Chapter 9

Advances in Language-Model-Informed Protein–Nucleic Acid Binding Site Prediction

Sumit Tarafder, Xinyu Wang, Rahmatullah Roche,
and Debswapna Bhattacharya

Abstract

Interactions between proteins and nucleic acids are essential for understanding a wide range of cellular and evolutionary processes. Recent advancements in protein language models (pLMs), trained on vast protein sequence data, have revolutionized various predictive modeling tasks, offering unprecedented scalability and generalizability. Consequently, a number of computational methods have been developed in the recent past for protein–nucleic acid binding site prediction powered by pLMs. To this end, we recently developed the EquiPNAS method that integrates pLM embeddings with E(3) equivariant deep graph neural networks for enhancing accuracy and robustness in predicting protein–DNA and protein–RNA binding sites, thereby reducing the dependency on evolutionary information. Here we present an overview of the recent protein–nucleic acid binding site prediction methods, emphasizing the recent advances in harnessing the potential of pLMs, and provide a detailed description of the EquiPNAS methodology as well as the necessary materials and procedures for the computational prediction of protein–DNA and protein–RNA binding sites.

Key words Protein–DNA binding site prediction, Protein–RNA binding site prediction, Language models, Graph neural networks

1 Introduction

Proteins interact with nucleic acids, including DNA and RNA, to perform critical cellular functions such as regulating gene expression, facilitating DNA replication, and mediating cellular signaling pathways. The precise identification of these interaction or binding sites where proteins interact with nucleic acids is fundamental to understanding these biological processes [1–5]. This understanding has significant implications not only for basic biological research but also for practical applications in drug design and therapeutic interventions. While traditional laboratory methods for identifying these binding sites are effective [6, 7], they are often time-consuming, expensive, and technically challenging, prompting

Dukka B. KC (ed.), *Large Language Models (LLMs) in Protein Bioinformatics*, Methods in Molecular Biology, vol. 2941,
https://doi.org/10.1007/978-1-0716-4623-6_9,

researchers to explore computational approaches for predicting these sites based on available data. Computational methods for predicting protein–nucleic acid binding sites can be categorized into two main types: sequence-based methods [8–15] and structure-aware methods [8, 16–20]. Sequence-based methods rely solely on the amino acid sequences of proteins to make predictions about binding affinities and interaction sites. On the other hand, structure-aware methods utilize 3D structural information derived from experiments or computational predictions, resulting in more precise binding site predictions. However, these approaches frequently depend on limited experimental data from structural databases such as the Protein Data Bank (PDB) [21], which may not provide comprehensive coverage for all protein–nucleic acid interactions. The introduction of AlphaFold2 [22] and AlphaFold3 [23] significantly expands this landscape by providing highly accurate structural models for a wide range of proteins and their interactions, including with nucleic acids. Additionally, the AlphaFold Database [24] offers a comprehensive repository of these predicted structures, allowing researchers to access valuable data that was previously inaccessible. These advancements alleviate the limitations associated with reliance on PDB and enhance the predictive capabilities of protein structure modeling [25] and binding site identification.

Currently, large language models (LLMs) have emerged as transformative tools in the field of artificial intelligence [26–30], possessing the capability to comprehend and generate human language through the analysis of extensive textual datasets. Their applications extend across numerous domains, encompassing natural language processing tasks such as sentiment analysis, language translation, and content generation [31–33], thereby reshaping industries ranging from healthcare [34–39] to finance [40–44]. Notably, in the area of protein–nucleic acid binding site predictions, LLMs demonstrate significant advancements by leveraging their understanding of biological sequences and molecular interactions. The advent of protein language models (pLMs) [45–50], which are analogous to natural language processing models but specifically trained on extensive datasets of protein sequences, has opened new avenues for predicting protein characteristics and functions. While the application of pLMs in predicting protein–nucleic acid binding sites is still in its formative stages, their potential for enhancing prediction accuracy is considerable [48, 50]. Research initiatives such as EquiPPIS [51] lay critical groundwork for advancements in predicting protein–RNA binding positions.

Although numerous methods have been developed for protein–nucleic acid binding site prediction, EquiPNAS [52] is one of the first methods to leverage pLM embeddings for the effective prediction of both DNA and RNA binding protein residues, consistently surpassing state-of-the-art techniques across an array of widely utilized benchmarking datasets for protein–DNA and protein–RNA binding site prediction tasks. This innovative approach employs a specific type of graph neural network known as E(3) Equivariant Graph Neural Network (EGNN) [53], which excels at processing the intricate three-dimensional (3D) nature of protein structures. By being sensitive to positional, rotational, and orientational transformations, this model maintains robustness and accuracy, even when utilizing protein structures predicted by AlphaFold2 [22] as opposed to those resolved through experimental methods (*see* **Notes 1** and **2**).

2 Method

2.1 Overview of Existing Protein–Nucleic Acid Binding Site Prediction Methods

Protein–nucleic acid binding site prediction is a well-studied problem with numerous methods available in the literature utilizing a combination of network architectures and features as listed in Table 1. Currently, the available protein–nucleic acid binding site prediction methods can be broadly divided into sequence-based and structure-based approaches. Sequence-based methods, such as NCBRPred [9], DNAPred [10], and RNABindRPlus [12], leverage the abundance of protein sequence data to predict nucleic

Table 1
A selection of state-of-the-art, deep learning–based frameworks for protein–nucleic acid binding site prediction, arranged in descending order by year of publication

Name	Year	Architecture	pLM_Method	Protein–DNA/RNA
EquiPNAS (Roche et al.)	2024	EGNN	ESM2	Both
CLAPE (Liu et al.)	2024	CNN	ProtBert	Protein–DNA
ULDNA (Zhu et al.)	2024	LSTM-attention	ESM2, ProtTrans	Protein–DNA
EGPDI (Zheng et al.)	2024	EGNN+GCN-II	ESM2, ProtTrans	Protein–DNA
ESM-NBR (Zeng et al.)	2023	BiLSTM + MLP	ESM2	Both
GLMsite (Song et al.)	2023	GVP-GNN	ProtTrans	Both
GraphSite (Yuan et al.)	2022	Graph transformer	–	Protein–DNA
bindEmbed21DL (Littmann et al.)	2021	CNN	ProtT5	Both
GraphBind (Xia et al.)	2021	GNN	–	Both
NCBRPred (Zhang et al.)	2021	BiGRU	–	Both
NucleicNet (Lam et al.)	2019	ResNet	–	Protein–RNA
aaRNA (Li et al.)	2014	DenseNet	–	Protein–RNA
RNABindRPlus (Walia et al.)	2014	SVM	–	Protein–RNA

acid binding sites. These methods often utilize machine learning techniques, such as hidden Markov models and bidirectional Gated Recurrent Units (BiGRUs), to capture patterns from sequence data. Although these methods are widely applicable, their reliance solely on sequence information limits their predictive accuracy, as they miss critical spatial details of protein–nucleic acid interactions. Structure-based methods, including COACH-D [16], NucBind [8], and GraphBind [19], incorporate three-dimensional structural information, which enhances prediction accuracy. By utilizing structural templates or advanced computational models like graph neural networks, structure-based methods can better capture the intricate spatial patterns essential for identifying binding sites.

In recent years, hybrid methods have emerged, combining both sequence-based and structure-based approaches to improve prediction outcomes. For instance, DNABind [54] and NABind [55] integrate machine learning with template-based methods, while NucBind [8] combines predictions from both COACH-D [16] and SVMnuc [8]. Graph-based approaches, such as GraphSite [20] and GLMSite [56], have introduced new levels of sophistication by encoding secondary structure and spatial positions of atoms. While these hybrid methods outperform some of the state-of-the-art deep learning-only based frameworks, their hybrid approaches and high dependence on the quality of the templates make them susceptible to orphan proteins with low homology depths. However, with the recent surge in the availability of various families of protein language models such as ESM and ProtTrans, an array of deep learning–based binding site prediction methods have emerged as shown in Table 1. Methods such as EquiPNAS [52], CLAPE [57], ULDNA [58], EGPDI [4], ESM-NBR [59], and bindEmbed21DL [60], utilizing these language model embeddings, have surpassed those without them in terms of prediction accuracy, reducing the overall dependency on evolutionary information for binding site prediction task [61].

2.2 Significance of pLMs for Protein–Nucleic Acid Binding Site Prediction

While computational methods have advanced in predicting protein and nucleic acid binding sites, significant challenges persist, particularly in effectively utilizing protein-related data. Sequence-based approaches, which depend heavily on evolutionary information, often struggle with orphan proteins and encounter difficulties with intrinsically disordered proteins (IDPs) and regions (IDRs) due to their unstable and dynamic structures [62]. Additionally, many existing techniques rely on manually curated features that demand domain-specific expertise and may miss critical biological insights [17].

In contrast, protein language models (pLMs) present notable advantages in protein–nucleic acid binding predictions. Using self-supervised learning, pLMs can capture long-range dependencies and structural information from sequences without requiring manual feature engineering [63]. These large language models, originally designed for natural language tasks like translation and question-answering, have evolved into biological language models due to the linguistic parallels between human and biological languages. Through transfer learning, these models excel at capturing the structure and function of biological molecules such as proteins, creating rich feature representation from large datasets of protein sequences, which proves to be significant for binding site prediction tasks (*see* **Notes 3** and **4**). This approach minimizes the reliance on evolutionary data and enables more accurate and generalizable embeddings [64].

However, in the context of binding site prediction, especially with nucleic acid ligands, the information derived from pLMs should account for the interactions between proteins and nucleic acids. A promising approach can be the potential incorporation of multimodal learning, where pLMs designed for DNA and RNA sequences [65, 66], or a platform like BioSeq-BLM, capable of analyzing DNA, RNA, and protein sequences using a range of biological language models [67] can be leveraged. This is especially valuable for improving prediction accuracy and efficiency when dealing with highly flexible molecules like RNA [61].

2.3 Overview of EquiPNAS Framework

The primary objective of the EquiPNAS method is to accurately predict nucleic acid (both DNA and RNA)-binding residues for a given input protein 3D structure as shown in Fig. 1. The prediction framework follows a structured methodology that consists of three key components. The first component is the graph representation of the protein, which involves constructing a graph in which nodes correspond to individual residues and edges denote interactions among them. This graph structure effectively captures the spatial and relational information inherent to the protein. The second component involves feature generation for both nodes (residues) and edges (interactions) within the graph. This includes the extraction of sequence-based and structure-based information, alongside the utilization of protein language model (pLM) embeddings derived from the ESM-2 model [47], thereby enriching the data set available for predictive purposes. The final component employs an E(3) equivariant graph neural network (EGNN) to process the graph representation and make binding site predictions. By leveraging the coordinate information from the input monomer in conjunction with the generated features, the EGNN performs graph node classification, estimating the probability of whether each residue acts as a binding site for the corresponding nucleic acid as annotated in Fig. 1. In the subsequent sections, each of these components will be explored in detail, illuminating their significance within the overarching EquiPNAS architecture and the enhanced predictive capabilities they provide.

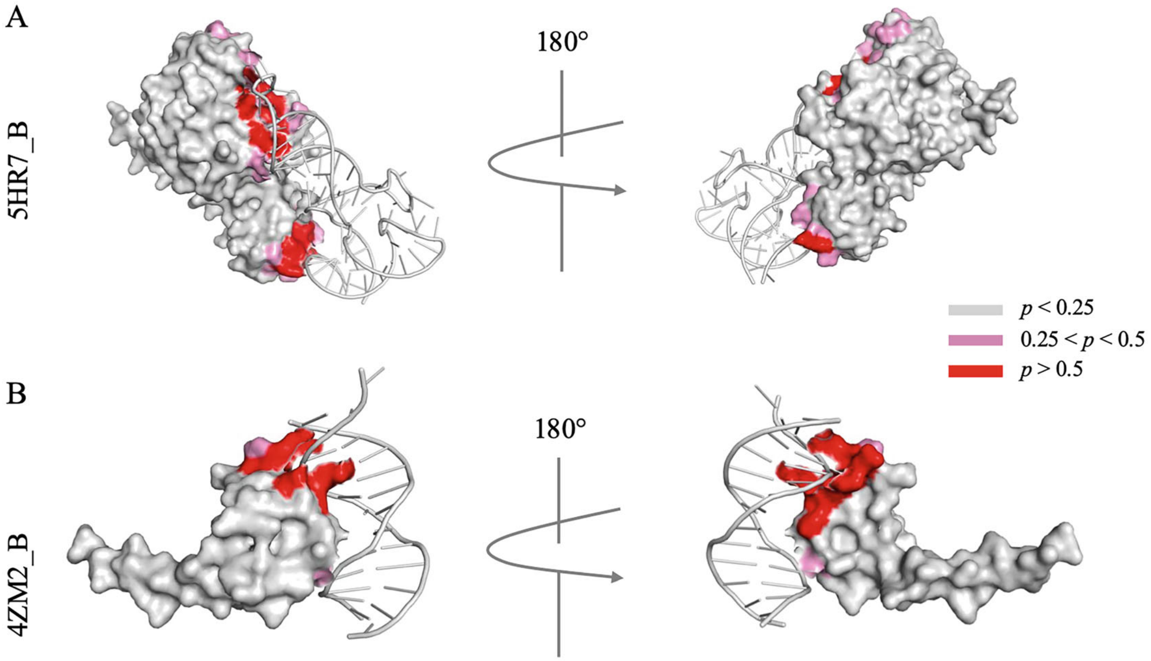

Fig. 1 Experimental structure of (**a**) protein–RNA complex (PDB ID: 5HR7_B) and (**b**) protein–DNA complex (PDB ID: 4ZM2_B) with proteins shown in surface representation. Each residue of the protein structure is color coded according to its predicted probability score (*p*) by EquiPNAS, visualizing the correlation between EquiPNAS predictions and experimentally derived binding sites

2.3.1 Protein Graph Representation

Protein as a Graph

In the EquiPNAS framework, proteins are represented as graphs wherein each residue is depicted as a node. An edge is established between two nodes if the Cα atoms of the corresponding residues are within a defined distance threshold, thus transforming the protein structure into an interactive network of components.

Interacting Residue Pairs

Residues are classified as interacting based on specific criteria. Notably, the distance between their Cα atoms must be less than or equal to 14 Å for DNA-binding interactions and 15 Å for RNA-binding interactions. Furthermore, the residues must be at least six positions apart within the protein sequence. This requirement ensures that only relevant interactions are captured, offering a focused representation of the protein's binding potential. Through this graph-based approach, EquiPNAS adeptly captures both spatial relationships and interaction dynamics, facilitating accurate predictions of P-NA binding sites.

2.3.2 Feature Generation

Sequence-Based Features

The generation of sequence-based features is essential for the effective representation of proteins. Each amino acid is encoded using a simple binary (one-hot) encoding scheme for the 20 standard amino acids. Additionally, a Position-Specific Scoring Matrix (PSSM) is obtained through PSI-BLAST [68], with subsequent normalization of values to maintain consistent scaling. To augment the feature set, a Multiple Sequence Alignment (MSA) is generated using MMseqs2 [69] and subsequently refined by ColabFold [70] pipeline to provide distilled evolutionary information to the model. Finally, the language model embeddings have been extracted from ESM-2, a pre-trained protein language model (pLM) containing 15 billion parameters represented by a feature dimension of 5120 for each residue in the protein sequence.

Structure-Based Features

Structure-based features capture critical aspects of the protein's topology and spatial orientation [71]. One-hot encoding is utilized for secondary structure and solvent accessibility, indicating the exposure of each residue to the surrounding environment. Furthermore, geometric relationships—such as bond angles and the positioning of residues in three-dimensional space—are meticulously recorded. The framework also tracks the number of residues in contact and computes their spatial positions relative to the protein's center, thus refining the understanding of interaction dynamics.

Edge Features

For edge features, the framework calculates the ratio of sequence distance to three-dimensional (3D) distance for each pair of interacting residues. This ratio provides valuable insights into the spatial relationships between residues, significantly contributing to the overall predictive capability of the framework.

Coordinate Features

Finally, coordinate features leverage the 3D coordinates (x, y, z) of each residue's Cα atom, effectively capturing the spatial arrangement within the protein structure. By integrating these diverse feature sets, EquiPNAS substantially enhances its ability to accurately predict protein–nucleic acid binding sites.

2.4 Architecture

The core architecture of EquiPNAS is constructed utilizing deep E (3)-equivariant graph neural networks (EGNNs), specifically tailored to predict protein–DNA and protein–RNA binding sites. These networks operate by leveraging features pertinent to the protein's structure and its spatial coordinates, such as the positions of Cα atoms in three-dimensional space. The architecture comprises multiple layers of equivariant graph convolution layers (EGCL), which simultaneously update both the node features, and the coordinates of the protein based on the interactions (edges) among residues.

The process can be outlined as follows: (1) *Node representation*: Each node, representing a residue in the graph, encompasses features (such as the amino acid type) and coordinates (indicating its spatial position). These attributes undergo transformation across several layers of the network (2) *Feature and coordinate updates*: The layers update the node features and coordinates by utilizing edge information, including the distances between residues.

(3) *Interaction learning*: Within each layer, the model learns how residues interact based on the provided features, continuously refining its predictions to enhance accuracy.

The EGNN employs 12 layers of transformations, each characterized by 768 hidden dimensions. To mitigate the risk of overfitting, dropout regularization is applied at every layer, ensuring the model generalizes effectively beyond the training data. Ultimately, the model predicts the likelihood of each residue acting as a binding site for nucleic acids (DNA or RNA), condensing all acquired information into a singular predictive score. The training procedure of this model is conducted using ADAM optimizer [72] and cosine annealing [73] for dynamic adjustment of learning rates. The training process extends for up to 40 epochs on a high-performance NVIDIA Graphics Processing Unit (GPU), ensuring robust learning from the datasets. To assess the efficacy of the EGNN structure in enhancing prediction accuracy, baseline models devoid of equivariant updates were also trained for comparative purposes. This allows for a thorough evaluation of the model's performance enhancements attributable to the incorporation of equivariant features.

3 Materials

To train and evaluate a machine learning–based model for protein–DNA binding site prediction, the GraphBind [19] study's Train_573 (573 protein chains) and Test_129 (129 chains) datasets are used, along with GraphSite's [20] Test_181 (181 chains). All datasets, curated from BioLiP [74] and filtered with CD-Hit [75] for non-redundancy, span different timeframes: Train_573 (pre-2016), Test_129 (2016–2018), and Test_181 (2018–2021), with varying counts of binding and non-binding residues. For protein–RNA binding site prediction, Train_495 (495 chains) and Test_117 (117 chains) from GraphBind are utilized, also processed from BioLiP and filtered with CD-Hit. These datasets contain a higher proportion of non-binding residues compared to binding residues, necessitating the use of specialized metrics for evaluation, as detailed below.

The predictive performance of EquiPNAS is compared with an array of existing methods for protein–nucleic acid binding site prediction. The structure-aware methods such as NucBind [8], DNABind [54], GraphBind [19], and GraphSite [20] represent the state-of-the-art methodologies in the literature, with GraphBind currently being the top-performing method for protein–RNA binding site prediction. Two key metrics Receiver Operating Characteristic–Area Under the Curve (ROC-AUC) and Precision-Recall–Area Under the Curve (PR-AUC) are used for this comparison, which provides a robust, threshold-independent assessment of classification models, especially when dealing with imbalanced datasets. The ROC curve plots the true positive rate (TPR) against the false positive rate (FPR), whereas the PR curve plots precision (positive predictive value) against recall (TPR) at different thresholds. Area Under the Curve (AUC) represents the overall ability of the model to distinguish between positive and negative classes, where a ROC-AUC of 1.0 indicates perfect classification. PR-AUC, which is the area under the Precision-Recall curve, is particularly useful for evaluating models on imbalanced datasets where the positive class is rarer than the negative class (*see* **Note 5**).

4 Notes

1. All the structure-aware methods, such as GraphSite, GraphBind, and EquiPNAS rely on 3D structures of proteins for binding site prediction. While the use of experimental structures shows better accuracy for all the competing methods, EquiPNAS experiences a minimal performance drop when using AlphaFold2 predicted structures highlighting its robustness and generalizability with predicted input structures.
2. The accuracy of EquiPNAS predictions correlates with AlphaFold2's self-estimated accuracy, measured by pLDDT with high confidence predictions leading to better ROC-AUC and PR-AUC scores. This suggests that the self-estimated accuracy of AlphaFold2 models can reliably predict the accuracy of EquiPNAS binding site predictions, especially for highly confident AlphaFold2 structures.
3. EquiPNAS incorporates pretrained pLM embeddings from the ESM-2 model as essential sequence-based features, which play a more critical role in performance compared to evolutionary features like PSSM and MSA. Excluding pLM features results in a significant performance drop; in contrast, removing evolutionary features leads to only minor reductions in accuracy. This highlights that pLM embeddings play a major role in EquiPNAS's improved performance, whereas evolutionary features contribute only modestly.
4. The ESM-2 provides a range of pre-trained protein language models (pLMs) with sizes from 8 million to 15 billion parameters. EquiPNAS uses the largest model, esm2_t48_15B_UR50D, with 15 billion parameters by default. Interestingly, EquiPNAS with the smallest model, esm2_t6_8M_UR50D, performed the worst and the performance consistently improved as the number of parameters increased, confirming that the largest model, esm2_t48_15B_UR50D, yields the best results for both protein–DNA and protein–

RNA binding site predictions, showing the significance of large LLMs in protein bioinformatics.

5. While the ROC-AUC achieved by EquiPNAS demonstrates high accuracy, the PR-AUC results indicate poorer performance overall in terms of binding site prediction. A model with high ROC-AUC but low PR-AUC suggests that although it is effective at distinguishing between positive and negative classes, it struggles with precision, leading to a higher number of false positives, in this case, binding residue prediction. This discrepancy highlights the need for further optimization in refining the precision of EquiPNAS, especially in handling imbalanced datasets, to ensure more reliable predictions.

Acknowledgments

This work was partially supported by the National Institute of General Medical Sciences (R35GM138146 to D.B.) and the National Science Foundation (DBI2208679 to D.B.). The authors gratefully acknowledge high-performance computing resources and technical support from the National AI Research Resource Pilot award (NAIRR240093 to D.B.).

References

1. Ferraz RAC, Lopes ALG, Da Silva JAF, Moreira DFV, Ferreira MJN, De Almeida Coimbra SV (2021) DNA–protein interaction studies: a historical and comparative analysis. Plant Methods 17:82
2. Ofran Y, Mysore V, Rost B (2007) Prediction of DNA-binding residues from sequence. Bioinformatics 23:i347–i353
3. Yesudhas D, Batool M, Anwar M, Panneerselvam S, Choi S (2017) Proteins recognizing DNA: structural uniqueness and versatility of DNA-binding domains in stem cell transcription factors. Genes 8:192
4. Zheng M, Sun G, Li X, Fan Y (2024) EGPDI: identifying protein–DNA binding sites based on multi-view graph embedding fusion. Brief Bioinform 25:bbae330
5. Zhou J, Xu R, He Y, Lu Q, Wang H, Kong B (2016) PDNAsite: identification of DNA-binding site from protein sequence by incorporating spatial and sequence context. Sci Rep 6:27653
6. Si J, Zhao R, Wu R (2015) An overview of the prediction of protein DNA-binding sites. IJMS 16:5194–5215
7. Mittler G, Butter F, Mann M (2009) A SILAC-based DNA protein interaction screen that identifies candidate binding proteins to functional DNA elements. Genome Res 19:284–293
8. Su H, Liu M, Sun S, Peng Z, Yang J (2019) Improving the prediction of protein–nucleic acids binding residues via multiple sequence profiles and the consensus of complementary methods. Bioinformatics 35:930–936
9. Zhang J, Chen Q, Liu B (2021) NCBRPred: predicting nucleic acid binding residues in proteins based on multilabel learning. Brief Bioinform 22:bbaa397
10. Zhu Y-H, Hu J, Song X-N, Yu D-J (2019) DNAPred: accurate identification of DNA-binding sites from protein sequence by ensembled hyperplane-distance-based support vector machines. J Chem Inf Model 59:3057–3071
11. Zhang J, Ghadermarzi S, Katuwawala A, Kurgan L (2021) DNAgenie: accurate prediction of DNA-type-specific binding residues in protein sequences. Brief Bioinform 22:bbab336
12. Walia RR, Xue LC, Wilkins K, El-Manzalawy Y, Dobbs D, Honavar V (2014) RNABindRPlus: a predictor that combines machine learning

and sequence homology-based methods to improve the reliability of predicted RNA-binding residues in proteins. PLoS One 9:e97725

13. Hu J, Li Y, Zhang M, Yang X, Shen H-B, Yu D-J (2017) Predicting protein-DNA binding residues by weightedly combining sequence-based features and boosting multiple SVMs. IEEE/ACM Trans Comput Biol Bioinf 14: 1389–1398
14. Zhang J, Kurgan L. SCRIBER: accurate and partner type-specific prediction of protein-binding residues from proteins sequences
15. Yu D-J, Hu J, Yang J, Shen H-B, Tang J, Yang J-Y (2013) Designing template-free predictor for targeting protein-ligand binding sites with classifier ensemble and spatial clustering. IEEE/ACM Trans Comput Biol Bioinf 10: 994–1008
16. Wu Q, Peng Z, Zhang Y, Yang J (2018) COACH-D: improved protein–ligand binding sites prediction with refined ligand-binding poses through molecular docking. Nucleic Acids Res 46:W438–W442
17. Li S, Yamashita K, Amada KM, Standley DM (2014) Quantifying sequence and structural features of protein–RNA interactions. Nucleic Acids Res 42:10086–10098
18. Lam JH, Li Y, Zhu L et al (2019) A deep learning framework to predict binding preference of RNA constituents on protein surface. Nat Commun 10:4941
19. Xia Y, Xia C-Q, Pan X, Shen H-B (2021) GraphBind: protein structural context embedded rules learned by hierarchical graph neural networks for recognizing nucleic-acid-binding residues. Nucleic Acids Res 49:e51–e51
20. Yuan Q, Chen S, Rao J, Zheng S, Zhao H, Yang Y (2022) AlphaFold2-aware protein–DNA binding site prediction using graph transformer. Brief Bioinform 23:bbab564
21. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. Nucleic Acids Res 28(1):235–242
22. Jumper J, Evans R, Pritzel A et al (2021) Highly accurate protein structure prediction with AlphaFold. Nature 596:583–589
23. Abramson J, Adler J, Dunger J et al (2024) Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature 630: 493–500
24. Varadi M, Anyango S, Deshpande M et al (2022) AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res 50:D439–D444
25. Moussad B, Roche R, Bhattacharya D (2023) The transformative power of transformers in protein structure prediction. Proc Natl Acad Sci USA 120:e2303499120
26. Minaee S, Mikolov T, Nikzad N, Chenaghlu M, Socher R, Amatriain X, Gao J (2024) Large language models: a survey
27. Zhao WX, Zhou K, Li J, et al (2023) A survey of large language models
28. Zhou C, Li Q, Li C, et al (2023) A comprehensive survey on Pretrained foundation models: a history from BERT to ChatGPT
29. Liu S, Guo B, Fang C, Wang Z, Luo S, Zhou Z, Yu Z (2024) Enabling resource-efficient AIoT system with cross-level optimization: a survey. IEEE Commun Surv Tutor 26:389–427
30. Dong Q, Li L, Dai D, et al (2024) A survey on in-context learning
31. Chowdhery A, Narang S, Devlin J, et al (2022) PaLM: scaling language modeling with pathways
32. Touvron H, Lavril T, Izacard G, et al (2023) LLaMA: open and efficient foundation language models
33. OpenAI, Achiam J, Adler S, et al (2024) GPT-4 Technical Report
34. Singhal K, Azizi S, Tu T, et al (2022) Large language models encode clinical knowledge
35. Esteva A, Chou K, Yeung S, Naik N, Madani A, Mottaghi A, Liu Y, Topol E, Dean J, Socher R (2021) Deep learning-enabled medical computer vision. npj Digit Med 4:5
36. Tomašev N, Harris N, Baur S et al (2021) Use of deep learning to develop continuous-risk models for adverse event prediction from electronic health records. Nat Protoc 16:2765–2787
37. Yim J, Chopra R, Spitz T et al (2020) Predicting conversion to wet age-related macular degeneration using deep learning. Nat Med 26:892–899
38. Aydın Ö, Karaarslan E (2022) OpenAI ChatGPT generated literature review: digital twin in healthcare. SSRN J. https://doi.org/10.2139/ssrn.4308687
39. Yang G, Liu X, Shi J, Wang Z, Wang G (2024) TCM-GPT: efficient pre-training of large language models for domain adaptation in traditional Chinese medicine. Comput Methods Prog Biomed Update 6:100158
40. Shah RS, Chawla K, Eidnani D, Shah A, Du W, Chava S, Raman N, Smiley C, Chen J, Yang D (2022) WHEN FLUE MEETS FLANG: benchmarks and large pre-trained language model for financial domain

41. Shah A, Chava S (2023) Zero is not hero yet: benchmarking zero-shot performance of LLMs for financial tasks
42. Chen ZZ, Ma J, Zhang X, Hao N, Yan A, Nourbakhsh A, Yang X, McAuley J, Petzold L, Wang WY (2024) A survey on large language models for critical societal domains: finance, healthcare, and law
43. Yang H, Liu X-Y, Wang CD (2023) FinGPT: Open-Source Financial Large Language Models
44. Kalyan KS (2024) A survey of GPT-3 family large language models including ChatGPT and GPT-4. Nat Lang Process J 6:100048
45. Elnaggar A, Heinzinger M, Dallago C, et al (2020) ProtTrans: towards cracking the language of life's code through self-supervised learning. https://doi.org/10.1101/2020.07.12.199554
46. Brandes N, Ofer D, Peleg Y, Rappoport N, Linial M (2022) ProteinBERT: a universal deep-learning model of protein sequence and function. Bioinformatics 38:2102–2110
47. Lin Z, Akin H, Rao R et al (2023) Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 379:1123–1130
48. Chowdhury R, Bouatta N, Biswas S et al (2022) Single-sequence protein structure prediction using a language model and deep learning. Nat Biotechnol 40:1617–1623
49. Rives A, Meier J, Sercu T et al (2021) Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc Natl Acad Sci USA 118:e2016239118
50. Ferruz N, Schmidt S, Höcker B (2022) ProtGPT2 is a deep unsupervised language model for protein design. Nat Commun 13:4348
51. Roche R, Moussad B, Shuvo MH, Bhattacharya D (2023) E(3) equivariant graph neural networks for robust and accurate protein-protein interaction site prediction. PLoS Comput Biol 19:e1011435
52. Roche R, Moussad B, Shuvo MH, Tarafder S, Bhattacharya D (2024) EquiPNAS: improved protein–nucleic acid binding site prediction using protein-language-model-informed equivariant deep graph neural networks. Nucleic Acids Res 52:e27–e27
53. Satorras, V. G., Hoogeboom, E., & Welling, M. (2021, July). E (n) equivariant graph neural networks. In International conference on machine learning (pp. 9323-9332). PMLR.
54. Liu R, Hu J (2013) DNABind: a hybrid algorithm for structure-based prediction of DNA-binding residues by combining machine learning-and template-based approaches. Proteins 81:1885–1899
55. Jiang Z, Shen Y-Y, Liu R (2023) Structure-based prediction of nucleic acid binding residues by merging deep learning- and template-based approaches. PLoS Comput Biol 19:e1011428
56. Song Y, Yuan Q, Zhao H, Yang Y (2023) Accurately identifying nucleic-acid-binding sites through geometric graph learning on language model predicted structures. Brief Bioinform 24:bbad360
57. Liu Y, Tian B (2024) Protein–DNA binding sites prediction based on pre-trained protein language model and contrastive learning. Brief Bioinform 25:bbad488
58. Zhu Y-H, Liu Z, Liu Y, Ji Z, Yu D-J (2024) ULDNA: integrating unsupervised multi-source language models with LSTM-attention network for high-accuracy protein–DNA binding site prediction. Brief Bioinform 25:bbae040
59. Zeng W, Lv D, Liu X, Chen G, Liu W, 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). pp 76–81
60. Littmann M, Heinzinger M, Dallago C, Weissenow K, Rost B (2021) Protein embeddings and deep learning predict binding residues for various ligand classes. Sci Rep 11:23916
61. Wang B, Li W (2024) Advances in the application of protein language modeling for nucleic acid protein binding site prediction. Genes 15:1090
62. Järvelin AI, Noerenberg M, Davis I, Castello A (2016) The new (dis)order in RNA regulation. Cell Commun Signal 14:9
63. Baek M, McHugh R, Anishchenko I, Baker D, DiMaio F (2022) Accurate prediction of nucleic acid and protein-nucleic acid complexes using RoseTTAFoldNA 2022.09.09.507333
64. Shen Y, Chen Z, Mamalakis M, He L, Xia H, Li T, Su Y, He J, Wang YG (2024) A fine-tuning dataset and benchmark for large language models for protein understanding. https://doi.org/10.48550/arXiv.2406.05540
65. Ji Y, Zhou Z, Liu H, Davuluri RV (2021) DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. Bioinformatics 37:2112–2120

66. Zhang Y, Lang M, Jiang J et al (2023) Multiple sequence alignment-based RNA language model and its application to structural inference. Nucleic Acids Res gkad1031:e103

67. Li H-L, Pang Y-H, Liu B (2021) BioSeq-BLM: a platform for analyzing DNA, RNA and protein sequences based on biological language models. Nucleic Acids Res 49:e129

68. Altschul S (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389–3402

69. Steinegger M, Söding J (2017) MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol 35:1026–1028

70. Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M (2022) ColabFold: making protein folding accessible to all. Nat Methods 19:679–682

71. Jing B, Eismann S, Suriana P, Townshend RJL, Dror R (2020) Learning from protein structure with geometric vector perceptrons. In: International conference on learning representations

72. Kingma DP, Ba J (2017) Adam: a method for stochastic optimization. https://doi.org/10.48550/arXiv.1412.6980

73. Loshchilov I, Hutter F (2017) SGDR: stochastic gradient descent with warm restarts. https://doi.org/10.48550/arXiv.1608.03983

74. Yang J, Roy A, Zhang Y (2013) BioLiP: a semi-manually curated database for biologically relevant ligand–protein interactions. Nucleic Acids Res 41:D1096–D1103

75. Huang Y, Niu B, Gao Y, Fu L, Li W (2010) CD-HIT suite: a web server for clustering and comparing biological sequences. Bioinformatics 26:680–682