Predicting the Protein-DNA Binding Residue Using Protein Language Model with Residual based CNN module

Surabhi Mishra
Information Technology
ABV-IIITM
Gwalior, India
surabhi@iiitm.ac.in

Oshin Misra
Information Technology
ABV-IIITM
Gwalior, India
oshin@iiitm.ac.in

Mahua Bhattacharya
Information Technology
ABV-IIITM
Gwalior, India
mb@iiitm.ac.in

2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 979-8-3503-8622-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/BIBM62325.2024.10821747

***Abstract*—Numerous life-sustaining processes such as transcription, replication, and splicing, regulate the biological systems in a complex way. Therefore, it is important to recognize the pivotal role of crucial protein and DNA interactions for understand the above life sustaining processes. Hence, Protein-DNA binding residues is essential to identify as mentioned above. However, the classical experimental methods are tedious and labor-intensive. The recent research address these challenges and focused on predicting binding residues from sequence data. This present model utilizes a protein language model to transform protein sequence data into vector formats, enabling efficient computation. The model comprises three main components: an embedding layer that provide conversion of protein sequences into continuous vectors in the form of feature embedding using Protein-Bert model and a four-layer residual based convolutional neural network (CNN) trained on these vectors for prediction. The Binary Cross Entropy and contrastive loss function is used for loss computation. The same model, then reformatted, could be used to predict binding residues of RNA and antibodies. the present model achieves 0.884 AUC score and 0.795 specificity as compare to previous models.**

***Index Terms*—Protein-DNA binding site, ProtBert, Convolutional Neural Networks(CNN), Binary Cross Entropy, contrastive Loss function**

I. Introduction

There are many tiny cellular machines that may control activation of human genes through interacting DNA, RNA and proteins. These machines are DNA-binding proteins that are crucial players in functions like copying DNA, making RNA blueprints, and fixing DNA damage. The human's DNA blueprint actually contains instructions for making a whopping 6-7% of proteins into these DNA-binding machines. Surprisingly, despite their vast numbers, these proteins come in surprisingly diverse structures, falling into 54 distinct families with little resemblance to each other. Because understanding these proteins is so important, scientists are constantly looking for better ways to identify and classify them [1], [2].Hence, the Identification of interaction regions of a protein with DNA is crucial for understanding gene regulation, including DNA transcription, replication, expression, signal transduction and metabolism [3].

Over the past decade, researchers have explored various methods for predicting DNA-binding proteins (DBPs). These methods can be fall into two categorizes based on techniques i.e. experimental techniques and computational models such as machine learning models [4]. Based on data and information utilization, various methods are further classified as structure based and sequence based models.

The recognition of DNA-binding proteins (DBPs) has gotten a lot easier in recent years due to involvement of computational and experimental approach [5]. There are various experimental techniques developed by researchers. These experimental techniques includes in-vivo and in-vitro researches such as systematic evolution of ligands by exponential enrichment and chromatin immunoprecipitation [6], [7]. However, experimental methods for prediction are time-consuming and expensive. Scientists used to rely on laborious and structure based experiments like filter binding tests, genetic analysis, nuclear magnetic resonance and X-ray crystallography [8]. However, with the technological advancements and a boom in protein sequence data, researchers are turning to machine learning for faster and more accurate DBP classification [9].

Machine learning now provides powerful ways of carrying out the analysis of large datasets for the identification of patterns that can be of help in the prediction of these interactions. At the core of predicting DNA-binding proteins is extracting features from the respective sequences. In this case, much credit goes to machine learning algorithms, from linear regression [10] to deep learning models [11].

In particular, deep learning models are capable to capture intricate patterns embedded in protein sequences, which help revolutionize the power of detecting DNA-binding proteins. All this has allowed progress toward more accurate and effective ways to predict DNA binding proteins. These deep learning models have succeeded in the identification of new protein-DNA interactions, enabled fundamental discoveries across a spectrum of scientific disciplines, and offered new ways forward toward drug discovery. Therefore, machine learning has revolutionized the prediction ability of DNA-binding proteins to understand the intricate relationship between proteins and DNA. The enormous number of potential interactions, as well as the imperfections in the available information, predicts DNA binding proteins a formidable challenge. With the fast growth of datasets and technology, such predictions are becoming even more accurate, thereby permitting new possibilities in healthcare. On the contrary, the traditional

methods dependent on protein structure are limited by the scarcity of accurate 3D data and the computationally expensive time needed for training [12]. Therefore, Protein sequence based learning methods are alternative way to predict the protein-DNA binding residue. A model developed through the algorithm can pick out distinct patterns in protein sequences, resulting in overall efficacy in the prediction of bindings while maintaining accuracy [11], [13], [14].

Various motivations have continued to drive the use of NLP based language models. Due to providing resource efficiency and better protein sequence representation,researchers have turned to pre-trained protein language models like ProtBert. The models, being highly efficient due to training on large datasets, can capture vital patterns in protein sequences, therefore enhancing the representation of the protein structures and enabling the making of interaction predictions with DNA. On the other hand, pre-trained protein language models can quickly help extract features from prior knowledge in large amounts of data, leading to reduced training time with more effective prediction models. These possible applications range from DNA binding to protein structure and drug discovery. Further benefits include that it is resthe presentce-efficient, allows for accurate feature extraction, and may even lead to generalization in the prediction of interaction beyond DNA to RNA to even antibody-protein binding [15].

II. Related Work

Traditionally, DNA-binding protein (DBP) prediction heavily relied on structural analysis methods, presumed to offer superior accuracy. However, the challenge of obtaining high-resolution protein crystal structures limited their practicality, prompting a shift towards sequence-based approaches [16]. These methods, gaining traction for their simplicity and convenience, extract features directly from protein sequences. These features encompass profile-based, composition-based, and autocorrelation-based categories, with profile-based features, notably Position-Specific Scoring Matrix (PSSM) and Hidden Markov Model (HMM), emerging as the most effective for DBP prediction. Research underscores the superiority of PSSM-based features in enhancing prediction accuracy, as evidenced by various innovative approaches integrating evolutionary information from PSSM profiles. As a result, recent years have witnessed the emergence of DBP prediction as a supervised learning problem. Support Vector Machines, Random Forests, classifiers based on the Naive Bayes approach, and ensemble classifiers, including Stacking, and Deep Learning, among others have been used. Stacking is an ensemble learning method for the optimization of predictions based on combining outputs from base classifiers. We have shown the successful prediction of DBPs using derived features from PSSM profiles effectively in this novel stacking-based approach [17].

A prevalent sequence-based approach involves leveraging Position-Specific Scoring Matrix (PSSM) or Position Specific Feature Matrix (PSFM) data [11], transforming protein sequences into feature vectors through various conversion techniques. Alternatively, models may exploit amino acid properties within the sequence, often employing Support Vector Machines (SVMs) for classification, a strategy further enhanced by the integration of PSSM data [13].

Recent advancements in DBP prediction have seen the emergence of sophisticated algorithms employing Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs) as classifiers, enabling the capture of intricate relationships within protein sequence data [11], [18]. However, each method has its drawbacks. Some experimental techniques are time-consuming and resthe presentce-intensive, while others may have limitations in their accuracy.

One of the key challenges in this field is the difficulty of accurately representing protein structures. Traditionally, this required complex calculations and data manipulation. This research proposes a solution to this problem by leveraging pre-trained models. These models can effectively represent protein structures while preserving the important sequence-specific information.

Various models like single-supervised models, unsupervised models, CNNs, or RNNs, have been explored for both structure-based and sequence-based predictions. However, the primary benefit of using machine learning models in sequence-based approaches is their ability to function even without detailed structural data for the protein. This allows researchers to analyze proteins where obtaining structural data might be difficult or impractical. Nonetheless, enhancing accuracy necessitates models adept at portraying protein characteristics solely from sequences, along with algorithms capable of learning from these portrayals to discern DNA-binding proteins.

This study delves into the possibilities offered by sequence-based strategies, emphasizing pivotal advancements in this domain. The proposed approach,try to apply a pre-trained model for feature extraction from protein sequences, offers several advantages over traditional techniques. Firstly, it can learn effectively even with limited datasets. Secondly, it can identify correlations within the protein structure based solely on the sequence information. Finally, by incorporating deep machine learning techniques, this approach has the potential to significantly improve the accuracy of DBP prediction.

the present contributions can be encapsulated as follows:

- To use ProtBert, a pre trained protein language model which helps in provide embedding of protein sequences.
- To propose a residual based CNN prediction model which will leverage a large language protein model to efficiently extract informative features from protein structures, aiding in the prediction of DNA-binding proteins.
- To leverage a loss function during model training which helps the model become more robust (resistant to errors) and adaptable (able to handle new data) by focusing on learning similarities and differences between protein sequences.
- To assess the model's effectiveness in identifying DNA-binding proteins, we will employ several evaluation metrics including specificity, precision, recall, F1-score, and MCC (Matthews correlation coefficient).

The paper is structured as follows: Section II introduces related work which have been done in this area. Section III discusses details of dataset which we are using in the present model. Section IV introduces methodology of the present model how we are implementing. Section V covers the results which we obtained. Section VI described the conclusion of the present work. Finally, Section VII highlights potential avenues for future research.

III. Dataset Details

TABLE I
SUMMARY OF DATA SETS USED

Dataset entities	count
DNA binding residues	16601
Non binding residues	308414
Percentage of binding residues	6.925

The dataset was formulated during the study of DBPred model [19] for prediction of protein DNA-binding residues. The dataset is a combination of two data sources, namely hybridNAP [20] and ProNA [18]. Dataset, shown in table I, consisted of data on 646 proteins, where the number of binding residues was listed as 16601 and the non binding residues were counted as 308414, the percentage of binding residues came out to be 6.925%.

IV. Methodology

This research work aims to find the DNA protein binding residues through sequence based data. In order to achieve the goal, the proposed approach considers two modules : an sequence embedding module and residual based CNN prediction module. The sequence embedding module utilize the a protein language model to get the embedded vector with respect to protein sequence. Whereas CNN based prediction module 1D Convolutional layer based architecture within the residual block for DNA protein binding residues prediction, shown in the figure 1. This section also provide loss computation that provide improvement in the results.

A. Module for Embedding Sequence

The embedding sequence module utilize a pre-trained protein language model [15], [21] to handle the complicated process of encoding protein sequences , thereby the present model avoid inherent complexities in this encoding task. While alternative methodologies may opt to refine this model, potentially enhancing accuracy, such endeavors entail heightened temporal complexity. Hence, in proposed approach, this model serves solely as a feature extractor. It provide an 1024, dimension for protein embedding. Noteworthy advantages of this architecture include its capability to encode protein sequences of varying lengths into fixed-size vectors, sans the need for parameter adjustments.

B. Prediction using Residual based CNN Module

This segment of the model employs a sequence of the present layers belonging to a specialized neural network architecture known as residual convolutional Neural Network (ResNet), shown in figure 1. Similar to general architecture, one residual block includes two 1D convolution layers followed by a skip connection that adds the input to the output shown in figure 2. a combination of 1D CNN layered architecture and These layers are designed to scrutinize localized segments of the protein sequence and their surrounding context to extract pertinent information. Additionally, they facilitate data compression to enhance computational efficiency.

At each iteration of the residual block, the input parameters encompass the dimensions of the input data, the longest protein sequence, and the dimensions of the preceding layer, while the output parameters encapsulate the dimensions of the input data, the longest protein sequence, and the dimensions of the current layer.

This module utilizes the present successive residual blocks (1024x1, 512x1, 256x1, 128x1) to provide prediction for discreet segments of the protein sequence, each characterized by two convolution layers with batch normalization to inspect the sequence and its neighboring elements. To ensure uniformity across all sequences, a padding technique is employed to adjust the boundaries. Subsequently, the present model employs an output layer that outputs 2 channels for binary classification (per position in the sequence). Ultimately, the last layer leverages a specialized function, namely Softmax, to ascertain the likelihood of the protein sequence binding to DNA.

C. Loss Computation Module

In order lead to better generalization and improve the performance for protein-DNA Binding residue prediction, the contrastive loss function is used. Contrastive loss module provide an improved learning with respect to learning space by consider pairs of amino acids within a single protein sequence [22]. It mathematically formulated as:

$$L = (y \cdot d^2) + ((1 - y) \cdot (\max(0, m - d))^2) \quad (1)$$

Where, y represents the label indicating whether the pair of data points is similar (y=1) or dissimilar (y=0). d represents the Euclidean distance between the representations of the two data points. m represents a margin parameter that defines the minimum distance between dissimilar pairs.

The present model also consider binary cross entropy based loss computation to improve prediction performance as it use gradient based optimization and effectively handle class imbalance in case of binary classification.

V. Result and Discussion

This section presents the performance of the proposed model in terms of the evaluation metrics like specificity, recall, precision, f1-score, MCC. A tabular comparison is also presented with the other existing methods in the literature. Also, other ligands like RNA and antibody's performance has been calculated too in terms of similar metrics.

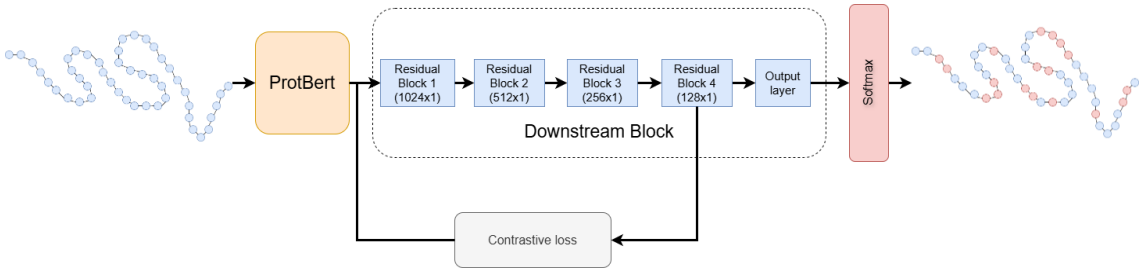

Fig. 1. Methodology of proposed scheme

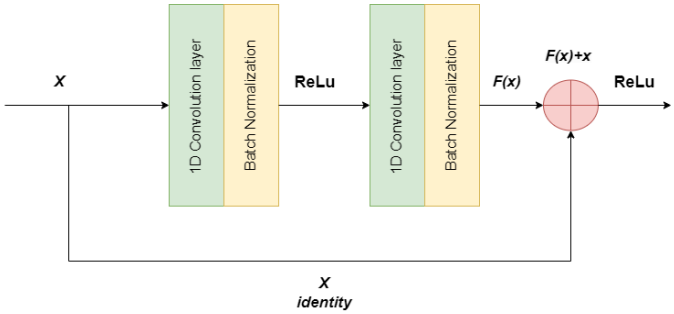

Fig. 2. Residual Module Architecture

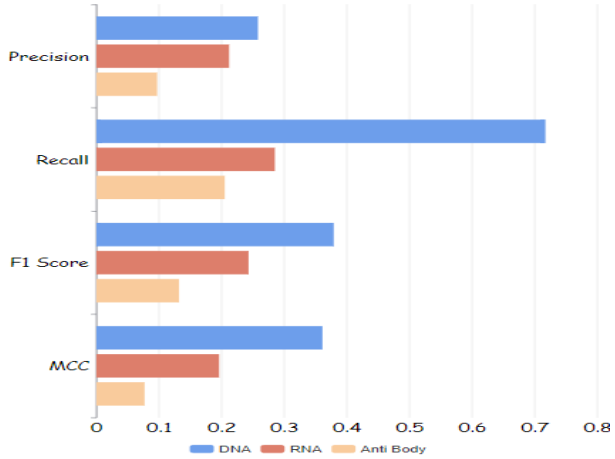

Fig. 3. All Metrics of the present model

A. Comparison with Previous work

In the table II, we have elucidated the comparative analysis conducted between the proposed model and other existing models that have been engineered for the purpose of predicting Protein-DNA binding site. It is evident from the comparison that the present model exhibits superior performance in comparison to other sequential methodologies that leverage machine learning techniques.

B. Model as a general framework

The versatile framework articulated by the present model holds considerable promise for the nuanced prediction of binding locales spanning a spectrum of ligand types, including intricate scenarios such as protein-RNA and antibody-antigen engagements. In pursuit of this endeavor, we meticulously curated benchmark datasets tailored to these specific binding modalities, facilitating the rigorous training and meticulous evaluation of the present model's predictive prowess. Here, the present model also achieves AUC score 0.552 for antibody and 0.775 for RNA binding residues. The other performance metric for RNA and anti body based predictions in comparison to DNA are shown in the figure 3.

Figure 3 depict the Precision, Recall, F1-score and MCC score, for DNA, RNA and antibodies site prediction for proposed algorithm. These graphical representations offer a visual depiction of the performance of the present model in discerning binding residues across various nucleic binding proteins. Additionally, the Precision, Recall, F1-score, and Matthews correlation coefficient (MCC) metrics are presented for each nucleic binding protein, providing a comprehensive assessment of the algorithm's predictive capabilities. Furthermore, the ROC curve is plotted individually for different nucleic binding proteins, offering insights into the discriminative power of the model across diverse protein-ligand interactions.

VI. Conclusions

Diverse prior investigations exhibit distinctive attributes, merits, demerits, and, notably, avenues for future exploration. The present approach utilizes a pre-trained model for feature extraction followed by a Residual based CNN. The model's primary contributions and goals include developing an efficient framework for DNA-binding protein prediction that eliminates the computationally expensive process of protein feature extraction. This research has successfully met all the objectives using the integration of embedding, predicting and computational loss modules. Consequently, the proposed model attained superior AUC scores.

The future scope of above research include expanded feature representation from larger outputs using protein language models and incorporating ensemble techniques, fine-tuning large-scale protein language models for improved performance and developing a unified framework capable of predicting residues across all nucleic acids using multi-class classification techniques. These advancements could further enhance the accuracy and applicability of DNA-binding protein prediction models.

TABLE II
RESULTS OF DIFFERENT METHODS

Models	Specificity	Recall	Precision	F1	AUC	MCC
DBPred [11]	0.784	0.708	0.243	0.362	0.794	0.320
DNAPred [17]	0.655	0.671	0.157	0.254	0.730	0.194
NCBRPred [23]	0.674	0.677	0.165	0.265	0.713	0.207
DRNAPred [24]	0.692	0.677	0.185	0.291	0.755	0.226
SVMnuc [13]	0.666	0.668	0.154	0.250	0.715	0.192
Proposed Model	0.795	0.717	0.258	0.379	0.884	0.361

REFERENCES

[1] H. P. Nasheuer, A. M. Meaney, T. Hulshoff, I. Thiele, and N. O. Onwubiko, "Replication protein a, the main eukaryotic single-stranded dna binding protein, a focal point in cellular dna metabolism," *International Journal of Molecular Sciences*, vol. 25, no. 1, p. 588, 2024.

[2] A. Krishnan, S. Nijmeijer, C. de Graaf, and H. B. Schiöth, "Classification, nomenclature, and structural aspects of adhesion gpcrs," *Adhesion G protein-coupled receptors: molecular, physiological and pharmacological principles in health and disease*, pp. 15–41, 2016.

[3] S. A. Lambert, A. Jolma, L. F. Campitelli, P. K. Das, Y. Yin, M. Albu, X. Chen, J. Taipale, T. R. Hughes, and M. T. Weirauch, "The human transcription factors," *Cell*, vol. 172, no. 4, pp. 650–665, 2018.

[4] Y.-H. Qu, H. Yu, X.-J. Gong, J.-H. Xu, and H.-S. Lee, "On the prediction of dna-binding proteins only from primary sequences: A deep learning approach," *PloS one*, vol. 12, no. 12, p. e0188129, 2017.

[5] Y. He, Q. Zhang, S. Wang, Z. Chen, Z. Cui, Z.-H. Guo, and D.-S. Huang, "Predicting the sequence specificities of dna-binding proteins by dna fine-tuned language model with decaying learning rates," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 20, no. 1, pp. 616–624, 2023.

[6] T. S. Furey, "Chip–seq and beyond: new and improved methodologies to detect and characterize protein–dna interactions," *Nature Reviews Genetics*, vol. 13, no. 12, pp. 840–852, 2012.

[7] R. A. C. Ferraz, A. L. G. Lopes, J. A. F. da Silva, D. F. V. Moreira, M. J. N. Ferreira, and S. V. de Almeida Coimbra, "Dna–protein interaction studies: a historical and comparative analysis," *Plant Methods*, vol. 17, pp. 1–21, 2021.

[8] R. Jaiswal, S. K. Singh, D. Bastia, and C. R. Escalante, "Crystallization and preliminary X-ray characterization of the eukaryotic replication terminator Reb1–Ter DNA complex," *Acta Crystallographica Section F*, vol. 71, no. 4, pp. 414–418, Apr 2015. [Online]. Available: https://doi.org/10.1107/S2053230X15004112

[9] B. Liu, J. Xu, S. Fan, R. Xu, J. Zhou, and X. Wang, "Psedna-pro: Dna-binding protein identification by combining chou's pseaac and physicochemical distance transformation," *Molecular Informatics*, vol. 34, no. 1, pp. 8–17, 2015. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/minf.201400025

[10] J. Wu, H. Liu, X. Duan, Y. Ding, H. Wu, Y. Bai, and X. Sun, "Prediction of DNA-binding residues in proteins from amino acid sequences using a random forest model with a hybrid feature," *Bioinformatics*, vol. 25, no. 1, pp. 30–35, 11 2008. [Online]. Available: https://doi.org/10.1093/bioinformatics/btn583

[11] R. G. Patiyal S, Dhall A, "A deep learning-based method for the prediction of dna interacting residues in a protein," *Brief Bioinform*, vol. 23, no. 5, pp. 339–349, 2022.

[12] M. Waris, K. Ahmad, M. Kabir, and M. Hayat, "Identification of dna binding proteins using evolutionary profiles position specific scoring matrix," *Neurocomputing*, vol. 199, pp. 154–162, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231216300455

[13] H. Su, M. Liu, S. Sun, Z. Peng, and J. Yang, "Improving the prediction of protein–nucleic acids binding residues via multiple sequence profiles and the consensus of complementary methods," *Bioinformatics*, vol. 35, no. 6, pp. 930–936, 2019.

[14] M. F. Hosen, S. H. Mahmud, K. Ahmed, W. Chen, M. A. Moni, H.-W. Deng, W. Shoombuatong, and M. M. Hasan, "Deepdnabp: A deep learning-based hybrid approach to improve the identification of deoxyribonucleic acid-binding proteins," *Computers in Biology and Medicine*, vol. 145, p. 105433, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482522002256

[15] D. O. Nadav Brandes and M. L. Yam Peleg, Nadav Rappoport, "Proteinbert: a universal deep-learning model of protein sequence and function," *Bioinformatics*, vol. 38, pp. 2102–2110, 2022.

[16] S. H. Mahmud, K. O. M. Goh, M. F. Hosen, D. Nandi, and W. Shoombuatong, "Deep-wet: a deep learning-based approach for predicting dna-binding proteins using word embedding techniques with weighted features," *Scientific reports*, vol. 14, no. 1, p. 2961, 2024.

[17] Y.-H. Zhu, J. Hu, X.-N. Song, and D.-J. Yu, "Dnapred: accurate identification of dna-binding sites from protein sequence by ensembled hyperplane-distance-based support vector machines," *Journal of chemical information and modeling*, vol. 59, no. 6, pp. 3057–3071, 2019.

[18] J. Qiu, M. Bernhofer, M. Heinzinger, S. Kemper, T. Norambuena, F. Melo, and B. Rost, "Prona2020 predicts protein–dna, protein–rna, and protein–protein binding proteins and residues from sequence," *Journal of Molecular Biology*, vol. 432, no. 7, pp. 2428–2443, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022283620302035

[19] S. Patiyal, A. Dhall, and G. Raghava, "A deep learning-based method for the prediction of dna interacting residues in a protein," *Briefings in Bioinformatics*, vol. 23, 08 2022.

[20] J. Zhang, Z. Ma, and L. Kurgan, "Comprehensive review and empirical analysis of hallmarks of dna-, rna- and protein-binding residues in protein chains," *Briefings in bioinformatics*, vol. 20, 12 2017.

[21] A. Elnaggar, M. Heinzinger, C. Dallago, G. Rehawi, Y. Wang, L. Jones, T. Gibbs, T. Feher, C. Angerer, M. Steinegger, D. Bhowmik, and B. Rost, "Prottrans: Toward understanding the language of life through self-supervised learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 7112–7127, 2022.

[22] Y. Liu and B. Tian, "Protein–dna binding sites prediction based on pre-trained protein language model and contrastive learning," *Briefings in Bioinformatics*, vol. 25, no. 1, p. bbad488, 2024.

[23] J. Zhang, Q. Chen, and B. Liu, "Ncbrpred: predicting nucleic acid binding residues in proteins based on multilabel learning," *Briefings in bioinformatics*, vol. 22, no. 5, p. bbaa397, 2021.

[24] 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. e84–e84, 2017.