MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **97** (2027) 873–899
ISSN: 0340–6253
doi: 10.46793/match.97-3.11726

DNCLA: A Deep Learning Model for TFBS Identification Based on Structural and Conformational Properties of Nucleotides and Dinucleotides

Jinjie Wei, Jie Feng*

School of Science, Minzu University of China, Beijing 100081, China

weijinjie9527@163.com, fengjie0536@163.com

(Received April 3, 2026)

Abstract

Identifying transcription factor binding sites (TFBSs) is fundamental to understanding complex gene regulatory mechanisms and the functions of non-coding regions. Although existing methods have achieved substantial strides, capturing both local structural features and long-range spatial dependencies within DNA sequences remains a major challenge for improving prediction accuracy. In this study, we propose DNCLA, a deep learning model that synergizes multisize convolutional fusion, Bidirectional Long Short-Term Memory (Bi-LSTM) networks, and a multi-head self-attention mechanism. At the feature extraction level, DNCLA breaks through the limitations of traditional single-sequence encoding by fusing Nucleotide Chemical Properties (NCP) with Dinucleotide Physicochemical Properties (DPCP). NCP provides a refined characterization of chemical differences between bases based on ring structures, hydrogen bond sites, and functional group properties, while DPCP introduces parameters such as local structural stability and geometric flexibility of the DNA. Subsequently, the model extracts spatial evolution from these high-dimensional features through a mul-

*Corresponding author: fengjie0536@163.com

This work is licensed under a Creative Commons "Attribution 4.0 International" license.

tisize convolutional module; captures long-range spatial dependencies using Bi-LSTM layers; and employs a multi-head self-attention mechanism to achieve adaptive weight distribution of global features, thereby enhancing the perception of key regulatory motifs. Results from training and testing the proposed model on 165 ChIP-seq datasets demonstrate that DNCLA possesses robust generalization capabilities and high predictive performance in TFBSs identification. This suggests that the incorporation of physicochemical features better elucidates the essence of interactions between transcription factors and DNA.

1 Introduction

Transcription factors (TFs) interact with specific DNA sequences known as transcription factor binding sites (TFBSs), which play a crucial role in regulating gene expression [1,2]. By recognizing and binding to these sites, TFs control mRNA transcription levels, thereby influencing cell differentiation, development, and the response to environmental signals [3,4]. Therefore, the accurate identification of TFBSs is essential for understanding physiological mechanisms, constructing metabolic regulatory systems, and revealing serious diseases caused by mutations in non-coding regions [5,6].

Numerous methods for TFBSs identification have been proposed over the past few decades. Early research primarily relied on low-throughput biochemical experiments. In terms of computation, scientists developed the Position Weight Matrix (PWM) [7], a foundational model for TFBSs identification that represents binding preferences by calculating the probability of each base occurring at a specific position. Later, with the advancement of next-generation sequencing technology, Chromatin Immunoprecipitation sequencing (ChIP-seq) [8,9] became the primary method for identifying TFBSs. Although these methods have achieved good results in TFBSs identification, they are characterized by stringent experimental requirements and high development costs.

With the rapid development of technology, prediction models based on machine learning and deep learning have been widely applied in TFBS identification. Early computational models primarily relied on traditional machine learning methods such as Support Vector Machines (SVMs) [10,

11] and Random Forest models (RFs) [12, 13]. However, traditional machine learning methods are highly dependent on designed feature engineering, which limits the generalization capability of the models [14]. In recent years, with the rapid rise of deep learning, TFBSs identification has witnessed breakthrough progress. Models represented by DeepBind [15] and DeepSEA [16] have demonstrated the powerful ability of Convolutional Neural Networks (CNNs) to extract local sequence motifs. Subsequently, researchers introduced Recurrent Neural Networks (RNN/LSTM) to model long-range dependencies in sequences, such as the typical DanQ [17] and MLSNet [18], improving the modeling accuracy of global sequence logic. Regarding feature representation, researchers have begun exploring the fusion of multimodal features. For instance, DLBSS [19] achieved more comprehensive binding preference modeling by extracting DNA sequence and DNA shape features in parallel. CRPTS [20] further employed a shared convolutional recurrent neural network architecture to deeply integrate DNA sequence information with physicochemical shape features, significantly enhancing predictive robustness across datasets. Later, with the migration of Natural Language Processing (NLP) technologies, attention mechanisms and Transformer architectures were introduced to the field. For example, DNABERT [21] and its improved version, BERT-TFBS [22], captured deep genomic semantic information through large-scale pre-training, augmenting prediction accuracy across cell lines. D-SSCA [23] and DeepSTF [24] introduced attention mechanisms, leading to further performance gains. Furthermore, existing studies have shown that the physicochemical properties of DNA sequences play a key role in the specific recognition of transcription factors [25], providing new insights for TFBSs identification.

Despite significant progress in current research, how to further refine feature extraction within complex genomic backgrounds and enhance model interpretability remains a core challenge. Building upon previous studies, we propose DNCLA, a novel deep learning model that fuses multi-dimensional physicochemical features. The specific framework of DNCLA is illustrated in Figure 1. The model employs a dual-branch parallel structure to simultaneously extract Dinucleotide Physicochemical Properties

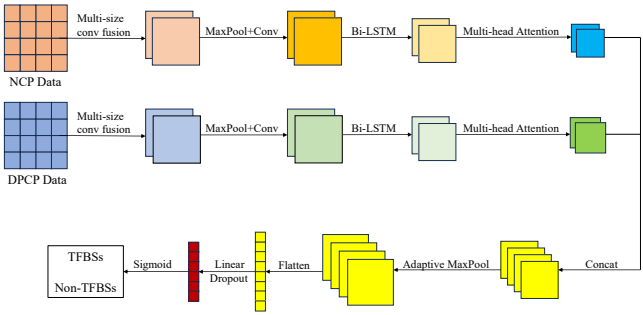

Figure 1. The workflow of the DNCLA model. The DNCLA framework fuses NCP and DPCP features, combining multi-size convolutional fusion, Bi-LSTM, and a multi-head self-attention mechanism to extract sequence and structural information.

(DPCP) and Nucleotide Chemical Properties (NCP) of DNA sequences. Regarding the core architecture, the model first utilizes multisize convolutional fusion to extract local motif features from the DNA sequences. Subsequently, a Bidirectional Long Short-Term Memory (Bi-LSTM) network is used to capture long-range spatial dependencies. Finally, a multi-head self-attention mechanism is introduced to re-weight global features, focusing on critical functional signaling regions. The results demonstrate that the model has superior prediction accuracy, achieving an ACC of 0.840, an ROC-AUC of 0.913, and a PR-AUC of 0.916.

2 Materials and methods

2.1 Benchmark dataset

We utilized 165 ChIP-seq datasets previously employed by BERT-TFBS to evaluate the model's performance. This dataset was originally filtered by Zeng et al. [26] from 690 ChIP-seq datasets in the ENCODE project, encompassing 29 unique transcription factors (TFs) across 32 different cell lines. Each dataset comprises training and testing subsets, consisting of a multitude of positive and negative instances. The DNA sequences are

represented as fixed-length strings of 101 base pairs, with binary labels (0 and 1) indicating the presence or absence of a transcription factor binding site (TFBS), respectively.

2.2 Feature extraction methods

2.2.1 Nucleotide chemical properties (NCP)

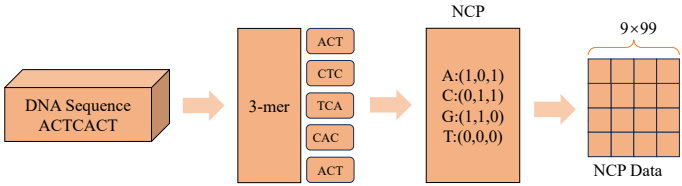

Figure 2. Acquisition process of the NCP feature matrix.

NCP encoding transforms discrete nucleotide sequences into high-dimensional numerical vectors by quantifying the physicochemical properties of bases, as illustrated in Figure 2. This method is primarily based on three ring structure characteristics and chemical properties of the bases: Ring Structure: Purines (A, G) possess a double-ring structure, while pyrimidines (C, T) possess a single-ring structure; Hydrogen Bonds: Strong hydrogen-bond bases (C, G) contain three hydrogen bonds, whereas weak hydrogen-bond bases (A, T) contain two; Chemical Groups: Based on the classification of amino versus keto groups, A and C belong to the amino group, while G and T belong to the keto group. Based on these properties, each base is mapped to a three-dimensional physicochemical property vector (x,y,z) via one-hot encoding technology: A: (1, 0, 1); C: (0, 1, 1); G: (1, 1, 0); and T: (0, 0, 0). By segmenting a nucleotide sequence of length n into $n-2$ overlapping 3-mer fragments, each trinucleotide unit is represented by a 9-dimensional column vector formed by concatenating the 3-dimensional physicochemical property vectors (x, y, z) of its constituent bases. Consequently, for a sequence with a length of 101 bp ($n = 101$), a 9×99 dimensional feature matrix is finally formed, represented as follows:

$$S_1 = [M_1, M_2, M_i, \ldots, M_{98}, M_{99}]$$

where M_i denotes the i-th trinucleotide fragment.

2.2.2 Dinucleotide physicochemical properties (DPCP)

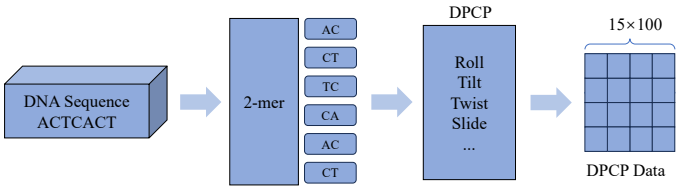

Figure 3. Acquisition process of the DPCP feature matrix.

DPCP encoding transforms biological sequences into numerical vectors reflecting the local structure and thermodynamic stability of DNA by quantifying the physicochemical parameters of adjacent bases (dinucleotides) within the DNA sequence. The specific process is illustrated in Figure 3. Unlike NCP, which is based on individual base properties, DPCP focuses on describing the geometric deformation and energy characteristics of double-stranded DNA. This method is based on 15 core physicochemical properties of dinucleotides, which are primarily categorized into three types: Geometric Structure: Roll, Tilt, Twist, Slide, Shift, and Rise; Energy and Stability: Free Energy, Enthalpy, and Entropy; Stiffness/Flexibility: F-roll, F-tilt, etc.

Due to the significant differences in scales among various physicochemical parameters [27], the raw data must undergo Z-score normalization prior to mapping:

$$z = \frac{x - \mu}{\sigma}$$

where x denotes the value of one specific physicochemical property among the 15 categories. For each property, the mean (μ) and standard deviation (σ) are calculated across the comprehensive set of all N=16 possible

dinucleotides (AA, AC, ..., TT) as follows:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

where x_i represents the value of that specific property for the i-th dinucleotide. This ensures that each feature carries equal weight during model computation.

Based on the 15 standardized properties mentioned above, each dinucleotide step is mapped to a 15-dimensional numerical vector. For a nucleotide sequence of length n, the adjacent dinucleotide steps are concatenated to form a $15\times(n-1)$ dimensional feature matrix. When $n = 101$, the dinucleotide steps are concatenated to form a 15×100 dimensional feature matrix, represented as follows:

$$S_2 = [N_1, N_2, N_i, \ldots, N_{99}, N_{100}]$$

where N_i represents the 15 physicochemical property values corresponding to the i-th dinucleotide position in the sequence.

2.3 NCP module

Utilizing 3-mer encoding enhances the representative breadth of the input sequence. However, since NCP encoding is based on physicochemical coordinates, it generates a sparse matrix composed solely of 0s and 1s, which may adversely affect model performance. To address this issue, we adopted a multisize convolutional strategy (using 3×3, 5×5, and 7×7 kernels) to strengthen multi-scale feature extraction. Subsequently, a Bi-LSTM layer is employed to capture long-range dependencies within the DNA sequence and further extract latent features. Finally, a multi-head self-attention mechanism is introduced to re-weight the global features, focusing on critical functional signaling regions. The NCP module is rep-

resented as follows:

$$M_1 = MaxPooling\left(Concat(Conv1(S_1), Conv2(S_1), Conv3(S_1))\right)$$

$$C_1 = BN\left(ReLU(Conv(M_1))\right)$$

$$P_1 = MultiHead\left(Bi\text{-}LSTM(C_1)\right)$$

where $Conv1(*)$, $Conv2(*)$, $Conv3(*)$ denotes the convolutional operation, utilizing kernels of sizes 3×3, 5×5, and 7×7, respectively; $Concat(*)$ represents feature concatenation, which merges the heterogeneous features extracted by multiple convolutional kernels along the channel dimension to enhance the representative breadth of the features; $MaxPooling(*)$ refers to maximum pooling; $BN(*)$ stands for Batch Normalization; $ReLU(*)$ is the activation function; $Bi\text{-}LSTM(*)$ represents the Bidirectional Long Short-Term Memory network, which captures bidirectional long-range temporal dependencies; $MultiHead(*)$ denotes the multi-head self-attention mechanism, which re-weights global features to focus on key regions.

2.4 DPCP module

The DPCP module adopts an architecture similar to that of the NCP module to mine the local geometric structure and thermodynamic stability features of DNA. It captures the spatial correlations of dinucleotide physicochemical parameters through multisize convolutional kernels and utilizes residual connections to ensure the integrity of feature transmission. Subsequently, Bi-LSTM is employed to extract the bidirectional long-range dependencies of physicochemical signals along the sequence, and a multi-head self-attention mechanism is introduced to re-weight global features, focusing on critical functional signaling regions. The DPCP module is represented as follows:

$$M_2 = MaxPooling\left(Concat(Conv1(S_2), Conv2(S_2), Conv3(S_2))\right)$$

$$C_2 = BN\left(ReLU(Conv(M_2))\right)$$

$$P_2 = MultiHead\left(Bi\text{-}LSTM(C_2)\right)$$

2.5 Output module

We fuse the previously obtained sequence features P_1 and P_2, represented as follows:

$$O = Sigmoid\left(Linear(Flatten(AdaptiveMaxPool(Concat(P_1, P_2))))\right)$$

where $Concat(*)$ denotes feature concatenation, which merges the global feature vectors extracted from the NCP and DPCP streams along the feature dimension to form a comprehensive feature representation; $AdaptiveMaxPool1d(*)$ refers to adaptive maximum pooling; $Flatten(*)$ represents the flattening operation; $Linear(*)$ denotes a linear transformation, which performs linear combination and dimensionality reduction on the integrated features through a weight matrix, mapping the high-dimensional feature space into the classification space; and $Sigmoid(*)$ is the activation function, which compresses the output of the fully connected layer into the (0,1) interval.

2.6 Model implementation

Our model is implemented based on the PyTorch framework using a dual-branch parallel architecture. Multi-scale one-dimensional convolutional layers (1D Conv) with kernel sizes of 3, 5, and 7 are employed to extract local biological motifs from NCP and DPCP features, respectively. Subsequently, a Bidirectional Long Short-Term Memory (Bi-LSTM) network with 128 hidden units is utilized to capture the global contextual information of the sequences. To further enhance features at key positions, the model introduces an 8-head self-attention mechanism for automated weight allocation, and the outputs of both branches are fused via an Adaptive Max Pooling layer. Regarding the training configuration, we utilize the AdamW optimizer with an initial learning rate of 0.0005 and a weight decay coefficient of 0.001. This optimizer improves the application of weight decay by applying it directly during the parameter update steps, which

typically leads to better generalization. Additionally, the learning rate is adjusted using cosine annealing technology, dynamically tuning the learning rate throughout the training cycle to help the model locate deeper and more stable local minima. The Binary Cross-Entropy Loss (BCELoss) is selected as the loss function. To enhance generalization and prevent overfitting, ReLU activation functions and a Dropout probability of 0.2 are applied in the intermediate layers. Finally, the output layer generates a prediction score through a Sigmoid activation function; samples with a prediction score greater than 0.5 are classified as transcription factor binding sites.

To ensure a rigorous performance evaluation, we split the original training data into training and validation sets at an 8:2 ratio. This partitioning ensures that the test set is used for final evaluation, preventing data leakage during hyperparameter tuning. Throughout the training process, we determine whether to optimize hyperparameters or adjust training strategies based on the loss performance on the validation set. Finally, the generalization performance of the model is verified using a dedicated benchmark test set that remains isolated from the parameter optimization process.

2.7 Evaluation metrics

TFBS identification based on DNA sequences is essentially a binary classification problem. We employ Accuracy (ACC) [28], ROC-AUC [29], and PR-AUC [30] to evaluate the model's performance.

ACC represents the proportion of correctly identified samples out of the total test samples. Its mathematical expression is as follows:

$$ACC = \frac{TN + TP}{TP + TN + FN + FP}$$

where TP, FN, TN, and FP represent the number of true positive, false negative, true negative, and false positive samples, respectively.

ROC-AUC refers to the Area Under the Receiver Operating Characteristic (ROC) curve. The ROC curve visually demonstrates the overall performance of a classifier across different thresholds by simultaneously examining the True Positive Rate (TPR) and the False Positive Rate (FPR).

A higher ROC-AUC score indicates superior classification performance. The mathematical definitions of TPR and FPR are as follows:

$$TPR = \frac{TP}{FN + TP}$$

$$FPR = \frac{FP}{FP + TN}$$

PR-AUC refers to the area under the Precision-Recall (PR) curve. The PR curve is designed to describe the performance of a classifier in the presence of imbalanced class distributions, measured by balancing the relationship between Precision and Recall. A higher PR-AUC score represents more outstanding classification performance. The definitions of Precision and Recall are as follows:

$$presision = \frac{TP}{FP + TP}$$

$$recall = \frac{TP}{FN + TP}$$

2.8 Baseline methods

In this study, the proposed model was compared with other deep learning methods. To ensure the fairness of the evaluation, all models were trained and performance-tested under unified experimental conditions. Particularly noteworthy is the DeepSTF model [24], which mines DNA structural data through a hybrid Transformer-Bi-LSTM architecture, significantly enhancing TFBS prediction performance. Among other comparative methods, DeepBind [15] and DanQ [17] focus on sequence modeling, while DLBSS, CRPTS, and D-SSCA [23] integrate DNA structural information using Siamese convolutions, convolutional recurrent networks, and MLPs, respectively. Given that DeepSTF exhibits the strongest performance, we utilized it as the primary performance benchmark.

3 Results and discussion

3.1 Ablation study

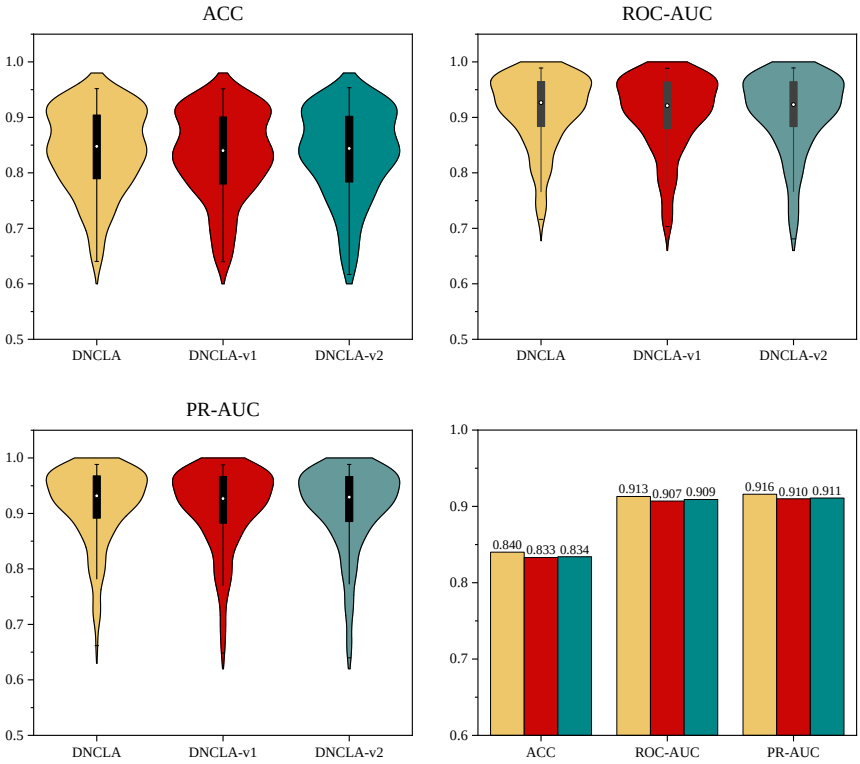

Figure 4. Performance comparison of ACC, ROC-AUC, and PR-AUC between DNCLA and its variant models across 165 ChIP-seq datasets. DNCLA-v1: without NCP feature matrix; DNCLA-v2: without DPCP feature matrix.

To evaluate the contributions of Nucleotide Chemical Properties (NCP) and Dinucleotide Physicochemical Properties (DPCP) to the TFBSs prediction task, this study designed two models: DNCLA-v1 and DNCLA-v2.

DNCLA-v1: This model utilizes only the DPCP feature matrix as input, focusing on evaluating the impact of spatial geometric configurations and local physical energy parameters at the dinucleotide level on the model's discriminative performance.

DNCLA-v2: This model retains only the NCP feature matrix at the input layer, aiming to explore the representative capability of functional group properties and hydrogen bond distribution at the single-nucleotide dimension for binding site recognition.

The experimental results are shown in Figure 4 and Table 1. The complete DNCLA model outperformed the control models that used only a single feature matrix across the ACC, ROC-AUC, and PR-AUC metrics. DNCLA achieved an ACC of 0.840, representing an increase of 0.7% and 0.6% compared to DNCLA-v1 and DNCLA-v2, respectively. In terms of ROC-AUC, DNCLA reached 0.913, an improvement of approximately 0.6% and 0.4% over DNCLA-v1 and DNCLA-v2, respectively. For the PR-AUC metric, DNCLA reached 0.916, which is approximately 0.6% and 0.5% higher than DNCLA-v1 and DNCLA-v2, respectively. Although the performances of the ablation models DNCLA-v1 and DNCLA-v2 were similar, they exhibited greater fluctuations on complex sequences. In contrast, the DNCLA model can capture the chemical heterogeneity of the base microenvironment, demonstrating stronger robustness when identifying DNA regions with specific topological structures. This proves that NCP and DPCP features are highly complementary in terms of biological information: the former focuses on chemical recognition at the molecular level, while the latter provides structural constraints at the physical level.

Table 1. Average values of ACC, ROC-AUC, and PR-AUC for DNCLA and its variant models across 165 ChIP-seq datasets.

Method	ACC	ROC-AUC	PR-AUC
DNCLA-v1	0.833	0.907	0.910
DNCLA-v2	0.834	0.909	0.911
DNCLA	0.840	0.913	0.916

To further investigate the contributions of the core internal components—namely the multisize convolutional fusion module, Bi-LSTM module, and multi-head self-attention mechanism—to the transcription factor binding site (TFBS) identification performance of the DNCLA model, this study constructed three variant models with missing components for com-

parative analysis.

DNCLA-v3: The multisize convolutional fusion module was removed to evaluate the fundamental impact of local feature extraction on the model's identification performance.

DNCLA-v4: The Bi-LSTM module was removed to assess the role of long-range dependency modeling in sequence analysis.

DNCLA-v5: The multi-head self-attention mechanism was removed to quantify the performance gain contributed by the automated allocation of feature weights.

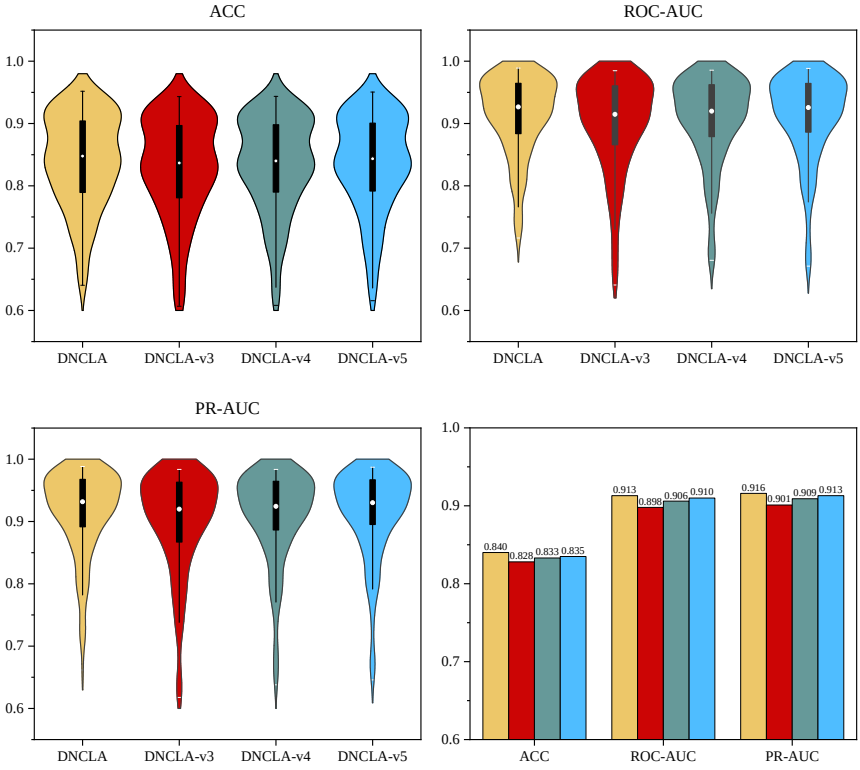

Figure 5. Performance comparison of ACC, ROC-AUC, and PR-AUC between DNCLA and its variant models across 165 ChIP-seq datasets. DNCLA-v3: without multisize convolutional fusion module; DNCLA-v4: without Bi-LSTM module; DNCLA-v5: without multi-head self-attention mechanism.

The experimental results are presented in Figure 5 and Table 2, demonstrating that the full DNCLA model achieved the superior performance across all evaluation metrics. Among the three ablation models, DNCLA-v3 exhibited the most pronounced performance degradation, with its ACC, ROC-AUC, and PR-AUC dropping to 0.828, 0.898, and 0.901, respectively. This indicates that the multisize convolutional fusion module plays an indispensable role in capturing local DNA sequence motifs; the absence of this module hinders the model from recognizing local patterns in base arrangements, thereby significantly impairing prediction accuracy.

Concurrently, a decline across all metrics was observed for DNCLA-v4, which yielded an ROC-AUC of 0.906 and a PR-AUC of 0.909. This underscores the advantage of the bidirectional Long Short-Term Memory network in capturing long-range dependencies and spatial evolutionary patterns within sequences. The Bi-LSTM is capable of integrating local features extracted by the convolutional layers to form a global representation with sequential logic, which is of paramount importance for understanding the complex functions of non-coding regions.

Although DNCLA-v5 outperformed the other two variants, a distinct performance gap remained compared to the full model, with its ACC and PR-AUC resting at 0.835 and 0.913, respectively. This implies that the multi-head self-attention mechanism enables adaptive allocation of feature weights, allowing the model to focus on critical signal regions that contribute most to the binding sites, thereby exerting a fine-grained regulatory effect on the predictions.

Table 2. Average values of ACC, ROC-AUC, and PR-AUC for DNCLA and its variant models across 165 ChIP-seq datasets.

Method	ACC	ROC-AUC	PR-AUC
DNCLA-v3	0.828	0.898	0.901
DNCLA-v4	0.833	0.906	0.909
DNCLA-v5	0.835	0.910	0.913
DNCLA	0.840	0.913	0.916

3.2 Model interpretability analysis based on SHAP

To further elucidate the internal decision-making mechanism of the DNCLA model and validate the biological relevance of the integrated feature representations, this study introduces the SHAP (SHapley Additive exPlanations) [33] framework for interpretability analysis. Distinct from conventional feature importance metrics, SHAP provides a rigorous attribution strategy grounded in game theory; by calculating the Shapley value for each input feature, it precisely quantifies its marginal contribution to the model's final predictive output.

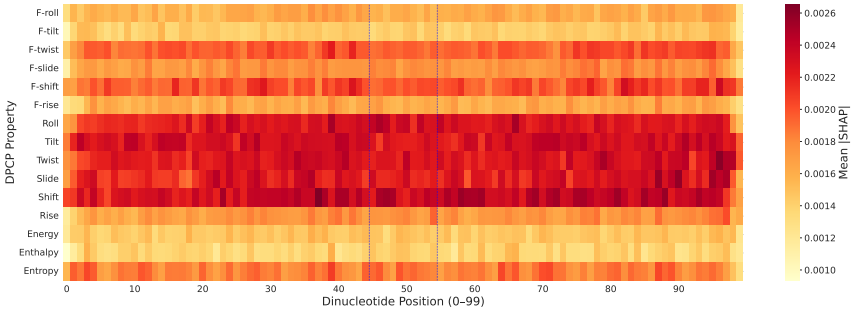

Figure 6. Global interpretability heatmap of DPCP features based on Mean |SHAP|values.

As illustrated in Figure 6, the global SHAP heatmap reveals the contribution landscape of 15 dinucleotide physicochemical properties (DPCP). Among the 15 parameters, geometric descriptors—specifically Roll, Tilt, and Twist—exhibit the highest mean absolute SHAP values. High-intensity signals are concentrated in the central region of the sequences (40–60 bp), indicating that the model relies on local DNA bending and helical twisting for the recognition of binding sites. This suggests that transcription factors (TFs) preferentially recognize specific three-dimensional structural motifs rather than relying solely on the primary sequence.

Thermodynamic parameters, including Energy and Enthalpy, exhibit prominent local feature contributions, with the model effectively capturing the "energy landscape" of DNA sequences. This indicates that TF binding is energetically favored in regions possessing specific stability characteris-

tics. The pronounced SHAP values observed within these thermodynamic channels further demonstrate that the DNCLA model can transform abstract thermodynamic stability into quantifiable predictive features.

Notably, the F-prefixed parameters (such as F-roll and F-twist), which represent the influence of the flanking environment, yield consistent contributions to the model's decision-making process. This not only validates the effectiveness of DPCP encoding in capturing long-range structural dependencies but also confirms that binding affinity is governed not only by the core motif but also by the topological constraints imposed by the secondary structure of the surrounding DNA backbone.

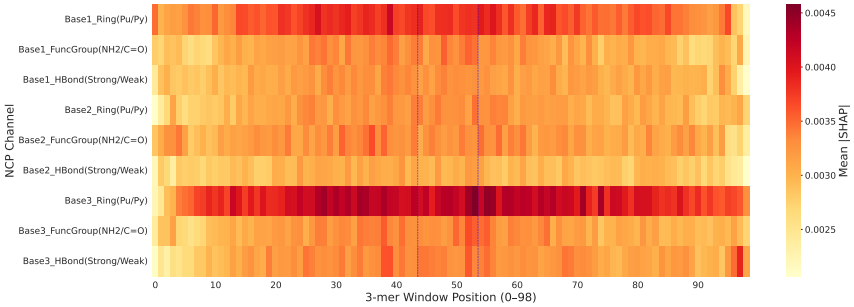

Figure 7. Global interpretability heatmap of NCP features based on Mean |SHAP|values.

As observed from the global attribution heatmap in Figure 7, the most significant feature contributions are clustered within the Base3_Ring(Pu/Py) channel. Given the fundamental geometric distinction between the double-ring structure of purines and the single-ring structure of pyrimidines, this result indicates that the DNCLA model relies heavily on steric occupancy characteristics during the recognition process. The model's preference for this "ring structure" arrangement essentially reflects its acute sensitivity in capturing the local geometric conformation of the DNA backbone, thereby modeling the biological mechanism of "indirect readout" at the computational level. Furthermore, alongside the dominant ring-structure features, Base2_HBond and Base2_FuncGroup, located at the window center, exhibit synergistic attribution patterns within the central sequence region. This demonstrates that during prediction,

the model not only references macro-scale steric hindrance constraints but also precisely identifies the distribution of hydrogen bond acceptors and donors along the base edges, thereby achieving a multi-scale cooperative decision-making process that integrates spatial topology with the chemical microenvironment.

3.3 Different cell lines and transcription factors

3.3.1 Cross-cell lines

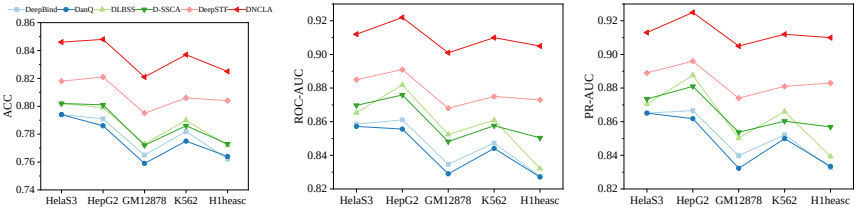

Figure 8. Performance comparison between the DNCLA model and other models (DeepBind, DanQ, DLBSS, D-SSCA, and DeepSTF) across five cell lines (HelaS3, HepG2, GM12878, K562, and H1-hESC).

The 165 ChIP-seq datasets we used encompass 32 cell lines. We selected the five cell lines containing the largest number of ChIP-seq datasets, namely HelaS3, HepG2, GM12878, K562, and H1-hESC, which contain 13, 26, 17, 32, and 16 ChIP-seq datasets, respectively, ensuring representativeness.

The experimental results are shown in Figure 8 and Table 3. Across all tested cell lines, the performance of DNCLA significantly exceeded that of the baseline model, DeepSTF. Particularly in the HepG2 cell line, which has a larger data volume, the PR-AUC of DNCLA reached 0.925, an improvement of 2.9% over DeepSTF. Although the epigenetic environments and transcription factor abundances vary greatly among different cell lines, DNCLA is still able to achieve high-precision predictions by capturing the intrinsic physicochemical properties of the base microenvironment. For instance, in the most challenging GM12878 cell line, the model maintained an ROC-AUC of 0.901. This further demonstrates the strong generaliza-

tion capability of DNCLA when processing complex genomic backgrounds.

Table 3. Comparison of average ACC, ROC-AUC, and PR-AUC between the DNCLA model and the DeepSTF model across five cell lines (HelaS3, HepG2, GM12878, K562, and H1-hESC).

Cell lines	ACC		ROC-AUC		PR-AUC	
	DeepSTF	DNCLA	DeepSTF	DNCLA	DeepSTF	DNCLA
HelaS3	0.818	0.846	0.885	0.912	0.889	0.913
HepG2	0.821	0.848	0.891	0.922	0.896	0.925
GM12878	0.795	0.821	0.868	0.901	0.874	0.905
K562	0.806	0.837	0.875	0.910	0.881	0.912
H1-hESC	0.804	0.825	0.873	0.905	0.883	0.910

3.3.2 Cross-transcription factors

Among the 165 ChIP-seq datasets used in this study, a series of transcription factors (TFs) highly representative of genomic research are included. These TFs represent distinct regulatory functions (such as insulators, promoter binding, and epigenetic modification) and diverse binding characteristics. We selected several core categories of TFs from the 165 ChIP-seq datasets for comparison.

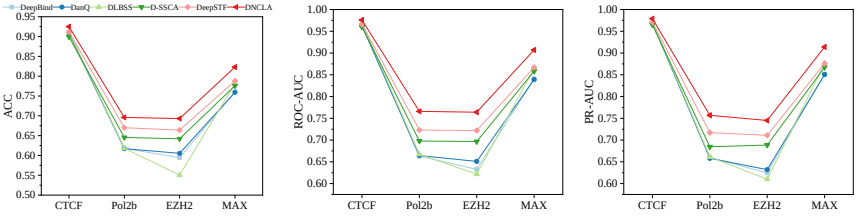

Figure 9. Performance comparison between the DNCLA model and other models (DeepBind, DanQ, DLBSS, D-SSCA, and DeepSTF) across four cross-transcription factors (CTCF, Pol2b, EZH2, and MAX).

CTCF belongs to the BORIS + CTCF family, and the transcription control factor it encodes possesses 11 highly conserved and consistent zinc

finger (ZF) motifs. This factor binds to diverse DNA sequences and proteins through different zinc finger configurations. By associating with histone acetyltransferases or deacetylases, it plays a dual role in gene regulation by either initiating or repressing transcription. As a genomic "insulator", CTCF can block the communication between enhancers and promoters, thereby influencing the expression of imprinted genes. In clinical research, alterations in CTCF are closely associated with various cancers, including invasive breast cancer, prostate cancer, and Wilms' tumor [31].

Pol2b is a core component of the eukaryotic transcription machinery, directly responsible for the transcription process of all protein-coding genes and various non-coding RNAs [32]. It is primarily enriched in the promoter regions and transcription start sites (TSS) of actively transcribed genes, and its binding level is generally regarded as a direct indicator of gene expression activity. Since the binding of Pol2b on the genome often manifests as broad peaks spanning promoters and gene bodies rather than sharp motif features, sequence-based binding site prediction for Pol2b is more challenging compared to traditional transcription factors.

EZH2, as a typical epigenetic regulator, plays a key role in maintaining cell stemness, regulating developmental processes, and repressing tumor suppressor genes [33]. In ChIP-seq datasets, the binding of EZH2 often depends on local chromatin topology rather than specific DNA sequences; therefore, the model's prediction accuracy for EZH2 is frequently influenced by physicochemical properties (such as DNA shape) features.

MAX is a member of the basic helix-loop-helix leucine zipper (bHLH-Zip) family. As the central hub of the MYC regulatory network, it is an essential partner for MYC proteins to exert transcriptional activity. MAX extensively regulates the expression of genes involved in the cell cycle, growth metabolism, and protein synthesis. Simultaneously, it can associate with members of the MAD family to form inhibitory complexes, playing a decisive switching role in maintaining the dynamic equilibrium between cell proliferation and differentiation [34].

The experimental results are shown in Figure 9 and Table 4. DNCLA consistently outperformed the baseline model DeepSTF across various transcription factor prediction tasks: for the insulator protein CTCF,

Table 4. Comparison of average ACC, ROC-AUC, and PR-AUC between the DNCLA model and the DeepSTF model across four cross-transcription factors (CTCF, Pol2b, EZH2, and MAX).

TFs	ACC		ROC-AUC		PR-AUC	
	DeepSTF	DNCLA	DeepSTF	DNCLA	DeepSTF	DNCLA
CTCF	0.912	0.925	0.966	0.976	0.971	0.979
Pol2b	0.670	0.696	0.723	0.766	0.717	0.757
EZH2	0.664	0.693	0.722	0.764	0.711	0.745
MAX	0.788	0.823	0.867	0.907	0.876	0.914

which exhibits strong sequence conservation, DNCLA achieved an ROC-AUC of 0.976 and a PR-AUC of 0.979. In the case of the epigenetic modification factor EZH2, which presents higher prediction difficulty due to significant influence from the chromatin environment, DNCLA still improved the ROC-AUC from 0.722 (DeepSTF) to 0.764, and the ACC to 0.693. Furthermore, in tests targeting basic transcriptional regulators MAX and Pol2b, DNCLA also demonstrated identification capabilities, with PR-AUC values reaching 0.914 and 0.757, respectively. This fully demonstrates that by fusing Nucleotide Chemical Properties (NCP) and Physicochemical Properties (DPCP), DNCLA can capture the chemical heterogeneity and physical structural constraints of the base microenvironment. Consequently, it maintains strong prediction robustness when facing transcription factors with diverse regulatory functions and binding characteristics.

3.4 Comparison with prior models

To comprehensively evaluate the advancement of the DNCLA model in the transcription factor binding site (TFBS) prediction task, we conducted comparative experiments between DNCLA and five mainstream deep learning models (DeepBind, DanQ, DLBSS, D-SSCA, and DeepSTF) across 165 ChIP-seq datasets. The experimental results, as shown in Figure 10 and Table 5, indicate that DNCLA achieved significant breakthroughs in all performance metrics. The average ACC of DNCLA reached 0.840,

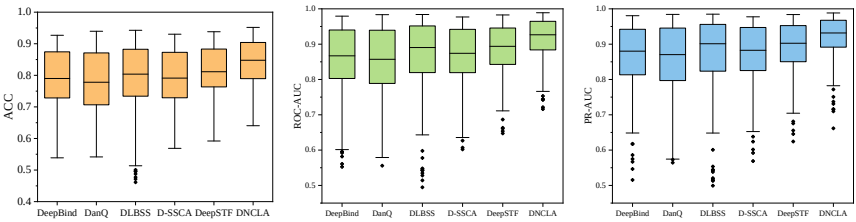

Figure 10. Performance comparison between the DNCLA model and other models (DeepBind, DanQ, DLBSS, D-SSCA, and DeepSTF) across 165 ChIP-seq datasets.

an increase of 2.8% compared to the baseline model DeepSTF (0.812). In terms of ROC-AUC, DNCLA reached 0.913, an improvement of 3.1% over DeepSTF (0.882). For the PR-AUC metric, DNCLA reached 0.916, which is 2.8% higher than DeepSTF (0.888).

In summary, by integrating Nucleotide Chemical Properties (NCP) and Physicochemical Properties (DPCP), DNCLA not only outperforms traditional methods such as DeepBind in average performance but also demonstrates superior generalization capability and robustness when addressing complex transcriptional regulatory logic. This provides a powerful computational tool for the precise mining of large-scale genomic regulatory elements.

Table 5. Comparison of average performance across 165 ChIP-seq datasets.

Method	**ACC**	**ROC-AUC**	**PR-AUC**
DeepBind	0.785	0.853	0.858
DanQ	0.782	0.849	0.855
DLBSS	0.793	0.865	0.871
D-SSCA	0.784	0.854	0.857
DeepSTF	0.812	0.882	0.888
DNCLA	**0.840**	**0.913**	**0.916**

4 Conclusion

To achieve accurate identification of transcription factor binding sites (TFBSs), this study proposes DNCLA, a deep learning model that integrates multi-dimensional physicochemical features. The model combines Nucleotide Chemical Properties (NCP) with Dinucleotide Physicochemical Properties (DPCP). Through multi-scale convolutional fusion, Bidirectional Long Short-Term Memory (Bi-LSTM) networks, and multi-head self-attention mechanisms, it effectively captures local DNA sequence motifs, long-range dependencies, and global key signals. Experimental results across 165 ChIP-seq datasets demonstrate that DNCLA outperforms mainstream models such as DeepBind, DanQ, and DeepSTF in key evaluation metrics, including average ACC (0.840), ROC-AUC (0.913), and PR-AUC (0.916). This strongly validates the critical role of physicochemical features in revealing the essence of transcription factor-DNA interactions.

Despite the significant improvements in performance and generalization, the current study still possesses certain limitations. First, the model currently relies primarily on static physicochemical features derived from DNA sequences and has not yet fully incorporated cell-line-specific dynamic epigenetic information, such as chromatin accessibility, histone modifications, or DNA methylation. This may constrain predictive performance on targets influenced by dynamic epigenetic regulation (e.g., EZH2), representing a key direction for future improvement. Second, although an attention mechanism was introduced for visualization analysis, there remains a lack of systematic explanation regarding the internal logic of how the deep neural network quantitatively decodes specific physicochemical parameters to achieve specific recognition.

In response to these limitations, future research can be conducted in the following two directions. On one hand, a multi-modal learning architecture could be constructed by introducing feature vectors that reflect the dynamic chromatin environment, thereby improving the model's context-awareness in complex genomic backgrounds and enhancing prediction robustness across different cellular environments. On the other hand, research into model interpretability can be further deepened. Through

feature masking experiments or sensitivity analysis, the specific contributions of functional group distributions in NCP and geometric parameters in DPCP (such as Roll and Tilt) to the binding of different transcription factor families can be quantified, leading to the development of regulatory element mining tools with greater biological significance.

In summary, by fusing multi-dimensional physicochemical features with an advanced deep learning architecture, the DNCLA model provides an efficient and computational framework for the identification of large-scale genomic regulatory elements. Its outstanding performance in cross-cell line and multi-type transcription factor prediction tasks not only verifies the high biological complementarity between the DNA chemical microenvironment and local structural deformation information but also provides important bioinformatics support for the in-depth analysis of gene regulatory logic in non-coding regions and the pathogenic mechanisms of related diseases.

References

[1] A. Farrel, J. T. Guo, An efficient algorithm for improving structure-based prediction of transcription factor binding sites, *BMC Bioinformatics* **18** (2017) #342.

[2] M. Hajheidari, S. S. C. Huang, Elucidating the biology of transcription factor–DNA interaction for accurate identification of cis-regulatory elements, *Curr. Opin. Plant Biol.* **68** (2022) #102232.

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

[4] F. Spitz, E. E. M. Furlong, Transcription factors: from enhancer binding to developmental control, *Nat. Rev. Genet.* **13** (2012) 613–626.

[5] A. Casamassimi, A. Ciccodicola, M. Rienzo, Transcriptional regulation and its misregulation in human diseases, *Int. J. Mol. Sci.* **24** (2023) #8640.

[6] M. T. Maurano, R. Humbert, E. Rynes, R. E. Thurman, E. Haugen, H. Wang, A. P. Reynolds, R. Sandstrom, H. Qu, J. Brody, A. Shafer, F. Neri, K. Lee, T. Kutyavin, S. Stehling-Sun, A. K. Johnson, T. K.

Canfield, E. Giste, M. Diegel, D. Bates, R. S. Hansen, S. Neph, P. J. Sabo, S. Heimfeld, A. Raubitschek, S. Ziegler, C. Cotsapas, N. Sotoodehnia, I. Glass, S. R. Sunyaev, R. Kaul, J. A. Stamatoyannopoulos, Systematic localization of common disease-associated variation in regulatory DNA, *Science* **337** (2012) 1190–1195.

[7] G. D. Stormo, DNA binding sites: representation and discovery, *Bioinformatics* **16** (2000) 16–23.

[8] D. S. Johnson, A. Mortazavi, R. M. Myers, B. Wold, Genome-wide mapping of in vivo protein-DNA interactions, *Science* **316** (2007) 1497–1502.

[9] P. J. Park, ChIP–seq: advantages and challenges of a maturing technology, *Nat. Rev. Genet.* **10** (2009) 669–680.

[10] C. Fletez-Brant, D. Lee, A. S. McCallion, M. A. Beer, kmer-SVM: a web server for identifying predictive regulatory sequence features in genomic data sets, *Nucleic Acids Res.* **41** (2013) W544–W556.

[11] M. W. Libbrecht, W. S. Noble, Machine learning applications in genetics and genomics, *Nat. Rev. Genet.* **16** (2015) 321–332.

[12] A. M. Khamis, O. Motwalli, R. Oliva, B. R. Jankovic, Y. A. Medvedeva, H. Ashoor, M. Essack, X. Gao, V. B. Bajic, A novel method for improved accuracy of transcription factor binding site prediction, *Nucleic Acids Res.* **46** (2018) #e72.

[13] B. Hooghe, S. Broos, F. Van Roy, P. De Bleser, A flexible integrative approach based on random forest improves prediction of transcription factor binding sites, *Nucleic Acids Res.* **40** (2012) #e106.

[14] M. Ghandi, D. Lee, M. Mohammad-Noori, M. A. Beer, Enhanced regulatory sequence prediction using gapped k-mer features, *PLoS Comput. Biol.* **10** (2014) #e1003711.

[15] B. Alipanahi, A. Delong, M. T. Weirauch, B. J. Frey, Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning, *Nat. Biotechnol.* **33** (2015) 831–838.

[16] J. Zhou, O. G. Troyanskaya, Predicting effects of noncoding variants with deep learning–based sequence model, *Nat. Methods* **12** (2015) 931–934.

[17] D. Quang, X. Xie, DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences, *Nucleic Acids Res.* **44** (2016) #e107.

[18] Y. Zhang, Z. Wang, F. Ge, X. Wang, Y. Zhang, S. Li, Y. Guo, J. Song, D.-J. Yu, MLSNet: a deep learning model for predicting transcription factor binding sites, *Brief. Bioinform.* **25** (2024) #bbae489.

[19] Q. Zhang, Z. Shen, D. S. Huang, Predicting in-vitro transcription factor binding sites using DNA sequence+ shape, *IEEE/ACM Trans. Comput. Biol. Bioinform.* **18** (2019) 667–676.

[20] S. Wang, Q. Zhang, Z. Shen, Y. He, Z. H. Chen, J. Li, D. S. Huang, Predicting transcription factor binding sites using DNA shape features based on shared hybrid deep learning architecture, *Mol. Ther. Nucleic Acids* **24** (2021) 154–163.

[21] Y. Ji, Z. Zhou, H. Liu, R. V. Davuluri, DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome, *Bioinformatics* **37** (2021) 2112–2120.

[22] K. Wang, X. Zeng, J. Zhou, F. Liu, X. Luan, X. Wang, BERT-TFBS: a novel BERT-based model for predicting transcription factor binding sites by transfer learning, *Brief. Bioinform.* **25** (2024) #bbae195.

[23] Y. Zhang, Z. Wang, Y. Zeng, Y. Liu, S. Xiong, M. Wang, J. Zhou, Q. Zou, A novel convolution attention model for predicting transcription factor binding sites by combination of sequence and shape, *Brief. Bioinform.* **23** (2022) #bbab525.

[24] P. Ding, Y. Wang, X. Zhang, X. Gao, G. Liu, B. Yu, DeepSTF: predicting transcription factor binding sites by interpretable deep neural networks combining sequence and shape, *Brief. Bioinform.* **24** (2023) #bbad231.

[25] X. Xiao, H. X. Ye, Z. Liu, J. H. Jia, K. C. Chou, iROS-gPseKNC: predicting replication origin sites in DNA by incorporating dinucleotide position-specific propensity into general pseudo nucleotide composition, *Oncotarget* **7** (2016) 34180–34189.

[26] H. Zeng, M. D. Edwards, G. Liu, D. K. Gifford, Convolutional neural network architectures for predicting DNA–protein binding, *Bioinformatics* **32** (2016) i121–i127.

[27] W. Chen, X. Zhang, J. Brooker, H. Lin, L. Zhang, K. C. Chou, PseKNC-General: a cross-platform package for generating various modes of pseudo nucleotide compositions, *Bioinformatics* **31** (2015) 119–120.

[28] J. A. Swets, Measuring the accuracy of diagnostic systems, *Science* **240** (1988) 1285–1293.

[29] J. Muschelli III, ROC and AUC with a binary predictor: a potentially misleading metric, *J. Classif.* **37** (2020) 696–708.

[30] B. Martin, T. D. Bennett, P. E. DeWitt, S. Russell, L. N. Sanchez-Pinto, Use of the area under the precision-recall curve to evaluate prediction models of rare critical illness events, *Pediatr. Crit. Care Med.* **26** (2025) e855–e859.

[31] S. M. Lundberg, S. I. Lee, A unified approach to interpreting model predictions, *Adv. Neural Inf. Process. Syst.* **30** (2017) 4765–4774.

[32] E. M. Liu, A. Martinez-Fundichely, B. J. Diaz, B. Aronson, T. Cuykendall, M. MacKay, P. Dhingra, E. W. Wong, P. Chi, E. Apostolou, N. E. Sanjana, E. Khurana, Identification of cancer drivers at CTCF insulators in 1,962 whole genomes, *Cell Syst.* **8** (2019) 446–455.

[33] P. Cramer, Organization and regulation of gene transcription, *Nature* **573** (2019) 45–54.

[34] K. H. Kim, C. W. M. Roberts, Targeting EZH2 in cancer, *Nat. Med.* **22** (2016) 128–134.

[35] C. Grandori, S. M. Cowley, L. P. James, R. N. Eisenman, The Myc/Max/Mad network and the transcriptional control of cell behavior, *Annu. Rev. Cell Dev. Biol.* **16** (2000) 653–699.