Supplementary Information

ESM-NBR: fast and accurate nucleic acid-binding residue prediction via protein language model feature representation and multi-task learning

Wenwu Zeng[1], Dafeng Lv [1], Xuan Liu [1], Guo Chen [1], Wenjuan Liu[1,*] and Shaoliang Peng[1,*]

[1] College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China;

[*] Address correspondence to S.L. Peng at slpeng@hnu.edu.cn or W.J. Liu at liuwenjuan89@hnu.edu.cn

Supporting Texts

Text S1. Benchmark Datasets

Two pairs of widely used mixed datasets of DBPs and RBPs are employed to evaluate the proposed methods fairly and comprehensively. The first one named YK17 is collected by Yan *et al*. [1] from the Protein Data Bank (PDB) database [2], which contains a training subset (denoted as YK17-Tr) and an independent test subset (denoted as YK17-Tst). The second one (denoted as DRNA-1314) is constructed by Xia *et al*. [3] from the BioLip database [4] which composed of four subsets, i.e., DNA-573_Train, RNA-495_Train, DNA-129_Test, and RNA-117_Test. In this study, DNA-573_Train and RNA-495_Train (or DNA-129_Test and RNA-117_Test) are combined as training data set abbreviated as DRNATr-1068 (or independent test set abbreviated as DRNATst-246) of multi-task model. In these two datasets, the sequence identities between the protein chains in the test set and the protein chains in the training set are less than 30% [1, 5]. The detailed compositions are listed in Table S1.

Table S1. Composition of the training and testing data sets

Data set	Subset	No. protein	No. DBR [a]	No. RBR [a]	No. non-NBR [a]
YK17	YK17-Tr	488	7,764	4,684	90,594
	YK17-Tst	82	955	807	17,119
DRNA-1314	DNA-573_Train	573	14,479	0	145,404
	RNA-495_Train	495	0	14,609	122,290
	DNA-129_Test	129	2,240	0	35,275
	RNA-117_Test	117	0	2,031	35,314

a. "DBR", "RBR", and "non-NBR" mean the residues binding to DNA, residues binding to RNA, and residues neither bind to nucleic acid residue nor are disordered residue, respectively.

Text S2. Architecture of ESM-NBR

In this study, based on the feature extracted from large protein language model ESM2 and the multi-task BiLSTM-based network, a novel nucleic acid-binding residues prediction method, named ESM-NBR, is proposed and implemented. The overall architecture of ESM-NBR is shown in Figure S1. It is easy to see that the workflow of ESM-NBR can be roughly divided into three steps:

Step 1: For each protein primary sequence in the dataset, features containing knowledge of important biochemical attributes are generated by feeding it into the ESM2 model;

Step 2: The well-generated ESM2 feature representation is first inputted into the network shared by DNA- and RNA-binding residues composed of stacked BiLSTM and MLP to learn common knowledge;

Step 3: Two private MLP blocks are employed to learn essential authentication information for DNA- and RNA-binding residues identification, respectively. The outputs of the final linear layers are used as the predictive probabilities to determine whether a residue is binding to DNA or RNA.

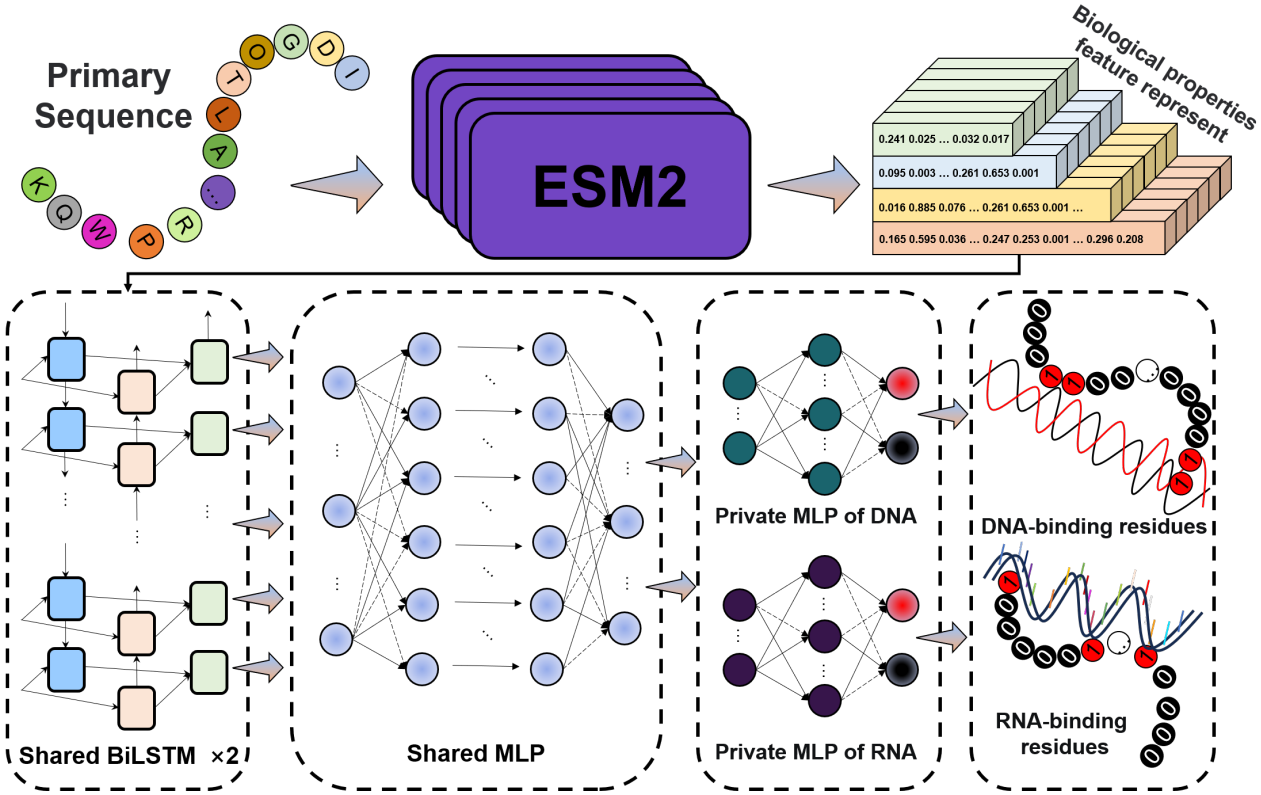

Figure. S1. Architecture of ESM-NBR.

Text S3. Architecture of multi-task neural network

Various neural networks like BiGRU, CNN, and graph transformer [6] have been utilized by researchers to identify nucleic acid-binding residue [7] [8] [9]. These networks are usually chosen to better match the input features. For example, in GraphSite [9], to capture spatial information, Yuan *et al.* employed a graph transformer model with predicted protein 3D structural from AlphaFold2 as feature, which takes the protein structural information into account. Here, considering the small training set and the long ESM2 features, we used the relatively lightweight BiLSTM and MLP as prediction model, which is suitable for capturing the long-term dependence of sequences and is not prone to overfitting. Concretely, for each input matrix of size $L \times M$ where L and M mean sequence length and feature dimension, respectively, we first feed it into two stacked BiLSTM layers whose size of hidden layer is set to 100; then, the output of each time step of second BiLSTM of length 200 is fed into the later three MLP layers shared by DNA and RNA; finally, the output of the shared layer is fed into two separate MLP blocks contained three linear layers, which are specifically designed to predict DNA- and RNA-binding residues, respectively. The final prediction results of DNA- and RNA-binding residues can be determined according to the outputs of two separate MLP blocks. Multi-task model is more space- and time-efficient than single-task model and is relatively less prone to overfitting due to the need to fit multiple labels simultaneously. The number of parameters need to be learned of the whole model are 872,404, 1,000,404, 1,128,404, 1,640,404, and 2,664,404 when lengths of input features are 320, 480, 640, 1280, and 2,560, respectively.

During the training phase of the model, the cross-entropy loss is used to calculated by follow functions:

$$loss = \frac{1}{N}\sum_{i=1}^{2}\sum_{j=1}^{N} L_{i,j} + \frac{\lambda}{2N}\sum_{t=1}^{M} {w_t}^2 \quad (S1)$$

$$L_{i,j} = \begin{cases} y_{i,j} \times \log(p_{i,j}) + (1 - y_{i,j}) \times \log(1 - p_{i,j}), & y_{i,j} \geq 0 \\ 0, & y_{i,j} < 0 \end{cases} \quad (S2)$$

where N means the total number of residues in the training data set; $y_{1,j}$ and $y_{2,j}$ are the true DNA and RNA labels of the j-th residue, respectively; $p_{1,j}$ and $p_{2,j}$ are the predicted probabilities that the target residue is predicted to be a positive sample; λ is the L2 regularization factor and is set to 0.0001; M represents the total number of model parameters. Since the unknown residues in the disordered region do not participate in the loss calculation, their labels are set to -1. The Adam algorithm [10] is used to optimize the loss function with a learning rate of 0.0001. To prevent overfitting and enhance fitting ability, the dropout algorithm [11] with a dropout rate of 0.5 and RLUE function are applied to all BiLSTM and MLP layers, respectively. All the training process was done on a Tesla V100 with 16G of memory.

Text S4. Evaluation Indexes

Four evaluation indexes, i.e., Matthew's correlation coefficient (MCC), the area under Receiver Operating Characteristic curve (AUC), the area under Precision-Recall curve (AP), and the area under Cross Prediction Rate-True Prediction Rate (CPR-TPR) curve (AURC) [12], are utilized to assess the proposed ESM-NBR. MCC is a threshold-related index for the comprehensive assessment of unbalanced datasets; AUC is threshold-independent and indicates overall prediction performance of negative and positive samples. In contrast to AUC, AP mainly focuses on prediction performance of positive samples; The AURC applied on cross-prediction problem is used to indicate the proportion of native RNA-binding residue that are mistakenly predicted to be DNA-binding residue or native DNA-binding residue that are mistakenly predicted to be RNA-binding residue. Out of four indexes, only the smaller the AURC the better the model prediction performance. Besides the above four indexes, the Pearson correlation coefficient (PCC) and p-value in Student's t-test are employed to indicate the linear correlation degree and the difference between ESM-NBR and other methods.

Text S5. The impact of features generated by ESM2 models with different parameter levels on performance

A number of models with different parameter levels were constructed and compared for extracting important biochemical property knowledge hidden in protein sequence as fully as possible in the study of ESM2 [13]. Since to the extremely complex mapping of protein sequence and structure, the proteins 3D structure at the atomic-level can only be predicted with high accuracy when the number of parameters of the ESM2 model reaches 15 billion. On the nucleic acid-binding residue prediction problem, we do not necessarily use such a large model-generated feature representation as input taking into account possible redundant information. Here, to select features at the appropriate scale, we perform the 10-fold cross validation test on the DRNATr-1068 and YK17-Tr using feature representations generated by ESM2 models with 8 million, 35 million, 150 million, 650 million, and 3 billion parameters on the model in Figure S1, respectively. By looking at Figure S2, it is easy to see that prediction performance is generally poor at lower parameter levels, such as 8 million and 35 million. Obviously small models do not contain enough capacity to learn the vast knowledge of protein sequence space. The features generated by the small model do not contain enough biochemical attributes to accurately predict nucleic acid-binding residues. The overall performance gets progressively better as the number of parameters increases, and the best results are achieved in several indexes when the number of parameters reaches 650 million. For example, in the term of RNA-binding residue prediction, both the MCC and AP values of the feature of model at 650 million level are greater than those of other features, despite the AUC value of the feature of model at 3 billion level is slightly higher than it. Such result suggests that when the number of model parameters is too large, the redundant information in them may lead to deterioration in the performance of the downstream prediction task. By looking at the AURC index marked in purple star, similarly, it reaches lowest on the RNA-binding residue prediction (or DNA-binding prediction) of YK17-Tr (DRNATr-1068) when parameters of ESM2 model at 650 million level. This suggests that DNA-binding residue and RNA-binding residue are well differentiated and that the network learns knowledge about RNA-specificity and DNA-specificity. In conclusion, the feature of ESM2 model at 650 million parameter level is able to predict NBR well in this study, and considering the computational power requirements of larger models, it is appropriate to use it as the input feature of ESM-NBR network.

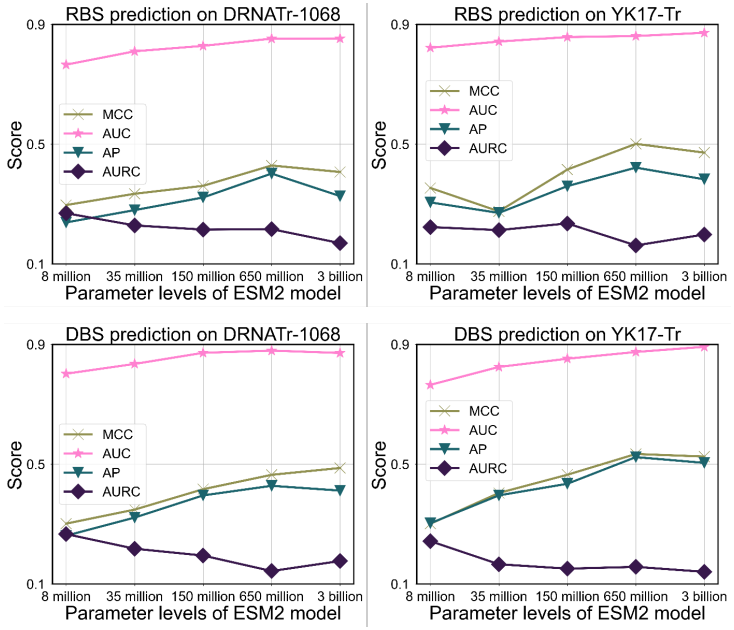

Figure. S2. Prediction performance changes of features generated by ESM2 models with different parameter levels on DRNATr-1068 and YK17-Tr over a 10-fold cross validation test.

Text S6. Performance comparison of ESM2 feature and evolution information feature

To demonstrate the efficacy of the ESM2 feature representation on nucleic acid-binding residue, the widely used evolutionary information feature of HMM (against to the Uniclust30 database with a e-value of 0.1) is utilized as a control. Specifically, the HMM feature, ESM2 feature generated by model contained 650 million parameters (Abbreviated as ESM2_650M), and their combination are severally as the input feature of ESM-NBR network for training. The prediction performance on DRNATr-1068 and YK17-Tr over a 10-fold cross validation test are shown in Figure S3 and Table S2. In Table S2, it is obvious that ESM2_650M comprehensively outperforms HMM on both DNA and RNA. Concretely, take results on DRNATr-1068 as an example, the values of MCC, AUC, and AP of ESM2_650M are 0.534, 0.874, and 0.523 (or 0.501, 0.861, and 0.421) considering DNA-binding residue (or RNA-binding residue), which are 90.71, 12.48, and 144.39% (or 187.93, 16.03, and 158.28%) higher than those of HMM respectively. Considering cross-prediction performance, the AURC values of DNA and RNA of ESM2_650M on YK17-Tr are 0.143 and 0.216, which are 148.95 and 115.27% lower than those of HMM separately, indicating ESM2_650 can better distinguish between DNA-binding residue and RNA-binding residue compared to HMM. It is worth that the p-values between these two features are so small that the computer cannot calculate it, which means there is a big difference between them. In addition, we can find that the degree of linear correlation between prediction results of these two features is low by visiting PCC values, which further proves the difference between them. By looking at the results of combination feature of ESM2_650M and HMM, the prediction performance does not necessarily get better by simply splicing these two features. For example, the MCC and AP values of DNA-binding residue of ESM2_650M on DRNATr-1068 are 0.534 and 0.523, which are 4.70 and 12.23% higher than those of combination feature respectively. The PCC and p-value between ESM2_650M and combination feature show much smaller differences compared to the single HMM. There is no doubt that ESM2_650M dominates in combination feature. Figure S3 shows the PR, CPR-TPR, and ROC curves with corresponding AP, AURC, and AUC values separately. The solid and dashed lines indicate the predicted results for DNA and RNA, respectively. It is intuitive that the PR and ROC curves of HMM are much lower compared to ESM2_650M, and the CPR-TPR curve is much higher especially on RNA-binding residue of YK17-Tr whose AURC is 0.659. Overall, since combination feature consists mainly of ESM2_650M, their curves do not differ much. To sum up, ESM2_650M feature far exceeds the HMM for nucleic acid-binding residue prediction and shows significant difference with it in this section. Since most of the previous methods rely on evolutionary information features heavily, this result provides a new thinking of studying protein-nucleic acid interactions.

Table S2. Performance of ESM2 feature and HMM on two data sets over 10-fold cross validation

Dataset	Feature	DNA-binding residue						RNA-binding residue					
		MCC	AUC	AP	AURC	*p*-value [a]	PCC [b]	MCC	AUC	AP	AURC	*p*-value	PCC
DRNATr-1068	ESM2_650M	**0.534**	0.874	**0.523**	0.157	-	-	**0.501**	0.861	0.421	**0.162**	-	-
	HMM	0.280	0.777	0.214	0.214	4.81e-04	3.56e-01	0.174	0.742	0.163	0.659	N/A [c]	2.94e-01
	Combination	0.510	**0.881**	0.466	**0.143**	9.04e-61	7.80e-01	0.460	**0.874**	**0.432**	0.164	7.82e-42	7.30e-01
YK17-Tr	ESM2_650M	0.464	**0.879**	0.427	**0.143** [d]	-	-	0.428	**0.852**	**0.402**	0.216	-	-
	HMM	0.180	0.745	0.127	0.356	N/A	3.04e-01	0.265	0.761	0.214	0.465	N/A	4.92e-01
	Combination	**0.475**	0.871	**0.437**	0.157	2.37e-35	5.49e-02	**0.438**	0.849	0.387	**0.203**	5.49e-02	8.27e-01

a. The *p*-values in Student's *t*-test are calculated for the differences between ESM2_650M and other features using the probabilities that each residue is predicted to be a positive sample.

b. The PCC values are calculated for the linear correlation coefficient between ESM2_650M and other features using the probabilities that each residue is predicted to be a positive sample.

c. "N/A" means the value is so small that our computer can't figure it out.

d. Bolded font indicates the best result.

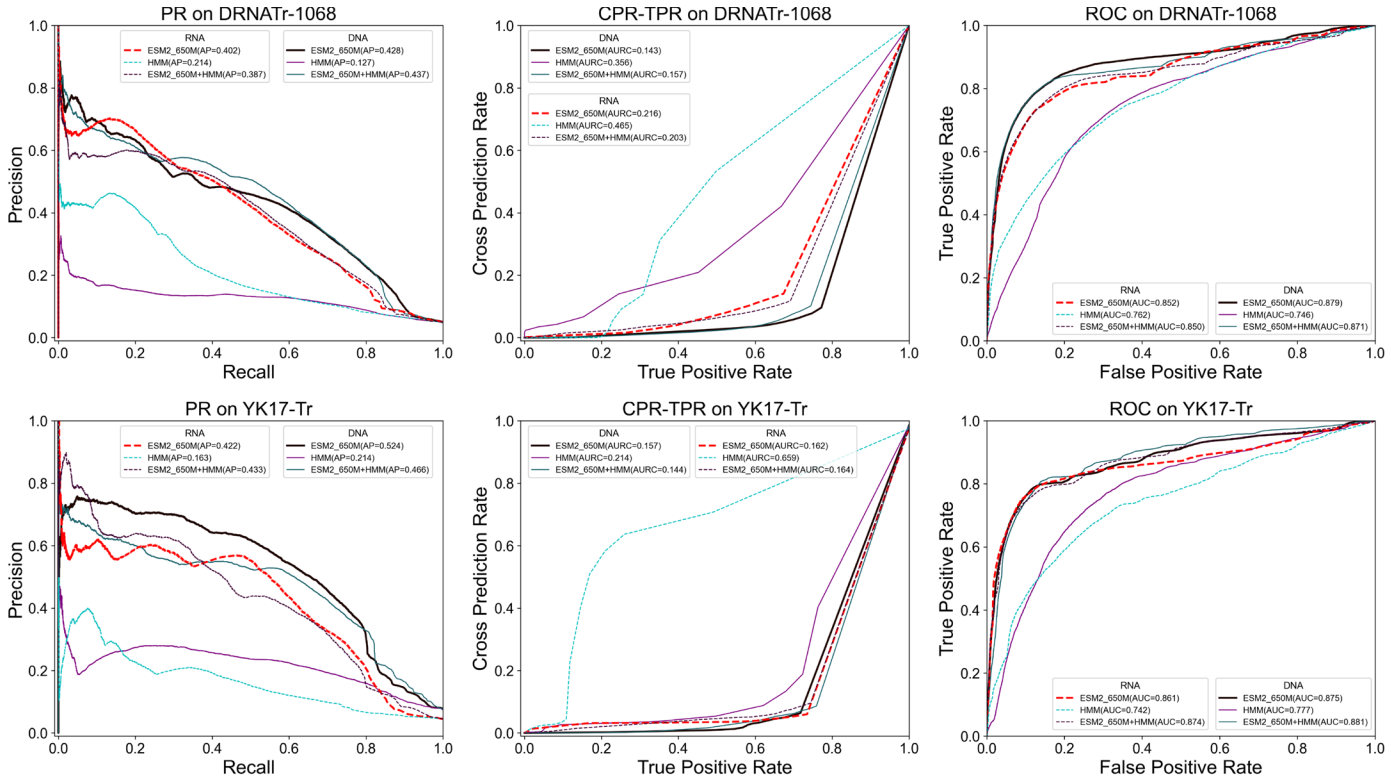

Figure. S3. RP, CPR-TPR, and ROC curves of ESM2_650M, HMM, and their combination on DRNATr-1068 and YK17-Tr over a 10-fold cross validation test. The solid and dashed lines indicate the predicted results for DNA- and RNA-binding residues, respectively. The higher the AUC and AP the better the prediction performance. The lower the AURC the better the prediction performance.

Text S7. Performance comparison of models for multi-task and single-task

There are shared and particular properties of DNA- and RNA-binding residues which should be able to be learnt simultaneously by the neural network. In this section, to investigate the effectiveness of multi-task network for nucleic acid-binding residues prediction by learning common and private knowledge, the prediction of multi-task and single-task models are performed on DRNATst-246 and YK17-Tst, respectively. Detail prediction results are shown in Table S3. By looking at the data on YK17-Tst, we can know that performance of multi-task model is outperforms that of single-task model both in terms of DNA- and RNA-binding residue prediction. Take DNA-binding residue for example, the MCC, AUC, and AP values of multi-task model are 0.391, 0.881, and 0.350, which are 7.41, 1.26, and 0.28 percent higher than those of single-task model respectively. In contrast, the predictive performance of the single-task model is better than that of the multi-task model across the board on the DRNATst-246. For example, the MCC, AUC, and AP values of single-task model on DNA-binding residue of DRNATst-246 are 0.474, 0.923, and 0.526, which are 4.17, 3.24, and 9.12 percent higher than those of multi-task model respectively. Completely opposite prediction performances on the two datasets caught our attention. By investigating the two training sets, i.e., DRNATr-1068 and YK17-Tr, we find that there are 12 proteins in YK17-Tr bind both DNA and RNA, containing 336 DNA-binding residues and 187 RNA-binding residues, whereas proteins in DRNATr-1068 bind only one of DNA and RNA. That is say, there more complementary information about DNA-binding and RNA-binding patterns in YK17-Tr, which is more suitable for multi-task model to learn. The *p*-value and PCC indexes also show also demonstrate the difference in prediction results between the two models, suggesting that both two tasks have learnt useful knowledge from each other to aid their own predictions compared to the single-task model. In addition, it is clear that the multi-task model is superior to the single-task model in both space and time, which means it can be more easily generalized to massive protein sequences, thus advancing the process of nucleic acid-protein interaction research.

Table S3. Performance comparisons of multi-task and single task models on DRNATst-246 and YK17-Tst over independent validation

Dataset	Model	DNA					RNA				
		MCC	AUC	AP	*p*-value [b]	PCC [c]	MCC	AUC	AP	*p*-value	PCC
DRNATst-246 [a]	single-task	**0.474**	**0.923**	**0.526**	1.36e-89	7.65e-01	**0.284**	**0.824**	**0.247** [d]	2.47e-92	6.31e-01
	multi-task	0.455	0.894	0.482	-	-	0.219	0.800	0.186	-	-
YK17-Tst	single-task	0.364	0.871	0.349	3.62e-82	7.64e-01	0.235	**0.793**	0.232	7.63e-10	6.28e-01
	multi-task	**0.391**	**0.881**	**0.350**	-	-	**0.275**	0.785	**0.233**	-	-

a. Since DRNATst-246 is divided into two subsets, i.e., DNA-129_Test and RNA-117_Test, and the single-task model can only target one task in either DNA or RNA, the experiments here are performed independently on the two subsets.

b. The *p*-values in Student's *t*-test are calculated for the differences between single-task model and multi-task model using the probability that the target residue is predicted to be a positive sample.

c. The PCC are calculated for the linear correlation coefficient between single-task model and multi-task model using the probability that the target residue is predicted to be a positive sample.

d. Bolded font indicates the best result.

Text S8. Case Studies

To visualize the advantages of the ESM-NBR, one native DNA-binding protein chain (PDB ID: 4OWX1) and one native RNA-binding protein chain (PDB ID: 6FQ3_A) are employed from YK17-Tst and RNA-117_Test for case studies. Figure S4 demonstrates the prediction results of ESM-NBR and four control methods on these two cases. The native DNA/RNA-binding residues are highlighted using sphere. Note that the DNA binding with 4OWX1 is a single-stranded DNA (ssDNA). By looking at the Figure S4, it is intuitive that ESM-NBR has better prediction performance on both protein chains. In particularly, most native DNA-binding residues are correctly predicted by the ESM-NBR on 4OWX1. In contrast, iDRNA-ITF and DRNAPred only correctly predicted a small percentage of DNA-binding residues. Even less favorable are the predictions of NCBRPred and GraphBind, who failed to predict even a single DNA-binding residue correctly. On 6FQ3_A, iDRNA-ITF and NCBRPred do not correctly predict any of the RNA-binding residues. GraphBind correctly predicted just a handful of RNA-binding residues. Although DRNAPred correctly identifies all positive samples, it also predicts a large number of negative samples as RNA-binding residues. On the other hand, the ESM-NBR shows a more accurate prediction and guarantees predictive performance for both positive and negative samples. The MCC of ESM-NBR on 6FQ3_A is 0.384 which are significantly higher than that of iDRNA-ITF, DRNAPred, GraphBind, and NCBRPred. The experimental results show that ESM-NBR has unique advantages on these two proteins.

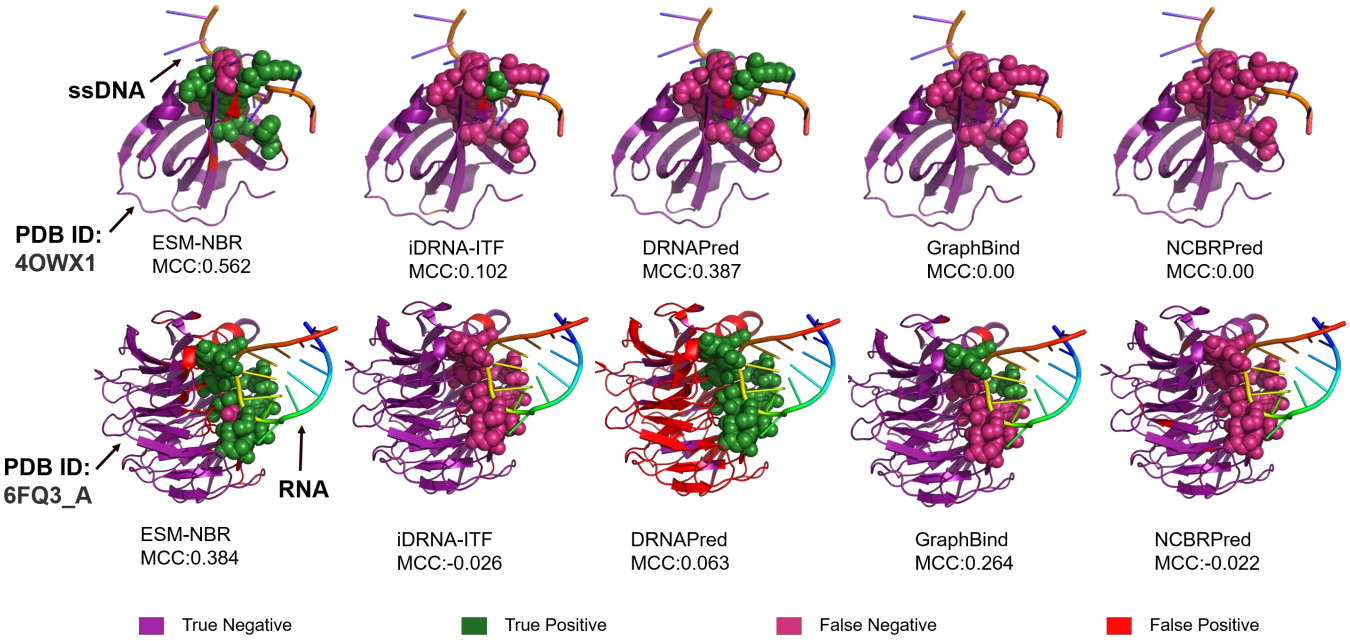

Figure. S4. Visualization of prediction results of ESM-NBR, iDRNA-ITF, DRNAPred, GraphBind, and NCBRPred on a DNA-binding protein chain (PDB ID: 4OWX1) and an RNA-binding protein chain (PDB ID: 6FQ3_A). The figures are plotted using pymol [14]. The spheres mean native DNA/RNA-binding residues.

REFERENCES

[1] J. Yan, and L. Kurgan, “DRNApred, fast sequence-based method that accurately predicts and discriminates DNA- and RNA-binding residues,” *Nucleic Acids Research,* vol. 45, no. 10, pp. 16, Jun, 2017.

[2] J. L. Sussman, D. Lin, J. Jiang, N. O. Manning, J. Prilusky, O. Ritter, and E. E. J. A. C. S. D. B. C. Abola, “Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules,” vol. 54, no. 6, pp. 1078-1084, 1998.

[3] Y. Xia, C. Q. Xia, X. Y. Pan, and H. B. Shen, “GraphBind: protein structural context embedded rules learned by hierarchical graph neural networks for recognizing nucleic-acid-binding residues,” *Nucleic Acids Research,* vol. 49, no. 9, pp. 17, May, 2021.

[4] J. Yang, A. Roy, and Y. Zhang, “BioLiP: a semi-manually curated database for biologically relevant ligand–protein interactions,” *Nucleic Acids Research,* vol. 41, no. D1, pp. D1096-D1103, 2012.

[5] N. Wang, K. Yan, J. Zhang, and B. Liu, “iDRNA-ITF: identifying DNA- and RNA-binding residues in proteins based on induction and transfer framework,” *Briefings in Bioinformatics,* vol. 23, no. 4, Jul, 2022.

[6] S. Yun, M. Jeong, R. Kim, J. Kang, and H. J. Kim, "Graph Transformer Networks," *Advances in Neural Information Processing Systems*, 2019.

[7] J. Hu, Y. S. Bai, L. L. Zheng, N. X. Jia, D. J. Yu, and G. J. Zhang, “Protein-DNA Binding Residue Prediction via Bagging Strategy and Sequence-Based Cube-Format Feature,” *Ieee-Acm Transactions on Computational Biology and Bioinformatics,* vol. 19, no. 6, pp. 3635-3645, Nov-Dec, 2022.

[8] J. Zhang, Q. C. Chen, and B. Liu, “NCBRPred: predicting nucleic acid binding residues in proteins based on multilabel learning,” *Briefings in Bioinformatics,* vol. 22, no. 5, Sep, 2021.

[9] Q. Yuan, S. Chen, J. Rao, S. Zheng, H. Zhao, and Y. Yang, “AlphaFold2-aware protein–DNA binding site prediction using graph transformer,” *Briefings in Bioinformatics,* vol. 23, no. 2, 2022.

[10] D. P. Kingma, and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

[11] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A Simple Way to Prevent Neural Networks from Overfitting,” *Journal of Machine Learning Research,* vol. 15, no. 1, pp. 1929-1958, 2014.

[12] J. Yan, S. Friedrich, and L. Kurgan, “A comprehensive comparative review of sequence-based predictors of DNA- and RNA-binding residues,” *Briefings in Bioinformatics,* vol. 17, no. 1, pp. 88-105, 2015.

[13] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, and A. Rives, “Evolutionary-scale prediction of atomic-level protein structure with a language model,” *Science,* vol. 379, no. 6637, pp. 1123-1130, 2023.

[14] S. G. Yuan, H. C. S. Chan, and Z. Q. Hu, “Using PyMOL as a platform for computational drug design,” *Wiley Interdisciplinary Reviews-Computational Molecular Science,* vol. 7, no. 2, Mar-Apr, 2017.