TBCA: Prediction of Transcription Factor Binding Sites Using a Deep Neural Network With Lightweight Attention Mechanism

Xun Wang, Qiao Lian, Peng Qu, and Qing Yang

***Abstract*—The identification of transcription factor binding sites (TFBSs) is crucial for understanding the regulatory mechanisms of gene expression, which contributes to unraveling cellular functions and disease development. Currently, the most common approach involves the use of deep learning techniques to predict TFBSs by combining sequence and shape features. Although significant progress has been made with these methods, the integration of local features extracted from DNA sequences and shapes with global features has not yet reached a sufficient level, and there is still significant room for improvement in the accuracy of prediction results. In this paper, we propose a novel framework based on convolution and attention mechanisms, referred to as TBCA, which combines DNA sequence information and shape information for predicting transcription factor binding sites. In this work, we employ a two-layer convolutional neural network (CNNs) and self-attention mechanism to extract complex sequence features from DNA. What's more, we utilize a Fourier-transform-enhanced multi-head attention along with channel attention to extract high-order shape features of DNA. Finally, these high-order sequence and shape features are integrated into the channel dimension to achieve accurate TFBSs prediction. Our research results demonstrate that TBCA exhibits superior predictive performance in 165 validated ChIP-seq datasets. Furthermore, the employed attention mechanisms can automatically learn important features at different positions and scales, enhancing the accuracy and robustness of feature representation. We also conduct an in-depth analysis of the contributions of five different shapes to site prediction, revealing that shape features can enhance the prediction of transcription factor DNA binding.**

***Index Terms*—Transcription factor binding sites prediction, shape feature, convolutional neural network, fourier transform, attention mechanism.**

Manuscript received 4 June 2023; revised 1 November 2023 and 4 January 2024; accepted 15 January 2024. Date of publication 18 January 2024; date of current version 5 April 2024. This work was supported in part by the National Natural Science Foundation of China under Grants 61972416 and 62272479, and in part by the Natural Science Foundation of Shandong Province under Grant ZR2022LZH009. *(Corresponding author: Xun Wang.)*

The authors are with the College of Computer Science and Technology, China University of Petroleum, Beijing 102299, China (e-mail: wangsyun@upc.edu.cn; s21070055@s.upc.edu.cn; s21070009@s.upc.edu.cn; s21070069@s.upc.edu.cn).

Digital Object Identifier 10.1109/JBHI.2024.3355758

I. Introduction

THE ability of proteins to recognize specific DNA sequences plays a fundamental role in the intricate regulatory networks that govern cellular processes [1], [2], [3], [4], [5]. Transcription factors (TFs) represent a class of DNA-binding proteins widely present in organisms. They play a crucial role in the regulation of downstream gene expression by binding to specific regions of DNA, either enhancing or inhibiting the activity of downstream genes, and exert significant influence on cell growth, differentiation, and function. The DNA sequence sites bound by TFs are referred to as transcription factor binding sites (TFBSs), which typically range from a few to about 5–20 base pairs (bps), relatively conserved over long-term evolution, and exhibit a certain degree of sequence specificity [6]. Studies have shown that mutations in TFs and TFBSs are one of the major factors in human pathogenesis, with some genomic variations potentially leading to the development of tumors or genetic disorders [7], [8], [9]. Consequently, accurately identifying and locating TFBSs have the potential to support us in understanding how transcription factors regulate gene expression and provide important foundations for developing therapeutic methods and strategies targeting these regulatory mechanisms better. This holds tremendous importance in the identification and subsequent management of cancer [10], [11].

With the advancement in high-throughput sequencing technologies [12], biologists have integrated chromatin immunoprecipitation analysis (ChIP) with high-throughput sequencing, a technique known as ChIP-seq [13], [14], [15], which can mark the binding relationship between transcription factors and upstream sequences of genes on the entire genome, this integration facilitates the construction of gene regulatory networks and significantly accelerates research on transcription factor binding sites. Up to now, many sequence-based prediction algorithms have been developed to identify potential binding sites for various TFs, with the most common methods being probabilistic models [16], [17], [18] and machine learning models [19], [20], [21], [22]. However, these methods highly rely on experimentally obtained data and are limited in their scalability, being applicable only to specific transcription factors. They also do not consider the position dependency between nucleotides, which is an obvious weakness.

In recent years, deep learning (DL) has rapidly developed and has been widely applied across various domains. Many

scholars have incorporated DL algorithms into the prediction of TFBSs [23]. For instance, the DeepBind model [24] utilized a single-layer convolutional neural network to identify sequence-specificity in TFBSs within DNA sequences. DanQ [25], on the other hand, added bidirectional long short-term memory recurrent neural network (Bi-LSTM) [26] in this approach is capable of learning long-range dependencies within sequences. DSAC [27] adopted a dual-branch model, combining self-attention and CNNs interactively to fuse local features and global representations. These models effectively leverage sequence features for TFBSs prediction.

With the deepening of research on DNA structure recognition, researchers have come to realize that there are complex relationships and dependencies between fundamental positions due to the shape and deformation nature of DNA [28], [29], [30], [31]. Currently, numerous methods have integrated both DNA sequence and shape information to predict TFBSs, achieving more accurate and interpretable prediction results. Examples of such methods include DLBSS [32], CRPTS [33], D-SSCA [34], and DeepSTF [35]. However, these approaches do not fully harness the feature information obtained from sequences and shapes, and there is still significant potential for improving predictive performance by enhancing the integration of local and global information extracted from these sources.

Recently, multi-head attention mechanisms [36], [37], [38], [39] have found extensive applications across various domains. They aiding models in concentrating their attention on relevant information when dealing with complex data and tasks, consequently enhancing overall performance. DeepSTF, for the first time, introduced multi-head attention mechanisms into models for predicting TFBSs and achieved significant improvements. However, multi-head attention requires the computation of attention weights multiple times, its performance relies on the selection of the number of heads and hyperparameter tuning. Furthermore, each attention head introduces extra parameters, necessitating more extensive data for training. Therefore, multi-head attention may not be well-suited for smaller datasets. Inspired by this, Lee-Thorp and colleagues proposed the FNet model [40], which substitutes a standard, non-parametric Fourier transform in place of the self-attention sub-layers found in multi-head attention. The FNet model offers outstanding performance in a compact model while maintaining low memory consumption. Under similar computational speed and performance budgets, small-scale FNet models outperform multi-head attention models.

Motivated by these insights, we have developed an approach that combines convolution with lightweight attention mechanisms, referred to as TBCA. TBCA leverages convolution for local feature extraction from DNA sequence data and employs different attention mechanisms for DNA sequence and shape data. It combines the local features extracted through convolution with the global context provided by attention mechanisms, thereby focusing on various aspects of the input data for multi-scale information fusion. This improves the perception of the model and ultimately enhances its performance. We employ self-attention mechanisms for DNA sequence data, allowing the model to assign different weights to each nucleotide, dynamically capturing correlations and dependencies at different positions in the sequence. For DNA shape data, we utilize a Fourier-transform-enhanced transformer encoder combined with channel attention, aiding the model in capturing both local and global features of DNA shape for a better understanding of its three-dimensional structure. Furthermore, channel attention enables the model to dynamically assign weights to different shape feature channels, facilitating a better capture of key aspects of shape information. To validate the utility of the proposed model, we use the same dataset as Zhang et al. [34]. Experimental results demonstrate that TBCA effectively combines contextual information to integrate DNA sequence and shape features, bringing about more accurate predictions.

The architecture of TBCA is illustrated in Fig. 1. We conducted a series of experiments on a total of 165 ChIP-seq datasets. The results indicate that the model proposed in this paper, TBCA, exhibits superior performance compared to several existing methods, particularly demonstrating significant improvements on smaller datasets. Furthermore, we delved deeper into the contributions of individual shape information and the impact of excluding shape data from the model. Our findings reveal that models incorporating individual shape features and the combination of shape features outperform models that do not incorporate shape information. Thus, the inclusion of shape information has a positive impact on TFBSs prediction.

II. Related Works

A. The Significance of DNA Structural Characteristics in Predicting TFBSs

The structural characteristics of DNA in three-dimensional space directly influence the binding and function of biological macromolecules such as proteins and small molecules that interact with it. Recent research has indicated that the specific binding of proteins to DNA is a result of the protein's ability to perceive local changes in DNA shape and electrostatic potential. These local variations assist proteins in locating specific binding sites on DNA, allowing them to base-specific hydrogen bonds, thereby achieving DNA binding specificity. Some researchers have introduced a novel approach, combining sliding windows with Monte Carlo (MC) simulations [41], to extract various shape features based on sequences. These features have been depicted to significantly influence the capacity of TF-DNA interactions.

DNA shape features encompass six inter-base-pair features, six intra-base-pair features, and two minor groove features. Inter-base-pair features elucidate the translational distance and rotational angles between adjacent base pairs. Specifically, the six inter-base-pair features comprise 'Shift', 'Slide', 'Rise', 'Tilt', 'Roll', and 'HelT'. Intra-base-pair features delineate the translational distances and rotational angles within individual base pairs. The six intra-base-pair features encompass 'Shear', 'Stretch', 'Stagger', 'Buckle', 'ProT', and 'Opening'. Minor groove features describe the geometric shape and electrostatic potential of the minor groove center, which include 'MGW' and 'EP'.

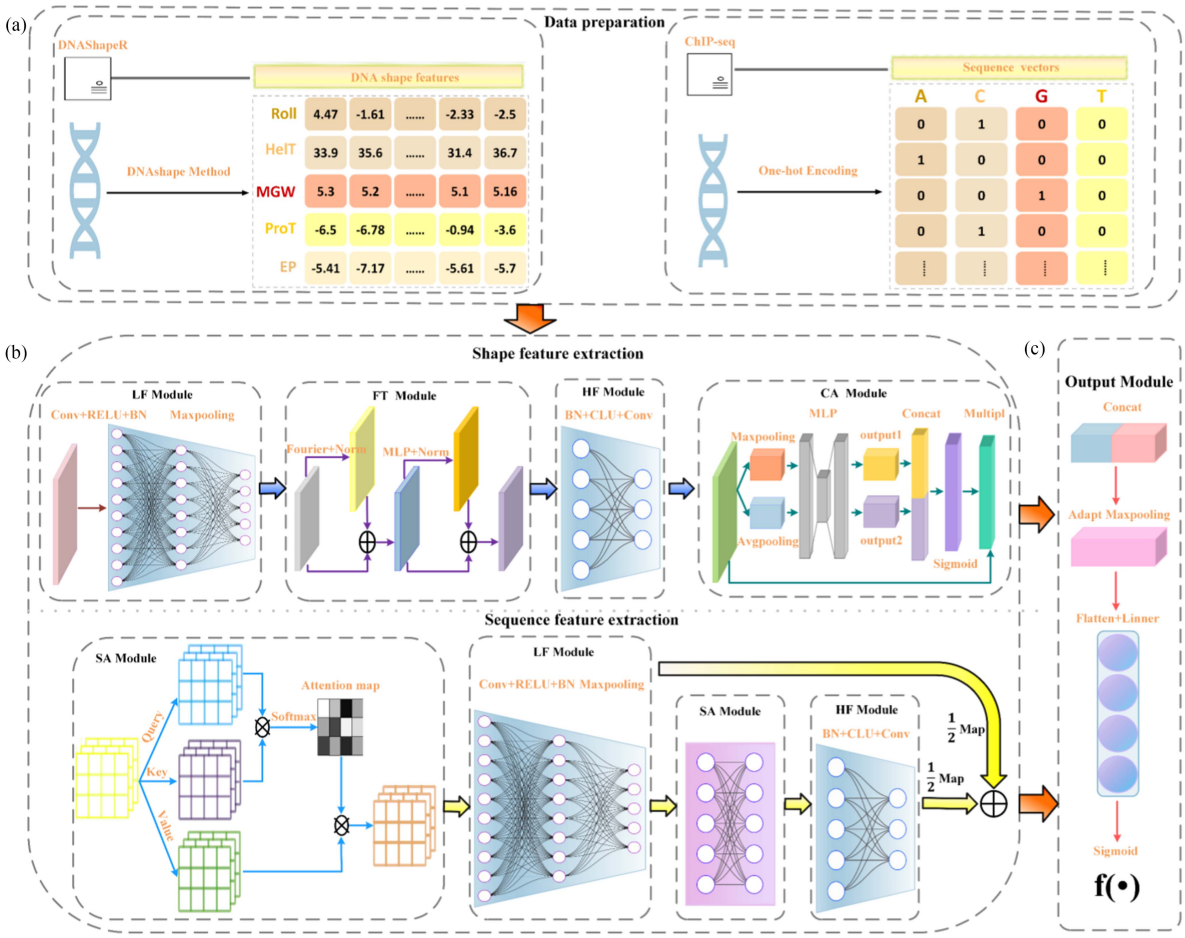

Fig. 1. Architecture of the TBCA model. (a) Data preprocessing: Conversion of DNA sequences and shape profiles into feature matrices. (b) Feature Extraction: Capturing multi-scale crucial features from DNA sequences and shapes. (c) Transformation of feature mappings into the final predictions.

B. Lightweight Attention Model

Attention mechanisms are a widely applied technique in the field of deep learning. They enable models to dynamically focus on different portions of input data, facilitating the dynamic capture of correlations and dependencies at various positions within sequences, thereby enhancing model performance. Besides that, channel attention allows models to dynamically assign weights to different shape feature channels, enabling better capture of critical shape information. By employing distinct attention mechanisms, a more comprehensive modeling and analysis of DNA sequence and shape data can be achieved. This, in turn, facilitates the extraction of key features relevant to transcription factor binding within DNA, ultimately leading to improved predictive performance.

III. Materials and Methods

A. Data Preprocessing

ChIP-seq Data: To assess the performance of TBCA, we employ the same dataset as used by Zhang et al.[34]. In detail, 165 sequence datasets are selected from the Encyclopedia of DNA Elements (ENCODE) project's 690 ChIP-seq datasets. These datasets contain TFBSs information obtained through ChIP-seq experimental techniques. They encompass 29 TFs from various cell lines. Each dataset extracts peak regions within a 101-bp window, annotating each nucleotide within this 101-bp region as either 0 or 1. Nucleotides labeled "1" are defined as positive sequences, indicating that they belong to TFBSs. Negative sequences are labeled "0" and consist of randomly generated sequences that maintain the dinucleotide frequency of the positive sequences. The GC content of these negative sample sequences matches that of the positive samples, and it is ensured "It is ensured that they are exclusive of any peaks in the positive data. These 165 datasets encompass a total of 7524836 samples. To facilitate model training and testing, these samples are divided into a training set (80% of the samples) and a test set (20% of the samples). The training set included 6019796 samples, while the test set consisted of 1505040 samples. Supplementary Table S1 provides detailed ChIP-seq statistics and specific counts of positive and negative samples for each dataset.

DNA shape data: We investigate numerous structural characteristics of DNA and select five representative shape features: Helical Twist (HelT), Minor Groove Width (MGW), Propeller

Twist (ProT), rolling (Roll), and Electrostatic Potential in the Minor Groove (EP). These features represent critical aspects of DNA shape, providing essential information about DNA structure and function. They play a pivotal role in the process of identifying and precisely regulating TFBSs. Through these features, we gain deeper insights into the three-dimensional structure of DNA molecules and their roles in gene expression and biological processes, which significantly contributes to research and exploration in the field of life sciences. DNAshapeR [28] stands as a powerful R/BioConductor package, specifically designed for the rapid and efficient prediction of DNA shape features, facilitating in-depth investigations into the structural characteristics of DNA.

B. The Model Architecture

Fig. 1 illustrates the structure of our proposed TBCA model. The model consists of three main components: the Data Preprocessing Module, the Feature Extraction Module, and the Output Module. The Data Preprocessing Module primarily focuses on preprocessing DNA sequence and shape data, converting them into corresponding feature matrices.The Feature Extraction Module is further divided into two main parts: Sequence Feature Extraction and Shape Feature Extraction. Each part employs a convolution combined with an attention pattern to capture multi-scale crucial features from DNA sequences and shapes. The Output Module is responsible for adaptive modeling, facilitating the prediction of TFBSs. Subsequently, we will provide a detailed explanation of these three major components in the following sections, in the order presented.

1) Data Preprocessing Module: The TBCA model requires two-dimensional vectors as input, comprising DNA sequence data and DNA shape data. For each input DNA sequence, the sequence information is initially transformed into a feature matrix S of dimension $1 \times 4 \times 101$ using one-hot encoding, where A = [1, 0, 0, 0], C = [0, 1, 0, 0], G = [0, 0, 1, 0], and T = [0, 0, 0, 1]. which can be seen as follows:

$$\mathrm{S}_1 = [o_1, o_2, \ldots o_i, \ldots, o_{101}] \tag{1}$$

where $\mathrm{o_i}$ denotes the one-hot vector of the ith nucleotide.

For the DNA shape data, we use the five types of nucleotide shape data generated by the DNAshapeR package based on Monte Carlo simulations, which are HelT, MGW, ProT, Roll, and EP. For each input DNA sequence was transformed into a 5×101 feature matrix.

$$\mathrm{S}_2 = [m_1, m, \ldots m_i, \ldots m_{101}] \tag{2}$$

where mi denotes the Monte Carlo simulation vector of the ith nucleotide. The data preprocessing procedure of the model is displayed in Fig. 1(a).

2) Feature Extraction Module:: For the extraction of sequence features, it is primarily composed of the following modules:

SA Module: Batch normalization is applied to the DNA sequence matrix S_1 to obtain the data X. This operation helps balance the distribution of input data, ensuring that the mean of input sequences is close to zero and the standard deviation is close to 1. This is beneficial for mitigating issues related to gradient explosion or vanishing during training and can also enhance training speed, facilitating faster model convergence. Subsequently, a ReLU [42] activation function is added to enhance the nonlinearity of the model, enabling it to better adapt to the data patterns. ReLU is a widely-used activation function in deep learning.

$$\mathrm{ReLU}\left(\mathrm{X}\right) = \max\left(0, \mathrm{X}\right) \tag{3}$$

Next, a linear transformation is applied to the preprocessed input data, resulting in the Query $\mathrm{Q} \in \mathrm{R}^{\mathrm{T}\times \mathrm{d_k}}$, Key $\mathrm{K} \in \mathrm{R}^{\mathrm{T}\times \mathrm{d_k}}$, and Value $\mathrm{V} \in \mathrm{R}^{\mathrm{T}\times \mathrm{d_v}}$ where T represents the sequence length, and $\mathrm{d_k}$ and $\mathrm{d_v}$ represent the hidden dimensions of the query or key and value, respectively. The calculations for query, key, and value are as follows:

$$\mathrm{Q} = \mathrm{W}_{\mathrm{Q}}^{\mathrm{T}}\,\mathrm{X} \tag{4}$$

$$\mathrm{K} = \mathrm{W}_{\mathrm{K}}^{\mathrm{T}}\,\mathrm{X} \tag{5}$$

$$V = W_v^T\, X \tag{6}$$

Where W_Q, W_K and W_V are the learned weight matrices for the query vector, key vector, and value vector, respectively. Subsequently, the similarity or correlation between Query and Key is computed, followed by normalization to obtain the matrix of attention weights.

$$\mathrm{Attention}\left(\mathrm{Q}, \mathrm{K}, \mathrm{V}\right) = \mathrm{softmax}\left(\frac{\mathrm{QK}^{\mathrm{T}}}{\sqrt{\mathrm{d_k}}}\right)\mathrm{V} \tag{7}$$

Where $\sqrt{\mathrm{d_k}}$ is the length of the vectors used for scaling the attention matrix, ensuring stable gradient values during the training process and preventing gradient vanishing.

LF Module: Convolutional operations are effective in extracting features from DNA sequence matrices and DNA shape matrices. By using a sliding window of filters, local features at different positions can be captured. We initially employ a single-layer convolutional neural network to extract low-level features from the input matrix. Subsequently, we apply batch normalization to normalize the extracted feature vectors to the range of 0–1, with ReLU serving as the activation function during this process. Next, a max-pooling layer is used to downsample the convolutional feature maps to reduce data dimensions, thereby decreasing model parameters and computational load, while enhancing model robustness. The max-pooling layer retains the most significant features and achieves translational invariance, allowing the model to capture relevant motifs at any position. Furthermore, we incorporate a dropout layer, which serves as a regularization technique that randomly sets some neuron outputs to zero during training, reducing interdependencies between neurons and preventing overfitting.

Technically, the module of LF operations on sequence feature maps are as follows:

$$\mathrm{L}_1 = \mathrm{Maxpool}\left(\mathrm{BN}\left(\mathrm{ReLU}\left(\mathrm{Conv}\left(\mathrm{P}_1, \mathrm{W_c}, \mathrm{b_c}\right)\right)\right)\right) \tag{8}$$

Where $\mathrm{W_c}$ signifies the weight matrix of the convolutional layer, $\mathrm{b_c}$ represents the bias of the convolutional layer, $\mathrm{Conv}(*)$

denotes the convolution process, BN$(*)$ stands for batch normalization operation, and Maxpool$(*)$ represents the local max-pooling operation.

In the first and third stages of feature extraction from the DNA sequence, two self-attention modules with dropout are employed. The use of a dropout rate of 0.2 in the first stage and 0.7 in the third stage effectively mitigates the occurrence of overfitting.

HF Module: To extract deeper high-level features, the HF module employs batch normalization and Exponential Linear Units (ELU) to accelerate neural network training, enhance network elasticity, and improve robustness. Subsequently, two-dimensional convolution operations are used to generate more complex high-level features.

$$\mathrm{H} = \mathrm{Conv}\left(\mathrm{ELU}\left(\mathrm{BN}\left(\mathrm{F}\right)\right)\right) \tag{9}$$

In a nutshell, the sequence feature extraction module begins with the SA module to capture global features of the sequence data, enhancing the model with an understanding of key information. Following this, the LF module conducts two-dimensional convolution operations and max-pooling to extract local patterns and structure within the sequence. The convolution operations help expand the receptive field of the model and extract information from different scales of features, while the max-pooling layer reduces the spatial dimension of features to preserve the most significant ones. This serves to reduce computational overhead while retaining crucial information. Subsequently, the extracted features are processed once again through the SA module to capture long-range dependencies within the sequence, extracting deeper essential features. The captured deep-level features are then passed to the HF module to extract high-level features weighted by attention for the second time, enhancing feature representation capabilities. To integrate features at different levels, the high-level features obtained from the second attention processing are linearly combined with the multi-scale features obtained through the second convolution processing using weighted parameters, resulting in a more representative feature representation. The sequence feature extraction process is illustrated in the lower part of Fig. 1(b).

Continuing, let me detail the composition of the shape feature extraction module.

FT Module: This module consists of an improved transformer encoder structure. This work draws inspiration from Fnet's practical experience and the concept of encoder modules [43]. Essentially, we replace the self-attention sub-layers in each transformer encoder layer with Fourier sub-layers. For the low-level shape feature map L_2 obtained after processing by the LF module, a Fourier transform is initially applied to convert the input into a frequency domain representation, which helps capture long-range dependencies of the input features. Unlike traditional self-attention mechanisms, the FT module has lower computational overhead due to the absence of parameters. Simultaneously, we apply a multi-layer perceptron to perform nonlinear transformations on the frequency domain representation of L_2, which helps capture complex features in the input. Finally, we merge the output of the Fourier transform module and the multi-layer perceptron module with the output of the residual connection and layer normalization modules to produce the final output of the FT module. The exact computational process is as follows:

$$\mathrm{F} = \mathrm{LN}(\mathrm{FFT}\left(\mathrm{L}_2\right) + \mathrm{LN}\left(\mathrm{MLP}\left(\mathrm{L}_2\right)\right) \tag{10}$$

Where FFT$(*)$ denotes the Fourier Transform operation, MLP$(*)$ represents the Multi-Layer Perceptron operation, and LN$(*)$ represents Layer Normalization operation.

CA Module: Subsequently, we further refine the deep-level feature information of DNA shape data obtained from the FT module using the Channel Attention (CA) module. First, we simultaneously employ global average pooling and global max pooling operations to obtain the overall statistics of the data as well as the most prominent local information.

$$\mathrm{C}_{\mathrm{avg}} = \mathrm{CAP}\left(\mathrm{H}_2\right) \tag{11}$$

$$\mathrm{C}_{\mathrm{max}} = \mathrm{CMP}\left(\mathrm{H}_2\right) \tag{12}$$

Where CAP$(*)$ and CMP$(*)$ define global average pooling and global max pooling operations, respectively.

Next, a shared fully connected layer with one hidden layer is applied to Cavg and Cmax, and the output vectors are integrated through element-wise summation to generate the attention map.

$$\begin{aligned}\mathrm{A} = \mathrm{Sigmoid}(&\mathrm{MLP}\left(\mathrm{C}_{\mathrm{avg}}, \mathrm{W}_{\mathrm{a}}^{(0)}, \mathrm{W}_{\mathrm{a}}^{(1)}\right)\\ &+ \mathrm{MLP}\left(\mathrm{C}_{\mathrm{max}}, \mathrm{W}_{\mathrm{a}}^{(0)}, \mathrm{W}_{\mathrm{a}}^{(1)}\right)\end{aligned} \tag{13}$$

Where Sigmoid$(*)$ indicates the Sigmoid function. $\mathrm{W}_{\mathrm{a}}^{(0)}$ and $\mathrm{W}_{\mathrm{a}}^{(1)}$ represent the weight matrices of the shared multi-layer perceptron.

Finally, the channel-wise attention map M is multiplied with the input feature map H_2 for adaptive feature refinement. Formally, the feature refinement operation is defined as follows:

$$\mathrm{M} = \mathrm{M} \otimes \mathrm{H}_2 \tag{14}$$

Where A stands for the optimized feature mapping. Similarly, an optimized feature map A is also generated. $\otimes$ defines element-wise multiplication operation.

To outline the main points, the shape feature extraction process commences with the LF convolution module, which initially captures patterns and structural information across different scales within shape data. Subsequent pooling layers serve to reduce feature dimension while retaining the most salient characteristics. Following pooling, the FT attention module allows the model to further understand the internal relationships and dependencies within the shape information, enabling better capture of abstract features in the shape information and weight assignment to different features. The attention module enhances the sensitivity of the model to shape information, facilitating the discrimination of distinct patterns. Consecutively, the HF module extracts features that have undergone attention adjustments, thereby enhancing feature expressiveness and discriminability. The convolution layers in this module help extract information at different hierarchical levels, including more abstract features. Finally, the features are passed through the channel-level attention

mechanism in the CA module. This mechanism weights the feature channels to improve the sensitivity of the model to different feature channels. It aids the model in better comprehending relationships between channels of shape information, consequently preserving critical shape features. The shape feature extraction process is illustrated in the upper portion of Fig. 1(b).

3) Output Module: To fully utilize the features extracted from sequence and shape data, we treat these two types of data as separate input channels and concatenate them along the channel dimension of a convolutional neural network. Later in the process, the concatenated feature maps undergo adaptive max-pooling, further enhancing the robustness of the extracted features. Adaptive max-pooling is an operation that dynamically adjusts the size and stride of the pooling kernel based on the size and shape of the input feature maps, ensuring that the size and shape of the output feature maps remain the same as the input feature maps. Finally, a Sigmoid function is applied to predict the features. The output process of the model is illustrated in Fig. 1(c).

From a technical standpoint, the operations can be described as follows:

$$\begin{aligned} \mathrm{Y} \;=\; & \mathrm{Sigmoid}(\mathrm{Linear}(\mathrm{Flatten} \\ & (\mathrm{AdaptMaxpool}\left(\mathrm{Concat}\left(\mathrm{H}_1, \mathrm{M}\right)\right))) \end{aligned} \tag{15}$$

C. Hyperparameter Setting

During the implementation of the convolutional module, it is necessary to determine parameters such as the number of convolutional layers, kernel size, stride, and the number of channels. To determine the optimal combination of parameters, multiple parameter samplings are usually performed, and the best parameter combination is selected based on experimental results. In the experiments described in this paper, convolutional operations are performed on both the DNA sequence and shape data. For convolution operations on DNA sequence data and shape data, we employ single-layer convolution, two-layer convolution, and three-layer convolution operations, with channel numbers of 32, 64, and 128. In addition, a kernel size of 4×16 (kernel width of 4 and height of 16) is used for sequence data, while a kernel size of 5×16 (kernel width of 5 and height of 16) is used for shape data. After conducting multiple comparative experiments, it is determined that a single-layer convolution with 128 channels represented the optimal parameter combination for the convolutional module.

In our training process, we employ the binary cross-entropy loss function (BCELoss) [44], which effectively mitigates the issue of gradient vanishing, ensuring more stable model training. We also utilize a relatively small batch size of 64 to prevent overfitting during model training. Furthermore, we employ the Adam optimizer [45], an adaptive learning rate algorithm that automatically adjusts the update step size for weights and bias parameters based on gradient magnitudes. This adaptation speeds up convergence to optimize the loss function efficiently. Gradient computations are performed using the backpropagation algorithm, efficiently computing the gradients of each parameter in the neural network concerning the loss function. The chain rule was applied to propagate gradients to each neuron, enabling efficient parameter updates. Moreover, we implement a learning rate with exponential decay, allowing the learning rate to gradually decrease as training progresses, facilitating faster convergence to the optimal solution. To determine the best hyperparameter set, we sample these hyperparameters ten times and perform a five-fold cross-validation approach on the training data to select the hyperparameter set corresponding to the highest average PR-AUC score. The best hyperparameter set is subsequently applied to train the final model on the entire training dataset. Each model is trained for a maximum of 15 epochs, and an early stopping strategy was used to prevent overfitting. Our models are implemented using PyTorch.

D. Performance Evaluation and Comparison

In this article, Accuracy(ACC), ROC-AUC, and PR-AUC are used to measure the performance of our proposed method. These metrics are summarized as follows.

ACC is used to measure the proportion of correctly predicted TFBSs and non-TFBSs among all samples. However, when there is an imbalance between positive and negative samples, the model tends to favor the dominant class, ultimately culminating in poorer performance in predicting the minority class. In such cases, using ROC-AUC and PR-AUC can be more informative for conducting a comprehensive evaluation of model performance, especially in the identification of minority classes.

ROC-AUC: This metric represents the area under the Receiver Operating Characteristic curve, which is the area under the curve of True Positive Rate (TPR) and False Positive Rate (FPR). It provides a thorough assessment of model performance across various thresholds, making it particularly suitable for assessing classification performance, regardless of whether the distribution of positive and negative samples is balanced.

PR-AUC: This metric represents the area under the Precision-Recall curve. It focuses more on the recognition performance of the model for minority classes because it is based on precision and recall and is often more informative when dealing with imbalanced data.

IV. Experimental Results

A. Model Ablation

The attention mechanism in the TBCA model improves predictive performance. This study establishes two sets of variable models to further understand the value of the attention mechanism.

1) To verify the effectiveness of the improved attention module, we develop a model, TBCNA, which uses only two layers of CNN to process feature information. The average prediction results on the test sets of 165 ChIP-seq datasets for this model are shown in Table I. For detailed calculation data, please refer to Supplementary Table S2. Compared to TBCNA, TBCA shows relative improvements of 3.2% in ACC, 6.06% in ROC-AUC, and 5.2% in PR-AUC. Additionally, the data distribution results of TBCA and its derived variant models on

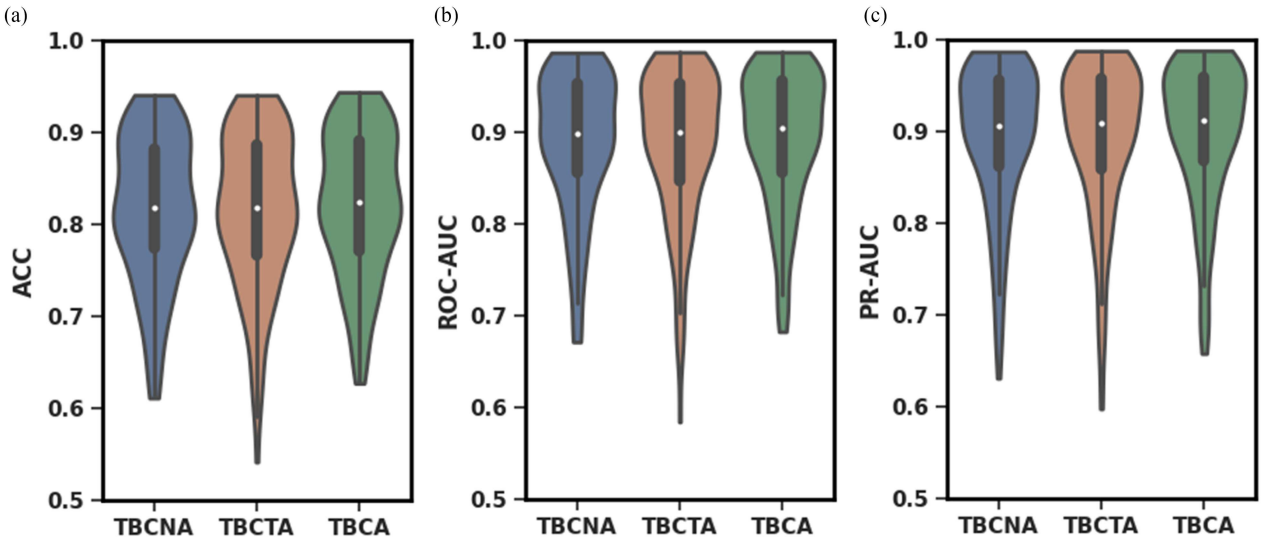

Fig. 2. Violin plots of ACC, ROC-AUC, and PR-AUC for TBCA and variant models. The larger the area of a certain region, the higher the probability distribution around a particular value.

TABLE I
PERFORMANCE COMPARISON

Method	ACC	ROC-AUC	PR-AUC
DeepBind	0.78	0.848	0.853
DanQ	0.778	0.847	0.853
DLBSS	0.784	0.859	0.864
CRPTS	0.787	0.858	0.621
D-SSCA	0.783	0.853	0.854
DeepSTF	0.811	0.882	0.888
DSAC	0.815	0.891	0.896
TBCA	**0.823**	**0.894**	**0.899**

The results comparing the ACC, ROC-AUC, and PR-AUC of TBCA with seven state-of-the-art methods on the test set of 165 ChIP-seq datasets.

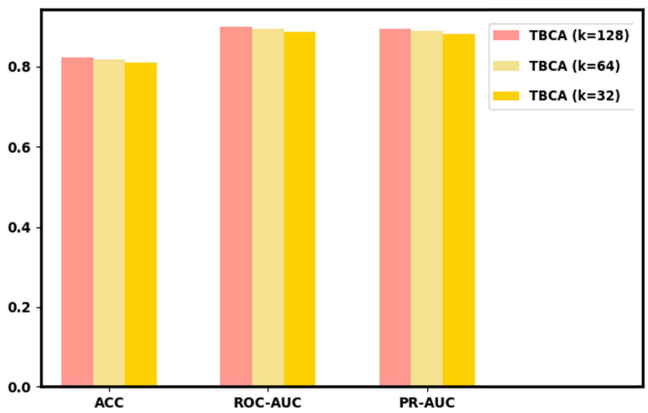

Fig. 3. Impact of different kernel sizes on the average ACC, ROC-AUC, and PR-AUC of the TBCA model. As the value of k increases, the performance of the model gradually improves, where k represents the number of convolutional kernels.

the test sets of 165 ChIP-seq datasets are depicted in Fig. 2. The results indicate that TBCA exhibits a more stable data distribution and overall performance. This is because using traditional convolution and max-pooling alone has limited capability to extract global features. They typically focus on the local features of input data and struggle to capture the overall information of the data. In contrast, the attention mechanism allows the model to capture global information and long-range dependencies in the input data. Combining convolution and attention mechanisms enables the model to incorporate both local and global features to some extent. Therefore, adding the attention mechanism to the model can effectively improve its predictive performance. Further, the performance of the TBCA model remains good on datasets of various sizes, demonstrating that the model structure constructed in this study has good generalization and robustness.

2) In the processing of DNA shape features, our study compares the performance of the FFT-enhanced unparameterized attention model (TBCA) with the combination of channel attention mechanism and multi-head attention with channel attention mechanism (TBCTA). The goal is to gain a better understanding of the advantages of the parameterless attention mechanism within the model. As displayed in Fig. 2, TBCA exhibits greater stability and outperforms the variant model, especially on smaller datasets when compared to TBCTA. At the same time, under the same dataset and computational environment, TBCA requires less time for training. Detailed computational data can be found in Supplementary Table S2. A reasonable explanation for this observation is that multi-head attention involves a relatively larger number of parameters, potentially demanding more computational resources and data for effective training. In contrast, the computational complexity of the Fourier Transform-based layer is independent of sequence length, making it more suitable for processing various types of sequences. Additionally, the Fourier Transform-based layer has lower computational complexity, effectively reducing the computational costs. This suggests that the use of the FFT-enhanced parameterless attention model (TBCA) along with the channel attention mechanism

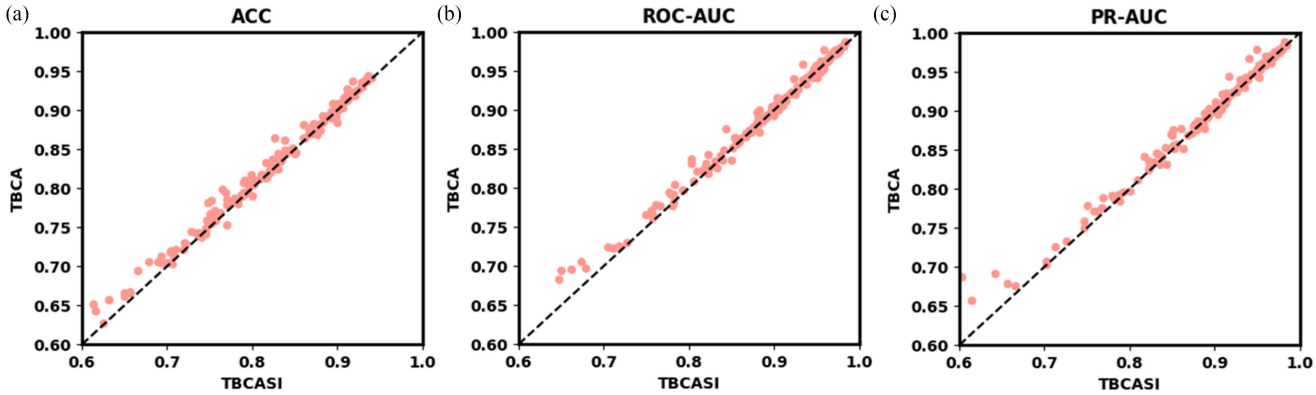

Fig. 4. ACC, ROC-AUC and PR-AUC comparison of TBCA and TBCASI on the test set of 165 ChIP-seq datasets.

significantly enhances the ability of the model to capture DNA shape features, making it a more suitable choice for the constructed TBCA model.

B. Analysis of the Contribution of DNA Shape

We extend the TBCA model with three sets of experiments to better understanding of the role of DNA shape information in predicting TFBSs. For detailed calculation data, please refer to Supplementary Table S3. Firstly, our study constructs three model structures by varying the number of convolutional kernels to assess the impact of model complexity on the ability to capture sequence and shape features. For each experiment, we evaluated the performance of models using different numbers of kernels (32, 64, and 128) in terms of ROC-AUC. The effect of different kernel sizes on the ACC, ROC-AUC, and ROC-AUC performance of the model is manifested in Fig. 3. The experimental results indicate that the model using 128 convolutional kernels performs the best, as it has a larger receptive field, allowing it to capture a broader range of input information.

Besides, we develop TBCASI, a model that solely employs independent DNA sequences as input. The comparative results of ACC, ROC-AUC and PR-AUC between TBCA and TBCASI on 165 ChIP-seq dataset test sets are illustrated in Fig. 4. It's easy to observe that in most test sets, TBCA outperforms TBCASI significantly. indicates a positive role of shape information in the interaction between DNA and transcription factors. Some TFBSs may highly dependent on shape information, making DNA shape data crucial in certain scenarios. However, some TFBSs might be more dependent on the [24] nucleotide composition of DNA sequences, in which case sequence information is likely to be more important. Overall, TBCA demonstrates a noticeable improvement in ACC, ROC-AUC and PR-AUC, highlighting the enhancement in TFBSs prediction accuracy due to the addition of shape information. Therefore, we believe that an appropriate combination of sequence and shape features can enhance the predictive capabilities of DL models.

Finally, to further investigate the contribution of each shape feature to TFBSs prediction, we sequentially exclude one of the

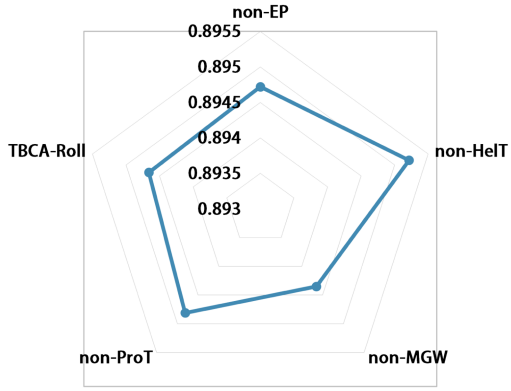

Fig. 5. Contribution distribution of Roll, HelT, MGW, ProT, and EP. The farther away from the center, the higher the contribution.

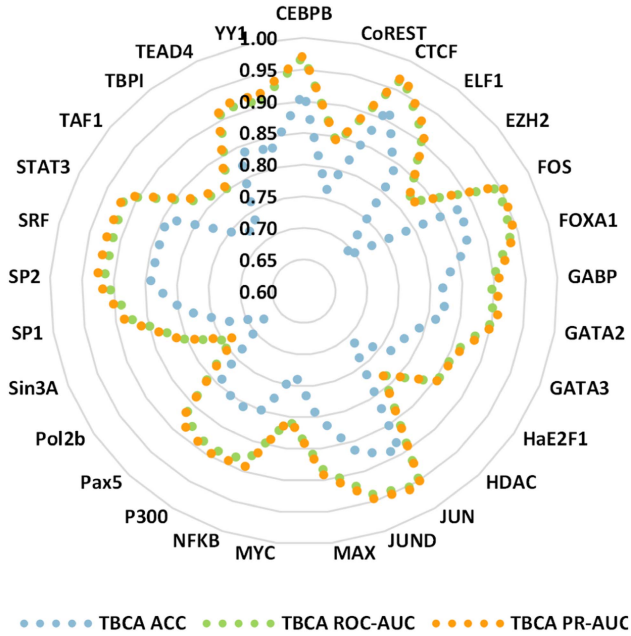

Fig. 6. Performance of TBCA on 29 TFs. The further the dots are from the center in the figure, the better the performance of the TFs.

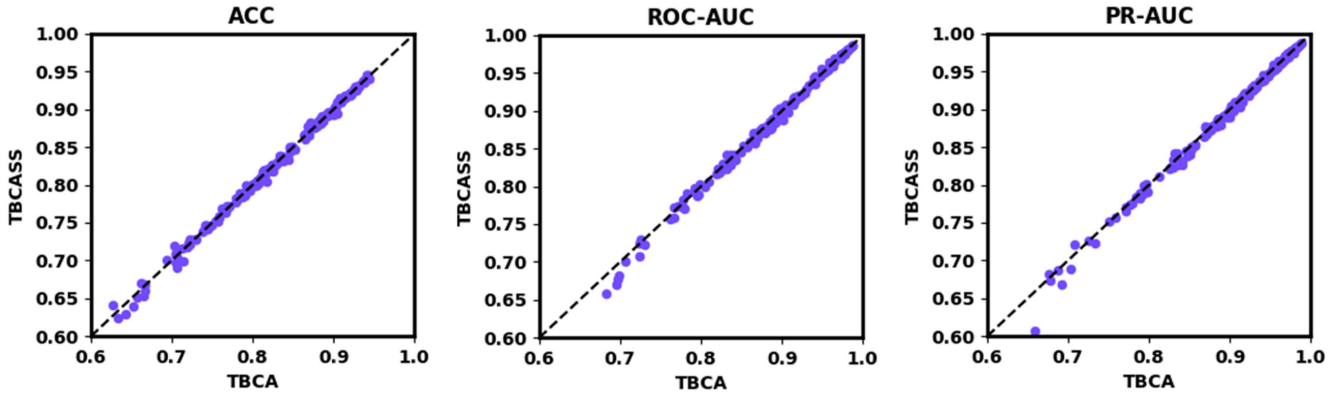

Fig. 7. Results comparing the usage of 14 DNA shape features to only 5 shape features across 165 test datasets are depicted in Fig. 7.

five shape features from the input DNA shape data, omitting it as input to observe changes in the performance of the model.

Fig. 5 illustrates the overall contributions of Roll, HelT, MGW, ProT, and EP in predictions across 165 datasets. It is evident that each DNA shape feature plays a positive role in the prediction results, with MGW and Roll exhibiting superior performance in TFBSs identification compared to the other three shape factors. This suggests that TFs with different structures may exhibit distinct shape recognition preferences when binding to DNA. In comparison to the TBCA model, the model utilizing a combination of all five shape features demonstrates the best performance. In other words, the average positive contribution of the five shape features surpasses the average negative contribution, indicating that the shapes can complement one another, thereby enhancing the predictive capabilities of the model. Therefore, we can infer that the convolution and attention model with added shape features is effective and has a positive impact on TFBSs prediction.

In this study, we conduct a comprehensive analysis of the performance of the TBCA model, with a particular focus on its average performance across 29 different TF datasets. The aim is to gain a deeper understanding of the model's applicability and its potential for broad generalization, especially concerning multiple distinct transcription factors. Fig. 6 illustrates the performance of TBCA across these 29 TFs.

Significant variations in the predictive results are observed among different TFs, which may be attributed to the specific binding specificity of each TF to nucleotides in DNA sequences, thus affecting their predictive outcomes. Furthermore, the differences in the quantity and quality of ChIP-seq experimental data available for various TFs can influence the accuracy of TFBSs predictions, with more extensive and high-quality data typically leading to improved predictive performance.

To investigate the impactof existing binding specificity models using nine additional DNA shape features (Rise, Shift, Slide, Tilt, Buckle, Opening, Shear, Stagger, Stretch), we develop TBCASS. We explore the performance of models utilizing 14 DNA shape features compared to those employing only five, conducting experiments on 165 test datasets. As illustrated in Fig. 7, models incorporating more shape features demonstrated slightly lower performance for 165 test datasets. Additionally, we analyze the predictive outcomes for each transcription factor using the 14 shape features. For most TFs, models integrating multiple shape features exhibited slight improvements in performance compared to TBCA. However, for smaller datasets of TFs, such as EZH2, TBCASS demonstrate poorer performance. Comprehensive computational details can be found in Supplementary Table S4. While integrating more DNA shape features might offer comprehensive insights, it could potentially augment model complexity, lengthen training times, and introduce noise or unnecessary intricacies. Therefore, leveraging a smaller yet more representative and predictive set of features may prove more effective.

C. Performance Comparison and Analysis

To further assess the performance of TBCA, we compare it with seven state-of-the-art TFBSs prediction methods. These seven methods can be categorized into two groups: (i) standalone sequence-based deep learning methods, such as DeepBind, DanQ, and DSAC, and (ii) deep learning methods that combine sequence and shape information, such as DLBSS, CRPTS, D-SSCA, and DeepSTF. To ensure a fair comparison, we carefully optimized the aforementioned methods and computed the average ACC, ROC-AUC, and PR-AUC on 165 datasets. The comparative results are presented in Fig. 8. Detailed computation data can be found in Supplementary Table S5. TBCA outperforms all competing methods in terms of average ACC, ROC-AUC, and PR-AUC, with a particularly noticeable improvement on smaller datasets. indicating the robustness of TBCA. Table I summarizes the average ACC, ROC-AUC, and PR-AUC results for our model and all competing methods across the 165 datasets.

Compared to DeepBind, DanQ, and DSAC models that utilize sequence information alone, TBCA exhibits significant improvement, as illustrated in Fig. 8. This performance enhancement may be attributed to the fact that the proposed model simultaneously considers both sequence and shape features. These two types of features are complementary, with sequence information better capturing base sequence patterns and shape information being more adept at capturing the three-dimensional

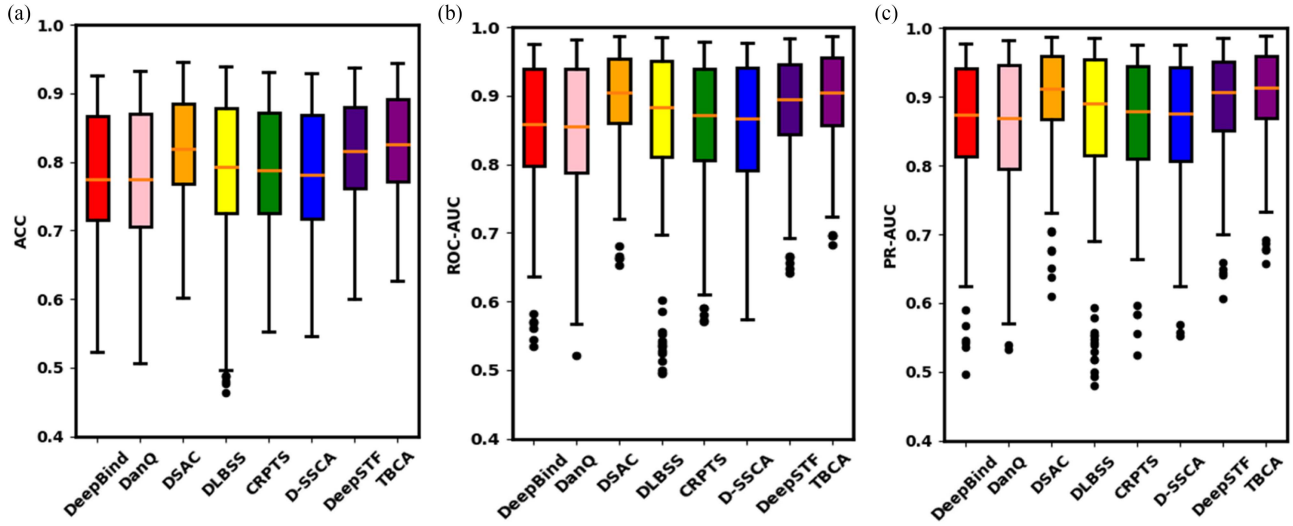

Fig. 8. Box plots comparing the ACC, ROC-AUC, and PR-AUC of TBCA with seven state-of-the-art methods on the test sets of 165 ChIP-seq datasets. The central orange line within the box represents the median, while the ends of the box represent the upper and lower quartiles. The lines extending to the highest and lowest observed values in the boxplot depict the upper and lower limits of the dataset, while dots signify outliers.

structure and spatial distribution characteristics of bases. This complementary advantage optimizes the predictive capabilities of the model for transcription factor binding sites. Hence, we believe that the appropriate combination of sequence and shape features can improve the predictive ability of deep learning models.

Compared to DLBSS, CRPTS, and D-SSCA models that incorporate both sequence and shape information, TBCA also shows obvious improvement, as depicted in Fig. 8. However, it is noticeable that models with a simple addition of LSTM perform worse than those using attention mechanisms. This is primarily due to the fact that LSTM often require a large amount of data for training to harness their potential advantages. With smaller datasets, LSTMs may tend to overfit, leading to decreased performance. Self-attention mechanisms, when combined with unparameterized attention along with channel attention, can better capture global dependencies, dynamically focus on various regions of the input data, and consequently, capture multi-scale features more effectively. This contributes to improved model generalization across different scales and complexities of data.

V. Discussion

We propose a novel convolution and attention-based framework named TBCA, which combines DNA sequence information and shape features for predicting TFBSs. The results indicate that the convolution with lightweight attention modules used in TBCA outperforms several common TFBSs prediction methods on a validated set of 165 datasets, especially excelling in scenarios with smaller data volumes. However, due to differences in data sizes among the 165 ChIP-seq sequence datasets from ENCODE, the performance of the model may be limited when training on datasets with fewer samples, resulting in reducing performance in such cases. Therefore, TBCA employs a convolution combined with an unparameterized attention mechanism to better capture global dependencies, focusing on multiple aspects of input sequences and enhancing the ability to capture multiscale features, ultimately improving the generalization of the model to data of varying scales and complexities.

Furthermore, we conduct individual analyses of each shape feature. The performance of combined DNA shape features showed only slight improvement compared to using individual DNA shape features, which suggests that TFs with different structures may exhibit distinct shape recognition preferences when binding to DNA. Moreover, integrating more DNA shape features can offer a more comprehensive set of information. However, they may also increase the complexity of the model and the training time.

We believe that advanced deep learning techniques combined with attention mechanisms for the in-depth analysis of sequence and shape motifs will assist in inferring gene expression regulation relationships, enabling more precise modeling of complex regulatory systems in the human genome. We hope that TBCA serves as a valuable tool for researchers in the genomics field, helping them gain a better understanding of the mechanisms governing gene expression regulation.

References

[1] T. I. Lee and R. A. Young, "Transcriptional regulation and its misregulation in disease," *Cell*, vol. 152, no. 6, pp. 1237–1251, 2013.
[2] P. J. Mitchell and R. Tjian, "Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins," *Science*, vol. 245, no. 4916, pp. 371–378, 1989.
[3] M. Ptashne, "Gene regulation by proteins acting nearby and at a distance," *Nature*, vol. 322, no. 6081, pp. 697–701, 1986.
[4] G. L. Semenza, "Transcriptional regulation of gene expression: Mechanisms and pathophysiology," *Hum. Mutat.*, vol. 3, no. 3, pp. 180–199, 1994.

[5] D. Wilson, V. Charoensawan, S. K. Kummerfeld, and S. A. Teichmann, "DBD—-Taxonomically broad transcription factor predictions: New content and functionality," *Nucleic Acids Res.*, vol. 36, no. suppl_1, pp. D88–D92, 2008.

[6] G. D. Stormo, "DNA binding sites: Representation and discovery," *Bioinformatics*, vol. 16, no. 1, pp. 16–23, 2000.

[7] K. J. Karczewski et al., "The mutational constraint spectrum quantified from variation in 141,456 humans," *Nature*, vol. 581, no. 7809, pp. 434–443, 2020.

[8] E. Khurana, Y. Fu, D. Chakravarty, F. Demichelis, M. A. Rubin, and M. Gerstein, "Role of non-coding sequence variants in cancer," *Nature Rev. Genet.*, vol. 17, no. 2, pp. 93–108, 2016.

[9] T. H. Kim and B. Ren, "Genome-wide analysis of protein-DNA interactions," *Annu. Rev. Genomic. Hum. Genet.*, vol. 7, pp. 81–102, 2006.

[10] P. Ulz et al., "Inference of transcription factor binding from cell-free DNA enables tumor subtype prediction and early detection," *Nature Commun.*, vol. 10, no. 1, 2019, Art. no. 4666.

[11] S. Van Dam, U. Vosa, A. van der Graaf, L. Franke, and J. P. de Magalhaes, "Gene co-expression analysis for functional classification and gene–disease predictions," *Brief. Bioinf.*, vol. 19, no. 4, pp. 575–592, 2018.

[12] Q. Pan, O. Shai, L. J. Lee, B. J. Frey, and B. J. Blencowe, "Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing," *Nature Genet.*, vol. 40, no. 12, pp. 1413–1415, 2008.

[13] J. Harrow et al., "GENCODE: The reference human genome annotation for The ENCODE Project," *Genome Res.*, vol. 22, no. 9, pp. 1760–1774, 2012.

[14] P. J. Park, "ChIP–Seq: Advantages and challenges of a maturing technology," *Nature Rev. Genet.*, vol. 10, no. 10, pp. 669–680, 2009.

[15] D. Schmidt, M. D. Wilson, C. Spyrou, G. D. Brown, J. Hadfield, and D. T. Odom, "ChIP-seq: Using high-throughput sequencing to discover protein–DNA interactions," *Methods*, vol. 48, no. 3, pp. 240–248, 2009.

[16] A. Mathelier and W. W. Wasserman, "The next generation of transcription factor binding site prediction," *PLoS Comput. Biol.*, vol. 9, no. 9, 2013, Art. no. e1003214.

[17] P. Mehta, D. J. Schwab, and A. M. Sengupta, "Statistical mechanics of transcription-factor binding site discovery using hidden Markov models," *J. Stat. Phys.*, vol. 142, pp. 1187–1205, 2011.

[18] G. D. Stormo and Y. Zhao, "Determining the specificity of protein–DNA interactions," *Nature Rev. Genet.*, vol. 11, no. 11, pp. 751–760, 2010.

[19] M. Djordjevic, A. M. Sengupta, and B. I. Shraiman, "A biophysical approach to transcription factor binding site discovery," *Genome Res.*, vol. 13, no. 11, pp. 2381–2390, 2003.

[20] B. Hooghe, S. Broos, F. Van Roy, and P. De Bleser, "A flexible integrative approach based on random forest improves prediction of transcription factor binding sites," *Nucleic Acids Res.*, vol. 40, no. 14, pp. e106–e106, 2012.

[21] Y. Xiao and M. R. Segal, "Identification of yeast transcriptional regulation networks using multivariate random forests," *PLoS Comput. Biol.*, vol. 5, no. 6, 2009, Art. no. e1000414.

[22] T. Zhou et al., "Quantitative modeling of transcription factor binding specificities using DNA shape," *Proc. Nat. Acad. Sci.*, vol. 112, no. 15, pp. 4654–4659, 2015.

[23] Y. Zeng, M. Gong, M. Lin, D. Gao, and Y. Zhang, "A review about transcription factor binding sites prediction based on deep learning," *IEEE Access*, vol. 8, pp. 219256–219274, 2020.

[24] B. Alipanahi, A. Delong, M. T. Weirauch, and B. J. Frey, "Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning," *Nature Biotechnol.*, vol. 33, no. 8, pp. 831–838, 2015.

[25] D. Quang and X. Xie, "DanQ: A hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences," *Nucleic Acids Res.*, vol. 44, no. 11, 2016, Art. no. e107.

[26] G. Liu and J. Guo, "Bidirectional LSTM with attention mechanism and convolutional layer for text classification," *Neurocomputing*, vol. 337, pp. 325–338, 2019.

[27] Y. Yu, P. Ding, H. Gao, G. Liu, F. Zhang, and B. Yu, "Cooperation of local features and global representations by a dual-branch network for transcription factor binding sites prediction," *Brief. Bioinf.*, vol. 24, no. 2, 2023, Art. no. bbad036.

[28] T. - P. Chiu, F. Comoglio, T. Zhou, L. Yang, R. Paro, and R. Rohs, "DNAshapeR: An R/Bioconductor package for DNA shape prediction and feature encoding," *Bioinformatics*, vol. 32, no. 8, pp. 1211–1213, 2016.

[29] J. Li, J. M. Sagendorf, T. - P. Chiu, M. Pasi, A. Perez, and R. Rohs, "Expanding the repertoire of DNA shape features for genome-scale studies of transcription factor binding," *Nucleic Acids Res.*, vol. 45, no. 22, pp. 12877–12887, 2017.

[30] R. Rohs, S. M. West, A. Sosinsky, P. Liu, R. S. Mann, and B. Honig, "The role of DNA shape in protein–DNA recognition," *Nature*, vol. 461, no. 7268, pp. 1248–1253, 2009.

[31] T. Zhou et al., "DNAshape: A method for the high-throughput prediction of DNA structural features on a genomic scale," *Nucleic Acids Res.*, vol. 41, no. W1, pp. W56–W62, 2013.

[32] Q. Zhang, Z. Shen, and D.-S. Huang, "Predicting in-vitro transcription factor binding sites using DNA sequence+ shape," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 18, no. 2, pp. 667–676, Mar./Apr. 2019.

[33] S. Wang et al., "Predicting transcription factor binding sites using DNA shape features based on shared hybrid deep learning architecture," *Mol. Ther.-Nucleic Acids*, vol. 24, pp. 154–163, 2021.

[34] Y. Zhang et al., "A novel convolution attention model for predicting transcription factor binding sites by combination of sequence and shape," *Brief. Bioinf.*, vol. 23, no. 1, 2022, Art. no. bbab525.

[35] P. Ding, Y. Wang, X. Zhang, X. Gao, G. Liu, and B. Yu, "DeepSTF: Predicting transcription factor binding sites by interpretable deep neural networks combining sequence and shape," *Brief. Bioinf.*, vol. 31, 2023, Art. no. bbad231.

[36] F. A. Acheampong, H. Nunoo-Mensah, and W. Chen, "Transformer models for text-based emotion detection: A review of BERT-based approaches," *Artif. Intell. Rev.*, vol. 9, pp. 1–41, 2021.

[37] L. A. Hendricks, J. Mellor, R. Schneider, J. - B. Alayrac, and A. Nematzadeh, "Decoupling the role of data, attention, and losses in multimodal transformers," *Trans. Assoc. Comput. Linguistics*, vol. 9, pp. 570–585, 2021.

[38] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Comput. Surv.*, vol. 54, no. 10s, pp. 1–41, 2022.

[39] W. Yan, B. Zhang, M. Zuo, Q. Zhang, H. Wang, and D. Mao, "AttentionSplice: An interpretable multi-head self-attention based hybrid deep learning model in splice site prediction," *Chin. J. Electron.*, vol. 31, no. 5, pp. 870–887, 2022.

[40] J. Lee-Thorp, J. Ainslie, I. Eckstein, and S. Ontanon, "Fnet: Mixing tokens with Fourier transforms," 2021, *arXiv:2105.03824*.

[41] R. Y. Rubinstein and D. P. Kroese, *Simulation and the Monte Carlo Method*. Hoboken, NJ, USA: Wiley, 2016.

[42] A. K. Dubey and V. Jain, "Comparative study of convolution neural network's relu and leaky-relu activation functions," in *Proc. Appl. Comput., Automat. Wireless Syst. Elect. Eng.*, 2019, pp. 873–880.

[43] H. Chen, H. Lin, and M. Yao, "Improving the efficiency of encoder-decoder architecture for pixel-level crack detection," *IEEE Access*, vol. 7, pp. 186657–186670, 2019.

[44] Y. Ho and S. Wookey, "The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling," *IEEE Access*, vol. 8, pp. 4806–4813, 2019.

[45] B. Cortiñas-Lorenzo and F. Pérez-González, "Adam and the ants: On the influence of the optimization algorithm on the detectability of DNN watermarks," *Entropy*, vol. 22, no. 12, 2020, Art. no. 1379.