






MLSNet: a deep learning model for predicting transcription factor binding sites

Yuchuan Zhang ^{1,†}, Zhikang Wang ^{2,†}, Fang Ge ^{3,†}, Xiaoyu Wang², Yiwen Zhang⁴, Shanshan Li⁴, Yuming Guo⁴, Jiangning Song ^{2,5,*}, Dong-Jun Yu ^{1,*}

¹School of Computer Science and Engineering, Nanjing University of Science and Technology, 200 Xiaolingwei, Nanjing 210094, China

²Biomedicine Discovery Institute and Department of Biochemistry and Molecular Biology, Monash University, Wellington Rd, Clayton, Melbourne, VIC 3800, Australia

³State Key Laboratory of Organic Electronics and Information Displays & Institute of Advanced Materials (IAM), Nanjing University of Posts & Telecommunications, 9 Wenyuan, Nanjing, 210023, China

⁴Climate, Air Quality Research Unit, School of Public Health and Preventive Medicine, Monash University, 553 St Kilda Road, Melbourne, VIC 3004, Australia

⁵Monash Data Futures Institute, Monash University, Wellington Rd, Clayton, Melbourne, VIC 3800, Australia

*Corresponding authors. Dong-Jun Yu, School of Computer Science and Engineering, Nanjing University of Science and Technology, 200 Xiaolingwei, Nanjing 210094, China. E-mail: njyudj@njust.edu.cn; Jiangning Song, Monash Biomedicine Discovery Institute and Department of Biochemistry and Molecular Biology, Monash University, Wellington Rd, Clayton, Melbourne, VIC 3800, Australia. E-mail: jiangning.song@monash.edu

[†]The first three authors contributed equally to this work and should be considered cofirst authors.

Abstract

Accurate prediction of transcription factor binding sites (TFBSs) is essential for understanding gene regulation mechanisms and the etiology of diseases. Despite numerous advances in deep learning for predicting TFBSs, their performance can still be enhanced. In this study, we propose MLSNet, a novel deep learning architecture designed specifically to predict TFBSs. MLSNet innovatively integrates multisize convolutional fusion with long short-term memory (LSTM) networks to effectively capture DNA-sparse higher-order sequence features. Further, MLSNet incorporates super token attention and Bi-LSTM to systematically extract and integrate higher-order DNA shape features. Experimental results on 165 ChIP-seq (chromatin immunoprecipitation followed by sequencing) datasets indicate that MLSNet consistently outperforms several state-of-the-art algorithms in the prediction of TFBSs. Specifically, MLSNet reports average metrics: 0.8306 for ACC, 0.8992 for AUROC, and 0.9035 for AUPRC, surpassing the second-best methods by 1.82%, 1.68%, and 1.54%, respectively. This research delineates the effectiveness of combining multi-size convolutional layers with LSTM and DNA shape-based features in enhancing predictive accuracy. Moreover, this study comprehensively assesses the variability in model performance across different cell lines and transcription factors. The source code of MLSNet is available at <https://github.com/minghaidea/MLSNet>.

Keywords: transcription factor binding sites; multisize convolutional fusion; super token attention and Bi-LSTM; DNA sequence; DNA shape

Introduction

Transcription factors (TFs) are proteins that bind to specific DNA sequences, regulating gene expression. These binding sites, known as TF binding sites (TFBSs), typically range from 5 to 20 bp in length [1–3]. Accurate TFBS identification is vital for understanding gene regulation and disease mechanisms [4]. Genomic variations in TFBSs have been linked to diseases such as cancer, autoimmune disorders, and neurological diseases [5]. High-throughput sequencing technologies like ChIP-seq [6] provide extensive experimental data on TF–DNA interactions. ChIP-seq, which combines chromatin immunoprecipitation with high-throughput sequencing, efficiently identifies TFBSs. However, due to the difficult-to-obtain reagents and materials, such as antibodies against certain TFs [7, 8], computational prediction of TFBSs from sequence data has become a preferred method.

In the past decades, a variety of computational methods have been developed to predict TFBSs. Machine learning-based

approaches [9, 10] have been commonly used for predicting TF recognition binding sites. These approaches include methods based on Support Vector Machines (SVMs) [11], Random Forest models [9], and Hidden Markov Models (HMMs) [12], among others. Despite their success in TFBSs prediction tasks, these traditional methods often depend on handcrafted features and may not fully exploit the information in the raw input sequences.

Recently, deep learning methods have emerged as powerful tools in bioinformatics, offering new possibilities for TFBS prediction. Many methods have been proposed to predict the TFBSs, such as Expecto [13], Sei [14], and Enformer [15]. These methods focus on genetic variant effects with additional identification of some specific TFBSs. They are more as studies of noncoding variant effect prediction and analyses of gene regulatory processes, lacking specialized mechanisms for specific TF recognition. To addition to these methods, deep learning-based methods like DeepBind [16], DeepSEA [17], and DanQ [18] utilized the power of neural networks to learn complex patterns, outperforming

Received: July 22, 2024. Revised: September 5, 2024. Accepted: September 16, 2024

© The Author(s) 2024. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited.

For commercial re-use, please contact journals.permissions@oup.com

methods based on handcrafted features [19–26]. These methods not only capture local features of DNA sequences but also manage to some extent distant dependencies. This represents a significant advancement over traditional methods, paving the way for more accurate and comprehensive TFBS predictions. Recent studies have underscored the importance of DNA shape in predicting TFBSs. Consequently, new methods such as DLBSS [27], CRPTS [28], D-SSCA [29], and DeepSTF [30] have been developed to improve TFBSs prediction by integrating both sequence and shape information. Compared with previous methods, these methods not only utilize the sequence-only features but also incorporate DNA shape information. This allows models to comprehensively consider the characteristics of TF-DNA binding, thereby enhancing prediction accuracy and interpretability. For instance, DeepSTF combines convolutional neural networks (CNNs), Bi-LSTM, and an enhanced Transformer [30]. By processing DNA sequences and shape contours jointly, it adaptively extracts high-order features, improving TFBSs prediction accuracy. Furthermore, in-depth analysis of shape contours has demonstrated the positive influence of shape [27–30] on TF-DNA binding, offering a fresh perspective for TFBSs prediction.

However, the deep learning methods described above also have some limitations, such as only capturing local sequence features, failing to extract more information from sparse sequence features, and overemphasizing shape features while neglecting sequence features. To address these issues, we propose a novel deep learning model, MLSNet. This model employs a multisize convolution fusion framework with LSTM [31] to effectively capture higher-order features from sparse sequence features and obtain global sequence features. Additionally, we process DNA shape data as supplementary features and incorporate them into the model to enhance prediction accuracy.

Materials and methods

Construction of the benchmark dataset

DNA sequence

We evaluated our model’s performance using the dataset previously employed by DeepSTF. Notably, this dataset comprises 165 ChIP-seq sub-datasets curated by Zeng et al. [32] from the larger pool of 690 ENCODE ChIP-seq datasets. The selection criteria included diversity across different cell lines. Detailed dataset descriptions are available in Supplementary Table S1 available online at <http://bib.oxfordjournals.org/>. Each dataset consists of training and testing subsets, comprising multiple positive and negative instances. DNA sequences are represented as fixed-length strings of 101 bp, with binary labels (0 and 1) indicating the absence or presence of TFBSs, respectively. The 165 ChIP-seq dataset is extensively utilized for studying the prediction of DNA-TF binding sites [29, 30]. Comprehensive details about the datasets are presented in Tables S1–S3 available online at <http://bib.oxfordjournals.org/>. Access to all datasets is provided at http://cnn.csail.mit.edu/motif_discovery/.

DNA shape

In light of model interpretability and data availability, we sought to optimize performance using the simplest feasible model. Through extensive analysis of DNA structures, we identified five critical shape features: helical twist (HelT), minor groove width (MGW), propeller twist (ProT), rolling (Roll), and electrostatic potential (EP). These features, which encapsulate essential information on DNA’s structural and functional properties, are crucial for enhancing TFBSs identification. By examining these shape

features, we gain a deeper understanding of TF–DNA interaction mechanisms. DNashapeR, a high-performance R/BioConductor package, can encode DNA shape features rapidly and effectively. Full documentation and further details are available at <http://www.bioconductor.org/> and in Table S4 available online at <http://bib.oxfordjournals.org/>.

Model description

The overall architecture of MLSNet is illustrated in Fig. 1.

Data preprocessing

DNA sequence

The DNA sequence was encoded employing both 3-mer encoding and one-hot encoding technologies. An initial 101-nucleotide sequence was segmented into 99 overlapping 3-mers. 3-mer encoding, a common k -mer encoding method, has been proven to be highly effective [33], as detailed in Text S1 available online at <http://bib.oxfordjournals.org/>. For instance, the nucleotide sequence ATGCCG is transformed into overlapping 3-mers: ATG, TGC, GCC, and CCG. Subsequent one-hot encoding converted A, T, C, and G into the vectors [1,0,0,0], [0,1,0,0], [0,0,1,0], and [0,0,0,1], respectively. Finally, this process yields a feature matrix with dimensions $1 \times 12 \times 99$, represented as follows:

$$S_1 = [M_1, M_2, M_i, \dots, M_{98}, M_{99}] \quad (1)$$

where M_i refers to the i -th 3-mer nucleotide fragment.

DNA shape

The DNashapeR toolkit processes DNA sequences to predict associated structural features: Helt, MGW, ProT, Roll, and EP. These features are concatenated to form matrix S_2 , serving as the input data for DNA shape analysis in $1 \times 5 \times 101$ format, defined as follows:

$$S_2 = [N_1, N_2, N_i, \dots, N_{100}, N_{101}] \quad (2)$$

where N_i represents the set of five spatial structural features predicted by the DNashapeR toolkit for the i -th nucleotide.

Sequence data processing

Utilizing 3-mer encoding can enhance the representational breadth of the input sequence matrix. However, a limitation of one-hot encoding is its production of a sparse matrix comprised solely of zeros and ones, which may hinder the performance of standard convolutional networks. To overcome this, we adopted a multiscale strategy (3×3 , 5×5 , and 7×7 kernels) with incremental channel scaling from 32 to 128 and mixed-channel integration, to enhance feature extraction across multiple scales. Padding was adjusted to maintain the original input’s dimensionality throughout the process and outputs were integrated along the channel dimension, followed by max-pooling to improve computational efficiency and network robustness. Besides, a dropout layer was also implemented to prevent overfitting, represented as follows:

$$M = \text{Concat}(\text{Conv}_1(S_1, W_1, b_1), \text{Conv}_2(S_1, W_2, b_2), \text{Conv}_3(S_1, W_3, b_3)) \quad (3)$$

$$C_1 = \text{Dropout}(\text{MaxPooling}(M)) \quad (4)$$

where $W_i, i = 1, 2, 3$ represents the weight matrices and $b_i, i = 1, 2, 3$ denotes the biases of the convolutional layers. The convolution operations, denoted as $\text{Conv}_1(*)$, $\text{Conv}_2(*)$, and $\text{Conv}_3(*)$,

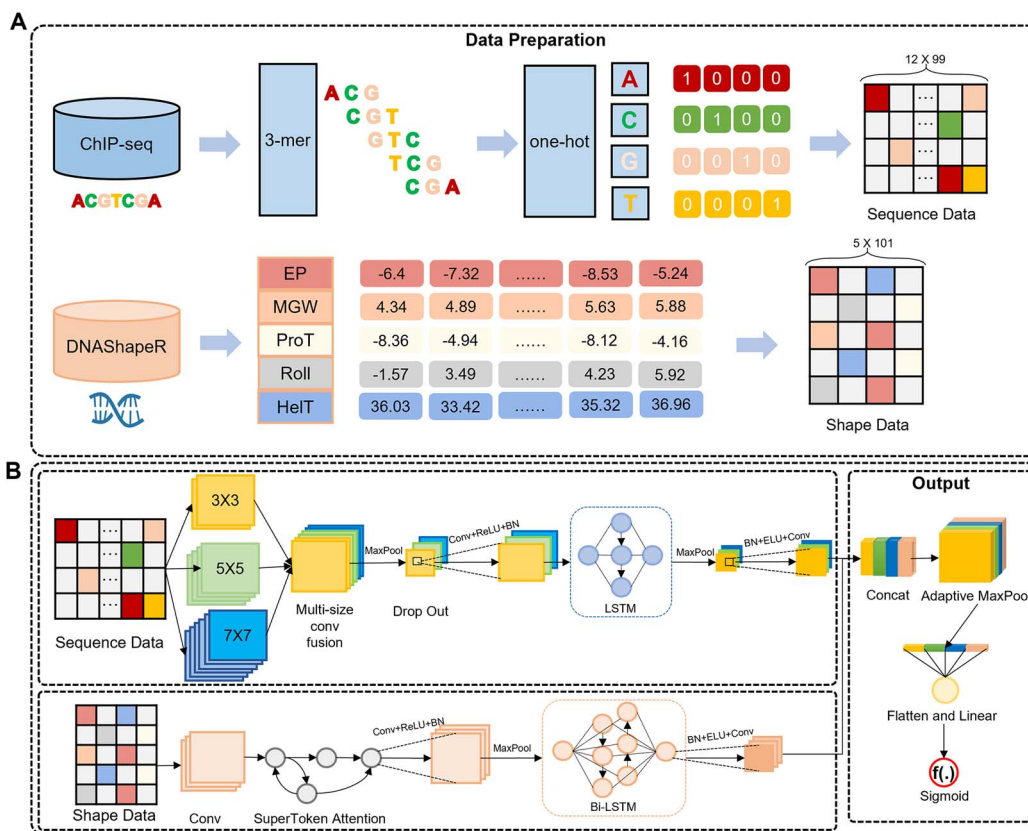


Figure 1. Overview of the MLSNet workflow. (A) Data preprocessing: This part involves the preparation of sequence data and shape data. (B) Deep learning framework: It consists of: sequence data processing flow (integrating multiscale convolutional fusion with LSTM), shape data processing flow (employing Super Token Attention and Bi-LSTM), and the output module. Note: “conv” means “convolution”.

employ kernels of sizes 3×3 , 5×5 , and 7×7 respectively, expanding the channel count to 32, 64, and 128. *Concat* (*) fuses channel outputs from these convolutions. *Maxpooling* (*) and *Dropout* (*) refer to the local max-pooling and dropout operations (to mitigate overfitting). The detailed architecture of the multiscale fusion channel convolution module is described in Table S5.

Subsequently, an LSTM layer is utilized to capture long-range dependencies within the DNA sequence, to further extract deeper hidden features, represented as follows:

$$L_1 = LSTM(BN(ReLU(Conv_{l_1}(C_1, W_{l_1}, b_{l_1})))) \quad (5)$$

where $ReLU(*)$ is an activation function, $BN(*)$ denotes batch normalization, and $LSTM(*)$ is LSTM, adept at capturing long-range dependencies in sequences.

Finally, we adopted a symmetric processing structure to obtain rich feature information. Specifically, this operation can be represented as:

$$P_1 = Conv(ELU(BN(MaxPooling(L_1)))) \quad (6)$$

where $ELU(*)$ is a variant of the ReLU operation, designed to mitigate the “dying ReLU” issue and provide enhanced activation upon subsequent use. The *Conv* (*) operation performs convolution on all features to extract the most abstract, high-level hidden feature information.

Shape data processing

In our approach to DNA shape data processing, we initially employ a tool known as SuperTokenViT [34]. This tool is instrumental in

discerning the global spatial dependencies inherent in the DNA shape. Once these dependencies have been successfully captured, our next step involves the extraction of high-level hidden features. This is achieved through the application of convolutional operations. The specific procedures involved in these operations are detailed in the following sections:

$$C_2 = MaxPooling(BN(ReLU(Conv(SuperTokenAttention(Conv(R_1)))))) \quad (7)$$

Our methodology begins with the application of the Super Token Attention module, a concept derived from STViT [34]. This module cleverly deconstructs conventional global attention into a series of multiplications involving a sparse association map and low-dimensional attention, as detailed in Text S2 available online at <http://bib.oxfordjournals.org/>. This innovative approach allows for the efficient capture of global dependencies. Huang et al. [34] have demonstrated the efficacy of this module in achieving remarkable results on various image datasets. In our research, we have adapted this concept for DNA shape data processing, with the anticipation of obtaining similarly promising results. Our experimental findings corroborate this expectation, demonstrating a noticeable improvement in performance.

For the DNA shape data processing, we adopt an approach akin to that used for sequence data. The key distinction lies in our choice of Bi-LSTM over the unidirectional LSTM. While LSTM has proven effective in handling sequence data, we contend that Bi-LSTM is more apt for dealing with DNA shape data. Bi-LSTM can capture the information in both directions. DNA has two strands in the shape; thus, using Bi-LSTM can better capture the features

Table 1. Hyperparameters of MLSNet and the corresponding search space

Calibration parameters	Search space	Sampling	Final settings
Learning rate	–	–	Auto-adjustment
Optimizer	Adam, RMSProp, SGD	Evaluate all	Adam
Kernel numbers	32, 64, 128, 256	Evaluate all	128
Batch size	32, 64, 128	Evaluate all	64
Dropout ratio	0.1, 0.2, 0.3, 0.4	Evaluate all	0.1, 0.2

of DNA shape in both directions. The specifics of this operation are as follows:

$$P_2 = \text{Conv}(\text{ELU}(\text{BN}(\text{Bi-LSTM}(C_2)))) \quad (8)$$

The processing workflow for DNA shape can be found in Fig. 1B.

Output processing

In this phase, we incorporate the DNA shape data as auxiliary features into the sequence data for processing and subsequent prediction. The primary methodology involves the fusion of the previously obtained higher-order sequence features, denoted as P_1 , with the higher-order shape features, denoted as P_2 . Specifically, the detailed operations involved in this process are outlined in the subsequent sections:

$$O = \text{Sigmoid}(\text{Linear}(\text{Flatten}(\text{AdaptiveMaxPool}(\text{Concat}(P_1, P_2)))))) \quad (9)$$

The comprehensive framework of MLSNet is illustrated in Fig. 1. The details of MLSNet are illustrated in Tables S5–S7 available online at <http://bib.oxfordjournals.org/>.

Model implementation

The proposed MLSNet model is based on PyTorch and follows the same training/testing process as the baseline. This approach ensures the reliability of our experimental comparisons. During the training phase, we employ the binary cross-entropy loss function (BCELoss) and the Adam optimizer [35], as detailed in Text S3 available online at <http://bib.oxfordjournals.org/>. To maintain the independence of the test set, we split each training set into training and validation subsets at 80:20, ensuring that the test set remained untouched during hyperparameter tuning. We assessed the model’s loss on the validation set to monitor its performance, which facilitates determining if adjustments to hyperparameters or modifications in training strategies are required. Ultimately, the model was evaluated using the test set. This methodology safeguards the independence of each test set, thereby facilitating a more precise evaluation of the model’s generalization capability. The batch size is set at 64, and the exact parameter settings of the model are specified in Table 1. The impact of various convolutional kernel sizes is shown in Text S4 available online at <http://bib.oxfordjournals.org/>. For each training and validation set, the model undergoes training for 15 epochs, and its performance is evaluated using the test set with a batch size of 1.

Evaluation metrics

This study focuses on TFBS identification based on DNA sequence, which is a binary classification problem. Therefore, model performance is evaluated using accuracy (ACC), receiver operating characteristic area under the curve (ROC-AUC), and precision–recall area under the curve (PR-AUC) [30]. ACC quantifies the overall prediction accuracy, while ROC-AUC and PR-AUC are employed to address potential imbalances in the data. ROC-AUC and PR-AUC are suitable for evaluating imbalanced datasets, and PR-AUC

measures the trade-off between precision and recall. Comprehensive details of these evaluation metrics are provided in Text S5 available online at <http://bib.oxfordjournals.org/>.

Baseline methods

Our model was compared with other state-of-the-art deep learning methods, including those leveraging DNA sequences alone and those integrating DNA sequences and shapes together. All models were trained and evaluated under the same conditions for fairness.

Among the selected baselines, DeepSTF is particularly noteworthy. It significantly enhances TFBSs prediction accuracy by incorporating an advanced Transformer coupled with a Bi-LSTM module specifically designed for DNA shape processing [30]. Other baselines include DeepBind [16], DanQ [18], DLBSS [27], CRPTS [28], D-SSCA [29], and DeepSTF [30]. DeepBind and DanQ use only DNA sequences for prediction, with DanQ adding a Bi-LSTM to DeepBind’s approach. DLBSS, CRPTS, and D-SSCA also use DNA shape data, processing features with Siamese convolution, convolutional recursive neural network, and multi-layer perception (MLP), respectively. Among these, DeepSTF demonstrates the most outstanding performance, making it our primary comparison target. The details of baseline methods are displayed in Text S6 available online at <http://bib.oxfordjournals.org/>.

Results and discussion

Multisize convolutional fusion with long short-term memory and shape data improving prediction performance

This study constructed two variant models to explore the improvement of the model studied in this paper regarding the multisize convolutional fusion with LSTM and shape data.

MLSNet-1

In our study, we developed MLSNet-1 to assess the efficacy of multisize convolutional fusion combined with LSTM for DNA sequence data. This model incorporates multisize convolutional fusion with LSTM and adjusts the remaining parameters for optimal performance. The model was trained on the 165 ChIP-seq datasets, and the results are presented in Table 2. Compared to the MLSNet-1 model, MLSNet demonstrated improvements in ACC, ROC-AUC, and PR-AUC by 1.67%, 1.54%, and 1.43%, respectively. The model’s performance on each dataset is visualized in Fig. 2C, showing that MLSNet consistently outperforms MLSNet-1 across almost all datasets, yielding more stable results. By employing multisize convolution fusion, we were able to capture DNA sequence information across various kernel sizes. The results of the multisize convolution are then fused along the channel dimension, mitigating some of the limitations of one-hot encoded DNA sequences, such as performance deficiencies or the risk of overfitting in models like LSTM when processing sparse

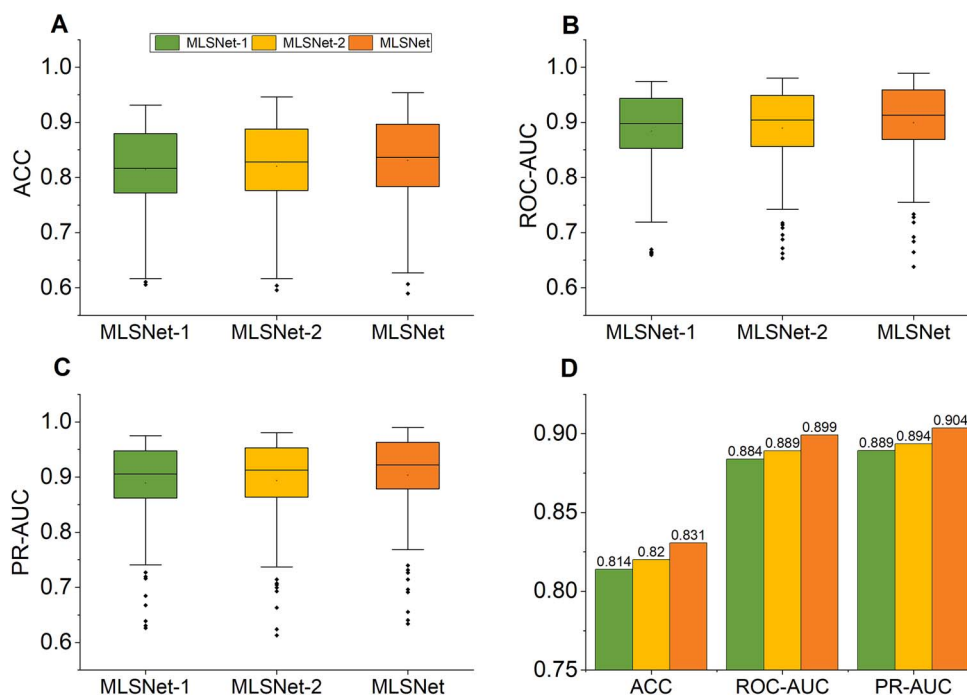


Figure 2. Performance comparison of MLSNet and variant models on 165 ChIP-seq datasets. (MLSNet-1: Without multisize convolutional fusion with LSTM; MLSNet-2: Without supplemental shape data). (A) ACC: This part involves the ACC comparison between MLSNet and variant models. (B) ROC-AUC: This part involves the ROC-AUC comparison between MLSNet and variant models. (C) PR-AUC: This part involves the PR-AUC comparison between MLSNet and variant models. (D) Average results: This part involves the average ACC, ROC-AUC, and PR-AUC comparison between MLSNet and variant models.

Table 2. The average ACC, ROC-AUC, and PR-AUC of the MLSNet model and its variants (MLSNet-1: without multisize convolutional fusion with LSTM; MLSNet-2: without supplemental shape data) on the test set of 165 ChIP-seq datasets

Method	ACC	ROC-AUC	PR-AUC
MLSNet-1	0.814	0.884	0.889
MLSNet-2	0.820	0.889	0.894
MLSNet	0.831	0.899	0.904

matrices. This LSTM processing also enhances the model’s robustness. The MLSNet-1 results underscore the necessity of multisize convolution fusion and LSTM in our experiment. These techniques significantly enhance the original model’s performance, enabling it to more effectively capture DNA sequence features and predict TFBSs.

MLSNet-2

To evaluate the contribution of DNA shape, we developed the MLSNet-2 model, which exclusively used sequence data while excluding the input of shape data. The relevant results and comparisons are displayed in Fig. 2 and Table 2. In our experiment, MLSNet-2 demonstrated lower performance than MLSNet in ACC, ROC-AUC, and PR-AUC by 1.05%, 1.02%, and 0.99%, respectively. This suggests that the inclusion of DNA shape data as supplementary information enhances MLSNet’s performance. We opted for STViT to better capture global latent features from DNA shape data. Shape data effectively capture the spatial attributes of DNA, enabling the model to acquire additional spatial features of DNA alongside sequence features. The chosen DNA spatial features have been corroborated by numerous deep learning methods [29, 30]. The final results confirm that the additional input of DNA shape data can further enhance MLSNet’s performance.

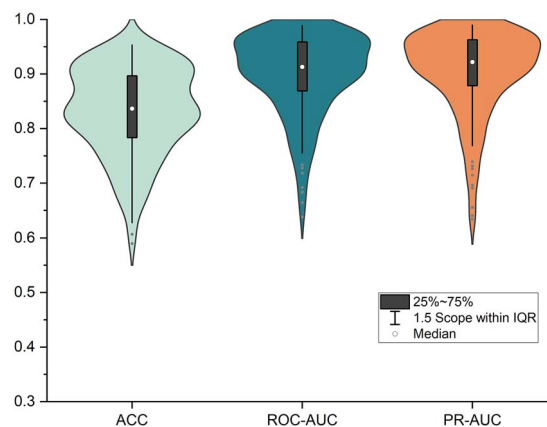


Figure 3. Distribution of MLSNet results on ACC, ROC-AUC, and PR-AUC on 165 ChIP-seq datasets.

We have also visualized the performance of MLSNet on the 165 ChIP-seq datasets. Please refer to Fig. 3 for details.

It can be observed that our model achieves excellent results on the vast majority of datasets. Compared to ACC, MLSNet demonstrates superior performance in ROC-AUC and PR-AUC. It can be seen that the vast majority of the datasets of MLSNet on ROC-AUC and PR-AUC are concentrated above 0.85, while ACC is mostly clustered above 0.75. Nevertheless, we have achieved noteworthy results with our MLSNet.

Results on different cell lines and transcription factors

In order to comprehensively assess the performance of MLSNet across various cell lines and TFs, we meticulously segmented and analyzed the results. The collection of 165 ChIP-seq datasets spans 32 cell lines and 29 TFs. The specific datasets related to

Table 3. The average ACC, ROC-AUC, and PR-AUC of the MLSNet model and the second-best model on several cell lines and TFs

Name		ACC		ROC-AUC		PR-AUC	
Cell lines	TFs	DeepSTF	MLSNet	DeepSTF	MLSNet	DeepSTF	MLSNet
Hepg2		0.821	0.839(+1.82%)	0.891	0.908(+1.73%)	0.896	0.911(+1.57%)
K562		0.806	0.825(+1.96%)	0.875	0.893(+1.80%)	0.881	0.897(+1.62%)
H1hesc		0.804	0.821(+1.67%)	0.873	0.889(+1.60%)	0.883	0.896(+1.25%)
Helas3		0.818	0.835(+1.72%)	0.885	0.902(+1.73%)	0.889	0.905(+1.63%)
GM12878		0.795	0.813(+1.80%)	0.868	0.886(+1.85%)	0.874	0.891(+1.74%)
	COREST	0.752	0.780(+2.74%)	0.827	0.857(+3.09%)	0.827	0.864(+3.72%)
	CTCF	0.912	0.929(+1.64%)	0.966	0.975(+0.89%)	0.971	0.979(+0.75%)
	EZH2	0.664	0.683(+1.92%)	0.722	0.742(+2.05%)	0.711	0.724(+1.27%)
	NFKB	0.784	0.805(+2.14%)	0.867	0.888(+2.15%)	0.879	0.901(+2.15%)

these cell lines and TFs are enumerated in Tables S1 and S2 available online at <http://bib.oxfordjournals.org/>. Unfortunately, most cell lines are represented by only one associated dataset when viewed from a cell line perspective. Consequently, we focused our analysis on the top five cell lines with the most datasets. The results concerning to all cell lines and TFs are catalogued in Tables S8 and S9 available online at <http://bib.oxfordjournals.org/>.

Table 3 presents the average prediction results of the MLSNet and DeepSTF models across different cell lines and TFs. The table illustrates the improvement of MLSNet compared to DeepSTF in each case. As indicated by the table results, MLSNet consistently DeepSTF for each cell line and TF selected.

Cross cell lines

Based on the number of datasets available in different cell lines, we selected Hepg2, K562, H1hesc, Helas3, and GM12878. These five cell lines contain 26, 32, 16, 13, and 17 ChIP-seq datasets, respectively, making them the most represented cell lines in terms of dataset count among the 32 cell lines. This selection strengthens the credibility of our comparative analysis. We aggregated the average results obtained from testing the models trained on all test sets within these cell lines. Evaluation metrics include ACC, ROC-AUC, and PR-AUC. The results of the analysis and comparison are depicted in Fig. 4. It can be observed that, except for MLSNet and DeepSTF, the results of the other models across these five cell lines are relatively similar, with the best performance observed in the Hepg2 cell line and the worst in the GM12878 dataset. Additionally, MLSNet consistently outperforms the competing models in all cell lines, with a significantly superior margin. These results demonstrate the strong generalization and robustness of our MLSNet model.

Cross-transcription factors

In contrast to the uneven distribution of datasets among cell lines, the number of datasets for different TFs is relatively balanced. Therefore, we selected several TFs for comparison:

- REST Corepressor (COREST): Contains two datasets, and all models exhibit notable deficiency in predicting this TF. Studying this TF can help us explore the robustness of the models. Given that this TF has only two datasets, both from K562 and HepG2 cell lines, we found that the models perform well on other TFs within these lines. However, the predictions on the two datasets differ significantly by over ten percentage points. This discrepancy, coupled with the limited data, suggests that the poor performance in predicting COREST is likely due to the dataset scarcity. The extreme lack of data also raises doubts about the value of further biological

exploration. Thus, this paper uses this TF as a case study to examine how different models perform with limited data.

- CCCTC-binding factor (CTCF): The CTCF gene, affiliated with the BORIS + CTCF group, produces a transcriptional controller endowed with 11 consistent zinc finger (ZF) motifs. It binds various DNA sequences and proteins using different ZF configurations, thereby modulating gene activity by either initiating or inhibiting transcription linked to histone acetyltransferases and deacetylases [36]. Furthermore, CTCF adjusts genomic interactions by blocking communication between enhancers and promoters, influencing the expression of imprinted genes [37]. Alterations in this gene are associated with a range of cancers, including invasive breast cancer, prostate cancer, and Wilms tumor [38]. Herein, there are 20 CTCF ChIP-seq datasets, and all models demonstrate superior performance in predicting this TF's binding sites. This aids in exploring the optimal performance of deep learning methods in predicting specific TFs.
- Enhancer of Zeste Homolog 2 (EZH2): EZH2, implicated in Weaver syndrome and lymphomas, is associated with the activation of RNA polymerase I promoters and PIP3-mediated AKT signaling. Its Gene Ontology(GO) annotations include sequence-specific DNA binding and chromatin binding, highlighting its critical roles in epigenetic regulation and cellular functions. These functions are crucial for transcriptional repression and cellular proliferation, as explored in depth by Kim [39]. In this study, there are six EZH2 ChIP-seq datasets, and almost all models obtain suboptimal results in predicting this TF's binding sites. The datasets of this TF have relatively little data. Investigating this TF helps us explore the bottleneck of data volume or specific TFs in deep learning methods for predicting TFBSs.
- Nuclear Factor kappa-B (NFKB): Contains six datasets, and both the data volume and prediction results are comparable to the average results of the models across all datasets. Studying this TF can better explore the generalization and robustness of the models.

Among the four TFs examined, CTCF demonstrates superior predictive performance, whereas EZH2 shows poorer outcomes. The analyses for these findings are as follows:

- CTCF: The effective prediction of CTCF binding sites is underpinned by several key factors. (i) High sequence conservation: CTCF binding sites are highly conserved evolutionarily, providing robust signals for sequence-based prediction algorithms [37]. (ii) Abundant experimental data: Comprehensive ChIP-seq datasets furnish extensive samples that enhance the precision and applicability of CTCF prediction models

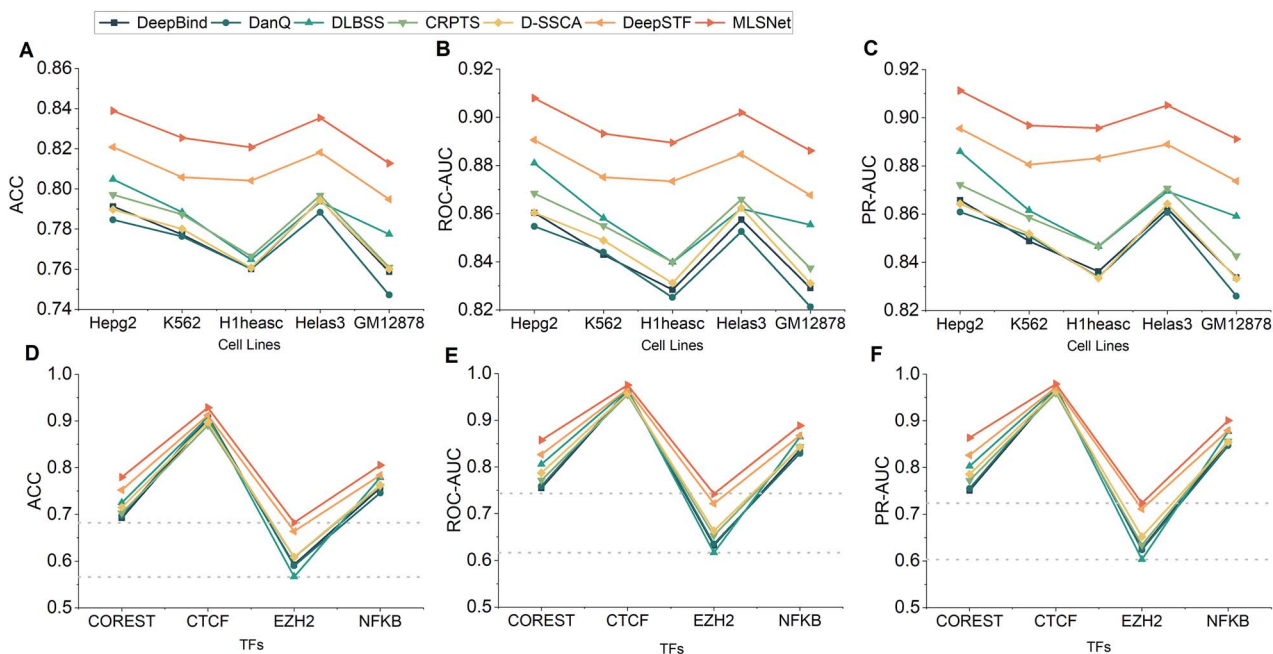


Figure 4. Performance comparison of MLSNet with competing models on selected cell lines and TFs. (A) ACC of cell lines: This part involves the ACC comparison between MLSNet and competing models on selected cell lines. (B) ROC-AUC of cell lines: This part involves the ROC-AUC comparison between MLSNet and competing models on selected cell lines. (C) PR-AUC of cell lines: This part involves the PR-AUC comparison between MLSNet and competing models on selected cell lines. (D) ACC of TFs: This part involves the ACC comparison between MLSNet and competing models on selected TFs. (E) ROC-AUC of TFs: This part involves the ROC-AUC comparison between MLSNet and competing models on selected TFs. (F) PR-AUC of TFs: This part involves the PR-AUC comparison between MLSNet and competing models on selected TFs.

[39]. (iii) Central regulatory role: CTCF is integral to regulating genomic structure and gene expression, contributing to the stability and recognizability of its binding sites [40].

- EZH2: In predicting TFBSs, models frequently underperform on the EZH2 gene due to its complex involvement in the Polycomb Repressive Complex 2, which governs a broad spectrum of epigenetic activities including histone methylation. This complexity leads to unique regulatory patterns that are challenging to decode using standard sequence-based features, underscoring the distinctive epigenetic roles of EZH2 [41]. Additionally, the evolutionary variability of some EZH2-associated TFBSs compromises the accuracy of prediction methods reliant on sequence homology, thereby reducing the efficacy of conventional prediction algorithms [42].

Our model stands out in predicting four crucial TFs, surpassing other models in accuracy. Refer to Fig. 4 for a visual comparison. Notably, COREST and EZH2 consistently exhibit lower performance across all models, primarily due to limited data, especially evident with EZH2, which has smaller datasets. Thanks to its unique combination of multisize convolutional and LSTM layers, MLSNet can achieve a good performance on the limited datasets and extract valuable insights from sparse data. On the other hand, when it comes to CTCF, our model slightly surpasses others, demonstrating deep learning’s strength with large datasets. However, MLSNet truly excels, excelling across datasets of all sizes. This versatility is crucial in bioinformatics, where data availability varies. Deep learning remains promising, especially with ample data, as seen with NFKB. These results highlight MLSNet’s potential in bioinformatics tasks. From a computational perspective, when data volume is sufficient, deep learning methods for predicting TFBSs still hold considerable potential.

Figure 5 illustrates the performance of MLSNet and other high-performing models on individual cell lines and TFs on the 165

Table 4. The average ACC, ROC-AUC, and PR-AUC of MLSNet model and several advanced methods on 165 ChIP-seq dataset test sets

Method	ACC	ROC-AUC	PR-AUC
DeepBind	0.784	0.851	0.857
DanQ	0.779	0.848	0.854
DLBSS	0.795	0.867	0.872
CRPTS	0.789	0.859	0.864
D-SSCA	0.784	0.854	0.857
DeepSTF	0.812	0.882	0.888
MLSNet	0.831	0.899	0.904

ChIP-seq datasets, where redder color indicates better results. As illustrated in the figure, MLSNet outperforms other models in almost all cell lines and TFs, further substantiating the excellent generalization performance of MLSNet. In particular, Fig. 5 reveals a consistent performance trend across all neural network models when applied to different cell lines and TFs. Specifically, certain TFs or cell lines yield high performance across nearly all models, whereas others result in uniformly poor performance. This observation, in conjunction with the specific TFs analyzed earlier, suggests that the prediction of TFBSs is influenced by factors such as dataset size, the biological characteristics of the TFs, and the specificity of the cell lines. We provide complimentary heatmaps on ROC-AUC in Fig. S1 available online at <http://bib.oxfordjournals.org/> and heatmaps on PR-AUC in Fig. S2 available online at <http://bib.oxfordjournals.org/> and corresponding analyses in Text S7 available online at <http://bib.oxfordjournals.org/>.

Comparing MLSNet with existing predictors

To more rigorously validate the superiority of MLSNet, we compared MLSNet with multiple models, as described in the previous

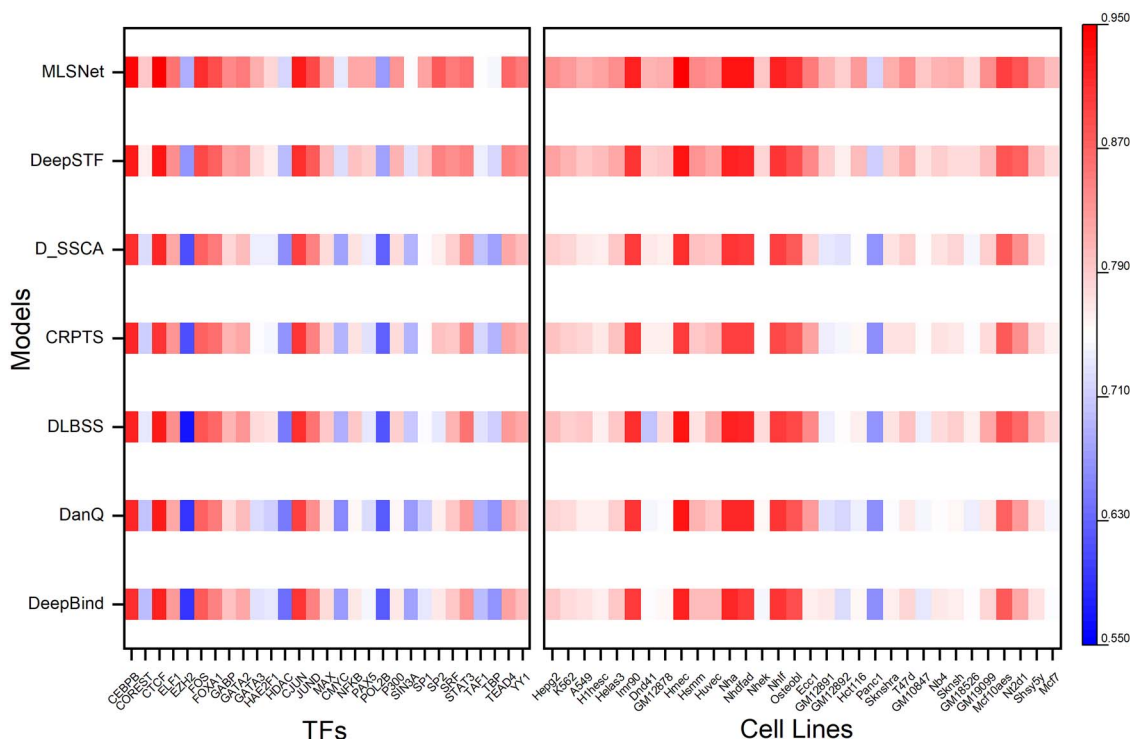


Figure 5. The heatmap of ACC results for MLSNet and other competing models, evaluated across all cell lines and transcription factors within 165 ChIP-seq datasets.

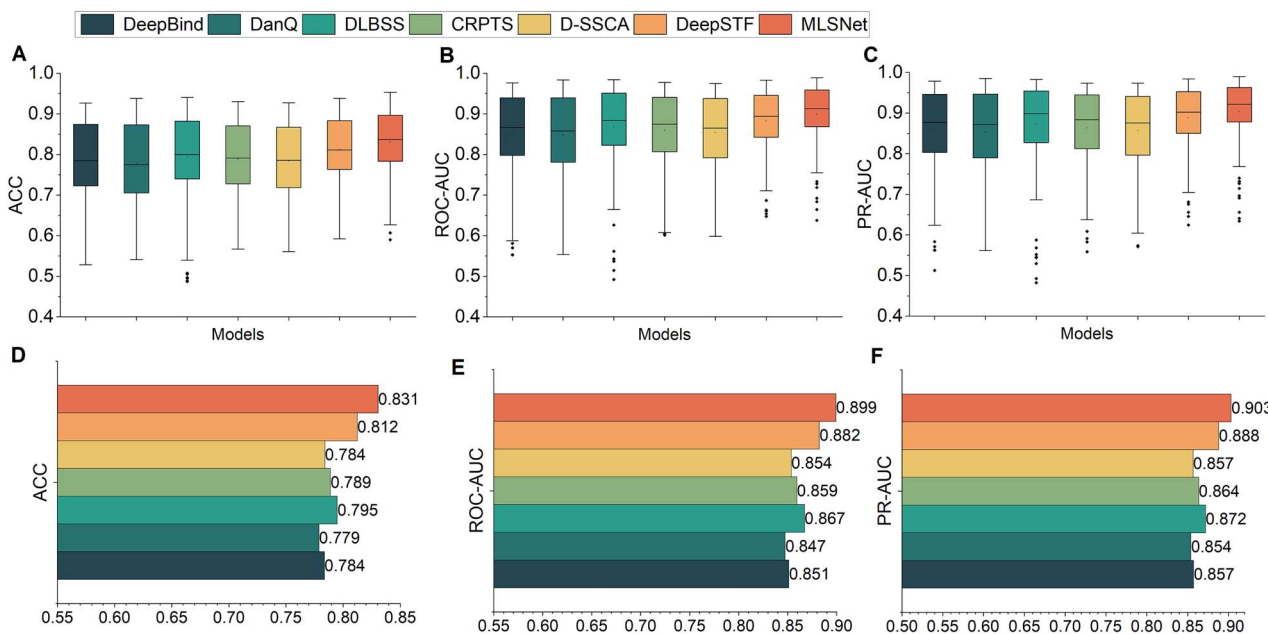


Figure 6. Overview of the comparative analysis and average results of MLSNet and competing models' results on 165 ChIP-seq datasets. (A) ACC: This part involves the ACC comparison between MLSNet and competing models. (B) ROC-AUC: This part involves the ROC-AUC comparison between MLSNet and competing models. (C) PR-AUC: This part involves the PR-AUC comparison between MLSNet and competing models. (D) Average ACC: This part involves the average ACC comparison between MLSNet and competing models. (E) Average ROC-AUC: This part involves the average ROC-AUC comparison between MLSNet and competing models. (F) Average PR-AUC: This part involves the average PR-AUC comparison between MLSNet and competing models.

text. All models were evaluated using the 165 ChIP-seq datasets used in Zhang et al. [32]. The comparison employed the evaluation standards mentioned earlier in this paper, including ACC, ROC-AUC, and PR-AUC. MLSNet achieved scores of 0.8306, 0.8992, and 0.9035 in these three metrics, respectively. These scores are 1.82%, 1.68%, and 1.54% higher than the second-best model, DeepSTF

(0.8124, 0.8824, and 0.8881, respectively). The detailed results are presented in Table 4.

MLSNet's superiority over many comparative models arises from its focus on optimizing sequence data processing while also incorporating shape data as supplementary features. Through the multisize convolutional fusion with LSTM architecture, MLSNet

effectively captures DNA sequence features across various scales and temporal relationships. Additionally, leveraging the Super Token ViT structure and Bi-LSTM enhances our ability to discern intricate patterns within DNA shape data. Compared to DeepSTF, which focuses on supplementary shape data, our MLSNet emphasizes the original sequence data, leading to better training efficiency and performance. Thanks to our multisize fusion convolutional framework, MLSNet captures a broader range of features before LSTM processing, enhancing prediction accuracy. In contrast, DeepSTF's basic convolutional operations result in some loss of feature information, as shown in our cross-cell line and cross-TF studies. This indicates that MLSNet is more effective and robust in complex prediction scenarios. The predicted results of DeepSTF and MLSNet on 165 ChIP-seq datasets are enumerated in Tables S10 and S11 available online at <http://bib.oxfordjournals.org/>. The results in Fig. 6 demonstrate MLSNet's performance on the 165 ChIP-seq datasets, consistently outperforming other methods in both average values and stability metrics. These findings underscore MLSNet's proficiency in extracting predictive insights regarding TFBSs, reaffirming its efficacy in bioinformatics.

Upon evaluating the prediction performance of MLSNet on different cell lines and TFs, it can be further considered for predicting binding sites in other cell lines or for specific TF's binding sites.

Conclusion

We introduced MLSNet, a deep learning framework for predicting TFBSs, utilizing multisize convolutional fusion with LSTM to adeptly capture intricate DNA sequence features, further complemented by the supplementation of DNA shape data supplementation. Our detailed analysis of MLSNet's mechanisms, the incremental value of shape data, and its robust performance across diverse cell lines and TFs highlight its efficacy. Analyses of the 165 ChIP-seq datasets validate MLSNet's exceptional TFBS prediction capabilities.

We also find that certain TFs or cell lines yield high performance across nearly all models, whereas others result in uniformly poor performance. Thus, we reveal that deep learning prediction of TFBSs is affected by dataset size, the biological characteristics of the TFs, and cell lines.

Despite its strengths, MLSNet faces limitations. It still grapples with occasional performance in predicting certain TF's binding sites, albeit showing significant improvement over competing deep learning methods. Moreover, MLSNet's potential extends beyond TFBSs prediction; it could also be leveraged for other bioinformatics challenges, including predictions specific to cell lines or TFs, as well as interactions between proteins and TFBSs.

Key Points

- MLSNet is a deep learning method for predicting transcription factor binding sites (TFBSs) that leverages multisize convolutional fusion with long short-term memory (LSTM) and convolutional neural networks to capture DNA-sparse higher-order sequence features. It also incorporates super token attention, Bi-LSTM, and convolutional neural networks to capture DNA shape features as supplementary features. Finally, it integrates these features to predict TFBSs. Benchmark experiments have shown that MLSNet surpasses several state-of-the-art prediction methods in TFBSs prediction.

- We have provided separate explanations for the roles of the modules in capturing dependencies and higher-order features of both DNA sequence and DNA shape.
- The multisize convolutional fusion with the LSTM module effectively captures the hidden information within DNA sequences, aiding in the extraction of higher-order hidden features and thereby enhancing prediction efficiency. We also analyzed the enhancements brought by the inclusion of shape data as supplementary features.
- Furthermore, we conducted an analysis of the prediction results of different models across various cell lines and transcription factors (TFs), which contributes to the effective utilization of deep learning methods in predicting TFBSs. Specifically, certain TFs or cell lines yield high performance across nearly all models, whereas others result in uniformly poor performance. Thus, we reveal that deep learning prediction of TFBSs is affected by dataset size, the biological characteristics of the TFs and cell lines.

Supplementary data

Supplementary data are available at *Briefings in Bioinformatics* online.

Funding

This work was supported by the National Natural Science Foundation of China (62372234, 62072243), the Natural Science Foundation of Jiangsu (BK20201304), Major Inter-Disciplinary Research project awarded by Monash University, and the Natural Science Research Start-up Foundation of Recruiting Talents of Nanjing University of Posts and Telecommunications (Grant No. NY223062).

References

1. Guo JT, Lofgren S, Farrel A. Structure-based prediction of transcription factor binding sites. *Tsinghua Sci Technol* 2014;**19**: 568–77.
2. Dunham I, Kundaje A, Aldred SF. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;**489**: 57–74.
3. Kaiser MI. ENCODE and the parts of the human genome. *Stud Hist Phil Biol Biomed Sci* 2018;**72**:28–37.
4. Chen X-F, Zhang Y-W, Xu H. *et al.* Transcriptional regulation and its misregulation in alzheimer's disease. *Mol Brain* 2013;**6**:1–9.
5. Stormo Gary D. [13] consensus patterns in dna. *Elsevier* 1990;211–21.
6. Rhee HS, Pugh BF. Comprehensive genome-wide protein DNA interactions detected at single-nucleotide resolution. *Cell* 2011;**147**:1408–19.
7. Han H, Li XL. Multi-resolution independent component analysis for high-performance tumor classification and biomarker discovery. *BMC Bioinform* 2011;**12**:S7.
8. Zheng CH, Zhang L, Ng VTY. *et al.* Molecular pattern discovery based on penalized matrix decomposition. *IEEE/ACM Trans Comput Biol Bioinform* 2011;**8**:1592–603.
9. Bernard S, Adam S, Heutte L. Dynamic random forests. *Pattern Recogn Lett* 2012;**33**:1580–6.

10. Antikainen AA, Heinonen M, Lahdesmaki H. Modeling binding specificities of transcription factor pairs with random forests. *BMC Bioinform* 2022;**23**:212.
11. Fletez-Brant C, Lee D, McCallion AS. et al. Kmer-SVM: a web server for identifying predictive regulatory sequence features in genomic data sets. *Nucleic Acids Res* 2013;**41**:W544–56.
12. Adams S, Beling PA, Cogill R. Feature selection for hidden Markov models and hidden semi-Markov models. *IEEE Access* 2016;**4**:1642–57.
13. Zhou J, Theesfeld CL, Yao K. et al. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nat Genet* 2018;**50**:1171–9.
14. Chen KM, Wong AK, Troyanskaya OG. et al. A sequence-based global map of regulatory activity for deciphering human genetics[J]. *Nat Genet* 2022;**54**:940–9.
15. Avsec Ž, Agarwal V, Visentin D. et al. Effective gene expression prediction from sequence by integrating long-range interactions[J]. *Nat Methods* 2021;**18**:1196–203.
16. Alipanahi B, DeLong A, Weirauch MT. et al. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol* 2015;**33**:831–8.
17. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model[J]. *Nat Methods* 2015;**12**:931–4.
18. Quang D, Xie XH. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Res* 2016;**44**:e107–e107.
19. Zhang Q, Zhu L, Huang D-S. High-order convolutional neural network architecture for predicting DNA-protein binding sites. *IEEE/ACM Trans Comput Biol Bioinform* 2018;**16**:1184–92.
20. Zhang Q, Zhu L, Bao W. et al. Weakly-supervised convolutional neural network architecture for predicting protein-DNA binding. *IEEE/ACM Trans Comput Biol Bioinform* 2018;**17**:679–89.
21. Zhang Y, Qiao S, Ji S. et al. Identification of DNA-protein binding sites by bootstrap multiple convolutional neural networks on sequence information. *Eng Appl Artif Intell* 2019;**79**:58–66.
22. Shen Z, Zhang Q, Han K. et al. A deep learning model for RNA-protein binding preference prediction based on hierarchical LSTM and attention network. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 2020;**19**:753–62. <https://doi.org/10.1109/TCBB.2020.3007544>.
23. Zhang Y, Qiao S, Ji S. et al. DeepSite: bidirectional LSTM and CNN models for predicting DNA-protein binding. *Int J Mach Learn Cybern* 2020;**11**:841–51.
24. Zhang Q, Wang S, Chen Z. et al. Locating transcription factor binding sites by fully convolutional neural network. *Brief Bioinform* 2021;**22**:bbaa435.
25. Avsec Z, Weilert M, Shrikumar A. et al. Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nat Genet* 2021;**53**:354–66.
26. Zheng A, Lamkin M, Zhao H. et al. Deep neural networks identify sequence context features predictive of transcription factor binding. *Nat Mach Intell* 2021;**3**:172–80.
27. Zhang QH, Shen Z, Huang DS. Predicting in-vitro transcription factor binding sites using DNA sequence plus shape. *IEEE/ACM Trans Comput Biol Bioinform* 2021;**18**:667–76.
28. Wang SG, Zhang QH, Shen Z. et al. Predicting transcription factor binding sites using DNA shape features based on shared hybrid deep learning architecture. *Mol Ther Nucleic Acids* 2021;**24**:154–63.
29. Zhang YQ, Wang ZX, Zeng YQ. et al. A novel convolution attention model for predicting transcription factor binding sites by combination of sequence and shape. *Brief Bioinform* 2022;**23**:bbab525.
30. Ding P, Wang Y, Zhang X. et al. DeepSTF: predicting transcription factor binding sites by interpretable deep neural networks combining sequence and shape. *Brief Bioinform* 2023;**24**:bbad231.
31. Graves A. Long short-term memory. *Supervised Sequence Labelling with Recurrent Neural Networks* 2012;37–45.
32. Zeng H, Edwards MD, Liu G. et al. Convolutional neural network architectures for predicting DNA-protein binding. *Bioinformatics* 2016;**32**:i121–7.
33. Wang W, Jiao X, Sun B. et al. DeepGenBind: a novel deep learning model for predicting transcription factor binding sites[C]//2022 IEEE international conference on bioinformatics and biomedicine (BIBM). *IEEE* 2022;3629–35.
34. Huang H, Zhou X, Cao J. et al. Vision transformer with super token sampling. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023; 22690–99.
35. Cortinas-Lorenzo B, Perez-Gonzalez F. Adam and the ants: on the influence of the optimization algorithm on the detectability of DNN watermarks. *Entropy* 2020;**22**:1379.
36. Ohlsson R, Renkawitz R, Lobanenkov V. CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease. *Trends Genet* 2001;**17**:520–7.
37. Phillips JE, Corces VG. CTCF: master weaver of the genome. *Cell* 2009;**137**:1194–211.
38. Liu EM, Martinez-Fundichely A, Diaz BJ. et al. Identification of cancer drivers at CTCF insulators in 1,962 whole genomes. *Cell Syst* 2019;**8**:446–455.e8. <https://doi.org/10.1016/j.cels.2019.04.001>.
39. Kim KH, Roberts CWM. Targeting EZH2 in cancer. *Nat Med* 2016;**22**:128–34.
40. Merckenschlager M, Nora EP. CTCF and cohesin in genome folding and transcriptional gene regulation. *Annu Rev Genomics Hum Genet* 2016;**17**:17–43.
41. Varambally S, Dhanasekaran SM, Zhou M. et al. The polycomb group protein EZH2 is involved in progression of prostate cancer. *Nature* 2002;**419**:624–9.
42. Ernst J, Kellis M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol* 2010;**28**:817–25.