Analytical Biochemistry 703 (2025) 115862

Contents lists available at ScienceDirect

Analytical Biochemistry

journal homepage: www.elsevier.com/locate/yabio

A comprehensive review of computational methods for Protein-DNA binding site prediction

Zi Liu [a], Wang-Ren Qiu [a], Yan Liu [b], He Yan [c], Wenyi Pei [d,*], Yi-Heng Zhu [e,**], Jing Qiu [f,***]

[a] School of Information Engineering, Jingdezhen Ceramic University, Jingdezhen, 333403, China
[b] Department of Computer Science, Yangzhou University, 196 Huayang West Road, Yangzhou, 225100, China
[c] College of Information Science and Technology & Artificial Intelligence, Nanjing Forestry University, 159 Longpanlu Road, Nanjing, 210037, China
[d] Geriatric Department, Shanghai Baoshan District Wusong Central Hospital, 101 Tongtai North Road, Shanghai, 200940, China
[e] College of Artificial Intelligence, Nanjing Agricultural University, 1 Weigang Road, Nanjing, 210095, China
[f] Information Department, The First Affiliated Hospital of Naval Medical University, 168 Changhai Road, Shanghai, 200433, China

ARTICLE INFO

Keywords:
Protein-DNA binding site
Template detection
Statistical machine learning
Deep learning
Large language model

ABSTRACT

Accurately identifying protein-DNA binding sites is essential for understanding the molecular mechanisms underlying biological processes, which in turn facilitates advancements in drug discovery and design. While biochemical experiments provide the most accurate way to locate DNA-binding sites, they are generally time-consuming, resource-intensive, and expensive. There is a pressing need to develop computational methods that are both efficient and accurate for DNA-binding site prediction. This study thoroughly reviews and categorizes major computational approaches for predicting DNA-binding sites, including template detection, statistical machine learning, and deep learning-based methods. The 14 state-of-the-art DNA-binding site prediction models have been benchmarked on 136 non-redundant proteins, where the deep learning-based, especially pre-trained large language model-based, methods achieve superior performance over the other two categories. Applications of these DNA-binding site prediction methods are also involved.

1. Introduction

Protein-DNA interactions participate in various critical biological processes, including DNA repair, replication and recombination, transcription regulation, and gene expression [1,2]. Accurately locating protein-DNA binding sites is vital for uncovering the molecular-level mechanisms of these processes, thereby advancing drug discovery and design [3–6]. In light of this, DNA-binding site identification has emerged as one of the most hot topics in the post-genomic era [7,8].

Protein-DNA binding sites were primarily recognized through biochemical experiments in the early stage, such as electrophoretic mobility shift assay [9], nuclear magnetic resonance spectroscopy [10], and Cryo-EM [11]. While these methods provide the highest identification accuracy for DNA binding sites, they are typically time-intensive, expensive, and incomplete. As a result, a significant proportion of sequenced proteins remain without DNA-binding annotations. As of December 2024, the UniProt database [12] had amassed ~249 million protein sequences, but less than 0.1 % of these had experimental records of DNA-binding sites. To bridge this gap, there is an urgent need to design efficient computational methods that can quickly and accurately predict DNA-binding sites from protein sequences [13,14].

In recent years, numerous computational methods have been developed for DNA-binding site prediction [15,16]. These methods often employ knowledge-based models trained on available protein data with experimental DNA-binding annotation, enabling the direct inference of DNA-binding sites from protein sequences. Their development is interdisciplinary, involving statistical mathematics, computer science, and molecular biology. Existing methods for DNA-binding site prediction could generally be divided into three main categories, including template detection, statistical machine learning, and deep learning-based methods. This work provides a comprehensive overview of representative methods within each category.

* Corresponding author.
** Corresponding author.
*** Corresponding author.
E-mail addresses: wenyi_1212@163.com (W. Pei), yihzhu@njau.edu.cn (Y.-H. Zhu), power_ko@126.com (J. Qiu).

https://doi.org/10.1016/j.ab.2025.115862
Received 11 December 2024; Received in revised form 20 March 2025; Accepted 6 April 2025
Available online 8 April 2025

2. Protein-DNA interaction data

2.1. Definition of Protein-DNA binding site

Protein-DNA binding sites are specific residues on a protein that directly interact with DNA molecules, as illustrated in Fig. 1. There are two main ways to define protein-DNA binding sites.

The first definition originates from Critical Assessment of Structure Prediction (CASP) [18,19]. Specifically, a protein residue is classified as a DNA binding site if it forms at least one inter-molecular atomic contact with a DNA molecule. Such a contact is defined as a non-hydrogen atom pair from the protein and DNA with a Euclidean distance less than the sum of their van der Waals radii plus 0.5 Å.

The second definition was proposed in Ahmad's work [20], which is the first work to predict DNA-binding sites from protein sequences, as far as we know. An amino acid residue in a protein-DNA complex is identified as a DNA-binding site if the distance between any of its atoms and any atom of the DNA molecule is less than a defined cut-off value, which is usually set to be 3.5 Å.

2.2. Public databases

BioLip [21,22] is the most commonly used protein-ligand interaction database, which has collected ~45000 protein-DNA interaction entries from the Protein Data Bank (PDB), as of December 2024. Each entry is a protein chain with the corresponding sequence, atom-level structure, function annotation, and DNA-interaction annotation, where the CASP criterion defines the protein's DNA-binding residues. There are other famous databases for protein-DNA interaction, including DNAproDB [23], hPDI [24], PDIdb [25], 3D-footprint [26], HOCOMOCO [27], and CIS-BP [28], with the details in Table 1.

2.3. Benchmark datasets

In this work, we utilize the PDNA-136 dataset, constructed in our previous work [15], to benchmark the start-of-the-art protein-DNA binding site prediction methods. The PDNA-136 dataset consists of 136 protein chains, with less than 30 % sequence identity, which was released in the PDB after January 1, 2023. There are 2193 DNA-binding sites and 47287 non-DNA-binding sites in total, where the criterion of

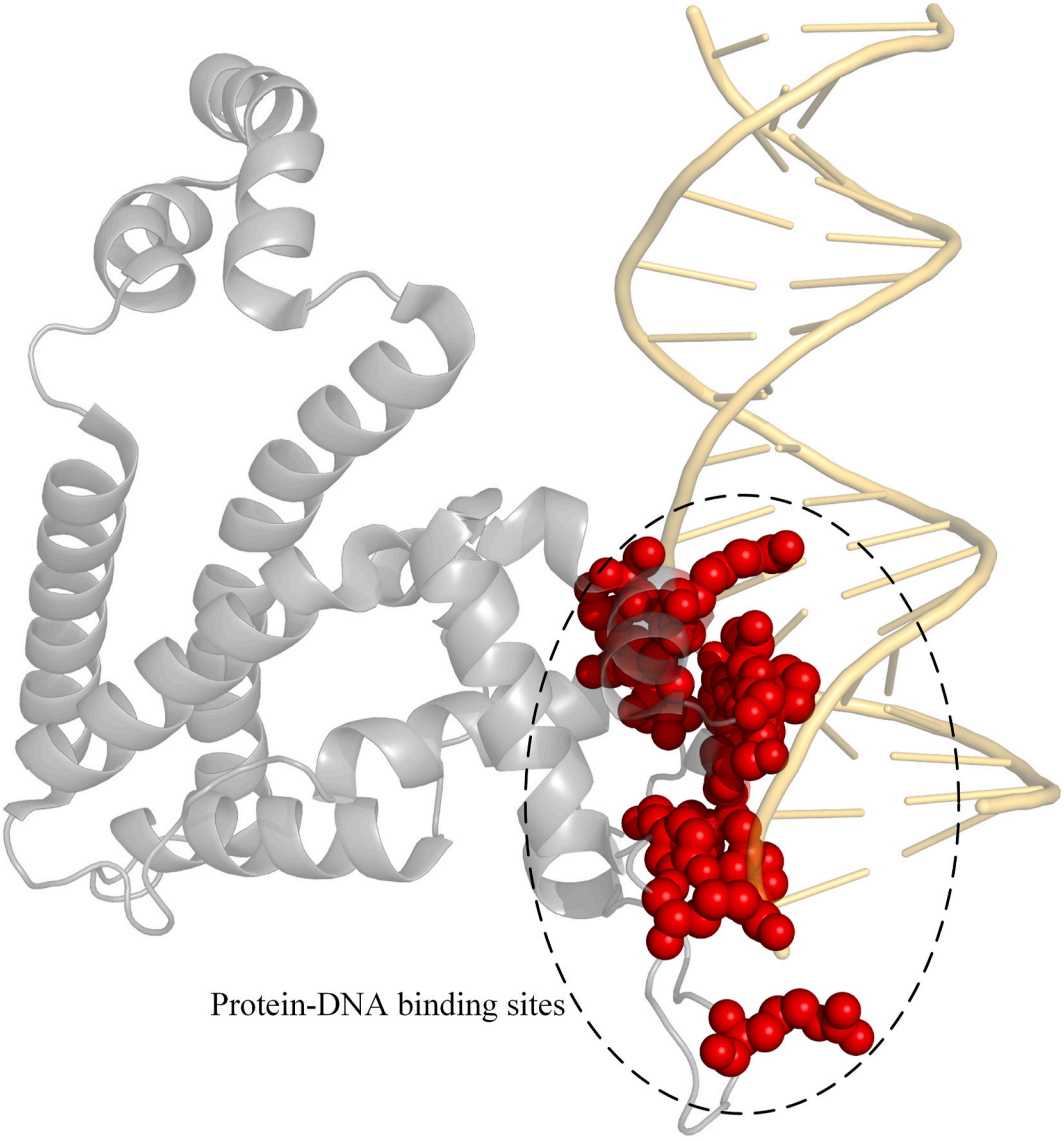

Fig. 1. The flowchart of protein-DNA complex (PDB ID: 3xmf). The atomic-level native structure for this case is downloaded from the PDB database and then visualized as a cartoon representation using PyMOL software [17]. The color scheme is used as follows: protein in gray, DNA in yellow, and DNA-binding site in red. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Table 1
Summary of 7 start-of-the-art protein-DNA interaction databases.

Database	Availability	Note
BioLip [21,22]	https://zhanggroup.org/BioLiP	~45000 interaction entries with sequence identity<90 %
DNAproDB [23]	https://dnaprodb.usc.edu	~6700 DNA-protein complexes
hPDI [24]	http://bioinfo.wilmer.jhu.edu/PDI/	~17000 human protein-DNA interaction entries
PDIdb [25]	http://melolab.org/pdidb	~900 protein-DNA interaction entries
3D-footprint [26]	http://floresta.eead.csic.es/3dfootprint	~11000 protein-DNA complexes
HOCOMOCO [27]	https://hocomoco12.autosome.org	~1400 DNA-binding proteins with specificity patterns
CIS-BP [28]	http://cisbp.ccbr.utoronto.ca/	~160000 DNA-binding proteins with binding motifs

CASP determines DNA-binding sites. There are other important benchmark datasets, including PDNA-960 [15], PDNA-543/PDNA-41 [29], PDNA-335/PDNA-52 [30], PDNA-573/PDNA-129 [31], PDNA-735/PDNA-180 [32], and TR646/TE46 [7], with the details in Table 2.

3. Computational methods for Protein-DNA binding site prediction

3.1. Template detection-based methods

In the early stage, template detection-based methods dominated in protein-DNA binding site prediction [33,34]. These methods are founded on the principle that proteins with similarities in biological attributes, such as sequence or structure, tend to capture similar binding patterns to DNA molecules. The principles of template detection-based methods are straightforward: for a query protein, the corresponding homology templates that share similar biological attributes with itself are detected from public databases; then, the available DNA-binding annotations could be transferred from templates to the query. Based on the involved biological attributes, template detection-based methods can be broadly classified into three categories: sequence alignment-based, structure alignment-based, and hybrid, as summarized in Table 3.

Sequence alignment-based methods employ sequence alignment tools (e.g., HHblits [42] and PSI-BLAST [43]) to assess the sequence similarity between proteins, using this similarity as a metric to detect homology templates with known DNA-binding annotations, with the typical examples of S-SITE [34], Rate4site [35], and alignment-based method in Yu's work [30].

Structure alignment-based methods utilize structure alignment tools (e.g., DALI [44] and TM-align [45]) to measure the similarity of atomic-level protein structures for selecting homology templates. When the query protein's structure is unavailable, tools such as AlphaFold2 [46] and I-TASSER [47] could be employed to predict its structure based on the sequence. The classical examples include PreDs [36], DR_bind [37], Pro-DNA [33], TM-SITE [34], and Morozov's method [38].

Hybrid methods first design sub-methods for DNA-binding site prediction through sequence or structure alignment and then fuse the confidence scores of DNA-binding sites for these sub-methods to generate consensus scores. For example, COACH [34] inherits the prediction scores of S-SITE and TM-SITE, which are driven by sequence and structure alignments, respectively, as mentioned above. Other notable cases in this category are PreDNA [48], FINDSITE [39], ConCavity [40], and COFACTOR [41].

An inherent limitation of template-based detection methods is their heavy reliance on the availability and quality of templates with experimentally DNA-binding annotations. When high-quality templates are unavailable, the prediction accuracy is likely to deteriorate significantly.

3.2. Statistical machine learning-based methods

To address the limitation of the template detection-based methods, machine learning algorithms provide an effective alternative for protein-DNA binding site prediction [49–51].

These approaches focus on representing proteins as feature vectors or

Table 2
Summary of 10 commonly used benchmark datasets for protein-DNA binding site prediction.

Dataset	N_{seq}, N_{bind}, $N_{non\text{-}bind}$	Note	Availability
PDNA-960	960, 18336, 271988,	Training/testing dataset in the ULDNA model [15]	https://github.com/yiheng-zhu/ULDNA
PDNA-136	136, 2193, 47287		
PDNA-543	534, 9549, 134995	Training/testing dataset in the DNAPred model [29]	https://csbioinformatics.njust.edu.cn/dnapred/
PDNA-41	41, 734, 14021		
PDNA-335	335, 6461, 71320	Training/testing dataset in the TargetS model [30]	http://www.csbio.sjtu.edu.cn/TargetS/
PDNA-52	52, 973, 16225		
DNA-573	573, 14479, 145404	Training/testing dataset in the GraphBind model [31]	http://www.csbio.sjtu.edu.cn/bioinf/GraphBind/
DNA-129	129, 2240, 35275		
DNA-735	735, 18611, 178125	Training/testing dataset in the GLMSite model [32]	https://github.com/biomed-AI/nucleic-acid-binding
DNA-180	180, 4255, 60964		
TR646	646, 15636, 298503	Training/testing dataset in the CLAPE model [7]	https://github.com/YAndrewL/clape/
TE46	46, 965, 9911		

Note: N_{seq}, N_{bind}, and $N_{non\text{-}bind}$ are the numbers of sequences, DNA-binding residues, and non-DNA binding residues.

Table 3
Summary of 13 state-of-the-art template detection-based methods for protein-DNA binding site prediction.

Type	Method	Ref[a]	Year	Availability
Sequence alignment	S-SITE	[34]	2013	https://zhanggroup.org/COACH/
	Rate4site	[35]	2004	https://www.tau.ac.il/~itaymay/cp/rate4site.html
	alignment-based method	[30]	2013	NA[b]
Structure alignment	PreDs	[36]	2005	NA
	DR_bind	[37]	2012	http://dnasite.limlab.ibms.sinica.edu.tw/
	Pro-DNA	[33]	2005	NA
	TM-SITE	[34]	2013	https://zhanggroup.org/COACH/
	Morozov's method	[38]	2005	NA
Hybrid	COACH	[34]	2013	https://zhanggroup.org/COACH/
	Pro-DNA	[33]	2013	NA
	FINDSITE	[39]	2008	http://cssb.biology.gatech.edu/findsite
	ConCavity	[40]	2009	https://compbio.cs.princeton.edu/concavity/
	COFACTOR	[41]	2012	https://zhanggroup.org/COFACTOR/

[a] Ref: Reference.
[b] NA: Not available.

matrices derived from various biological views, which are then fed to machine learning algorithms for training models for protein-DNA binding site prediction.

Early prediction methods relied on manually designed protein features, such as position-specific scoring matrix (PSSM) [52], secondary structure matrix (SSM) [30], relative solvent accessibility (RSA) [53], and physicochemical property vector (PPV) [54], which were processed by statistical machine learning algorithms like support vector machines (SVM) [55] and random forest (RF) [56] to build DNA-binding site prediction models. Taking DNAPred [29] as an example, it extracts four types of features (i.e., PSSM, SSM, SASA, and DNA-binding frequency) from the protein sequence, which are then captured by ensemble hyperplane-distance-based support vector machines for training prediction models. Other elegant examples include DBS-PSSM [52], DNABindR [57], DP-Bind [58], BindN-RF [59], BindN + [60], MetaDBSite [50], DNABR [54], TargetS [30], DNABind [61], PDNAsite [62], TargetDNA [53], EC-RUS [51], funDNApred [63], HybridNAP [64], SVMnuc [65], DNAgenie [66], and DRBpred [67], with the details in Table 4.

While machine learning methods serve as a complement to template-based detection approaches, their prediction accuracy can sometimes be suboptimal. This is primarily due to poorly designed or overly simplistic feature representations, which may fail to capture relevant and valuable information from the input sequences.

3.3. Deep learning-based methods

Recently, deep learning techniques inspired by developments in computer vision have been applied to address the weakness of manually crafted feature representations [68,69]. A key advantage of deep learning models is their capacity to build complex neural network architectures specifically designed for various data structures representing proteins, including one-dimensional sequences, two-dimensional contact maps, and three-dimensional atomic coordinates. This capability enables deep-level and comprehensive information extraction from input sequences/structures, significantly expanding the potential and richness of feature representations. Deep learning-based methods can be broadly divided into two categories based on the use of pre-training: direct training-based methods and pre-trained large language model-based methods.

Early deep learning approaches trained prediction models directly on protein sequences with annotated DNA-binding sites by integrating deep neural networks, such as recurrent neural network (RNN) and convolutional neural network [70], with sequence encoding strategies. Taking CNNsite [68] as an example, the first deep-learning model for DNA-binding site prediction (to our best knowledge), it extracts evolutionary and motif features from the sequences, then processed by the CNN model. Other notable examples include EL_LSTM [71], iProDNA-CapsNet [69], NCBRPred [72], PredDBR [73], iDRNA-ITF [74], DeepDISOBind [75], Guan's method [76], DBpred [77], Zhao's method [78], and DeepDBS [79]. Additionally, several methods integrate structural knowledge with sequence data to implement prediction models, with typical examples of DeepDISE [80], GraphBind [31], BindWeb [81], and HybridDBRpred [82] as listed in Table 5.

The aforementioned deep-learning models generally deliver more accurate predictions compared to statistical machine learning-based methods. However, there remains significant potential for further improvement. Specifically, the performance of these models, trained on protein data with experimental DNA-binding annotations, largely hinges on the scale of the training datasets. When the training data is limited, deep learning models may struggle to fully capture the relationships between protein sequences/structures and DNA-binding patterns, which may result in suboptimal prediction performance. As of December 2024, there are only ~11000 experimental protein-DNA complexes with full length in the PDB database. Such limited training data may be inadequate for training high-accuracy deep learning models, particularly when the neural networks are overly large.

To tackle the challenges caused by insufficient training data, leveraging pre-trained large language models has emerged as a promising approach. First, deep learning techniques are employed to pre-train an unsupervised large language model on a huge protein sequence dataset without any DNA-binding annotations, leveraging evolutionary, structural, and ligand-binding patterns. Then, this language model could encode the input sequences as a feature embedding

Table 4
Summary of 16 statistical machine learning-based methods for protein function prediction.

Method	Ref[a]	Year	Feature	Classifier	Availability
DBS-PSSM	[52]	2005	PSSM	MLP	NA[b]
DNABindR	[57]	2006	AAC	NB	NA
DP-Bind	[58]	2007	PSSM	PLR	http://lcg.rit.albany.edu/dp-bind/
BindN-RF	[59]	2009	PSSM + PPV	RF	NA
BindN+	[60]	2010	PSSM + PPV	SVM	http://bioinfo.ggc.org/bindn+/
MetaDBSite	[50]	2011	Meta[c]	SVM	http://sysbio.zju.edu.cn/metadbsite
DNABR	[54]	2012	PSSM + AAC + OBV	RF	http://www.cbi.seu.edu.cn/DNABR
TargetS	[30]	2013	PSSM + SSM	SVM	http://www.csbio.sjtu.edu.cn/TargetS/
DNABind	[61]	2013	PSSM + RSA + AAC + PPV	SVM	http://mleg.cse.sc.edu/DNABind/
PDNAsite	[62]	2016	PSSM + SSM + RSA + AAC	SVM	NA
TargetDNA	[53]	2016	PSSM + RSA	SVM	http://csbio.njust.edu.cn/bioinf/targetdna/
EC-RUS	[51]	2017	PSSM + RSA	WSRC	NA
funDNApred	[63]	2018	RAA + RSA + ECO	FCM	http://biomine.cs.vcu.edu/servers/funDNApred/
HybridNAP	[64]	2019	RAA + RSA + ECO	LCM	http://biomine.cs.vcu.edu/servers/hybridNAP/
SVMnuc	[65]	2019	PSSM + SSM + HMMP	SVM	NA
DNAPred	[29]	2019	PSSM + SSM + RSA + DBF	SVM	https://csbioinformatics.njust.edu.cn/dnapred/
DNAgenie	[66]	2021	PSSM + SSM + RSA + AAC +PPV + ECO	LR + SVM + RF +KNN + NB	http://biomine.cs.vcu.edu/servers/DNAgenie/
DRBpred	[67]	2024	PSSM + SSM + RSA +AAC + PPV + SD	LR + SVM + RF +KNN + ET	https://bmll.cs.uno.edu

[a] Ref: Reference.
[b] NA: Not available.
[c] Meta means that the input of the model is the confidence score outputted by existing protein-DNA binding site predictors. PSSM: Position-specific scoring matrix; AAE: Amino acid coding; PPV: Physicochemical property vector; OBV: Orthogonal binary vector; RSA: Relative solvent accessibility; RAA: Relative propensity of specific amino acids for the DNA-binding; ECO: Evolutionary conservation; HMMP: Hidden Markov model profile; DBF: DNA-binding frequency; SD: Structure descriptor; MLP: Multi-layer perceptron; NB: Naïve Bayes; PLR: Penalized logistic regression; WSRC: Weighted sparse representation based classifier; FCM: Fuzzy cognitive map; LCM: Linear scoring function; KNN: K-nearest neighbors; ET: Extra tree; LGBM: Light gradient boosting machine; CGB: Categorical gradient boosting.

Table 5
Summary of 15 popular direct training-based methods in deep learning-based DNA-binding site prediction.

Method	Ref[a]	Year	network model	Availability
CNNsite	[70]	2016	CNN	NA
EL_LSTM	[71]	2018	LSTM	NA
iProDNA-CapsNet	[69]	2019	CNN	https://github.com/ngphubinh/iProDNA-CapsNet
NCBRPred	[72]	2021	GRU	http://bliulab.net/NCBRPred/
PredDBR	[73]	2021	CNN	https://jun-csbio.github.io/PredDBR/
iDRNA-ITF	[74]	2022	CA + GRU	http://bliulab.net/iDRNA-ITF/
DeepDISOBind	[75]	2022	CNN	https://www.csuligroup.com/DeepDISOBind/
Guan's method	[76]	2022	AN + CNN	https://github.com/ShixuanGG/DNA-protein_binding_residues
DBpred	[77]	2022	CNN	https://webs.iiitd.edu.in/raghava/dbpred/
Zhao's method	[78]	2023	AN + CNN	https://github.com/HaipengZZhao/Prediction-of-Residues
DeepDBS	[79]	2024	CNN + LSTM + RF	https://github.com/BioMaticsDotPK/DeepDBS
HybridDBRpred	[82]	2024	AN	http://biomine.cs.vcu.edu/servers/hybridDBRpred/
DeepDISE	[80]	2021	CNN	NA
GraphBind	[31]	2021	GNN	http://www.csbio.sjtu.edu.cn/bioinf/GraphBind/
BindWeb	[81]	2022	GNN + LSTM	http://www.csbio.sjtu.edu.cn/bioinf/BindWeb/

[a] Ref: Reference. CNN: Convolutional neural network; LSTM: Long short-term memory network; GNN: Graph neural network; GRU: Gated recurrent units; CA: Convolution attention; AN: Attention network; RF: Random forest.

matrix, in which the complex DNA-binding patterns are buried. Finally, the encoded feature embedding is processed by a supervised neural network for decoding the corresponding DNA-binding patterns.

Recently, a variety of biological large language models have been developed, demonstrating exceptional performance across numerous bioinformatics tasks, including protein structure and function prediction [46,83], with representative examples of SeqVec [84], TAPE [85], ESM-1b [86], ProtTrans [87], ESM2 [88], SaProt [89], Ankh [90], and CARP [91]. Their superior performance mainly stems from the use of large-scale training datasets and highly sophisticated neural network architectures, with the details in Table 6.

Table 6
Summary of 8 state-of-the-art biological large language models.

Model	Ref[a]	Year	(Layers, Params)[b]	Availability
SeqVec	[84]	2019	(3, 93 M)	https://github.com/Rostlab/SeqVec
TAPE	[85]	2019	(12, 38 M)	https://github.com/songlab-cal/tape
ESM-1b	[86]	2021	(33, 650 M)	https://github.com/facebookresearch/esm
ProtTrans	[87]	2021	(24, 3B)	https://github.com/agemagician/ProtTrans
ESM2	[88]	2023	(48, 15B)	https://github.com/facebookresearch/esm
SaProt	[89]	2023	(33, 650 M)	https://github.com/westlake-repl/SaProt
Ankh	[90]	2023	(48, 1.15B)	https://github.com/agemagician/Ankh/tree/main
CARP	[91]	2024	(56, 640 M)	https://github.com/microsoft/protein-sequence-models

[a] Ref: Reference.
[b] Layers and params: The number of layers and hyper-parameters for neural networks in biological large language models.

In light of the superior performance, the above-mentioned biological large language models have been widely used in protein-DNA binding site prediction. Taking ULDNA [15] as an example, it inherits the sequence feature embeddings from three large language models (i.e., ESM2 [88], ProtTrans [87], and ESM-MSA [92]), which are then fed to an LSTM-attention architecture for implementing a high-accuracy DNA-binding site prediction model. Other elegant examples includes bindEmbed21 [93], GraphSite [94], GLMSite [32], NABind [95], Shan's method [96], GraphPBSP [97], PDNAPred [67], EquiPNAS [98], CLAPE [7], and EGPDI [99], as summarized in Table 7.

In summary, deep learning-based (especially pre-trained large language model-based) methods have emerged as the leading approach for protein-DNA binding site prediction, often outperforming template-based detection and statistical machine-learning methods. However, their drawback is the heavy dependency on large-scale training data and huge computational resources. Nevertheless, their limitation lies in their strong dependence on large-scale training datasets and substantial computational resources.

3.4. Evaluation metric

In evaluating protein-DNA binding site prediction models [14,101–103], six key metrics are commonly utilized, including Sensitivity (Sen), Specificity (Spe), Accuracy (Acc), Precision (Pre), F1-Score (F1), and Mathew's Correlation Coefficient (Mcc), with the following definitions:

$$Sen = TP/(TP + FN) \tag{1}$$

$$Spe = TN/(TN + FP) \tag{2}$$

Table 7
Summary of 11 competing pre-trained large language model-based methods in deep learning-based DNA-binding site prediction.

Method	Ref[a]	Year	Network models[b]	Availability
bindEmbed21	[93]	2021	ProtTrans + CNN	https://github.com/Rostlab/bindPredict
GraphSite	[94]	2022	AlphaFold2 [46] +GAT	https://biomed.nscc-gz.cn/apps/GraphSite
GLMSite	[32]	2023	ProtTrans + GVP-GNN	https://github.com/biomed-AI/nucleic-acid-binding
NABind	[95]	2023	(ESM-MSA, ESM-1F [100]) + EGAT	http://liulab.hzau.edu.cn/NABind/
Shan's method	[96]	2024	(ESM2, ProtBert)+ (CNN, LSTM, AN)	NA
GraphPBSP	[97]	2024	ProtTrans + GNN	https://github.com/ChunhuaLab/GraphPBSP
PDNAPred	[67]	2024	(ESM2, ProtTrans) + CNN-GRU	https://github.com/zlr-zmm/PDNAPred
EquiPNAS	[98]	2024	ESM2+EGNN	https://github.com/Bhattacharya-Lab/EquiPNAS
CLAPE	[7]	2024	ProtBert+(MLP, CNN, LSTM)	https://github.com/YAndrewL/clape
EGPDI	[99]	2024	(ESM2, ProtTrans)+ (EGNN + GCN)	https://github.com/HaaZheng/EGPDI
ULDNA	[15]	2024	(ESM2, EMS-MSA, ProtTrans) + LSTM-AN	https://github.com/yiheng-zhu/ULDNA

[c]NA: Not available. CNN: Convolutional neural network; GAT: Graph attention network; LSTM: Long short-term memory network; AN: Attention network; GRU: Gated recurrent units; EGNN: Equivariant graph neural network; GVP-GNN: Geometric vector perceptron-based graph neural network; EGAT: Edge aggregated graph attention network; MLP: Multi-layer perceptron; GCN: Graph convolution network; GNN: Graph neural network.
[a] Ref: Reference.
[b] Network models consist of a biological large language model for feature embeddings and a supervised training model for DNA-binding site prediction.

$$Acc = (TP + TN)/(TP + FP + TN + FN) \quad (3)$$

$$Pre = TP/(TP + FP) \quad (4)$$

$$F1 = 2 \times (Pre \times Sen)/(Pre + Sen) \quad (5)$$

$$Mcc = (TP \times TN - FP \times FN) \Big/ \sqrt{(TP + FP)(TN + FN)(TP + FN)(TN + FP)} \quad (6)$$

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

The abovementioned six indices are threshold-dependent, making it critical to select an appropriate threshold for fair comparisons between different protein-DNA binding site prediction models. In this study, the threshold is set to be 0.5 in all performance benchmarking. Additionally, to assess the overall prediction performance of the models, a threshold-independent metric, i.e., the area under the receiver operating characteristic curve (AUROC), is employed [101,104,105].

3.5. Performance comparision between existing models for Protein-DNA binding site prediction

We have benchmarked 14 competing protein-DNA binding site prediction methods on the PDNA-136 test dataset, as summarized in Table 8. There are one template-detection-based methods (i.e., sequence-alignment-base method, SABM), five statistical machine learning-based methods (i.e., DP-Bind [58], TargetS [30], TargetDNA [53], DNAPred [29], and DNAgenie [66]), five directly training-based deep learning methods (i.e., NCBRPred [72], PredDBR [73], iDRNA-ITF [74], hybridDBRpred [82], and GraphBind [31]), and three pre-trained large language model-based deep learning methods (i.e., GraphSite [94], CLAPE [7], and ULDNA [15]).

In SABM, we perform DNA-binding site prediction as follows. Given a query sequence in the PDNA-136 test dataset, it is aligned to all training sequences in the PDNA-960 dataset (see details in Table 2) using the Needleman-Wunsch algorithm [106], where the training sequence that shares the highest sequence identity to the query is selected as the homology template; then, the residues in the query sequence aligned to the DNA-binding sites in the selected template are predicted as DNA-binding sites. For the other 13 DNA-binding site prediction methods, we implemented the standalone software or accessed the webserver platforms with the default parameter settings on the PDNA-136 test dataset.

In Table 8, the deep-learning methods, especially pre-trained large language model-based methods, achieve significantly better performance than statistical machine learning-based and template detection-based methods. Specifically, from the view of F1, MCC, and AUROC values, the best three performers, i.e., ULDNA, GraphSite, and GraphBind, both employ deep learning techniques, where the top two methods (ULDNA and GraphSite) involve biological large language models. The significant advantage of ULDNA and GraphSite stems from utilizing large language models, which effectively capture the complex DNA-binding patterns from millions of sequences. Moreover, ULDNA achieves 2.7 % and 1.6 % increases in MCC and AUROC values compared to GraphSite. This may be because ULDNA employs three large language models pre-trained on different database sources, which could extract the complementary feature embeddings with DNA-binding patterns. It cannot be denied that SABM achieves the worst performance of all the methods. The underlying reason is that there are fewer high-quality templates in the training dataset for most query sequences. However, the template detection-based methods currently cannot be completely replaced by machine learning-based methods due to the following reasons. First, template detection-based methods are computationally inexpensive and well suited for resource-limited environments. Moreover, with highly homologous templates available, these methods still achieve outstanding predictive performance in DNA-binding prediction. Additionally, the integration of template detection and machine learning-based methods could further improve prediction accuracy, as demonstrated in NABind [95].

3.6. Applications of Protein-DNA binding site prediction

Protein-DNA binding site prediction models are essential for applications in the following areas.

(1) **Protein function annotation.** The identification of DNA-binding sites is pivotal for functional annotation, as it highlights specific regions where proteins interact with DNA. This insight aids in uncovering the biological roles of these proteins and their participation in molecular processes, thereby enhancing our understanding of their functions [107–109].
(2) **Drug Discovery and Design.** DNA-binding site prediction plays a vital role in drug discovery and design by pinpointing specific protein-DNA interaction regions that can be targeted for therapeutic intervention. This facilitates the development of drugs that regulate these interactions, providing innovative approaches to treat diseases linked to abnormal protein-DNA binding [110–112].
(3) **Gene Regulation Study.** Understanding DNA-binding sites is fundamental to gene regulation studies by specifying the locations and mechanisms through which proteins, including

Table 8
The performance comparison between 14 competing DNA-binding site prediction methods on the PDNA-136 test dataset.

	Method	Sen	Spe	Acc	Pre	F1	MCC	AUROC
Template detection-based method	SABM	0.053	0.976	0.935	0.092	0.068	0.038	-[b]
Statistical machine learning-based method	DP-Bind[a]	0.622	0.787	0.779	0.119	0.200	0.199	–
	TargetS[a]	0.266	0.959	0.929	0.233	0.248	0.211	–
	TargetDNA[a]	**0.738**	0.717	0.718	0.108	0.188	0.204	0.802
	DNAPred[a]	0.704	0.766	0.763	0.122	0.209	0.222	0.820
	DNAgenie	0.672	0.659	0.658	0.079	0.140	0.124	0.678
Directly training-based deep learning method	NCBRPred[a]	0.307	0.966	0.936	0.293	0.300	0.267	0.799
	PredDBR[a]	0.323	0.954	0.926	0.247	0.280	0.244	0.775
	iDRNA-ITF[a]	0.325	0.966	0.937	0.304	0.314	0.282	–
	hybridDBRpred	0.365	0.896	0.873	0.140	0.202	0.168	0.716
	GraphBind[a]	0.411	0.969	0.944	0.377	**0.393**	0.364	0.898
Pre-trained large language model-based deep learning method	GraphSite[a]	0.335	**0.981**	0.952	0.445	0.382	0.361	0.908
	CLAPE	0.226	0.980	0.946	0.339	0.271	0.250	0.815
	ULDNA[a]	0.271	0.992	**0.960**	**0.607**	0.375	**0.388**	**0.924**

[a] The DNA-binding site prediction methods have been benchmarked with other thresholds in our previous work [15].
[b] "-" means that the AUROC value is unavailable.

transcription factors and other DNA-binding proteins, interact with the genome to regulate gene expression [113–116].

(4) **Epigenetics and Chromatin Dynamics.** DNA-binding site prediction is critical to advance epigenetics and chromatin dynamics by facilitating the identification and functional characterization of protein-DNA interactions, which are essential for understanding gene regulation and chromatin remodeling [117–119].

(5) **Evolutionary Study.** Analyzing protein-DNA binding interaction contributes to evolutionary studies by allowing researchers to analyze and compare regulatory mechanisms across species, identify conserved or divergent patterns, and understand the evolutionary pressures that influence gene regulation [120,121].

4. Discussions

Protein-DNA binding site prediction can be framed as a binary classification problem that can be addressed using machine learning techniques. In recent years, machine learning-based, especially deep learning-based, methods have achieved remarkable progress in predicting DNA-binding sites. The work presents a comprehensive overview of state-of-the-art DNA-binding site prediction models, with the performance benchmarking on 136 non-redundant proteins, covering the following observations.

(1) Template detection-based methods were the predominant approach in the early DNA-binding site prediction studies, involving sequence alignment-based, structure alignment-based, and hybrid methods. A common weakness of these methods is their heavy reliance on the availability and quality of homology templates.

(2) Statistical machine learning methods could be an effective complement to template detection-based methods. However, they sometimes achieve suboptimal prediction performance due to the overly simplistic feature representations that cannot explore the deep-level relationship between sequence/structure and DNA-binding pattern.

(3) Deep-learning methods have emerged in recent years, including directly train-based and pre-trained large language model-based methods. They usually achieve superior performance than template detection-based and statistical machine learning-based methods, becoming the main driving force in DNA-binding site prediction.

Despite substantial advancements, several challenges remain. First, while most deep learning methods rely solely on primary sequences for DNA-binding site prediction, the rapid advancements in protein structure prediction models, such as AlphaFold2 [46] and ESMFold [88], highlight the need to more deeply integrate structural information, which is often crucial for exploring DNA-binding patterns [16]. Moreover, when structures predicted by AlphaFold2 or other methods are inaccurate, they can be further refined and corrected using cryo-EM data [122]. Additionally, since protein-DNA interactions are influenced not only by the protein but also by the DNA molecule, integrating DNA-specific information into binding site prediction could be a promising strategy to enhance accuracy. Studies along these lines are in progress [123,124].

CRediT authorship contribution statement

Zi Liu: Writing – review & editing, Writing – original draft, Methodology, Data curation. **Wang-Ren Qiu:** Writing – review & editing. **Yan Liu:** Writing – review & editing. **He Yan:** Writing – review & editing. **Wenyi Pei:** Writing – review & editing, Supervision, Conceptualization. **Yi-Heng Zhu:** Writing – review & editing, Supervision, Conceptualization. **Jing Qiu:** Writing – review & editing, Supervision, Conceptualization.

Declaration of competing interest

The authors have declared no competing interests.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (No. 62306142 and 62402227), Fundamental Research Funds for the Central Universities (YDZX2025024), the Department of Education of Jiangxi Province (GJJ2400905), and the Jiangsu Funding Program for Excellent Postdoctoral Talent (No. 2023ZB224).

Data availability

No data was used for the research described in the article.

References

[1] G.D. Stormo, Y. Zhao, Determining the specificity of protein–DNA interactions, Nat. Rev. Genet. 11 (11) (2010) 751–760.
[2] L.A. Gallagher, E. Velazquez, S.B. Peterson, J.C. Charity, M.C. Radey, M. J. Gebhardt, et al., Genome-wide protein–DNA interaction site mapping in bacteria using a double-stranded DNA-specific cytosine deaminase, Nature microbiology 7 (6) (2022) 844–855.
[3] R. Esmaeeli, A. Bauzá, A. Perez, Structural predictions of protein-DNA binding: meld-dna, Nucleic Acids Res. 51 (4) (2023) 1625–1636.
[4] Y. Hua, J. Li, Z. Feng, X. Song, J. Sun, D.-J. Yu, Protein drug interaction prediction based on attention feature fusion, J. Comput. Res. Dev. 59 (9) (2022) 2051–2065.
[5] E. Kim, Y.-J. Kim, Z. Ji, J.M. Kang, M. Wirianto, K.R. Paudel, et al., ROR activation by Nobiletin enhances antitumor efficacy via suppression of IκB/NF-κB signaling in triple-negative breast cancer, Cell Death Dis. 13 (4) (2022) 374.
[6] H. Shao, T. Peng, Z. Ji, J. Su, X. Zhou, Systematically studying kinase inhibitor induced signaling network signatures by integrating both therapeutic and side effects, PLoS One 8 (12) (2013) e80832.
[7] Y. Liu, B. Tian, Protein–DNA binding sites prediction based on pre-trained protein language model and contrastive learning, Briefings Bioinf. 25 (1) (2024) bbad488.
[8] M.W.U. Kabir, D.M. Alawad, P. Pokhrel, M.T. Hoque, DRBpred: a sequence-based machine learning method to effectively predict DNA-and RNA-binding residues, Comput. Biol. Med. 170 (2024) 108081.
[9] L.M. Hellman, M.G. Fried, Electrophoretic mobility shift assay (EMSA) for detecting protein–nucleic acid interactions, Nat. Protoc. 2 (8) (2007) 1849–1861.
[10] G. Otting, Y. Qian, M. Billeter, M. Müller, M. Affolter, W.J. Gehring, et al., Protein–DNA contacts in the structure of a homeodomain–DNA complex determined by nuclear magnetic resonance spectroscopy in solution, EMBO J. 9 (10) (1990) 3085–3092.
[11] Y. Yu, S. Li, Z. Ser, H. Kuang, T. Than, D. Guan, et al., Cryo-EM structure of DNA-bound Smc5/6 reveals DNA clamping enabled by multi-subunit conformational changes, Proc. Natl. Acad. Sci. 119 (23) (2022) e2202799119.
[12] U. Consortium, UniProt: a worldwide hub of protein knowledge, Nucleic Acids Res. 47 (D1) (2019) D506–D515.
[13] Q. Yuan, S. Chen, J. Rao, S. Zheng, H. Zhao, Y. Yang, AlphaFold2-aware protein-DNA binding site prediction using graph transformer, Briefings Bioinf. 23 (2) (2022) bbab564.
[14] K. Qu, L. Wei, Q. Zou, A review of DNA-binding proteins prediction methods, Curr. Bioinf. 14 (3) (2019) 246–254.
[15] Y.-H. Zhu, Z. Liu, Y. Liu, Z. Ji, D.-J. Yu, ULDNA: integrating unsupervised multi-source language models with LSTM-attention network for high-accuracy protein–DNA binding site prediction, Briefings Bioinf. 25 (2) (2024) bbae040.
[16] R. Mitra, J. Li, J.M. Sagendorf, Y. Jiang, A.S. Cohen, T.-P. Chiu, et al., Geometric deep learning of protein–DNA binding specificity, Nat. Methods 21 (9) (2024) 1674–1683.
[17] S. Yuan, H.S. Chan, Z. Hu, Using PyMOL as a platform for computational drug design, Wiley Interdiscip. Rev. Comput. Mol. Sci. 7 (2) (2017) e1298.
[18] T. Gallo Cassarino, L. Bordoli, T. Schwede, Assessment of ligand binding site predictions in CASP10, Proteins: Struct., Funct., Bioinf. 82 (2014) 154–163.
[19] T. Schmidt, J. Haas, T.G. Cassarino, T. Schwede, Assessment of ligand-binding residue predictions in CASP9, Proteins: Struct., Funct., Bioinf. 79 (S10) (2011) 126–136.
[20] S. Ahmad, M.M. Gromiha, A. Sarai, Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information, Bioinformatics 20 (4) (2004) 477–486.
[21] J. Yang, A. Roy, Y. Zhang, BioLiP: a semi-manually curated database for biologically relevant ligand–protein interactions, Nucleic Acids Res. 41 (D1) (2012) D1096–D1103.
[22] C. Zhang, X. Zhang, P.L. Freddolino, Y. Zhang, BioLiP2: an updated structure database for biologically relevant ligand–protein interactions, Nucleic Acids Res. 52 (D1) (2024) D404–D412.

[23] J.M. Sagendorf, N. Markarian, H.M. Berman, R. Rohs, DNAproDB: an expanded database and web-based tool for structural analysis of DNA–protein complexes, Nucleic Acids Res. 48 (D1) (2020) D277–D287.
[24] Z. Xie, S. Hu, S. Blackshaw, H. Zhu, J. Qian, hPDI: a database of experimental human protein–DNA interactions, Bioinformatics 26 (2) (2010) 287–289.
[25] T. Norambuena, F. Melo, The protein-DNA interface database, BMC Bioinf. 11 (2010) 1–12.
[26] B. Contreras-Moreira, 3D-footprint: a database for the structural analysis of protein–DNA complexes, Nucleic Acids Res. 38 (suppl_1) (2010) D91–D97.
[27] I.E. Vorontsov, I.A. Eliseeva, A. Zinkevich, M. Nikonov, S. Abramov, A. Boytsov, et al., HOCOMOCO in 2024: a rebuild of the curated collection of binding models for human and mouse transcription factors, Nucleic Acids Res. 52 (D1) (2024) D154–D163.
[28] M.T. Weirauch, A. Yang, M. Albu, A.G. Cote, A. Montenegro-Montero, P. Drewe, et al., Determination and inference of eukaryotic transcription factor sequence specificity, Cell 158 (6) (2014) 1431–1443.
[29] Y.-H. Zhu, J. Hu, X.-N. Song, D.-J. Yu, DNAPred: accurate identification of DNA-binding sites from protein sequence by ensembled hyperplane-distance-based support vector machines, J. Chem. Inf. Model. 59 (6) (2019) 3057–3071.
[30] D.-J. Yu, J. Hu, J. Yang, H.-B. Shen, J. Tang, J.-Y. Yang, Designing template-free predictor for targeting protein-ligand binding sites with classifier ensemble and spatial clustering, IEEE ACM Trans. Comput. Biol. Bioinf 10 (4) (2013) 994–1008.
[31] Y. Xia, C.-Q. Xia, X. Pan, H.-B. Shen, GraphBind: protein structural context embedded rules learned by hierarchical graph neural networks for recognizing nucleic-acid-binding residues, Nucleic Acids Res. 49 (9) (2021) e51, e51.
[32] Y. Song, Q. Yuan, H. Zhao, Y. Yang, Accurately identifying nucleic-acid-binding sites through geometric graph learning on language model predicted structures, Briefings Bioinf. 24 (6) (2023) bbad360.
[33] N. Bhardwaj, R. E. Langlois, G. Zhao, H. Lu, Structure based prediction of binding residues on DNA-binding proteins, in: 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference: IEEE, 2006, pp. 2611–2614.
[34] J. Yang, A. Roy, Y. Zhang, Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment, Bioinformatics 29 (20) (2013) 2588–2595.
[35] T. Pupko, R.E. Bell, I. Mayrose, F. Glaser, N. Ben-Tal, Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues, Bioinformatics 18 (suppl_1) (2002) S71–S77.
[36] Y. Tsuchiya, K. Kinoshita, H. Nakamura, PreDs: a server for predicting dsDNA-binding site on protein molecular surfaces, Bioinformatics 21 (8) (2005) 1721–1723.
[37] Y.C. Chen, J.D. Wright, C. Lim, DR_bind: a web server for predicting DNA-binding residues from the protein structure based on electrostatics, evolution and geometry, Nucleic Acids Res. 40 (W1) (2012) W249–W256.
[38] A.V. Morozov, J.J. Havranek, D. Baker, E.D. Siggia, Protein-DNA binding specificity predictions with structural models, Nucleic Acids Res. 33 (18) (2005) 5781–5798.
[39] M. Brylinski, J. Skolnick, A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation, Proc. Natl. Acad. Sci. 105 (1) (2008) 129–134.
[40] J.A. Capra, R.A. Laskowski, J.M. Thornton, M. Singh, T.A. Funkhouser, Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure, PLoS Comput. Biol. 5 (12) (2009) e1000585.
[41] A. Roy, J. Yang, Y. Zhang, COFACTOR: an accurate comparative algorithm for structure-based protein function annotation, Nucleic Acids Res. 40 (W1) (2012) W471–W477.
[42] M. Remmert, A. Biegert, A. Hauser, J. Söding, HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment, Nat. Methods 9 (2) (2012) 173–175.
[43] S.F. Altschul, T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller, et al., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (17) (1997) 3389–3402.
[44] L. Holm, L.M. Laakso, Dali server update, Nucleic Acids Res. 44 (W1) (2016) W351–W355.
[45] Y. Zhang, J. Skolnick, TM-align: a protein structure alignment algorithm based on the TM-score, Nucleic Acids Res. 33 (7) (2005) 2302–2309.
[46] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, et al., Highly accurate protein structure prediction with AlphaFold, Nature 596 (7873) (2021) 583–589.
[47] J. Yang, R. Yan, A. Roy, D. Xu, J. Poisson, Y. Zhang, The I-TASSER Suite: protein structure and function prediction, Nat. Methods 12 (1) (2015) 7–8.
[48] T. Li, Q.-Z. Li, S. Liu, G.-L. Fan, Y.-C. Zuo, Y. Peng, PreDNA: accurate prediction of DNA-binding sites in proteins by integrating sequence and geometric structure information, Bioinformatics 29 (6) (2013) 678–685.
[49] L. Wang, S.J. Brown, BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences, Nucleic Acids Res. 34 (suppl_2) (2006) W243–W248.
[50] J. Si, Z. Zhang, B. Lin, M. Schroeder, B. Huang, MetaDBSite: a meta approach to improve protein DNA-binding sites prediction, BMC Syst. Biol. 5 (2011) 1–7.
[51] Y. Ding, J. Tang, F. Guo, Identification of protein–ligand binding sites by sequence information and ensemble classifier, J. Chem. Inf. Model. 57 (12) (2017) 3149–3161.
[52] S. Ahmad, A. Sarai, PSSM-based prediction of DNA binding sites in proteins, BMC Bioinf. 6 (2005) 1–6.
[53] J. Hu, Y. Li, M. Zhang, X. Yang, H.-B. Shen, D.-J. Yu, Predicting protein-DNA binding residues by weightedly combining sequence-based features and boosting multiple SVMs, IEEE ACM Trans. Comput. Biol. Bioinf 14 (6) (2016) 1389–1398.
[54] X. Ma, J. Guo, H.-D. Liu, J.-M. Xie, X. Sun, Sequence-based prediction of DNA-binding residues in proteins with conservation and correlation information, IEEE ACM Trans. Comput. Biol. Bioinf 9 (6) (2012) 1766–1775.
[55] M.A. Hearst, S.T. Dumais, E. Osuna, J. Platt, B. Scholkopf, Support vector machines, IEEE Intell. Syst. Their Appl. 13 (4) (1998) 18–28.
[56] G. Biau, E. Scornet, A random forest guided tour, Test 25 (2016) 197–227.
[57] C. Yan, M. Terribilini, F. Wu, R.L. Jernigan, D. Dobbs, V. Honavar, Predicting DNA-binding sites of proteins from amino acid sequence, BMC Bioinf. 7 (2006) 1–10.
[58] S. Hwang, Z. Gou, I.B. Kuznetsov, DP-Bind: a web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins, Bioinformatics 23 (5) (2007) 634–636.
[59] L. Wang, M.Q. Yang, J.Y. Yang, Prediction of DNA-binding residues from protein sequence information using random forests, BMC Genom. 10 (2009) 1–9.
[60] L. Wang, C. Huang, M.Q. Yang, J.Y. Yang, BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features, BMC Syst. Biol. 4 (2010) 1–9.
[61] R. Liu, J. Hu, DNABind: a hybrid algorithm for structure-based prediction of DNA-binding residues by combining machine learning-and template-based approaches, Proteins: Struct., Funct., Bioinf. 81 (11) (2013) 1885–1899.
[62] J. Zhou, R. Xu, Y. He, Q. Lu, H. Wang, B. Kong, PDNAsite: identification of DNA-binding site from protein sequence by incorporating spatial and sequence context, Sci. Rep. 6 (1) (2016) 27653.
[63] A. Amirkhani, M. Kolahdoozi, C. Wang, L.A. Kurgan, Prediction of DNA-binding residues in local segments of protein sequences with fuzzy cognitive maps, IEEE ACM Trans. Comput. Biol. Bioinf 17 (4) (2018) 1372–1382.
[64] J. Zhang, Z. Ma, L. Kurgan, Comprehensive review and empirical analysis of hallmarks of DNA-, RNA-and protein-binding residues in protein chains, Briefings Bioinf. 20 (4) (2019) 1250–1268.
[65] H. Su, M. Liu, S. Sun, Z. Peng, J. Yang, Improving the prediction of protein–nucleic acids binding residues via multiple sequence profiles and the consensus of complementary methods, Bioinformatics 35 (6) (2019) 930–936.
[66] J. Zhang, S. Ghadermarzi, A. Katuwawala, L. Kurgan, DNAgenie: accurate prediction of DNA-type-specific binding residues in protein sequences, Briefings Bioinf. 22 (6) (2021) bbab336.
[67] L. Zhang, T. Liu, PDNAPred: interpretable prediction of protein-DNA binding sites based on pre-trained protein language models, Int. J. Biol. Macromol. 281 (2024) 136147.
[68] J. Zhou, Q. Lu, R. Xu, L. Gui, H. Wang, CNNsite: Prediction of DNA-Binding Residues in Proteins Using Convolutional Neural Network with Sequence Features, in: 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM): IEEE, 2016, pp. 78–85.
[69] B.P. Nguyen, Q.H. Nguyen, G.-N. Doan-Ngoc, T.-H. Nguyen-Vo, S. Rahardja, iProDNA-CapsNet: identifying protein-DNA binding residues using capsule neural networks, BMC Bioinf. 20 (2019) 1–12.
[70] Z. Li, F. Liu, W. Yang, S. Peng, J. Zhou, A survey of convolutional neural networks: analysis, applications, and prospects, IEEE Transact. Neural Networks Learn. Syst. 33 (12) (2021) 6999–7019.
[71] J. Zhou, Q. Lu, R. Xu, L. Gui, H. Wang, EL_LSTM: prediction of DNA-binding residue from protein sequence by combining long short-term memory and ensemble learning, IEEE ACM Trans. Comput. Biol. Bioinf 17 (1) (2018) 124–135.
[72] J. Zhang, Q. Chen, B. Liu, NCBRPred: predicting nucleic acid binding residues in proteins based on multilabel learning, Briefings Bioinf. 22 (5) (2021) bbaa397.
[73] J. Hu, Y.-S. Bai, L.-L. Zheng, N.-X. Jia, D.-J. Yu, G.-J. Zhang, Protein-DNA binding residue prediction via bagging strategy and sequence-based cube-format feature, IEEE ACM Trans. Comput. Biol. Bioinf 19 (6) (2021) 3635–3645.
[74] N. Wang, K. Yan, J. Zhang, B. Liu, iDRNA-ITF: identifying DNA-and RNA-binding residues in proteins based on induction and transfer framework, Briefings Bioinf. 23 (4) (2022) bbac236.
[75] F. Zhang, B. Zhao, W. Shi, M. Li, L. Kurgan, DeepDISOBind: accurate prediction of RNA-, DNA-and protein-binding intrinsically disordered residues with deep multi-task learning, Briefings Bioinf. 23 (1) (2022) bbab521.
[76] S. Guan, Q. Zou, H. Wu, Y. Ding, Protein-dna binding residues prediction using a deep learning model with hierarchical feature extraction, IEEE ACM Trans. Comput. Biol. Bioinf 20 (5) (2022) 2619–2628.
[77] S. Patiyal, A. Dhall, G.P. Raghava, A deep learning-based method for the prediction of DNA interacting residues in a protein, Briefings Bioinf. 23 (5) (2022) bbac322.
[78] H. Zhao, B. Zhu, T. Jiang, Z. Cui, H. Wu, A Transformer-Based Deep Learning Approach with Multi-Layer Feature Processing for Accurate Prediction of Protein-DNA Binding Residues, in: International Conference on Intelligent Computing, Springer, 2023, pp. 556–567.
[79] Y.D. Khan, T. Alkhalifah, F. Alturise, A.H. Butt, DeepDBS: identification of DNA-binding sites in protein sequences by using deep representations and random forest, Methods 231 (2024) 26–36.
[80] S.G. Hendrix, K.Y. Chang, Z. Ryu, Z.-R. Xie, DeepDISE: DNA binding site prediction using a deep learning method, Int. J. Mol. Sci. 22 (11) (2021) 5510.
[81] Y. Xia, C. Xia, X. Pan, H.B. Shen, BindWeb: a web server for ligand binding residue and pocket prediction from protein structures, Protein Sci. 31 (12) (2022) e4462.
[82] J. Zhang, S. Basu, L. Kurgan, HybridDBRpred: improved sequence-based prediction of DNA-binding amino acids using annotations from structured complexes and disordered proteins, Nucleic Acids Res. 52 (2) (2024) e10, e10.

[83] Y.-H. Zhu, C. Zhang, D.-J. Yu, Y. Zhang, Integrating unsupervised language model with triplet neural networks for protein gene ontology prediction, PLoS Comput. Biol. 18 (12) (2022) e1010793.
[84] M. Heinzinger, A. Elnaggar, Y. Wang, C. Dallago, D. Nechaev, F. Matthes, et al., Modeling aspects of the language of life through transfer-learning protein sequences, BMC Bioinf. 20 (2019) 1–17.
[85] D.D. Lee, P. Pham, Y. Largman, A. Ng, Advances in neural information processing systems 22, Tech Rep 1 (1) (2009) 1–11.
[86] A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, et al., Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences, Proc. Natl. Acad. Sci. 118 (15) (2021) e2016239118.
[87] A. Elnaggar, M. Heinzinger, C. Dallago, G. Rehawi, Y. Wang, L. Jones, et al., ProtTrans: toward understanding the language of life through self-supervised learning, IEEE Trans. Pattern Anal. Mach. Intell. 44 (10) (2021) 7112–7127.
[88] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, et al., Evolutionary-scale prediction of atomic-level protein structure with a language model, Science 379 (6637) (2023) 1123–1130.
[89] J. Su, C. Han, Y. Zhou, J. Shan, X. Zhou, F. Yuan, SaProt: protein language modeling with structure-aware vocabulary, bioRxiv 2023 (2023), 10.01.560349.
[90] A. Elnaggar, H. Essam, W. Salah-Eldin, W. Moustafa, M. Elkerdawy, C. Rochereau et al., Ankh: Optimized Protein Language Model Unlocks General-Purpose Modelling, arXiv preprint, 2023, arXiv:2301.06568.
[91] K.K. Yang, N. Fusi, A.X. Lu, Convolutions are competitive with transformers for protein sequence pretraining, Cell Systems 15 (3) (2024) 286–294, e2.
[92] R. M. Rao, J. Liu, R. Verkuil, J. Meier, J. Canny, P. Abbeel, et al., MSA Transformer, International Conference on Machine Learning: PMLR (2021) 8844–8856.
[93] M. Littmann, M. Heinzinger, C. Dallago, K. Weissenow, B. Rost, Protein embeddings and deep learning predict binding residues for various ligand classes, Sci. Rep. 11 (1) (2021) 23916.
[94] Q. Yuan, S. Chen, J. Rao, S. Zheng, H. Zhao, Y. Yang, AlphaFold2-aware protein–DNA binding site prediction using graph transformer, Briefings Bioinf. 23 (2) (2022) bbab564.
[95] Z. Jiang, Y.-Y. Shen, R. Liu, Structure-based prediction of nucleic acid binding residues by merging deep learning-and template-based approaches, PLoS Comput. Biol. 19 (9) (2023) e1011428.
[96] K. Shan, X. Zhang, C. Song, Prediction of Protein-DNA Binding Sites Based on Protein Language Model and Deep Learning, in: International Conference on Intelligent Computing: Springer, 2024, pp. 314–325.
[97] X. Sun, Z. Wu, J. Su, C. Li, GraphPBSP: protein binding site prediction based on Graph Attention Network and pre-trained model ProstT5, Int. J. Biol. Macromol. (2024) 136933.
[98] R. Roche, B. Moussad, M.H. Shuvo, S. Tarafder, D. Bhattacharya, EquiPNAS: improved protein–nucleic acid binding site prediction using protein-language-model-informed equivariant deep graph neural networks, Nucleic Acids Res. 52 (5) (2024) e27, e27.
[99] M. Zheng, G. Sun, X. Li, Y. Fan, EGPDI: identifying protein–DNA binding sites based on multi-view graph embedding fusion, Briefings Bioinf. 25 (4) (2024) bbae330.
[100] C. Hsu, R. Verkuil, J. Liu, Z. Lin, B. Hie, T. Sercu, et al., Learning Inverse Folding from Millions of Predicted Structures." International conference on machine learning, PMLR (2022) 8946–8970.
[101] F.S. Nahm, Receiver operating characteristic curve: overview and practical use for clinicians, Korean journal of anesthesiology 75 (1) (2022) 25–36.
[102] J. Si, R. Zhao, R. Wu, An overview of the prediction of protein DNA-binding sites, Int. J. Mol. Sci. 16 (3) (2015) 5194–5215.
[103] Y. Zhang, W. Bao, Y. Cao, H. Cong, B. Chen, Y. Chen, A survey on protein–DNA-binding sites in computational biology, Briefings in functional genomics 21 (5) (2022) 357–375.
[104] S. Xie, X. Xie, X. Zhao, F. Liu, Y. Wang, J. Ping, et al., HNSPPI: a hybrid computational model combing network and sequence information for predicting protein–protein interaction, Briefings Bioinf. 24 (5) (2023) bbad261.
[105] J.N. Mandrekar, Receiver operating characteristic curve in diagnostic test assessment, J. Thorac. Oncol. 5 (9) (2010) 1315–1316.
[106] V. Likic, The Needleman-Wunsch algorithm for sequence alignment. Lecture Given at the 7th Melbourne Bioinformatics Course, Bi021 Molecular Science and Biotechnology Institute, University of Melbourne, 2008, pp. 1–46.
[107] S.E. Halford, J.F. Marko, How do site-specific DNA-binding proteins find their targets? Nucleic Acids Res. 32 (10) (2004) 3040–3052.
[108] G.D. Stormo, DNA binding sites: representation and discovery, Bioinformatics 16 (1) (2000) 16–23.
[109] B. Ren, F. Robert, J.J. Wyrick, O. Aparicio, E.G. Jennings, I. Simon, et al., Genome-wide location and function of DNA binding proteins, Science 290 (5500) (2000) 2306–2309.
[110] M. Radaeva, A.-T. Ton, M. Hsing, F. Ban, A. Cherkasov, Drugging the 'undruggable'. Therapeutic targeting of protein–DNA interactions with the use of computer-aided drug discovery methods, Drug Discov. Today 26 (11) (2021) 2660–2679.
[111] D.R. Boer, A. Canals, M. Coll, DNA-binding drugs caught in action: the latest 3D pictures of drug-DNA complexes, Dalton Trans. (3) (2009) 399–414.
[112] J.B. Chaires, Drug—DNA interactions, Curr. Opin. Struct. Biol. 8 (3) (1998) 314–320.
[113] N.C. Smith, J.M. Matthews, Mechanisms of DNA-binding specificity and functional gene regulation by transcription factors, Curr. Opin. Struct. Biol. 38 (2016) 68–74.
[114] A.J. Walhout, Unraveling transcription regulatory networks by protein–DNA and protein–protein interaction mapping, Genome Res. 16 (12) (2006) 1445–1454.
[115] A.-L. Todeschini, A. Georges, R.A. Veitia, Transcription factors: specific DNA binding and specific gene regulation, Trends Genet. 30 (6) (2014) 211–219.
[116] M. Slattery, T. Zhou, L. Yang, A.C.D. Machado, R. Gordân, R. Rohs, Absence of a simple code: how transcription factors read the genome, Trends Biochem. Sci. 39 (9) (2014) 381–399.
[117] B. Fierz, M.G. Poirier, Biophysics of chromatin dynamics, Annu. Rev. Biophys. 48 (1) (2019) 321–345.
[118] S. Rao, T.-P. Chiu, J.F. Kribelbauer, R.S. Mann, H.J. Bussemaker, R. Rohs, Systematic prediction of DNA shape changes due to CpG methylation explains epigenetic effects on protein–DNA binding, Epigenetics Chromatin 11 (2018) 1–11.
[119] O. Cuvier, B. Fierz, Dynamic chromatin technologies: from individual molecules to epigenomic regulation in cells, Nat. Rev. Genet. 18 (8) (2017) 457–472.
[120] W.H. Hudson, E.A. Ortlund, The structure, function and evolution of proteins that bind DNA and RNA, Nat. Rev. Mol. Cell Biol. 15 (11) (2014) 749–760.
[121] A.C. Dantas Machado, T. Zhou, S. Rao, P. Goel, C. Rastogi, A. Lazarovici, et al., Evolving insights on how cytosine methylation affects protein–DNA binding, Briefings in functional genomics 14 (1) (2015) 61–73.
[122] S.H. Scheres, Processing of structurally heterogeneous cryo-EM data in RELION, Methods Enzymol. 579 (2016) 125–157.
[123] S. Wu, J.-t. Guo, Improved prediction of DNA and RNA binding proteins with deep learning models, Briefings Bioinf. 25 (4) (2024).
[124] J.M. Sagendorf, R. Mitra, J. Huang, X.S. Chen, R. Rohs, Structure-based prediction of protein-nucleic acid binding using graph neural networks, Biophysical Reviews 16 (3) (2024) 297–314.