu, T.; Goyal, S.; Lin, Z. M.; Liu, J. S.; Guo, D. M.; Ott, M.; Zitnick, C. L.; Ma, J., et al. 2021. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. In *Proceedings of the National Academy of Sciences of the United States of America,* 118(15), e2016239118.

Roche, R.; Moussad, B.; Shuvo, M. H.; Tarafder, S., and Bhattacharya, D. 2024. EquiPNAS: improved protein-nucleic acid binding site prediction using protein-language-model-informed equivariant deep graph neural networks. *Nucleic Acids Research,* 52(5), e27.

Satorras, V. G.; Hoogeboom, E., and Welling, M. 2021. E (n) equivariant graph neural networks. In *International conference on machine learning,* 9323-9332.

Tunyasuvunakool, K.; Adler, J.; Wu, Z.; Green, T.; Zielinski, M.; Žídek, A.; Bridgland, A.; Cowie, A.; Meyer, C., and Laydon, A. 2021. Highly accurate protein structure prediction for the human proteome. *Nature,* 596(7873), 590-596.

Wang, M.; Yu, L.; Zheng, D.; Gan, Q.; Gai, Y.; Ye, Z.; Li, M.; Zhou, J.; Huang, Q.; Ma, C., et al. 2019. Deep Graph Library: Towards Efficient and Scalable Deep Learning on Graphs. *ArXiv, abs/1909.01315*.

Wang, N.; Yan, K.; Zhang, J., and Liu, B. 2022. iDRNA-ITF: identifying DNA- and RNA-binding residues in proteins based on induction and transfer framework. *Briefings in Bioinformatics,* 23(4), bbac236.

Xia, Y.; Xia, C. Q.; Pan, X. Y., and Shen, H. B. 2021. GraphBind: protein structural context embedded rules learned by hierarchical graph neural networks for recognizing nucleic-acid-binding residues. *Nucleic Acids Research,* 49(9), e51.

Yan, J., and Kurgan, L. 2017. DRNApred, fast sequence-based method that accurately predicts and discriminates DNA- and RNA-binding residues. *Nucleic Acids Research,* 45(10), e84.

Yang, J.; Roy, A., and Zhang, Y. 2012. BioLiP: a semi-manually curated database for biologically relevant ligand–protein interactions. *Nucleic Acids Research,* 41(D1), D1096-D1103.

Ying, Z.; Bourgeois, D.; You, J.; Zitnik, M., and Leskovec, J. 2019. Gnnexplainer: Generating explanations for graph neural networks. In *Advances in Neural Information Processing Systems,* 32.

Yuan, Q.; Chen, S.; Rao, J.; Zheng, S.; Zhao, H., and Yang, Y. 2022. AlphaFold2-aware protein-DNA binding site prediction using graph transformer. *Briefings in Bioinformatics,* 23(2), bbab564.

Zeng, W.; Dou, Y.; Pan, L.; Xu, L., and Peng, S. 2024. Improving prediction performance of general protein language model by domain-adaptive pretraining on DNA-binding protein. *Nature Communications,* 15(7838).

Zeng, W.; Lv, D.; Liu, X.; Chen, G.; Liu, W., and Peng, S. 2023. ESM-NBR: fast and accurate nucleic acid-binding residue prediction via protein language model feature representation and multi-task learning. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 76-81.

Zhang, C.; Zhang, X.; Freddolino, P. L., and Zhang, Y. 2024. BioLiP2: an updated structure database for biologically relevant ligand-protein interactions. *Nucleic Acids Research,* 52(D1), D404-D412.

Zhang, F.; Li, M.; Zhang, J., and Kurgan, L. 2023. HybridRNAbind: prediction of RNA interacting residues across structure-annotated and disorder-annotated proteins. *Nucleic Acids Research,* 51(5), e25-e25.

Zhang, J.; Basu, S., and Kurgan, L. 2024. HybridDBRpred: improved sequence-based prediction of DNA-binding amino acids using annotations from structured complexes and disordered proteins. *Nucleic Acids Research,* 52(2), e10-e10.

Zhu, Y.-H.; Liu, Z.; Liu, Y.; Ji, Z., and Yu, D.-J. 2024. ULDNA: integrating unsupervised multi-source language models with LSTM-attention network for high-accuracy protein-DNA binding site prediction. *Briefings in Bioinformatics,* 25(2), bbae040.

Zhu, Y. H.; Hu, J.; Song, X. N., and Yu, D. J. 2019. DNAPred: Accurate Identification of DNA-Binding Sites from Protein Sequence by Ensembled Hyperplane-Distance-Based Support Vector Machines. *Journal of Chemical Information and Modeling*, 59(6), 3057-3071.