


Received November 24, 2020, accepted December 2, 2020, date of publication December 7, 2020, date of current version December 17, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3042903

# A Review About Transcription Factor Binding Sites Prediction Based on Deep Learning

YUANQI ZENG<sup>1</sup>, MEIQIN GONG<sup>2</sup>, MENG LIN<sup>1</sup>, DONGRUI GAO<sup>1</sup>, AND YONGQING ZHANG<sup>1,3</sup> 

<sup>1</sup>School of Computer Science, Chengdu University of Information Technology, Chengdu 610225, China

<sup>2</sup>West China Second University Hospital, Sichuan University, Chengdu 610041, China

<sup>3</sup>School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China

Corresponding author: Yongqing Zhang (zhangyq@cuit.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61702058, in part by the China Postdoctoral Science Foundation under Project 2017M612948, in part by the Scientific Research Foundation for Education Department of Sichuan Province under Grant 18ZA0098, in part by the Scientific Research Foundation of the Chengdu University of Information Technology under Grant KYTZ201717, and in part by the Scientific Research Foundation for Young Academic Leaders of the Chengdu University of Information Technology under Grant J201706.

**ABSTRACT** Transcription factors (TFs) recognize and bind to specific DNA sequences, thereby altering the chromatin structure and regulating transcription. TFs aid in the formation of a guide genome that facilitates the expression of genes under complex regulation. Understanding the underlying mechanism that mediates the TF-led regulation of gene expression is a popular topic in current genomic research. However, identifying the precise TF binding site (TFBS) and the specific role of the TFs in transcriptional regulation is challenging. This article summarizes the status of research concerning the prediction of TFBS. First, the experimental methods for identifying TFBS have been summarized by accessing related databases. Second, the machine learning methods for predicting TFBS, especially deep learning, have been summarized. Finally, the study elaborates on the main challenges faced in TFBS prediction. The purpose of this article is to provide researchers with a comprehensively understand the prediction of TFBS and to promote further development in this field.


**INDEX TERMS** Transcription factor binding site, genomic research, deep learning.

## I. INTRODUCTION

Transcription factors (TFs) are proteins involved in the regulation of gene expression at the transcription level [1]. Their functions include initiation and regulation of transcription, which depends on cell type, development stage and disease status. TFs establish direct contact with the DNA in a sequence-specific manner through their DNA binding domain (DBD). TF binding sites (TFBS) is the combining position between TF and their DBD and the length of TFBS usually is 6-20 bp long with variable sequences. Genome-wide identification of TFBSs is the key to better understand transcriptional regulation. However, it is impossible to experimentally determine all TFBSs for each cell type and cell condition; therefore, a calculation model based on determining the TF binding specificity helps to predict TFBSs that have not been identified by experimentation. These models can be used to not only predict the precise location of TF interaction

within the genome, but also forecast the effects of TFBSs in a set of sequences [2], and the effect of mutations on TF binding [3]. Although there has been significant progress in elucidating TFBSs, recent research methods for accurate identification of TFBS have increasingly shown that TF binding is much more complex and involves multiple regulatory and structural changes than was originally known.

In the process of machine learning for data analysis, feature extraction and feature representation are fundamental step for successful data prediction. The whole-genome sequencing methods employed for TFBS discovery can be divided into two types: genome-based comparison methods and motif search-based methods. The comparative genomics method is based on the assumption that the functional elements (such as motif) are inherited from the common ancestor comparing to the non-functional elements. These conserved functional elements can be identified using pairwise and multiple sequence alignment techniques. The commonly used alignment tools for conservation analysis between the sequences of orthologous or paralogous species,

The associate editor coordinating the review of this manuscript and approving it for publication was Mostafa M. Fouda .

including GenomeVISTA [5], LAGAN/MLAGAN [6], and AVID and MULAN [7]. However, pairwise and multiple sequence alignment methods are inefficient in terms of the process speed. In order to accelerate the genome alignment process, heuristic techniques such as anchoring, threaded block set and greedy search algorithm are used [7]. Although comparative genomic methods are effective in recognizing conserved motifs, they often overlook some conserved functional motifs. On the other hand, the motif search-based method uses the annotated topic profiles database to detect related TFs in the input dataset. Topic are usually represented as position weight matrices (PWM) [8] or variants [9]. The sequence-specificity in the identification of TFBS by the TFs has been demonstrated by utilizing next-generation sequencing (NGS) techniques such as systematic evolution of ligands by exponential enrichment sequencing (SELEX-seq) [10], chromatin immunoprecipitation sequencing (ChIP-seq) [11], ChIP-exo [12], and ChIP-nexus [13] in several structural studies involving protein-DNA complexes. The advent of chromatin immunoprecipitation (ChIP) technology has made genome-wide sequence analysis highly feasible. Many computing tools have been proposed, especially for deep learning methods. Since deep learning can automatically perform feature extraction for the input data, it has been widely used by researchers from multiple fields.

The unprecedented success of deep learning can be attributed to the following factors: (1) the development of a graphics processing unit (GPU), (2) availability of large amounts of data, and (3) the development of a learning algorithm. Deep learning has been associated with breakthrough achievements because it can learn the good characteristics of a feature representation from the data. The known applications of deep learning include gene and splicing regulation [14], DNA methylation [15], protein classification [16], gene recombination [17], nucleic acid sequence analysis [18], molecular evolution analysis [19], molecular immunology [20], gene cloning [21], genomic diagnosis [22], gene network construction [23], and TFBS prediction [24]. When the number of samples is large, the deep learning method is highly effective. However, for medical and genetic engineering applications, the number of samples is limited, with less than 1000 sequence samples. Therefore, one of the main challenges in the application of deep learning to medical and genetic engineering is the limited availability of training samples to build a deep model without the influence of overfitting. Researchers over the years have designed various strategies to alleviate the challenge of limited sample availability, such as (1) considering the structural information in addition to the sequence information, so as to increase the input dimension and expand the feature information; (2) generating some samples manually through data enhancement to expand the data set.

In this review, we have summarized the recent research progress on TF binding sites, focusing on the methods based on deep learning. We began by defining the basic concepts about TFBS and experimental methods used for

their identification. We then review a few TFBS databases and explore the basic theory behind the different deep learning models, such as CNN, LSTM GAN, embedding. Next, we introduce the application of machine learning and deep learning in the prediction of TFBS. We also describe the basic process of predicting TFBS by deep learning. In addition, we discuss the factors influencing TF-DNA binding. Finally, the challenges and potential limitations of TFBS prediction based on deep learning are discussed. Figure 1 represents TFBS analysis in a flow chart format, including the biological experiment analysis, data acquisition from the database, building of a computational model, and visualization of results.

## II. PROBLEM FORMULATION AND RESEARCH TECHNOLOGY

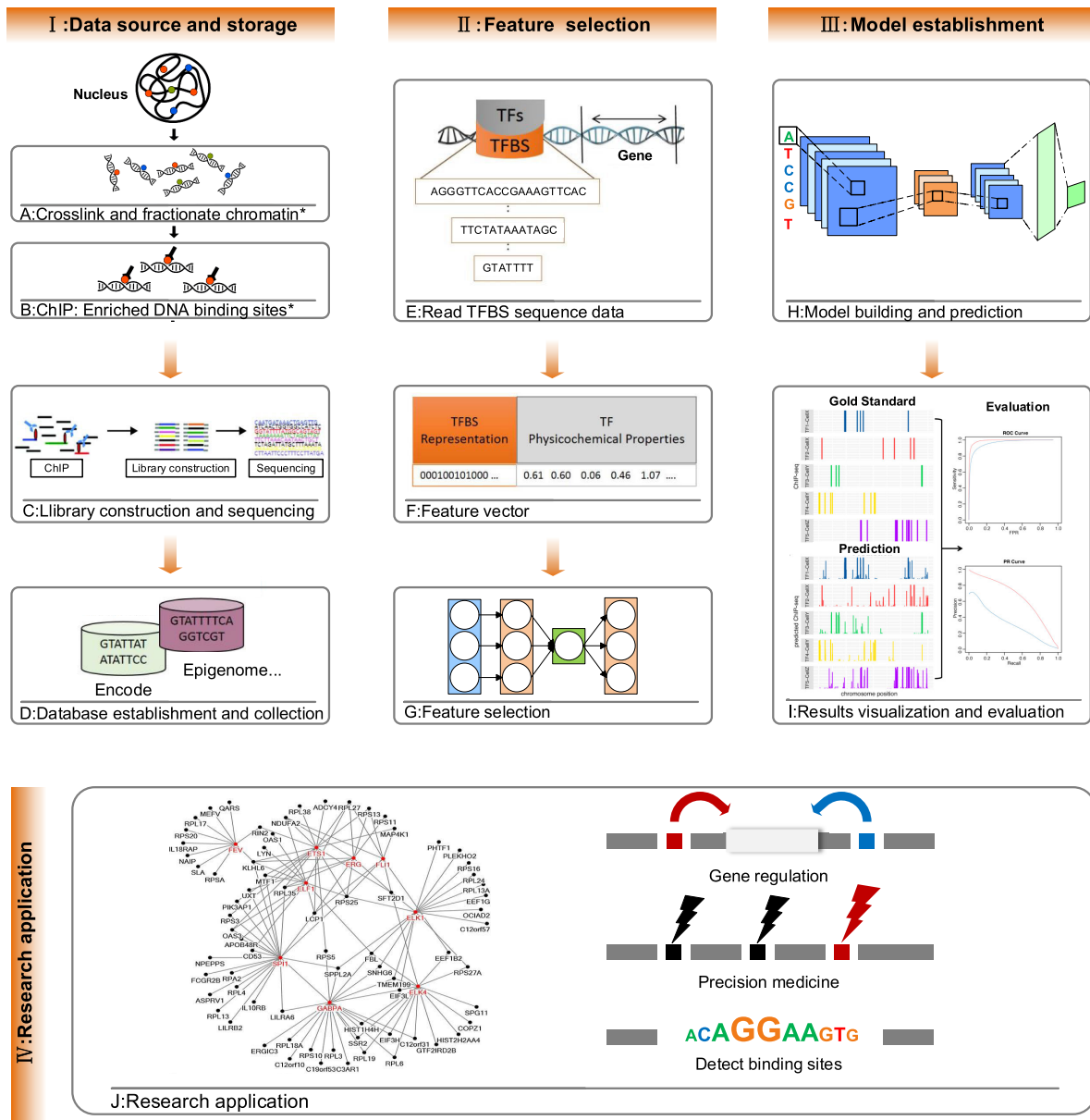
### A. PROBLEM FORMULATION

The issue of TFBS prediction can be defined as the manner in which appropriate modeling of TFs and the local chromatin structure at the TFBS can be performed. Efficient modeling necessitates development of a computable and physically reasonable model. According to the various histochemical data available for the TFBSs, the input information can be genomic DNA sequence, chromatin structure and accessibility, and protein sequence and their antibody. The method used for the prediction of TFBSs involves multiple learning models run on different input data modes. The output involves the assessment of existing TFBSs. The TFBS is primarily assessed for the presence of a motif, without judging the boundary and location of the motif. Research applications of TFBS prediction models include accurate recognition of the individual or multiple TFBSs, identification of mutations at TFBSs, and promotion of research concerning targeted drug delivery.

### B. DEVELOPMENT OF HIGH-THROUGHPUT SEQUENCING TECHNOLOGY

With the emergence of high-throughput techniques for measuring protein-DNA binding (Table 1 and Figure 2), designing of complex TF and DNA binding models through machine learning has made identification of TFBS more feasible. The experimental dataset for complex modeling of the TFBS contains both noise and bias. However, compared with the simple PWM model, complex models can easily fit noise and bias. Recently, some studies have focused on designing TFBS specific models from high-throughput data [25]. However, even the best-performing models with specific in vitro datasets do not always perform well on independent in vivo data.

High-throughput technology provides a wealth of datasets to enhance our ability to study the binding specificity of proteins to DNA. For example, many prediction methods, such as SELEX and protein binding microarray (PBM) have been developed that rely on in vitro sequence data. SELEX [37] and PBM methods are cheaper and faster than ChIP-seq. Furthermore, they do not rely on highly specific antibodies. The



**FIGURE 1.** TFBS modeling and application diagram. (I) Representatives of TFBS obtained through the high-throughput sequencing technology, ChIP, and collected from major databases. These include cross-linking, sequence truncation, enrichment, and storage sites. (II) Represents the data required for querying and forecasting from the database, expressing the data with vectors, and finally extracting the features through the vectorized data. (III) Represents the modeling and prediction TFBS (the modeling method can be machine learning, deep learning, etc.), and the prediction results are used for visual analysis. (IV) Predicting TFBS will be helpful for the construction of gene regulatory networks, drug research, and mutation detection [4].

PBM data measured TF binding specificity of all the possible 8 base pair (bp) sequences and facilitated the characterization of low-affinity TFBSs on the DNA, which are usually not captured by simple DNA binding models.

**C. DEVELOPMENT OF ChIP TECHNOLOGY**

ChIP is an effective method to study the mechanisms of gene regulation by selectively enriching the DNA fragments interacting with given proteins in living cells. ChIP-based methods for detecting the protein-DNA interaction site have witnessed

significant developments from ChIP-polymerase chain reaction (ChIP-PCR) for single locus detection to ChIP-ChIP microarray and ChIP-seq, which involve ChIP followed by a microarray hybridization and high-throughput sequencing, respectively [11].

Owing to the similarities between TFBS and enhancers, conventional protein-DNA localization methods lack the resolution to distinguish the spatial arrangement of the TFs. Therefore, the resolution of ChIP-exo close to a single base is essential for understanding the molecular mechanism

TABLE 1. In vivo high flux DNA determination.

Experimental name	Experimental description	Year	Reference
ChIP-chip	Chromatin immunoprecipitation and microarray hybridization	2000	[26]
ChIP-seq	Chromatin immunoprecipitation followed by high-throughput sequencing	2007	[11]
ChIP-exo	Chromatin immunoprecipitation and exonuclease digestion, followed by high-throughput sequencing	2017	[12]
DamID	Identification of DNA adenine methyltransferase	2006	[27]
DNase-seq	DNase cutting and high-throughput sequencing	2009	[28]
FAIRE-seq	Formaldehyde-assisted separation of regulatory elements, followed by high-throughput sequencing	2007	[29]
ATAC-seq	High throughput determination of chromatin accessible by transposase	2013	[30]
esATAC	Makes ATAC-seq data analysis easy to be widely used	2018	[31]
epic2	Quickly and expertly find distributed domains in ChIP-seq data at low cost	2019	[32]
PAtCh-Cap	Nonspecific capture of chromatin-binding protein by its carboxylate group	2016	[33]
lobChIP	lobchip is completed in the chip step, followed by sample elution and de-crosslinking	2015	[34]
ChIPmentation	This method combines chromatin immunoprecipitation and Tn5 transposase to prepare a sequencing library	2015	[35]
ChIP-nexus	Nucleic acid exonuclease, unique bar code, and single connection for ChIP experiment with high nucleotide resolution	2015	[13]
HT-ChIPmentation	High-throughput ChIPmentation: freely scalable, single day ChIPseq data generation from very low cell-numbers	2019	[36]

DNA binding specificity models

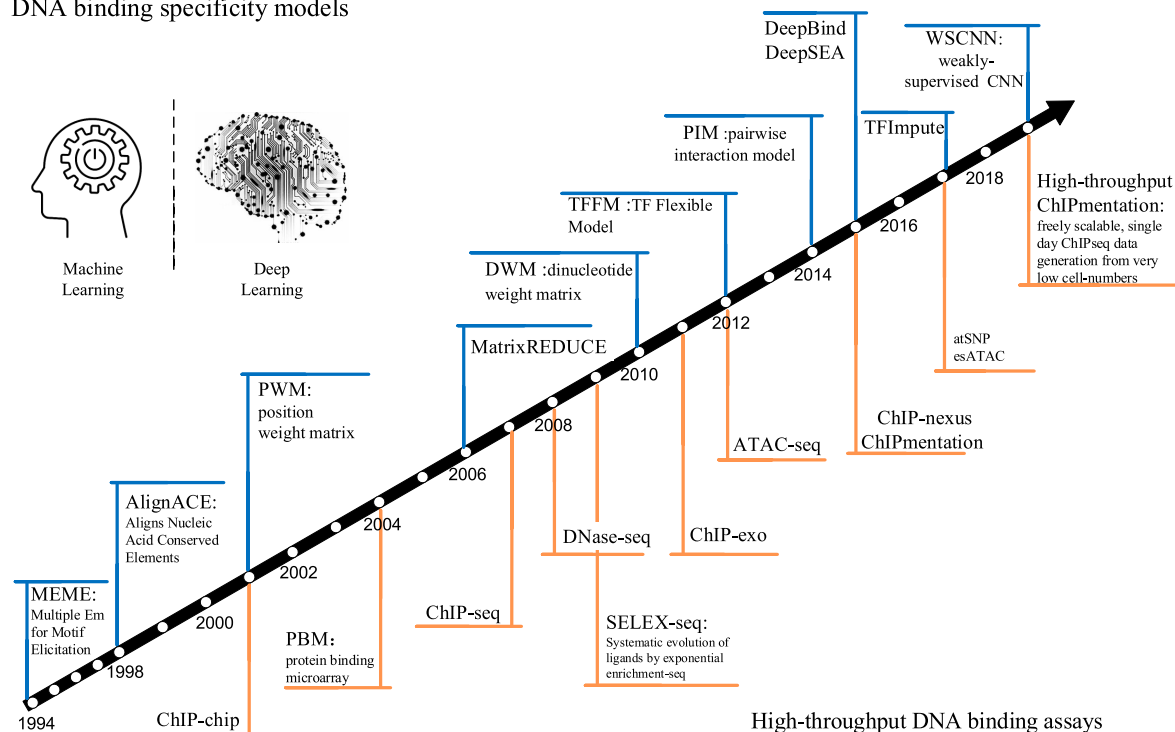


FIGURE 2. Based on the TF-DNA binding specificity experiment and the calculation of research genome method schedule. Develop an experimental high-throughput DNA binding assay (below the timeline) and computational DNA binding-specific models and algorithms (above the timeline). These experimental and computational methods are provided in Table 1. Considering the appearance of ChIP-seq as the boundary, the specific model algorithm changed from machine learning to a deep learning algorithm.

underlying TF binding [38]. ChIP-exo can significantly improve the resolution and reduce noise at the same time [39], [40]. Although mastering the ChIP-exo technology is challenging compared to ChIP-seq, it has now been widely adopted, aiming to obtain unique ultra-high resolution using multiple biological systems. In addition, ChIP-nexus developed as a variant of the ChIP-exo method, attempting to increase the complexity of the ChIP-exo library by replacing the traditional double-stranded DNA linear ligation with a circular ligation step [13].

The protein-attached chromatin capture (PAtCh-Cap) method involves bead-bound processing steps in addition to the ChIP-based methods. It relies on the nonspecific capture

of chromatin-bound proteins via their carboxylate groups, leaving DNA accessible to subsequent chemical treatments in parallel, such as chromatin immunoprecipitation for the target protein. Application of PAtCh-Cap includes enhanced artifact removal from ChIP-exo data, increasing confidence in peak identification and facilitating de novo motif search. The PAtCh-Cap method also mediated the discovery of a novel CCCTC-binding factor (CTCF) binding motif [33]. In order to overcome the problem of low throughput of standard protocols for ChIP and library preparation, Wallerman *et al.* developed a bead library for ChIP-seq protocol (lobChIP). The applications of the proposed input strategy include enhanced removal of artifacts from ChIP-exo data, accurate peak

identification, and search for motifs from scratch. The study by [36] proved that labeling immunoprecipitated chromatin to bead-bound chromatin directly in a robust one-step reaction is a fast and cost-effective ChIP-seq workflow. It has been reported to produce some excellent results for histone labeling and TFs [41].

### D. CONTRIBUTION AND LIMITATION OF HIGH THROUGHPUT TECHNOLOGY

High-throughput protein-DNA binding technology has revealed that a large number of proteins can bind to DNA using two or more different modes. In order to fathom the precise biochemical mechanism guiding the interaction of the TF at TFBSs, it is necessary to study TFs with multiple binding modes.

Although there is an increase in the use of ChIP-exo/nexus methods, these technologies have some limitations. ChIP-exo and ChIP-nexus are more complex than ChIP-seq, making the experiments both expensive and time consuming. For example, a single ChIP-seq or ChIP-exo experiment may contain multiple types of binding events generated by the variable protein-DNA interaction patterns. To systematically detect the multiple protein-DNA interaction patterns in a single ChIP-exo experiment, Yamada *et al.* introduced the ChIP-exo hybrid model (ChExMix). ChExMix uses the ChIP-exo tag distribution pattern and DNA pattern to model the genome location and subtype members of the combined event. Yamada *et al.* proved that ChExMix could accurately detect and classify the combined event subtypes using computer technology [12]. For a detailed description of the ChIP-exo/nexus methods, please refer to [33].

### III. DATABASE OF TRANSCRIPTION FACTOR BINDING SITES

Over the past decades, our ability to generate motifs and genomic-binding sites simultaneously has improved significantly, resulting in unprecedented amounts of data on TF-DNA interactions. Motif was first discovered through experimental methods. In other words, it is not that motif analysis was only possible with ChIP-seq. The motif research has been studied for a long time. For example, the ‘TATAAT’ box was discovered by pribnow in 1975 [42], and the upstream ‘TTGACA’ motif is the specific sequence of the RNA polymerase binding site. Moreover, not all the binding sites necessarily matched the motif perfectly, and most of them only matched 7-9 of the 12 bases. The matching degree between the binding site and the motif is often related to the binding strength between the protein and DNA. At present, there are more and more motifs recognized being discovery. For example, TRANSFAC [43] and JASPAR [44] databases have a large number of motifs for transcription factors. With the massive output of ChIP-seq data, motif research will be further in-depth. Some research groups integrate existing ChIP-seq data to provide a more comprehensive and accurate motif database. To develop the current TF catalog, we have completely utilized

TF directories such as TRANSFAC [43], JASPAR [44], SELEX\_DB [10], high-throughput (HT)-SELEX [37], UNIPROBE [45], Cis-BP [46], and previous human TF catalogs [1].

With the accumulation of TFBSs verified by biological experiments, there are various databases that collate this information. For example, the Homo sapiens comprehensive model (HOCOMOCO) [47] is a collection of selected entries from various sources dedicated about human TFs. Some major updates have been made in the latest version of HOCOMOCO V11. The latest HOCOMOCO contains binding models for 453 mouse and 680 human transcription factors, including 1302 single nucleotide and 576 dinucleotide position weight matrices, which describe the main binding preference and reliable alternative binding specificity of each transcription factor. For fruit flies, a large number of patterns can be found in the FlyfactorSurvey [48] and in OnTheFly [49]. For yeast, there is a database called ScerTF [50]. Until recently, the known motifs of TFs in *C. elegans* were relatively few; however, at present, about 40% of the PWM has been determined or inferred [8], which can be accessed at the CisBP database. PlantTFDB [25], the plant TF database, contains information on 26402 TFs from 22 plants. International system for agricultural science and technology (AGRIS) [51] contains information on Arabidopsis TFs and their corresponding binding sites. Transcriptional regulatory element database (TRED) [52] is a collection of mammalian transcription regulatory elements. The promoter regions for human, mouse, rat, and other species are almost completely annotated in this database. The regulatory relationship between mammalian TFs and target genes has been collated in Integrated TF platform (ITFP) [53].

The TFBSshape database can be used to generate heat maps and quantitative data from TF datasets for 23 different species for DNA structural features, that is, DNA data for minor groove widths (MGW), rolls, propeller twist (ProT), and helix twists (HelT) [54]. In the latest TFBSshape database [55], the data content has been increased to 2428 structural profiles, with 1900 TFs from 39 different species. The structure profile of each TFBS entry now includes 13 shape features of standard DNA and micro groove electrostatic potential, and 4 shape features of methylated DNA. TFBSshape has improved the flexibility and accuracy of shape-based transcription factor binding, and designed a new tool to compare the methylation and non-methylated structure of transcription factors, and deduced the method of DNA shape keeping nucleotide mutation in transcription factor binding. The construction of these databases greatly promotes our understanding of the TF and TFBS interaction in multiple species and tissues during different developmental stages. Table 2 provides brief information on some commonly used databases for the reference of the reader.

### IV. INTRODUCTION OF DEEP LEARNING METHODS

The basic unit of the neural network in deep learning is a node, which has been inspired by the biological neurons in

**TABLE 2. Database of transcription factor binding sites.**

Database	Description	Scale	Reference
TRANSFAC	TRANSFAC is database of genomic binding sites and DNA-binding properties.	gene data of more than 300 species; 23000 TF (and 1200 miRNA) reports; 2000 high-throughput TFBS chip experimental reports; 360000 promoter reports.	[43]
JASPAR	JASPAR is a set of transcription factor DNA-binding preferences, which is modeled as a matrix.	Includes 1646 non-redundant PFMs (746 vertebrates, 530 plants, 183 fungi, 143 insects, 43 nematodes, and 1 caudate).	[44]
HOCOMOCO	To establish a complete set of transcription factor binding models for humans and mice through large-scale ChIP-seq analysis	(HOCOMOCO) V11 provides a TF binding model for 680 individuals and 453 mice.	[47]
UniPROBE	UniPROBE is a repository of experimental data from universal protein binding microarray (PB-M) experiments.	Hosts 709 non-redundant proteins and complexes, including <i>Vibrio harveyi</i> , <i>Plasmodium falciparum</i> , <i>Cryptosporidium aphid</i> , <i>Saccharomyces cerevisiae</i> , <i>Caenorhabditis elegans</i> , mice, and humans.	[45]
PRODORIC	PRODORIC is a database of annotation information about gene expression regulation in prokaryotes.	More than 2500 TFBS. A total of 120 expression profiles have been stored, including links to about 9000 genes.	[56]
HTPSELEX	The HTPSELEX database contains a sequence set of TFBS selected in vitro by SELEX and high-throughput SELEX methods.	Contains 12 separate SELEX libraries for transcription factors CTF / NF1 and lef/TCF families, covering more than 40000 loci.	[9]
FlyfactorSurvey	FlyfactorSurvey is a DNA binding-specific database of transcription factors (TF) in <i>Drosophila</i> .	311 <i>Drosophila</i> transcription factors with at least one b1H motif. 23 alternative splicing isomers with new specificities.	[48]
OnTheFly	The OnTheFly databases include a systematic collection of transcription factors (TFs) and their DNA binding sites in <i>Drosophila</i> .	The TF structure was obtained from PDB (65 TF) and homology model (1489 TF, 1171 of modbase, and 318 of pipeline using Pudge homology modeling).	[49]
PlantTFDB	PlantTFDB covers the main genealogy of green plants and provides the TF genealogy of the whole green plant genome.	PlantTFDB identified 320370 TF from 165 species and classified them into 58 families.	[25]
TFBSshape	TFBSshape is a motif database for analyzing structural profiles of transcription factor binding sites (TFBSs).	TFBSshape content has increased to 2428 structural profiles for 1900 TFs from 39 different species. The structural profiles for each TFBS entry now include 13 shape features.	[55]

the mammalian brain. Deep learning can be defined as a neural network with a large number of parameters and layers, which can be roughly divided into (1) convolutional neural network, (2) long short-term memory, (3) generative adversarial network, (4) Word2vec, (5) attention mechanism, and (6) graph convolutional networks. The following paragraphs briefly introduce each of them.

#### A. CONVOLUTIONAL NEURAL NETWORK (CNN)

CNN a well-known deep learning framework, has been widely applied in image recognition [57], speech recognition [58], computer vision [59], natural language processing [60], bioinformatics [24], and other artificial intelligence research fields [61]. Wang [62] investigated essential relationships between generalization capabilities and fuzziness of fuzzy classifiers. The study makes a claim and offers sound evidence behind the observation that higher fuzziness of a fuzzy classifier may imply better generalization aspects of the classifier. The components of CNN include convolutional, pooling, and fully connected layers. The convolutional layer is proposed to extract and represent the local information of original features through several feature maps and kernels. The pooling layer is employed to compress the resolution of the feature maps to achieve spatial invariance. After several convolution and pooling operations, there may be one or more fully connected layers to perform advanced reasoning. The output of the last fully connected layer is transferred

to an output layer. For a classifier or regression task, softmax regression is commonly used because it can produce a well-formed probability distribution corresponding to the outputs.

#### B. LONG SHORT-TERM MEMORY (LSTM)

Recurrent neural network is a neural network that recurs linearly in time. Compared with the general fully connected neural network, the structure of the recurrent neural network has one or several memory units, and this memory unit is the key to the recurrent neural network. The input of RNN includes two parts: one is the current input  $x_t$ , which is used to update the state in real time, and the other is the state  $h_{t-1}$  of the hidden layer at the previous moment, which is used to remember the state, while the network at different times shares the same set of parameters. However, the back-propagation calculation process used is time-dependent in RNN optimization. When updating the gradient of the parameter  $W$ , the gradient at the current time and the gradient at the next time must be considered, the derivative at time  $t$  will propagate to  $t_1, t_2, \dots, t_n$  time, so there is a coefficient of continuous multiplication. Multiplication has always brought two problems: gradient explosion and disappearance. Moreover, during the forward process, the influence of the former input will be less and less on the later nodes, which is the long-distance dependence problem. In this way, the ability of the “memorize” is lost. It is necessary to know that biological neurons have a strong ability to remember past sequential states.

Long short-term memory (LSTM) can solve the aforementioned problems about RNN, which introduce several gates about the cell state. The cell state carries the information of the previous states. Every time a new moment comes, there are corresponding operations to decide what old information to discard and what new information to add. This state is different from the hidden layer state  $h$ . During the update process, its update is slow, while the hidden layer state  $h$  is updated quickly. LSTM is well suitable for capturing the long and short dependency information in sequence [63]. A memory mechanism is applied in LSTM to replace the hidden function in the traditional RNN. The functional unit in LSTM consists of a memory cell, a forget gate, an input gate, and an output gate, which is designed to enhance the ability of LSTM to model long-range dependence. LSTM memory cell is given in the following equations:

$$f_t = \sigma(W_{xf} * x_t + W_{hf} * h_{t-1} + b_f) \quad (1)$$

$$i_t = \sigma(W_{xi} * x_t + W_{hi} * h_{t-1} + b_i) \quad (2)$$

$$c_t = f_t \otimes c_{t-1} + i_t \otimes \tanh(W_{xc} * x_t + W_{hc} * h_{t-1} + b_c) \quad (3)$$

$$o_t = \sigma(W_{xo} * x_t + W_{ho} * h_{t-1} + b_o) \quad (4)$$

$$h_t = o_t \otimes \tanh(c_t) \quad (5)$$

where  $\sigma$  is the logistic Sigmoid function,  $\tanh$  is a function to confine the values between  $-1$  and  $1$ ,  $f$ ,  $i$ ,  $c$ ,  $o$  represent the forget gate, input gate, cell vectors and output gate, respectively, which are specified to be the same value as given in the hidden vector  $h$ ,  $W_{xf}$  is the input-forget gate matrix, and  $W_{hf}$  is the hidden-forget gate matrix. The index  $t$  refers to the time step.  $\otimes$  represents the vector product. It is worthwhile to note that the initial values of  $c_0 = 0$  and  $h_0 = 0$ .

### C. THE GENERATIVE ADVERSARIAL NETWORK (GAN)

GAN learns by letting two neural networks play against each other. GAN mediates the generation of new and fake data based on the original data set. The model generates a fairly good output through the mutual game learning of at least two modules in the framework: the generative model and the discriminative model. The discriminator differentiates the generated false target from the real one, while the generator cheats the discriminator in generating false targets [64], [65].

### D. Word2vec

The word2vec model is essentially a simplified neural network. The input is a one-hot vector, and the hidden layer, which is a linear unit, has no activation function. The dimension of the output layer is the same as that of the input layer, and is obtained using Softmax regression. However, when the model is trained; we will not use the trained model to handle new tasks. What we really need are the parameters learned by the model through the training data, such as the weight matrix of the hidden layer. How does this model define data input and output? The word2vec model architecture can be divided into two types: Continuous Bag-of-Words (CBOW) and

continuous Skip-Gram. In case of CBOW, multiple context-related words are used as the input information to get the target word as the output. Whereas, in the Skip-Gram model, the concept is reversed, that is, the input is the word vector of a specific word, and the output is the context word vector corresponding to the specific word. CBOW is more suitable for small databases, while Skip-Gram performs better for large ones [66].

### E. ATTENTION MECHANISM

Recently, the concept of attention mechanism has achieved great success in neural machine translation and sentiment analysis [67]. It enhances the ability of RNNs by focusing on information that is highly valuable for successful prediction within the input [68]. Combined with RNNs, it allows models to learn high-level representations of the input sequences with long-range dependencies. In addition, attention mechanism makes the RNN models more interpretable by assigning attention weights based on importance to different positions of the input. Different visualization methods have also been developed to explore the relationship between the input and output sequences using the attention mechanism, the alignment view [69], and the extra layer of interpretability. It is anticipated that introducing the attention mechanism to the prediction of binding sites would enhance the prediction accuracy as well as the level of interpretability for existing CNN-RNN architecture models [70].

### F. GRAPH CONVOLUTIONAL NETWORKS (GCN)

Interpreting complex graphs and extracting potential knowledge from them is a challenging task. Graphs are the storage medium for knowledge, and deep learning is an important tool for extracting graphical information. The combination of the two is an inevitable trend. Many data in the real world are stored in the form of graphs, such as social networks, knowledge graphs, and protein-DNA interaction networks [71]. Recently, some researchers have developed a general neural network model that could process graphical data. Since majority of the approaches are associated with correlating CNN to graph, the resulting structures often have certain commonalities. Using ideas similar to convolution weight sharing, this type of network can be referred to as graph convolutional networks (GCNs) [72].

## V. APPLICATION OF TRADITIONAL MACHINE LEARNING METHODS

### A. REPRESENTATION OF CONSENSUS SEQUENCE

“Consensus” is used to indicate the most frequently occurring nucleotide at each position of the transcription factor binding site. In practical, we found that the frequency of certain two or three nucleotides at some binding sites is relatively close, or even completely equal. In this case, only using a single nucleotide to represent the position cannot fully reflect the degree of conservation. Therefore, the expression “degenerate consensus” is also used in the description of

**TABLE 3. IUPAC degenerate codes for representing nucleotide sequence patterns.**

IUPAC code	Nucleotide	IUPAC code	Nucleotide
W	A or T	B	C,G or T
R	A or G	D	A,G or T
K	G or T	H	A,C or T
S	C or G	V	A,C or G
Y	C or T	N	A,C,G or T
M	A or C		

transcription factor binding sites. The degenerate consensus sequence uses symbols to indicate different nucleotides that occur at the same position, not just a single nucleotide. International union of pure and applied chemistry (IUPAC) codes [73] is a widely used symbolic representation of degenerate consensus sequences, as shown in Table 3. The representation method based on the consensus sequence is simple and easy to understand, but it cannot reflect the probability of different bases at each position. This way of expression sacrifices specificity and sensitivity.

Selective microfluidics-based ligand enrichment (SMiLE) [74] and WEEDER [75] were used to solve the problem of random replacement of the bases at each position of the motif during pattern matching. This kind of algorithm uses a suffix tree structure to build the initial index of the sequence, and thereafter exhaustively searches all consistent candidate sequences. The SMiLE algorithm compares the number of subsequences in the input sequence set with the number of subsequences in a negative or random sequence set. The WEEDER algorithm compares the actual number of occurrences of a subsequence with the expected number of occurrences of the subsequence in all promoter regions in the same set. It uses a measurement function similar to the amount of information, to measure the entire subsequence rather than a single base. The target motif is obtained on completion of the measurement [75]. In order to overcome the rough description of the motif by the consistent sequence, the algorithm based on the consistent sequence selects a better subsequence to form the profile and then extracts the corresponding motif instance of the sequence from the spectrum. In this way, not only can the predicted model instances be arranged more carefully, but the real model instances can also be filtered and the threshold can be found. The total number of permutation methods increases exponentially with the number of input sequences. Therefore, heuristic and clustering methods could effectively reduce the search space and be widely used in the motif discovery problem to obtain the statically significant sequence.

### B. REPRESENTATION OF POSITION WEIGHT MATRIX

Compared with the consensus sequence method, the position frequency matrix can reflect the probability of different bases at each position. The hypothesis of this model is that the probability of the occurrence of bases at each site is independent of each other. Some studies have shown that there is a correlation between the bases of transcription factor

binding sites. The most commonly used statistical model is the position frequency matrix (PFM), which is used to represent the frequency of each character in the character set  $\Omega = A, C, G, T$  at each position in the motif (transcription factor binding site). PFM from position count matrix (PCM). At the same time, the PFM is often transformed into position weight matrix (PWM) considering the bias of base composition in DNA sequence.

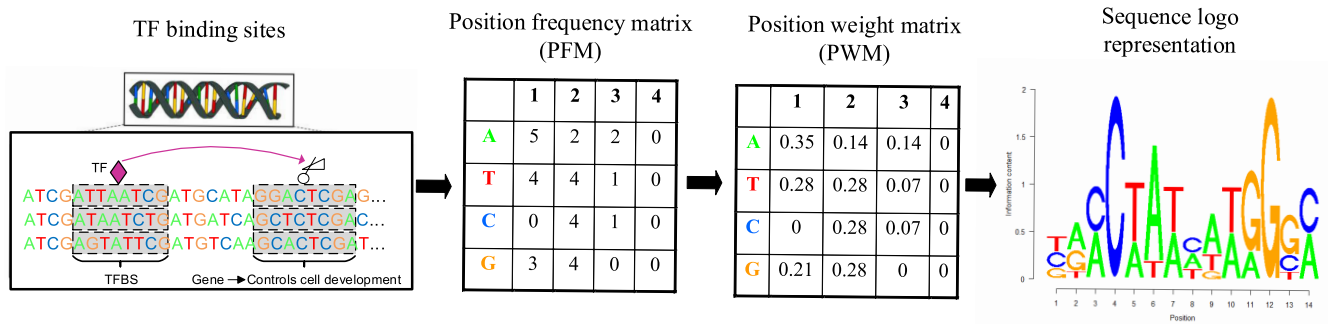
PWM [9] is sometimes called a position-specific scoring matrix (PSSM) or a weighing matrix, and it is a special expression matrix used to express TF. When using position weights to predict the TFBS, there are certain considerations: (1) how to obtain the PWM, (2) how to combine the search algorithm with PWM to predict TFBS, and (3) the limitations of using PWM to predict TFBS.

First, PWM provides a score for all possible bases at each position in the binding site. Any specific TF has its own PWM to describe its characteristics, and there are many ways to determine the PWM elements, as shown in Figure 3. A counting matrix, obtained from the aligned positions, simply records the numbers of each base at each position. Through the generated probability matrix, the statistical measures of these subsequences relative to the background sequence are calculated, to evaluate conservatism and specificity. The probability matrix with the highest score is the corresponding motif component.

Second, after determining the PWM, the motif discovery problem is transformed into a "combination optimization" problem, wherein, all possible subsequence combinations are explored to find the combination with the highest measurement value. Note that the size of the solution space will increase exponentially with the increase in the input sequence. Therefore, in order to avoid the huge overhead cost, in terms of time and space, involved in the exhaustive search, heuristic methods are often used to explore the solution space. At the same time, a large number of combinatorial optimization techniques are also widely used to design motif recognition algorithms such as greedy algorithm, local search, random search, and genetic algorithm. Another type of algorithm uses different strategies to explore the solution space: first, it selects some subsequences from the input sequence to form the initial state (probability spectrum); then, at each step of the algorithm, it replaces some of the subsequences to update the probability spectrum to obtain a higher measurement value. The training iterations continue until the highest measurement value is attained. Lawrence and Reilly first used a probabilistic algorithm based on training iterations for the motif discovery problem [76]. The algorithm is considered a well-known expectation maximization (EM) algorithm.

Finally, in the probability model using PWM, it is assumed that the independently distributed background components will have a huge impact on the phantom recognition signal. However, as not all nucleotides in the input sequence are affected by adjacent nucleotides, this assumption may seem impractical. Therefore, some algorithms improve the background model and use more complex high-order Markov





**FIGURE 3.** Position weight matrix (PWM) is a widely used representation of transcription factor recognition motif. It is a matrix of N rows and four columns, in which the frequency of each base at each position is described. The sequence score can also be interpreted in a physical framework as the binding energy for that sequence.

models as the background to improve the accuracy of identifying the phantom. The limitation of the PWM model, when used to express TF binding specificity, is the assumption that the position in the binding site independently affects the binding affinity. This is usually a good assumption, but it does not hold true for every TF. There are also some examples where the assumption is violated, and the TF binds in different ways at different locations. Hence, a single PWM cannot accurately capture the TF binding specificity [25].

**C. STATISTICAL LEARNING METHOD**

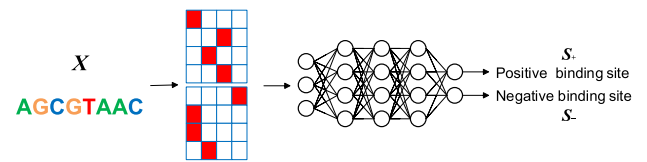
Widely used computational methods for TFBS prediction based on PWMs usually have high rate of false positives. Moreover, computational studies of transcription regulation in eukaryotes frequently require numerous PWM models of TFBSs because of the large number of TFs involved. To overcome these problems, Khamis developed DRAF models, a novel method for TFBS prediction DRAF [77]. DRAF models use more features than PWM models, by combining information from TFBS sequences and physicochemical properties of TF DNA-binding domains into machine learning models.

Because PWM assumes an independent distribution of sequence bases, which has a huge impact on the recognition of phantom signals, Gao and Ruan [78] proposed a novel algorithm based on the so-called multi-instance learning (MIL) paradigm. MIL divides each DNA sequence into multiple overlapping subsequences, and models each subsequence separately. It not only implicitly identifies the location of the binding site, but also maps sequence level features (k-mers) to binding events. Since MIL takes into account the dependence between the bases, the model has higher accuracy and better interpretability.

**VI. APPLICATIONS OF DEEP LEARNING IN TFBS**

**A. VARIOUS APPLICATION BASED ON DEEP LEARNING**

In recent years, deep learning finds application in various fields. In TFBS prediction-based research, there are some benchmark deep learning models. It is natural to regard a DNA sequence as an input sentence with four characters A, C, G, and T instead of an image; however, research on



**FIGURE 4.** The prediction model of TFBS using deep learning. The input is the DNA sequence. Following this, the input DNA sequence is encoded to a 4-dimensional matrix by one-hot encoding. Next, deep learning is employed to build the prediction model. Finally, the output indicates whether there are binding sites.

image classification and natural language processing provides valuable experience for DNA, for example, DeepBind [24] and IDeep [79], the recently developed applications for TF-DNA binding prediction. Figure 4 shows a conceptual model using deep learning for TFBS prediction. Using the DNA sequence as the input information, the neural network is trained autonomously to adjust the network parameters and results. Furthermore, Table 4 summary the TFBS prediction methods based on deep learning.

**B. APPLICATION OF CNN**

DeepBind [24] shows that deep learning technology can be used to determine sequence specificity from experimental data. It provides a scalable, flexible, and unified calculation method for pattern discovery. In addition, DeepBind is the first ever method that addresses the need for accurate representation of protein target binding motifs. The deep convolution method has ushered in the upsurge of deep learning applications to process biological information.

Due to the lack of effective methods to extract higher-order dependence, most of the proposed TFBS prediction methods use only low-order dependence for prediction. In this work, the author proposed a novel method to extract high-order dependence by applying CNN to histone modification features. Then, a novel TFBS prediction method called CNN\_TF that combines low-order and high-order dependence is reported [94].

In addition to the dependence between nucleotides, the variable binding lengths of different TFs are also

**TABLE 4.** Summary of TFBS prediction method based on deep learning.

Method	Description	Time	Technology	Reference
DeepBind	Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning	2015	CNN	[24]
DeeperBind	Enhancing the prediction of sequence specificities of DNA-binding proteins	2016	CNN-LSTM	[67]
DeepSNR	Deep learning model for predicting transcription factor binding location at single nucleotide resolution	2017	CNN-Deconv	[80]
TFImpute	Imputation for transcription factor binding predictions based on deep learning	2017	CNN	[81]
CKN	Biological sequence modeling with convolutional kernel networks	2017	Kernel + CNN	[82]
Sequence2Vec	A novel embedding approach for modeling transcription factor binding affinity landscape	2017	Markov + CNN	[83]
gkm-DNN	Probe efficient feature representation of gapped Kmer frequency vectors from sequences using deep neural networks	2018	DNN	[84]
HOCNN	High-Order Convolutional Neural Network Architecture for Predicting DNA-Protein Binding Sites	2018	High-Order + CNN	[85]
WSCNN	Weakly-Supervised Convolutional Neural Network Architecture for Predicting Protein-DNA Binding	2018	MIL + CNN	[86]
KEGRU	Recurrent Neural Network for Predicting Transcription Factor Binding Sites	2018	Word2vec + GRU	[87]
DeepGRN	Interpretable attention model in transcription factor binding site prediction with deep neural networks	2019	CNN-BiLST	[70]
MTTFsite	Cross-cell type TF binding site prediction using multitask learning	2019	CNN	[88]
WSCNN-LSTM	Modeling in vivo protein-DNA binding by combining multiple-instance learning with a hybrid deep neural network	2019	MIL + CNN-BiLSTM	[89]
DeepTF	Accurate Prediction of Transcription Factor Binding Sites by Combining Multi-scale Convolution and Long Short-Term Memory Neural Network	2019	MCNN-LSTM	[90]
DLBSS	Predicting in vitro transcription factor binding sites using DNA sequence + shape	2019	CNN	[91]
DeepRAM	Comprehensive evaluation of deep learning architectures for prediction of DNA/RNA sequence binding specificities	2019	CNN-LSTM	[92]
DESSO	Prediction of regulatory motifs from human chip-sequencing data using a deep learning framework	2019	BD + CNN	[93]
CNN_TF	Prediction of TF-binding site by inclusion of higher-order position dependencies	2019	CNN	[94]

considered in CNN. Zhang proposed [85] a high-order convolutional neural network architecture (HOCNN), to overcome the limitations of conventional CNN. HOCNN uses a high-order coding method to construct the high-order dependence between nucleotides, and a multi-scale convolutional layer to capture the pattern features of different binding lengths.

CNN usually distributes the learned topics among multiple filters, which makes the learned topic difficult to explain. In addition, the network trained on small datasets cannot be extended to new larger sets of sequences. Blum and Markus [95] introduced a circular filter, which can convolute the sequences with circular permutation variants of the same filter. They studied CNN's filters that correspond to shifting and truncating variants of real topics to activate loop filters. The circular filter not only learns the full-length pattern and explains the learned filter easily, but also improves the performance of base order reasoning in a wide range of super parameters and sequence lengths. In addition, for inferring DNA binding sites from ChIP-seq data, CNN with a circular filter is superior to conventional CNN in most cases. Chen *et al.* [82] proposed a hybrid method between the kernel method and CNN-seq method that retains the neural network to provide good representation for learning problems and defines a Hilbert space with clear characteristics to describe prediction functions.

### C. APPLICATION OF LSTM

Some studies have shown that RNN and their variants have better performance in processing time series data.

Huang *et al.* proposed a model named KEGRU, which identifies TFBS by combining a two-way gating recurrent unit (GRU) network with k-mer embedding. First, the DNA sequences are divided into k-mer sequences of specific length and span windows. Second, the word2vec algorithm is used to treat each k-mer as a word representation model. Third, a deep bidirectional GRU model is build to feature learning and classification. The robustness of KEGRU is attributed to the different lengths of the k-mer, stride window, and embedded vector size [87].

Despite its clever design, DeepBind lacks the ability to capture the dynamics of the probe sequence by indirectly assuming that there is at most one motif in each probe. This preconceived notion can mislead the training process. For instance, commonly, several moderately good motifs in the probe have high binding affinity because of their respective contributions. In this case, the TFBS prediction model tries to adjust the weight of the motif detector (kernel) so that all motifs except one are punished, while the remaining motifs are overweighed because of the high affinity of the entire probe. Even for a probe with only one pattern, its position may be important due to technical limitations. Wang *et al.* proposed DeeperBind, a novel double-deep model that can address the deficiencies of the DeepBind model. They added the location dimension to the core design of DeepBind by incorporating recursion into the model. CNN and LSTM are complementary in modeling capabilities. Therefore, it is desirable to combine these two to achieve synergistic improvement in the prediction of protein-DNA binding specificity [67].

#### D. APPLICATION OF HYBRID NEURAL NETWORK

In order to explore the impact of deep learning architecture on predicting the DNA- and RNA-binding specificity of the proteins, Trabelsi *et al.* proposed deepRAM, an end-to-end deep learning tool, which can provide the realization of multiple architectures; its fully automated model selection programs enable a fair and just comparison of deep learning architectures. After research, it was found that when the data samples are sufficient, the deeper and more complex architecture clearly has advantages, and the CNN-based RNN hybrid architecture is superior to other methods in terms of accuracy. Trabelsi *et al.* [92] provided insights into the differences between the models of convolutional networks and cyclic networks. In particular, they found that although recursive networks improve the accuracy of the model, they sacrifice the interpretability of the characteristics of model learning.

The in-depth learning method successfully simulates protein-DNA binding *in vivo*, but it usually follows a fully supervised learning framework and ignores the weak supervision information about the genome sequence. The combined DNA sequence may have multiple TFBSs that could be coded using single heat coding ((s) and (b)). The dependence between nucleotides can be ignored. Huang *et al.* proposed a weakly-supervised convolutional neural network (WSCNN) architecture that puts forward a weak supervision framework, combining multi-instance learning with a hybrid deep neural network [86]. After achieving good results, they continued to optimize the model and used k-mer code to transform DNA sequences for the *in vivo* modeling of protein-DNA binding. First, the frame uses a sliding window to segment the sequence into multiple overlapping instances, and then uses k-mer coding to encode all instances as high-order dependent image classes input. Second, it uses the hybrid deep neural network of integrated convolution and RNN to calculate the scores of all instances in the same package. It uses the noisy method to integrate the prediction values of all instances into the final prediction of the package. The improved method is termed as WSCNNLSTM [89].

#### E. APPLICATION OF OTHER METHODS

In addition to CNN, RNN, and their variants, an increasing number of efficient models that have achieved remarkable results in other fields have been applied to the prediction of TFBSs. The attention mechanism in deep learning has shown the ability to learn from long-term dependent input features. To date, this mechanism has not been applied to the deep neural network model for input data from large-scale parallel sequencing. In this study [70], the author established a model for TF binding site prediction by combining the attention mechanism with traditional deep learning techniques. The performance of the method was evaluated using the challenge dataset of TFBS prediction in ENCODE DREAM. Benchmark tests show that incorporation the focus mechanism (called deepGRN) improves the performance of the deep learning model. Visualization of attention weights extracted

from the trained models reveal the mechanisms by which these weights move when the combined signal peaks move along the genome sequence. This explains the method of prediction. Case studies show that the attention mechanism can help extract useful features by focusing on areas that are critical to successful prediction while ignoring the irrelevant signals from input information.

Abdollahyan *et al.* [96] proposed a graph-based method to detect TFBS that often occurs simultaneously with evolutionarily conserved non-coding elements (CNEs). They presented a graphical representation of the TFBS sequences recognized in CNE, which enables handling of overlapping binding sites. They used a dynamic programming algorithm to align these graphs and determine the relative enrichment of short TFBS sequences in alignment. In addition, Song *et al.* [71] proposed a new graph-based feature extraction algorithm that can accurately extract the features of TFBSs. The obtained features describe the pairwise correlation of the different positions of binding sites. Based on these characteristics, two correlations can be integrated into a statistical model, which describes the TFBS. The test results show that this method can recognize the important features of TFBSs, and that the statistical model based on these features can achieve a prediction accuracy higher than or comparable to other feature extraction methods.

We have tabulated the prediction methods, both described above and otherwise, as shown in Table 4. The results of this analysis can be used for planning and also as a guide for researchers who intend to predict TFBSs.

### VII. TFBS PREDICTION PROCESS BY DEEP LEARNING

The prediction of TFBSs mainly includes data pre-processing, feature extraction, model building, and research application. A graphical representation is shown in Figure 5, which we will introduce in detail.

#### A. DATA PRE-PROCESSING

In this study, data pre-processing is aimed at DNA sequence, protein sequence, and DNA shape analysis. Among them, sequence processing includes high-throughput data selection, data standardization, transcription factor selection, sequence truncation, and negative sample generation. In the following sections, each part has been briefly introduced.

##### 1) HIGH THROUGHPUT DATA SELECTION

It is known from the earlier discussion that high-throughput sequencing provides high-resolution TF binding datasets *in vivo* and *in vitro*. Despite the increasing number of such datasets, our ability to predict the location of TFBS on genomic DNA is still insufficient. Predicting the TFBS in a certain species needs contextual information as the basis for selection.

##### 2) DATA STANDARDIZATION

Because there are many redundant TFBSs in the forward and reverse directions of each DNA strand, we choose either

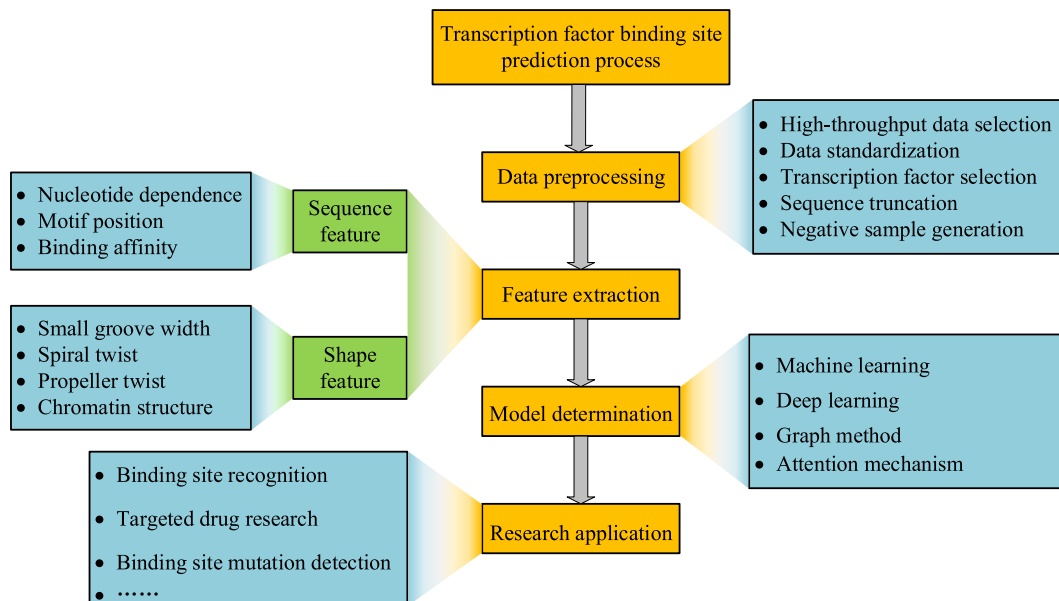


FIGURE 5. Basic flow chart for predicting transcription factor binding sites using Intelligent Computing.

the forward or the reverse chain for each TFBS. In addition, the method for generating background (negative) sequence is very important. It is well known that the background sequence must be selected to match the statistical properties of the foreground set; otherwise, the resulting motif may be inaccurate.

### 3) TRANSCRIPTION FACTOR SELECTION

There are multiple TFs in a cell, each binding to one or more chromosomes. Our research may focus on only one TF binding to a single chromosome or several TFs in a certain cell line. For example, the Sp1 factor on chromosome 1, a well-known TF family, is related to important biological processes and has a significant role in cell growth and differentiation.

### 4) SEQUENCE TRUNCATION

Many existing computational methods are either tissue-specific and species-specific or limited to short DNA sequences; therefore, they cannot be used to identify potential TFBSs in long DNA sequences without knowledge of the tissues or species. Owing to these limitations, predicting long series using the model reduces its reliability. The existing methods are divided into two categories according to the input. One truncates the sequence to a fixed length while the other accepts sequences of indefinite length and preprocesses it before sending as an input to the model.

### 5) NEGATIVE SAMPLE GENERATION

One of the difficulties in developing a calculation method for predicting TFBS is building a negative dataset. Unlike positive datasets, which are usually constructed from TFBS, negative datasets can be very unreliable. The performance of the classifier is certainly affected by the negative datasets

used in the training classifier. Several methods have been proposed for preparation of negative datasets with low expected TF binding. They are: (a) sequences not labeled as TFBS [85], (b) downstream random exons [97], (c) random selection of non-coding sequences [81], and (d) DNA regions far away from genes.

### B. FEATURE EXTRACTION

The comprehensive feature extraction is of great significance for the prediction of TFBSs. Whether it can completely extract all the features of DNA or protein sequence will directly determine the accuracy of TFBS prediction, since different features have different importance. The sequence and shape feature extraction are aimed at DNA and protein. We divided the features into sequence features and shape features. The sequence features include nucleotide dependence, phantom position, and binding affinity, while the shape features include MGW, rolling, ProT, and HelT. Owing to the lack of consideration of shape features in the existing literature, we will not expand on the description.

#### 1) NUCLEOTIDE DEPENDENCE

Identifying the interactions between regulatory proteins and DNA, especially between TFs and their corresponding binding sites, is an important step in predicting the binding sites. The dependence of nucleotide positions in TFBSs can be clearly indicated by k-mers such as dinucleotides or trinucleotides.

#### 2) MOTIF POSITION

TFBSs are usually unique, but a TFBS that does not show a one-to-one correspondence, may not be unique. The input sequence of the model usually includes the sequence of the

binding site and the sequences of the non-coding regions on both sides of the binding site. In case of high-resolution data, the boundary position can be identified by the model [80].

### 3) BINDING AFFINITY

In addition to the binding motifs, other sequence characteristics, such as low-affinity binding sites, flanking DNA, and specific targets for flanking symmetry of some repeat sequences, also affect the binding affinity of the TFs. For better TF-DNA binding specificity, it is important to also consider the length of the binding site sequence.

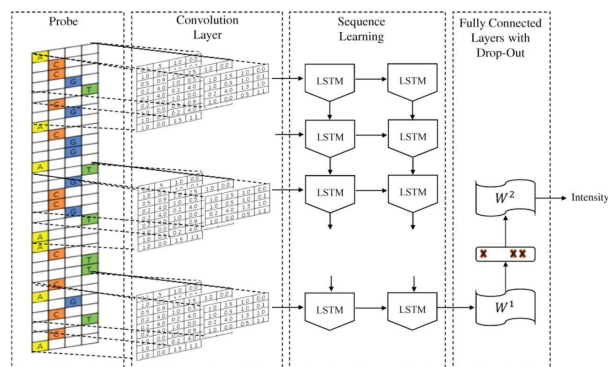
### 4) SHAPE FEATURES

An important challenge in TFBS prediction is attaining the correct modeling of the binding involving multiple TFs and local chromatin structure effects. Recent studies have shown that the interaction between TF, co-binding factors, and the local chromatin structure affects the TF-DNA binding [98]. On the other hand, because of the stacking interaction between the neighboring base pairs, a three-dimensional (3D) DNA structure is generated, and the DNA shape features represent an alternative method to implicitly code nucleotide dependence. It is known that the local structural features of the double helix (such as MGW, Roll, ProT, and HelT) greatly affect the TF-DNA interaction.

### C. MODEL ESTABLISHMENT

The task of model building begins with the input of the processed data; then, we analyze the data using the corresponding methods to determine whether there are TFBSs in the different input sequences. Finally, if a TFBS is reported, the existing site boundary or the relationship between the multiple points is determined. The overall accuracy of the model depends on the data location, accuracy of feature extraction, and selection of the learning model. To perform this task, we need various learning algorithms to integrate and analyze the data clues of gene expression. In recent years, prediction of TFBSs based on sequence data has been achieved in some studies. Among these, the methods to determine the presence of motifs are mainly based on traditional machine learning methods such as EM and support-vector machines [99], [100], and deep learning methods such as CNN and LSTM [24], [89]. Due to the abstract and sequential nature of DNA sequences, prediction method of TFBSs solely based on sequence is not suitable for complex sequence data. Scholars have, therefore, proposed hybrid methods for predicting TFBSs in complex sequence data. Examples include combination models such as CNN + LSTM [89], sequence2vec + CNN [83], word2vec + GRU [87], and attention + CNN [70]. Using the combination models, researchers can identify TFBSs in complex sequences, thereby allowing for a better understanding of the transcription process. Shown in Figure 6 is a model for predicting TFBSs based on CNN + LSTM [67].

In the first step, each probe sequence is converted into a  $4l$  one-hot coded binary matrix ( $l$  is the probe length)



**FIGURE 6.** Deeperbind block diagram. First, the input sequence is represented as a 2D binary matrix by an one-hot coding. The convolution layer generates the feature map by applying a number of pulse-width modulation (PWM) filters and rectifying linear elements. Pool layer is not used. Following this, two LSTM layer stacks capture the order dependence of the suborders on the probe [67].

and the intensity value is normalized. Next, we input the pre-processed probes into the convolution layer, followed by the corrected linear elements to map them into the intermediate feature vectors through parameterized nonlinear transformation. The pooling layer is often used in convolution network architecture; however, it is omitted to avoid losing location information in this way. Subsequently, one or two layers of LSTM are used, in which each LSTM block in the first layer receives the local features extracted from the location of interest on the DNA and encodes its interpretation of the overall contribution of the history in the hidden state. This interpretation is passed on to the next LSTM block above and to the right, and so on. Once the last nucleotide is observed, the last expanded LSTM module makes the final decision on the probe's merits and demerits based on the processed feedback from the nearest neighbor, which is the integration of all history in an attractive way. Finally, the results of the LSTM network are presented to a fully connected network, which contains at the most one hidden layer and can predict the binding preference of each probe by packet loss regularization.

### D. EVALUATION INDEX

The prediction of TFBSs mainly includes five basic evaluation indices: accuracy, precision, recall, F1 measure, and area under the receiver operating characteristic (ROC) curve (AUC). AUC metric has been widely used to rank the performance of excitation motifs in the literature. However, existing approaches for motif refinement choose to directly maximize the non-convex and discontinuous AUC itself, which is known to be difficult and may lead to suboptimal solutions. Therefore, researchers propose a method to optimize motif search based on AUC combined with deep learning technology. For example, De-Shuang Huang *et al.* propose a novel approach named Discriminative Motif Learning via AUC (DiscMLA) to discover motifs on high-throughput datasets [101] and Lin Zhu *et al.* propose Large Margin

**TABLE 5. Confusion matrix.**

Confusion matrix		Actual result	
		TRUE	FALSE
Prediction results	Positive	TP	FP
	Negative	TN	FN

Motif Optimizer (LMMO), a large-margin-type algorithm for refining regulatory motifs [102].

Confusion matrix is a situation analysis table that summarizes the prediction results of classification model in machine learning. The records in the data set are summarized in the form of matrix according to the two standards of real category and predicted category. For k-ary classification, it is actually a  $k \times k$  table to record the prediction results of the classifier. The row of the matrix represents the real value, and the column of the matrix represents the predicted value. Let's take the dichotomy as an example to see the matrix representation, as shown in the following Table 5:

where t (true) represents right, f (false) represents error, P (positive) represents 1 (positive sample), and n (negative) represents 0 (negative sample). In our task, TP represents the number of positive samples with correct classification, FP represents the number of positive samples with wrong classification, TN represents the number of negative samples with correct classification, and FN represents the number of negative samples with wrong classification.

### 1) ACCURACY

The so-called accuracy rate is the proportion of predicted correct results in the total sample. The calculation formula is as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

Accuracy, though a commonly used evaluation index, it is not a very good indicator because the results are not accurate when the samples are unbalanced. Therefore, balanced accuracy is introduced, and the calculation formula is as follows:

$$BalancedAccuracy = c * \frac{TP}{TP + FN} + (1 - c) * \frac{TN}{TN + FP} \quad (7)$$

Among them,  $c$  belongs to  $[0, 1]$ , and the value of  $c$  depends on the relative importance of sensitivity and specificity, usually  $1/2$ .

### 2) PRECISION

The so-called precision rate is the proportion of real positive samples to all positive samples in the prediction results. The calculation formula is as follows:

$$Precision = \frac{TP}{TP + FP} \quad (8)$$

The precision rate is the overall prediction evaluation, while the accuracy rate is only for local evaluation, which includes only the prediction evaluation of positive samples.

### 3) RECALL RATE

The so-called recall rate is the proportion of the number of predicted positive samples to the number of real samples that are positive. The calculation formula is as follows:

$$Recall = \frac{TP}{TP + FN} \quad (9)$$

### 4) F1 VALUE

The F1 value is based on the precision rate and recall rate, which have mutual influence and a complementary relationship. The F1 value not only considers the accuracy rate, but also the recall rate. Both these parameters are considered equally important, and F1 is expected to reach the highest value. The calculation formula is as follows:

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (10)$$

The above F1 values apply mainly for binary classification; however, when encountering multiple classification problems, we need to use macro F1 ( $F1_{macro}$ ) and micro F1 ( $F1_{micro}$ ), which can be regarded as multiple binary classification problems when dealing with multiple classification problems. F1\_macro and F1\_micro are based on the F1 value, and there are differences in the methods of calculation of the two indicators. F1\_macro calculates and averages the F1 value of each category, among which, the F1 value weight of each category is the same, while F1\_micro calculates the TP, FN, and FP as a whole and then calculates F1. The calculation formula of the two is

$$F1_{macro} = \frac{1}{C} \sum_{i=1}^C F1(i), \quad (11)$$

$$F1_{micro} = \frac{2 * \sum_{i=1}^C TP(i)}{2 * \sum_{i=1}^C TP(i) + \sum_{i=1}^C FP(i) + \sum_{i=1}^C FN(i)}, \quad (12)$$

where  $C$  is the number of categories.

### 5) ROC (RECEIVER OPERATING CHARACTERISTIC)

ROC focuses on two indicators:

- True positive rate (TPR) =  $TP / [TP + FN]$ , TPR represents the probability that positive examples can be paired.
- False positive rate (FPR) =  $FP / [FP + Tn]$ , FPR represents the probability of dividing negative cases into positive cases.

In the ROC space, the abscissa of each point is the FPR, and the ordinate is the TPR, which describes the tradeoff between the true positives and the false positives. For binary classification problems, the value of an instance is often a continuous value. By setting a threshold value, we can classify an instance into positive or negative class (for example, if the observed value is greater than the threshold value, and when the condition is vice versa, into a negative class). We can change the threshold value, classify according to

different thresholds, calculate the corresponding points in the ROC space according to the classification results, and connect these points to form the ROC curve. The ROC curve is an appropriate method to express the performance of a classifier. However, there is always the need for a number that marks the quality of the classifier. Therefore, the AUC is introduced. The value of AUC is the area below the ROC curve. In general, AUC values range from 0.5 to 1.0, with larger AUCs representing better performance.

## VIII. FACTORS INFLUENCING PROTEIN-DNA INTERACTION

Recent studies have revealed that TFs recognize a subset of putative DNA binding sites beyond the core binding site, which contributes to TF-DNA binding specificity. Several features contribute to TF-DNA readout at multiple levels, including the nucleotide sequence, 3D structure of the binding site, binding of TF-DNA with cofactors, chromatin accessibility and nucleosome occupancy, indirect co-operativity with nucleosomes, DNA methylation and so on. Additionally, interactions exist among all of these factors, which might alter the TFBS in a specific cell type. TFs can also recognize the structural features of their DNA binding sites, such as sequence-dependent DNA bending and unwinding. To fully understand the determinants of TFBS specificity in gene regulation, it is necessary to collectively understand all the factors that affect TFBS in cells [37].

### A. DNA ACCESSIBILITY

DNA accessibility *in vivo* is commonly measured through DNase I hypersensitive site sequencing (DNase-seq), formaldehyde-assisted isolation of regulatory elements (FAIRE-seq), and assay for transposase-accessible chromatin using sequencing (ATAC-seq). DNase-seq is based on the differential DNase I sensitivity of nucleosome-associated and nucleosome-free DNA. DNase I selectively cleaves DNA that is not protected by nucleosome association; therefore, accessible DNA regions manifest as DNase I-hypersensitive sites. TF binding protects DNA from cleavage by DNase I. Consequently, footprints of TF-DNA binding can be identified within hypersensitive regions.

### B. NUCLEOSOMES

Nucleosomes are the basic structural units of the chromatin and are composed of DNA and histone proteins. Each nucleosome comprises 146 base pairs of DNA wound 1.75 times around a histone octamer. Nucleosome-DNA interactions cause steric impediment to TF binding and increase the rates of TF-DNA dissociation. Consistent with this concept, most of the TFBSs identified by ENCODE consortium fall within the highly accessible DNA regions.

### C. HISTONE PROTEINS AND CHROMATIN STATUS

Histones and their variants exchange labeled promoter regions with the DNA in nucleosomes [103]. This proves that histones play an important role in the transcription process.

Histones are subjected to extensive post-translational modifications (PTMs), which regulate chromatin compaction and affect the recruitment of certain transcriptional regulators. With more than 100 possible histone PTMs and a tremendous possibility for combinatorial PTM interactions, the burgeoning field of epigenomics is rapidly defining genome-wide chromatin states (i.e., distinct combinations of histone modifications and other chromatin-associated factors at a given locus) across many cellular contexts [104]. Based on the integration of data for chromatin state and TF binding, it is observed that many TFs have specific histone PTM preferences that are consistent across multiple cell types. However, it is often unclear whether a specific chromatin state is simply permissive to TF binding, actively directs TF binding, or is a result of TF binding. Further mechanistic elucidation of the relationships between TFs and histone PTMs will likely influence our model targeting TF-DNA binding interactions.

### D. DNA 3D SHAPE

The DNA shape feature represents an alternative method of implicitly encoding nucleotide dependencies. The DNA shape integrates the complex interdependence between multiple positions of the TFBS. This integration is implemented implicitly without any explicit knowledge of personal dependencies. The combination of DNA shapes reduces the number of required parameters, while providing a convincing explanation for the mechanism that explains why dinucleotides and trinucleotides can improve the accuracy of motif descriptions [105]. Recent evidence suggests that key aspects of TF binding can be explained by the DNA shape at the selected target site [53].

### E. ENHANCER AND SNP

Enhancer is a key determinant of cell identity, and together with tissue-specific TFs, maintains gene expression patterns for a given cell type. Enhancer is a high-level complex of multiple TFs that are tightly linked to each other to regulate gene expression. Single nucleotide polymorphisms (SNPs) affect gene regulation by altering the TF-DNA binding. A recent study has shown that only a few regulatory SNPs can act through TF. The inheritance of SNPs contributes to the genetic diversity between humans and, in some cases, development of diseases. However, most disease-related SNPs are located in non-coding regions of the genome [79], including many enhancers and TFBSs [82]. However, the extent to which SNPs alter TF binding is still poorly understood.

## IX. SUMMARY AND CONCLUSION

In this review, we elaborate on the development of deep learning techniques and some state-of-the-art applications for the prediction of TFBSs. First, we introduce related research techniques and databases. Following this, we describe the development of deep learning and its applications in TFBS prediction. Next, we introduce the workflow of deep learning in the TFBS prediction problem. Finally, we conclude this article by summarizing the research trends and suggesting

directions for further improvements. Although deep learning techniques have improved the performance of TFBS prediction research, there are still significant challenges for its application in TFBS prediction data analysis.

#### A. DATA RESOLUTION AND NOISE ISSUES

An increasing number of databases have been collecting data regarding TFBSs. It is understood that these data will have problems with quality and even resolution. Different tasks have different resolution requirements. Resolution of the most common data is mostly around 10-100 bp at present. Data based on the resolution of a single base pair is relatively small, and the access method is not easy; therefore, it also limits the scope of the research. The experimental data obtained by high-throughput sequencing has the limitation of high false positives. The performance of the models is also limited by the generation of negative samples for model training. DNA is a double-stranded structure, and the current sequence data do not indicate the strand on to which the TF binds. The performance of the same model on the data complementary to the training data may be inconsistent [106].

#### B. THE DESIGN CONSIDERATIONS OF THE MODEL

Motif is a relatively conserved short fragment that remains diffusely distributed within the long DNA sequence. Identifying the position of the motif within the DNA sequence is challenging. The assumptions of existing models often limit the prediction of TFBSs. The most suited prediction model should ideally have good compatibility for both long and short sequences. Training a long sequence using a model with poor robustness will seriously affect the efficiency of the model. Models based on traditional deep learning, such as CNN and LSTM, have certain limitations on the input. Therefore, most researchers have come up with improved combination models that make up for this deficiency. Combining two technologies to develop a hybrid model enhances the compatibility and robustness of the model. Usually, there is an interaction between motifs, and the prediction ability can be improved by identifying or assisting each other. The same TF may have different TFBSs [106]. In other words, there is no one-to-one correspondence between TFs and TFBSs. There may be more than one binding site for the same TF in different tissues or a few with different structures, which greatly increases the difficulty of prediction modeling. To address these challenges, complex models have been developed, for example, combining preferences for dinucleotides and higher-order k-mers, and improving accuracy based on TF and its series. However, in many cases, this improvement is small and undetectable.

#### C. THE PROBLEM OF OBTAINING AND USING MULTI-SOURCE DATA

Data on TF are highly complex as they contain information on sequence, structural features, DNA-binding domains, and cell type. However, the factors that the prediction model can consider are limited. Since transcriptional regulation is

a highly dynamic and complex process that occurs in a cell- and tissue-specific manner, unbiased quantitative modeling of TFBS should be combined with TF to improve the predictability. This includes considering variables such as nucleosome localization, chromatin state, methylation patterns, and the 3D genome structure. All of these variables greatly affect TF binding and in a subset of these binding events, affect gene expression. Therefore, these variables should be incorporated into any model for better description of the in vivo functional TF binding and the concomitant gene regulation. Thus, it can be concluded that prediction based on the first-order sequence is, to an extent, unreliable. However, in addition to the DNA sequence, the direct sequence characteristics of protein-DNA complexes, histone proteins, chromatin accessibility, and DNA shape information are currently limited, and far from being complete. Hence, there is a lack of multi-source data for the prediction of TFBSs.

#### Key Points

1. The databases commonly used in TFBS prediction research at present are summarized.
2. The methods for using computer algorithms to establish models that help predict the TFBS, and the process of using them to predict TFBSs are elaborated.
3. The list of data sources pertaining to transcription regulation that can be used for the development of a computational model for the prediction of TFBS is collated. This will serve as a useful reference for interested researchers to access all available data sources for comprehensively understanding the research status on TFBS prediction. This study would also promote further development in this research field.
4. The limitations and challenges of research at the present stage and general directions for future development are discussed.

#### ACKNOWLEDGMENT

(Yuanqi Zeng and Meiqin Gong contributed equally to this work.)

#### REFERENCES

- [1] S. A. Lambert, A. Jolma, L. F. Campitelli, P. K. Das, Y. Yin, M. Albu, X. Chen, J. Taipale, T. R. Hughes, and M. T. Weirauch, "The human transcription factors," *Cell*, vol. 172, no. 4, pp. 650–665, Oct. 2018.
- [2] H. Imrichová, G. Hulselmans, Z. Kalender Atak, D. Potier, and S. Aerts, "I-cisTarget 2015 update: Generalized cis-regulatory enrichment analysis in human, mouse and fly," *Nucleic Acids Res.*, vol. 43, no. W1, pp. W57–W64, Apr. 2015.
- [3] O. Fornes, M. Gheorghe, P. A. Richmond, D. J. Arenillas, W. W. Wasserman, and A. Mathelier, "MANTA2, update of the mongo database for the analysis of transcription factor binding site alterations," *Sci. Data*, vol. 5, no. 1, Dec. 2018, Art. no. 180141.
- [4] S. S. Nishizaki, N. Ng, S. Dong, R. S. Porter, C. Morterud, C. Williams, C. Asman, J. A. Switzenberg, and A. P. Boyle, "Predicting the effects of SNPs on transcription factor binding affinity," *Bioinform.*, vol. 36, no. 2, pp. 364–372, 2020.



- [5] A. Poliakov, J. Foong, M. Brudno, and I. Dubchak, "GenomeVISTA—An integrated software package for whole-genome alignment and visualization," *Bioinformatics*, vol. 30, no. 18, pp. 2654–2655, May 2014.
- [6] M. Brudno, C. Do, G. Cooper, M. F. Kim, E. Davydov, E. D. Green, A. Sidow, and S. Batzoglou, "LAGAN and multi-LAGAN: Efficient tools for large-scale multiple alignment of genomic DNA," *Genome Res.*, vol. 13, no. 4, pp. 31–721, May 2003.
- [7] L. Ait, Z. Yamak, and B. Morgenstern, "DIALIGN at GOBICS—Multiple sequence alignment using various sources of external information," *Nucleic Acids Res.*, vol. 41, no. W1, pp. W3–W7, Jul. 2013.
- [8] K. Narasimhan, S. A. Lambert, A. W. Yang, J. Riddell, S. Mnaimneh, H. Zheng, M. Albu, H. S. Najafabadi, J. S. Reece-Hoyes, J. I. F. Bass, A. J. M. Walhout, M. T. Weirauch, and T. R. Hughes, "Mapping and analysis of *Caenorhabditis elegans* transcription factor sequence specificities," *Elife*, vol. 4, p. e06967, Apr. 2015.
- [9] G. D. Stormo and Y. Zhao, "Determining the specificity of protein-DNA interactions," *Nat. Rev. Genet.*, vol. 11, no. 11, pp. 751–760, Nov. 2010.
- [10] G. Badis-Breard, M. Berger, A. Philippakis, S. Talukder, A. R. Gehrke, S. A. Jaeger, E. T. Chan, G. Metzler, A. Vedenko, X. Chen, H. Kuznetsov, C.-F. Wang, D. Coburn, D. E. Newburger, Q. Morris, T. R. Hughes, and M. L. Bulyk, "Diversity and complexity in DNA recognition by transcription factors," *Science*, vol. 324, no. 5935, pp. 1720–1723, Jun. 2009.
- [11] D. Johnson, A. Mortazavi, R. Myers, and B. Wold, "Genome-wide mapping of *in vivo* protein-DNA interactions," *Science*, vol. 316, no. 5830, pp. 1497–1502, Jun. 2007.
- [12] B. J. Venters, "Insights from resolving protein-DNA interactions at near base-pair resolution," *Briefings Funct. Genomics*, vol. 17, no. 2, pp. 80–88, Dec. 2018.
- [13] Q. He, J. Johnston, and J. Zeitlinger, "ChIP-nexus enables improved detection of *in vivo* transcription factor binding footprints," *Nature Biotechnol.*, vol. 33, no. 4, pp. 395–401, Mar. 2015.
- [14] S. J. Smith, C. R. Nemr, and S. O. Kelley, "Chemistry-driven approaches for ultrasensitive nucleic acid detection," *J. Amer. Chem. Soc.*, vol. 139, no. 3, pp. 1020–1028, Jan. 2017.
- [15] A. Kiesel, C. Roth, W. Ge, M. Wess, M. Meier, and J. Söding, "The BAMB Web server for *de-novo* motif discovery and regulatory sequence analysis," *Nucleic Acids Res.*, vol. 46, no. W1, pp. W215–W220, May 2018.
- [16] A. Vidaki, C. D. López, E. Carnero-Montoro, A. Ralf, K. Ward, T. Spector, J. T. Bell, and M. Kayser, "Epigenetic discrimination of identical twins from blood under the forensic scenario," *Forensic Sci. Int. Genet.*, vol. 31, pp. 67–80, Nov. 2017.
- [17] V. Srivastava, "Gene stacking in plants through the application of site-specific recombination and nuclease activity: Methods and protocols," *Methods Mol. Biol.*, vol. 1864, pp. 267–277, Jan. 2019.
- [18] D. Subedi, A. K. Vijay, G. S. Kohli, S. A. Rice, and M. Willcox, "Nucleotide sequence analysis of NPS-1B-lactamase and a novel integron (in1427)-carrying transposon in an MDR *Pseudomonas aeruginosa* keratitis strain," *J. Antimicrobial Chemotherapy*, vol. 73, pp. 172–179, Mar. 2018.
- [19] Z. Zhang, Y. Zhao, X. Feng, Z. Luo, S. Kong, C. Zhang, A. Gong, H. Yuan, L. Cheng, and X. Wang, "Genomic, molecular evolution, and expression analysis of NOX genes in soybean (glycine max)," *Genomics*, vol. 111, no. 4, pp. 619–628, Jul. 2019, doi: [10.1016/j.ygeno.2018.03.018](https://doi.org/10.1016/j.ygeno.2018.03.018).
- [20] B. Xiang, C. Xiao, T. Shen, and X. Li, "Anti-inflammatory effects of anisalcohol on lipopolysaccharide-stimulated BV2 microglia via selective modulation of microglia polarization and down-regulation of NF- $\kappa$ B p65 and JNK activation," *Mol. Immunol.*, vol. 95, pp. 39–46, Mar. 2018, doi: [10.1016/j.molimm.2018.01.011](https://doi.org/10.1016/j.molimm.2018.01.011).
- [21] K.-Y. Ma, S.-F. Zhang, S.-S. Wang, and G.-F. Qiu, "Molecular cloning and characterization of a gonadotropin-releasing hormone receptor homolog in the Chinese mitten crab, *Eriocheir Sinensis*," *Gene*, vol. 665, pp. 111–118, Jul. 2018, doi: [10.1016/j.gene.2018.05.006](https://doi.org/10.1016/j.gene.2018.05.006).
- [22] M. Rama et al., "A decision tree for the genetic diagnosis of deficiency of adenosine deaminase 2 (DADA2): A French reference centres experience," *Eur. J. Hum. Genet.*, vol. 26, pp. 960–971, Apr. 2018.
- [23] H. Wang, P. Sham, T. Tong, and H. Pang, "Pathway-based single-cell RNA-seq classification, clustering, and construction of gene-gene interactions networks using random forests," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 6, pp. 1814–1822, Jun. 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, Aug. 2015.
- [25] M. T. Weirauch, D. Consortium, A. Cote, R. Norel, M. Annala, Y. Zhao, T. R. Riley, J. Saez-Rodriguez, T. Cokelaer, A. Vedenko, S. Talukder, H. J. Bussemaker, Q. D. Morris, M. L. Bulyk, G. Stolovitzky, and T. R. Hughes, "Evaluation of methods for modeling transcription factor sequence specificity," *Nature Biotechnol.*, vol. 31, no. 2, pp. 126–134, Feb. 2013.
- [26] B. Ren, "Genome-wide location and function of DNA binding proteins," *Science*, vol. 290, no. 5500, pp. 2306–2309, Dec. 2000.
- [27] F. Greil, C. Moorman, and B. Steensel, "DamID: Mapping of *in vivo* protein-genome interactions using tethered DNA adenine methyltransferase," *Methods Enzymol.*, vol. 410, pp. 342–359, Feb. 2006.
- [28] J. R. Hesselberth, X. Chen, Z. Zhang, P. J. Sabo, R. Sandstrom, A. P. Reynolds, R. E. Thurman, S. Neph, M. S. Kuehn, W. S. Noble, S. Fields, and J. A. Stamatoyannopoulos, "Global mapping of protein-DNA interactions *in vivo* by digital genomic footprinting," *Nature Methods*, vol. 6, no. 4, pp. 283–289, Apr. 2009.
- [29] P. G. Giresi, J. Kim, R. M. McDaniell, V. R. Iyer, and J. D. Lieb, "FAIRE (formaldehyde-assisted isolation of regulatory elements) isolates active regulatory elements from human chromatin," *Genome Res.*, vol. 17, no. 6, pp. 877–885, Jun. 2007.
- [30] J. D. Buenrostro, P. G. Giresi, L. C. Zaba, H. Y. Chang, and W. J. Greenleaf, "Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position," *Nature Methods*, vol. 10, no. 12, pp. 1213–1218, Dec. 2013.
- [31] Z. Wei, W. Zhang, H. Fang, Y. Li, and X. Wang, "EsATAC: An easy-to-use systematic pipeline for ATAC-seq data analysis," *Bioinformatics*, vol. 34, no. 15, pp. 2664–2665, Mar. 2018.
- [32] E. B. Stovner and P. Sætrom, "Epic2 efficiently finds diffuse domains in ChIP-seq data," *Bioinformatics*, vol. 35, no. 21, pp. 4392–4393, Mar. 2019.
- [33] T. W. Terooatea, A. Pozner, and B. A. Buck-Koehntop, "PATCh-cap: Input strategy for improving analysis of ChIP-exo data sets and beyond," *Nucleic Acids Res.*, vol. 44, no. 21, Aug. 2016, Art. no. gkw741.
- [34] O. Wallerman, H. Nord, M. Bysani, L. Borghini, and C. Wadelius, "LobChIP: From cells to sequencing ready ChIP libraries in a single day," *Epigenetics Chromatin*, vol. 8, no. 1, p. 25, Jul. 2015.
- [35] C. Schmidl, A. F. Rendeiro, N. C. Sheffield, and C. Bock, "ChIPmentation: Fast, robust, low-input ChIP-seq for histones and transcription factors," *Nature Methods*, vol. 12, no. 10, pp. 963–965, Aug. 2015.
- [36] C. Gustafsson, A. De Paepe, C. Schmidl, and R. Månsson, "High-throughput ChIPmentation: Freely scalable, single day ChIPseq data generation from very low cell-numbers," *BMC Genomics*, vol. 20, no. 1, p. 59, Dec. 2019.
- [37] M. Slattery, T. Zhou, L. Yang, A. C. Dantas Machado, R. Gordân, and R. Rohs, "Absence of a simple code: How transcription factors read the genome," *Trends Biochem. Sci.*, vol. 39, no. 9, pp. 381–399, Aug. 2014.
- [38] H. S. Rhee and B. F. Pugh, "Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution," *Cell*, vol. 147, no. 6, pp. 1408–1419, Dec. 2011.
- [39] D. Matteau and S. Rodrigue, "Precise identification of DNA-binding proteins genomic location by exonuclease coupled chromatin immunoprecipitation (ChIP-exo)," *Methods Mol. Biol.*, vol. 1334, pp. 173–193, Dec. 2015.
- [40] S. Barfeld and I. Mills, "Mapping protein-DNA interactions using ChIP-exo and illumina-based sequencing," *Methods Mol. Biol.*, vol. 1443, pp. 119–137, Jun. 2016.
- [41] M. Levo, E. Zalckvar, E. Sharon, A. C. Dantas Machado, Y. Kalma, M. Lotam-Pompan, A. Weinberger, Z. Yakhini, R. Rohs, and E. Segal, "Unraveling determinants of transcription factor binding outside the core binding site," *Genome Res.*, vol. 25, no. 7, pp. 1018–1029, Mar. 2015.
- [42] D. Pribnow, "Nucleotide sequence of an RNA polymerase binding site at an early t7 promoter," *Proc. Nat. Acad. Sci. USA*, vol. 72, no. 3, pp. 784–788, Mar. 1975. [Online]. Available: <https://www.pnas.org/content/72/3/784>
- [43] G. D. Stormo, "DNA binding sites: Representation and discovery," *Bioinformatics*, vol. 16, no. 1, pp. 16–23, Feb. 2000.
- [44] O. Fomes, J. A. Castro-Mondragon, A. Khan, R. van der Lee, X. Zhang, P. A. Richmond, B. P. Modi, S. Correard, M. Gheorghe, D. Baranašić, W. Santana-Garcia, G. Tan, J. Chèneby, B. Ballester, F. Parcy, A. Sandelin, B. Lenhard, W. W. Wasserman, and A. Mathelier, "JASPAR 2020: Update of the open-access database of transcription factor binding profiles," *Nucleic Acids Res.*, vol. 48, no. D1, pp. D87–D92, Nov. 2019.

- [45] M. A. Hume, L. A. Barrera, S. S. Gisselbrecht, and M. L. Bulyk, "UniPROBE, update 2015: New tools and content for the online database of protein-binding microarray data on protein-DNA interactions," *Nucleic Acids Res.*, vol. 43, no. D1, pp. D117–D122, Nov. 2015.
- [46] M. T. Weirauch et al., "Determination and inference of eukaryotic transcription factor sequence specificity," *Cell*, vol. 158, no. 6, pp. 1431–1443, Sep. 2014.
- [47] I. Kulakovskiy, I. Vorontsov, I. Yevshin, R. N. Sharipov, A. D. Fedorova, E. I. Rumynskiy, Y. A. Medvedeva, A. Magana-Mora, V. B. Bajic, and D. A. Papatsenko, "HOCOMOCO: Towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-seq analysis," *Nucleic Acids Res.*, vol. 46, no. D1, pp. D252–D259, Nov. 2017.
- [48] L. J. Zhu, R. G. Christensen, M. Kazemian, C. J. Hull, M. S. Enuameh, M. D. Basciotta, J. A. Brasefield, C. Zhu, Y. Asriyan, D. S. Lapointe, S. Sinha, S. A. Wolfe, and M. H. Brodsky, "FlyFactorSurvey: A database of drosophila transcription factor binding specificities determined using the bacterial one-hybrid system," *Nucleic Acids Res.*, vol. 39, no. 1, pp. D111–D117, Jan. 2011.
- [49] S. Shazman, H. Lee, Y. Socol, R. S. Mann, and B. Honig, "OnTheFly: A database of drosophila melanogaster transcription factors and their binding sites," *Nucleic Acids Res.*, vol. 42, pp. D167–D171, Nov. 2013.
- [50] A. T. Spivak and G. D. Stormo, "ScerTF: A comprehensive database of benchmarked position weight matrices for *saccharomyces* species," *Nucleic Acids Res.*, vol. 40, no. D1, pp. D162–D168, Jan. 2012.
- [51] A. Jolma, J. Yan, T. Whittington, J. Toivonen, K. R. Nitta, P. Rastas, E. Morgunova, M. Enge, M. Taipale, G. Wei, K. Palin, J. M. Vaquerizas, R. Vincentelli, N. M. Luscombe, T. R. Hughes, P. Lemaire, E. Ukkonen, T. Kivioja, and J. Taipale, "DNA-binding specificities of human transcription factor," *Cell*, vol. 152, nos. 1–2, pp. 327–339, Jan. 2013.
- [52] S. H. Meijnsing, M. A. Pufall, A. Y. So, D. L. Bates, L. Chen, and K. R. Yamamoto, "DNA binding site sequence directs glucocorticoid receptor structure and activity," *Science*, vol. 324, no. 5925, pp. 407–410, May 2009.
- [53] R. Rohs, S. West, A. Sosinsky, P. Liu, R. S. Mann, and B. Honig, "The role of DNA shape in protein-DNA recognition," *Nature*, vol. 461, pp. 1248–1253, Oct. 2009.
- [54] L. Yang, T. Zhou, I. Dror, A. Mathelier, W. W. Wasserman, R. Gordân, and R. Rohs, "TFBSshape: A motif database for DNA shape features of transcription factor binding sites," *Nucleic Acids Res.*, vol. 42, no. D1, pp. D148–D155, Nov. 2014.
- [55] T.-P. Chiu, B. Xin, N. Markarian, Y. Wang, and R. Rohs, "TFBSshape: An expanded motif database for DNA shape features of transcription factor binding sites," *Nucleic Acids Res.*, vol. 48, no. D1, pp. D246–D255, Dec. 2019, doi: [10.1093/nar/gkz970](https://doi.org/10.1093/nar/gkz970).
- [56] D. Eckweiler, C.-A. Dudek, J. Hartlich, D. Brötje, and D. Jahn, "PRODORIC2: The bacterial gene regulation database in 2018," *Nucleic Acids Res.*, vol. 46, no. D1, pp. D320–D326, Nov. 2018.
- [57] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4694–4702.
- [58] S. Ganapathy and V. Peddinti, "3-D CNN models for far-field multi-channel speech recognition," in *Proc. ICASSP*, Apr. 2018, pp. 5499–5503.
- [59] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1725–1732.
- [60] Z. Jizhao, J. Qiao, X. Dai, and X. Cheng, "Relation classification via target-concentrated attention CNNs," in *Proc. Int. Conf. Neural Inf. Process.*, Oct. 2017, pp. 137–146.
- [61] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 6232–6251, Oct. 2016.
- [62] X. Z. Wang, "Fuzziness based sample categorization for classifier performance improvement," *J. Intell. Fuzzy Systems*, vol. 29, no. 3, pp. 1185–1196, 2015.
- [63] Y. Zhang, G. Chen, D. Yu, K. Yao, S. Khudanpur, and J. Glass, "Highway long short-term memory RNNs for distant speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Shanghai, China, Mar. 2016, pp. 5755–5759, doi: [10.1109/ICASSP.2016.7472780](https://doi.org/10.1109/ICASSP.2016.7472780).
- [64] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE Signal Process.*, vol. 35, no. 1, pp. 53–65, Jan. 2017.
- [65] H. Yu, T. Qian, Y. Liang, and B. Liu, "AGTR: Adversarial generation of target review for rating prediction," *Data Sci. Eng.*, vol. 5, no. 4, pp. 346–359, Dec. 2020.
- [66] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proc. 1st Int. Conf. Word Represent.*, May 2013. [Online]. Available: <http://arxiv.org/abs/1301.3781>
- [67] H. R. Hassanzadeh and M. D. Wang, "DeeperBind: Enhancing prediction of sequence specificities of DNA binding proteins," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Nov. 2016, pp. 178–183.
- [68] S. Choi, H. Park, and S.-W. Hwang, "Meta-supervision for attention using counterfactual estimation," *Data Sci. Eng.*, vol. 5, no. 2, pp. 193–204, Jun. 2020.
- [69] T. Alkhouli and H. Ney, "Biasing attention-based recurrent neural networks using external alignment information," in *Proc. 2nd Conf. Mach. Transl.*, 2017, pp. 108–117. [Online]. Available: <https://www.aclweb.org/anthology/W17-4711>
- [70] C. Chen, J. Hou, X. Shi, H. Yang, J. A. Birchler, and J. Cheng, "Interpretable attention model in transcription factor binding site prediction with deep neural networks," *bioRxiv*, pp. 648–691, Jan. 2019.
- [71] Y. Song, A. Chi, and J. Qu, "A graph theoretic approach for the feature extraction of transcription factor binding sites," in *Proc. Int. Conf. Intell. Comput.*, vol. 9226, pp. 445–455, Aug. 2015.
- [72] T. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. 5th Int. Conf. Learn. Represent.*, Apr. 2017, pp. 1–14.
- [73] A. Cornish-Bowden, "Nomenclature for incompletely specified bases in nucleic acid sequences: Recommendations 1984," *Nucleic Acids Res.*, vol. 13, no. 9, pp. 3021–3030, May 1985, doi: [10.1093/nar/13.9.3021](https://doi.org/10.1093/nar/13.9.3021).
- [74] A. Isakova, R. Groux, M. Imbeault, P. Rainer, D. Alpern, R. Dainese, G. Ambrosini, D. Trono, P. Bucher, and B. Deplancke, "SMILE-seq identifies binding motifs of single and dimeric transcription factors," *Nature Methods*, vol. 14, no. 3, pp. 316–322, Mar. 2017, doi: [10.1038/nmeth.4143](https://doi.org/10.1038/nmeth.4143).
- [75] G. Pavesi, G. Mauri, and G. Pesole, "An algorithm for finding signals of unknown length in DNA sequences," *Bioinformatics*, vol. 17, pp. S207–S214, Jun. 2001.
- [76] C. E. Lawrence and A. A. Reilly, "An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences," *Proteins Struct., Function, Genet.*, vol. 7, no. 1, pp. 41–51, Jan. 1990.
- [77] A. M. Khamis, O. Motwalli, R. Oliva, B. R. Jankovic, Y. A. Medvedeva, H. Ashoor, M. Essack, X. Gao, and V. B. Bajic, "A novel method for improved accuracy of transcription factor binding site prediction," *Nucleic Acids Res.*, vol. 46, no. 12, pp. e72–e72, Apr. 2018.
- [78] Z. Gao and J. Ruan, "Computational modeling of *in vivo* and *in vitro* protein-DNA interactions by multiple instance learning," *Bioinformatics*, vol. 33, no. 14, pp. 2097–2105, Mar. 2017.
- [79] X. Pan and H.-B. Shen, "RNA-protein binding motifs mining with a new hybrid deep learning based cross-domain knowledge integration approach," *BMC Bioinf.*, vol. 18, no. 1, p. 136, Dec. 2017.
- [80] S. Salekin, J. M. Zhang, and Y. Huang, "Base-pair resolution detection of transcription factor binding site by deep deconvolutional network," *Bioinformatics*, vol. 34, no. 20, pp. 3446–3453, Feb. 2018.
- [81] Q. Qin and J. Feng, "Imputation for transcription factor binding predictions based on deep learning," *PLOS Comput. Biol.*, vol. 13, no. 2, Feb. 2017, Art. no. e1005403.
- [82] D. Chen, L. Jacob, and J. Mairal, "Biological sequence modeling with convolutional kernel networks," *Bioinformatics*, vol. 35, no. 18, pp. 3294–3302, Sep. 2019.
- [83] H. Dai, R. Umarov, H. Kuwahara, Y. Li, L. Song, and X. Gao, "Sequence2Vec: A novel embedding approach for modeling transcription factor binding affinity landscape," *Bioinformatics*, vol. 33, no. 22, pp. 3575–3583, Jul. 2017.
- [84] Z. Cao and S. Zhang, "Probe efficient feature representation of gapped K-mer frequency vectors from sequences using deep neural networks," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 17, no. 2, pp. 657–667, Mar. 2020.
- [85] Q. Zhang, L. Zhu, and D.-S. Huang, "High-order convolutional neural network architecture for predicting DNA-protein binding sites," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 16, no. 4, pp. 1184–1192, Jul. 2019.

- [86] Q. Zhang, L. Zhu, W. Bao, and D.-S. Huang, "Weakly-supervised convolutional neural network architecture for predicting protein-DNA binding," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 17, no. 2, pp. 679–689, Apr. 2020.
- [87] Z. Shen, W. Bao, and D.-S. Huang, "Recurrent neural network for predicting transcription factor binding sites," *Sci. Rep.*, vol. 8, no. 1, p. 15270, Oct. 2018.
- [88] J. Zhou, Q. Lu, L. Gui, R. Xu, Y. Long, and H. Wang, "MTTFsite: Cross-cell type TF binding site prediction by using multi-task learning," *Bioinformatics*, vol. 35, no. 24, pp. 5067–5077, Dec. 2019.
- [89] Q. Zhang, Z. Shen, and D.-S. Huang, "Modeling *in-vivo* protein-DNA binding by combining multiple-instance learning with a hybrid deep neural network," *Sci. Rep.*, vol. 9, no. 1, p. 8484, Dec. 2019.
- [90] X.-R. Bao, Z. Yi-Heng, and D.-J. Yu, "DeepTF: Accurate prediction of transcription factor binding sites by combining multi-scale convolution and long short-term memory neural network," in *Proc. Int. Conf. Intell. Sci. Big Data Eng.*, Nov. 2019, pp. 126–138.
- [91] 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.*, early access, Oct. 15, 2019, doi: [10.1109/TCBB.2019.2947461](https://doi.org/10.1109/TCBB.2019.2947461).
- [92] A. Trabelsi, M. Chaabane, and A. Ben-Hur, "Comprehensive evaluation of deep learning architectures for prediction of DNA/RNA sequence binding specificities," *Bioinformatics*, vol. 35, no. 14, pp. i269–i277, Jul. 2019, doi: [10.1093/bioinformatics/btz339](https://doi.org/10.1093/bioinformatics/btz339).
- [93] J. Yang, A. Ma, A. D. Hoppe, C. Wang, Y. Li, C. Zhang, Y. Wang, B. Liu, and Q. Ma, "Prediction of regulatory motifs from human ChIP-seq data using a deep learning framework," *Nucleic Acids Res.*, vol. 47, no. 15, pp. 7809–7824, Aug. 2019.
- [94] J. Zhou, Q. Lu, R. Xu, L. Gui, and H. Wang, "Prediction of tf-binding site by inclusion of higher order position dependencies," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 17, no. 4, pp. 1383–1393, Jul./Aug. 2020.
- [95] C. F. Blum and M. Kollmann, "Neural networks with circular filters enable data efficient inference of sequence motifs," *Bioinformatics*, vol. 35, no. 20, pp. 3937–3943, Oct. 2019.
- [96] M. Abdollahyan, G. Elgar, and F. Smeraldi, "Identifying potential regulatory elements by transcription factor binding site alignment using partial order graphs," *Int. J. Found. Comput. Sci.*, vol. 29, no. 08, pp. 1345–1354, Dec. 2018.
- [97] D. Hombach, J. M. Schwarz, P. N. Robinson, M. Schuelke, and D. Seelow, "A systematic, large-scale comparison of transcription factor binding site models," *BMC Genomics*, vol. 17, no. 1, p. 388, Dec. 2016.
- [98] Z. Shen, S.-P. Deng, and D.-S. Huang, "Capsule network for predicting RNA-protein binding preferences using hybrid feature," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 17, no. 5, pp. 1483–1492, Sep. 2020.
- [99] W. Lee, B. Park, and K. Han, "Sequence-based prediction of putative transcription factor binding sites in DNA sequences of any length," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 15, no. 5, pp. 1461–1469, Nov. 2018.
- [100] A. Yousefian-Jazi and J. Choi, "Sequential integration of fuzzy clustering and expectation maximization for transcription factor binding site identification," *J. Comput. Biol.*, vol. 25, no. 11, pp. 1247–1256, Nov. 2018.
- [101] H. Zhang, L. Zhu, and D.-S. Huang, "DiscMLA: An efficient discriminative motif learning algorithm over high-throughput datasets," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 15, no. 6, pp. 1810–1820, Nov. 2018.
- [102] L. Zhu, H.-B. Zhang, and D.-S. Huang, "LMMO: A large margin approach for refining regulatory motifs," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 15, no. 3, pp. 913–925, May 2018.
- [103] W. Xu, L. Zhu, and D.-S. Huang, "DCDE: An efficient deep convolutional divergence encoding method for human promoter recognition," *IEEE Trans. Nanobiosci.*, vol. 18, no. 2, pp. 136–145, Apr. 2019.
- [104] J. Ernst and M. Kellis, "Interplay between chromatin state, regulator binding, and regulatory motifs in six human cell types," *Genome Res.*, vol. 23, no. 7, pp. 1142–1154, Jul. 2013.
- [105] T. Zhou, N. Shen, L. Yang, N. Abe, J. Horton, R. S. Mann, H. J. Bussemaker, R. Gordán, and R. Rohs, "Quantitative modeling of transcription factor binding specificities using DNA shape," *Proc. Nat. Acad. Sci. USA*, vol. 112, no. 15, pp. 4654–4659, Mar. 2015.
- [106] A. Shrikumar, P. Greenside, and A. Kundaje, "Reverse-complement parameter sharing improves deep learning models for genomics," *bioRxiv*, pp. 1–9, Jan. 2017, doi: [10.1101/103663](https://doi.org/10.1101/103663).



**YUANQI ZENG** received the B.S. degree from the School of Computer Science, Chengdu University of Information Technology, Chengdu, China, in 2018. He is currently pursuing the master's degree in computer science with the Chengdu University of Information Technology. His research interests include bioinformatics and deep learning algorithms.



**MEIQIN GONG** received the master's degree from the West China Hospital, Sichuan University, Chengdu, China, in 2015. She is currently working with the West China Second University Hospital, Sichuan University. Her research interests include obstetrics, gynecology, and bioinformatics.



**MENG LIN** received the B.S. degree from the School of Computer Science, Chengdu University of Information Technology, Chengdu, China, in 2019. He is currently pursuing the master's degree in computer science with the Chengdu University of Information Technology. His research interests include bioinformatics and deep learning algorithms.



**DONGRUI GAO** received the Ph.D. degree from the University of Electronic Science and Technology, Chengdu, China, in 2016. He is currently a Lecturer with the School of Computer Science, Chengdu University of Information and Technology, Chengdu. His research interests include signal processing and neural regulation.



**YONGQING ZHANG** received the Ph.D. degree from Sichuan University, Chengdu, China, in 2016. He is currently an Associate Professor with the School of Computer Science, Chengdu University of Information and Technology, Chengdu. His research interests include machine learning and bioinformatics.

...