


A Comprehensive Review for methods in Transcription Factor Binding Site Prediction

Xinhui Du, Yiheng Zhu, , and Zhiwei Ji 

Corresponding author. Zhiwei Ji, College of Artificial Intelligence, Nanjing Agricultural University, No. 1 Weigang Rd., Nanjing, Jiangsu 210095, China.

E-mail: Zhiwei.Ji@njau.edu.cn

Abstract

Transcription factors are key regulatory proteins in gene expression control, and transcription factor binding sites are specific locations where transcription factors uniquely recognize DNA sequences. Transcription factor binding sites can be detected through various experimental methods, but can also be predicted computationally. However, experimental methods are often expensive and time-consuming, which is why computational predictions of transcription factor binding sites have rapidly developed over the past few decades. Recently, with the advancement of deep learning technologies, the accuracy of predicting transcription factor binding sites has significantly improved. In this article, we review the key computational methods for identifying transcription factor binding sites developed over the decades, the databases used, and perform experimental evaluations of selected algorithms. We also discuss the future prospects and directions of this field.

Keywords:transcription factor binding sites; deep neural network; machine learning; motifs

1 Introduction

Transcription factors are proteins within cells whose main function is to regulate gene expression. They do this by recognizing and binding to specific DNA sequences through their DNA-binding domains, such as zinc fingers, helix-turn-helix, and leucine zippers. These sequences, generally 5 to 20 nucleotides long, are known as transcription factor binding sites[1,2]. Besides the DNA-binding domain, transcription factors also contain transcription regulatory domains, which can activate or inhibit gene expression. These functions allow transcription factors to play a crucial role in cellular communication networks and respond to various biological signals during processes like cell proliferation and differentiation[3].

Transcription factor binding sites are typically found in the promoter regions of genes but can also exist in enhancers and silencers. These sites consist of short, specific nucleotide sequences known as motifs, which are highly conserved and determine the binding affinity and specificity of transcription factors. Each transcription factor usually recognizes a specific motif, using it to locate the correct DNA regions[4,5]. By interacting with the DNA-binding domains of transcription factors, these motifs regulate gene expression. This interaction can directly regulate the activity of nearby genes or indirectly by affecting chromatin structure, allowing cells to adapt to various physiological and environmental changes.

The main experimental methods for studying transcription factor binding sites encompass a variety of techniques, each tailored to specific research needs. Electrophoretic mobility shift assays (EMSA) provide an intuitive way to observe the interaction between transcription factors and DNA. Chromatin immunoprecipitation (ChIP) and its derivative technique, ChIP-Seq, utilize specific antibodies to capture transcription factor-DNA complexes and identify their precise locations at the cellular level through high-throughput sequencing. DNA footprinting techniques use DNase I enzymes to reveal areas of DNA protected by proteins, thereby accurately mapping the binding sites of transcription factors. Reporter gene analysis evaluates the impact of specific DNA sequences on reporter gene expression to assess the regulatory activity of transcription factors. Finally, the SELEX technique involves screening a vast array of random sequences to identify DNA sequences that have high affinity for specific transcription factors[6-9]. These methods each have their unique characteristics and are often used in combination to thoroughly analyze and deeply understand how transcription factors regulate gene expression, revealing their role in cellular functions.

Due to the high cost, lengthy timeframes, and technical complexity of experimental methods for determining transcription factor binding sites, computational methods have been developed for prediction. Early attempts to use computational approaches for predicting protein-DNA interactions include the use of simple machine learning algorithms, like the perceptron, by Stormo and others to analyze and predict translation initiation sites in *E. coli*[10-14]. This marked an important shift from laboratory experiments to computational predictions, laying the groundwork for later predictions of transcription factor binding sites.

Over the past decades, computational methods for predicting transcription factor binding sites have evolved significantly. From initial statistical learning methods like position-specific scoring matrices (PSSM) and consensus sequences[15,16], which assume that each base in the genome independently contributes to binding, to the use of machine learning techniques that can efficiently handle large datasets and identify patterns, speeding up research and improving prediction accuracy. The advent and development of deep learning techniques have significantly enhanced the accuracy and efficiency of predictions for transcription factor binding sites. Deep learning models automate the process of learning

sequence motifs and complex internal connections at binding sites, not only improving prediction accuracy but also revealing previously unknown DNA sequence regulatory features, such as chromatin structure and DNA shape.

This paper's main contributions are as follows: (1) Classifying transcription factor binding site prediction algorithms based on the technology used, introducing the development of algorithms over time within each major category, and discussing the advantages and disadvantages of each algorithm. (2) Introducing the basic databases used. (3) Drawing conclusions about the content of the paper and discussing the future development directions and challenges of transcription factor binding site prediction.

2 Methods for Predicting TF Binding Sites

2.1 Traditional Statistical Learning Methods

In recent decades, there has been substantial advancement and widespread progress in traditional statistical learning techniques used to predict transcription factor binding sites (TFBS). The primary approaches employed in this study encompass consensus sequences, position-specific scoring matrices (PSSM), and methodologies that compute the mean number of nucleotide matches between a probable site and all established sites[17]. Despite the existence of research indicating interdependent effects between bases, these foundational methods are based on the assumption that each base pair's contribution to binding is independent. Despite this assumption, they provide a natural extension by considering pairwise nucleotide dependencies and per-position information content.

The initial statistical learning methods, although demonstrated to be practical and offering a satisfactory estimation of the energetics involved in DNA-protein binding, have the potential for further enhancement through the incorporation of pairwise correlations and the utilization of information content at each site. The utilization of pairwise correlations and information content has demonstrated its utility in the representation and visualization of binding sites, as well as in the identification of motifs. The principles introduced in MatInspector and MATCHM were subsequently employed to improve the accuracy of predictions, with a particular focus on differentiating between conserved and non-conserved sections within sequences[18,19].

Over the course of time, tools such as COTRASIF and TFinder were created, which integrate position weight matrix search techniques and hidden Markov models, providing search interfaces that are easy for users to navigate[20,21]. Phylogenetic footprinting approaches have become more popular because to the growing availability of whole-genome data and improved computing capabilities. These methodologies facilitate the identification of functional elements through the comparison of conserved non-coding DNA sequences across several species, hence opening up novel avenues for the detection of regulatory elements that span multiple species[22].

To summarize, the progression from using traditional statistical techniques such as consensus sequences and PSSM to more advanced methods that take into account nucleotide interdependencies and information content, and ultimately to whole-genome analysis using phylogenetic footprinting, has greatly enhanced the accuracy of predictions and created new opportunities for identifying regulatory elements across different species. Notwithstanding advancements, disparities in experimental validation underscore the

significance of additional verification, indicating potential avenues for future research. These include exploring the optimal integration of conventional statistical learning approaches with contemporary machine learning and deep learning methodologies to elucidate intricate gene regulatory networks.

2.2 Machine Learning Methods

Comprehending the mechanism by which transcription factors identify and attach to particular DNA sequences is of utmost importance in contemporary biology, as it directly impacts the regulation of genes. The proliferation of biological data has led to a significant rise in the time and cost associated with conventional experimental methods, hence necessitating the exploration of more efficient alternatives. Machine learning technology has emerged as a potent tool for the prediction of transcription factor binding sites (TFBS) due to its capacity to effectively process large volumes of data and identify recurring trends. These strategies have the dual effect of expediting the research process and improving the accuracy of predictions, so providing novel avenues for uncovering the mechanisms behind gene expression control. The following will introduce key machine learning algorithms used in recent years for TFBS prediction and their development.

Machine learning techniques have made substantial advancements in the domain of TFBS prediction in recent years. The gkmSVM approach, introduced in 2016, greatly improved prediction efficiency and accuracy by utilizing an upgraded gap k-mer support vector machine classifier[23]. This algorithm is particularly effective in processing intricate regulatory DNA sequences. Mathelier's team subsequently proposed a method that relies on DNA shape features. This method effectively enhanced prediction accuracy by integrating DNA sequence information and shape features. This finding highlights the crucial role of shape features in transcription factor recognition, extending beyond the scope of sequence information[24]. The Mocap method was introduced in the same year, employing a sparse logistic regression model to incorporate genomic characteristics and sequence context information. This approach offers a robust tool for making predictions across various cell types and transcription factor circumstances[25].

In 2017, the TEPIC method demonstrated its ability to accurately predict gene expression by integrating transcription factor affinity and open chromatin signals. This achievement highlights the promise of machine learning techniques in maintaining prediction accuracy while simultaneously reducing experimental expenses[26]. The prediction accuracy and specificity of the DRAF algorithm were enhanced in 2018 through the integration of physicochemical properties of transcription factor DNA binding domains with target DNA sequence information. This integration was achieved by employing a random forest model and highlighting the significance of biophysical properties in augmenting the performance of machine learning models[27].

In 2019, the "Anchor" method shown notable advancements in enhancing the accuracy and generalization capabilities of TFBS prediction. This progress was particularly evident in its ability to effectively handle cell-type-specific information. These improvements were achieved by the integration of a comprehensive feature set and precise preprocessing of DNase-seq data[28]. In the aforementioned year, the TEPIC2 algorithm demonstrated its efficacy and precision as an analytical framework by including epigenetic data with TFBS predictions. This advancement underscores the significance of epigenetic information in enhancing TFBS predictions[29]. The Catchitt method demonstrated exceptional

performance in predicting cell-type-specific Transcription Factor Binding Sites (TFBS) by effectively integrating features and employing advanced machine learning techniques[30]. It achieved joint first place in the "ENCODE-DREAM in vivo Transcription Factor Binding Site Prediction Challenge."

Subsequently, machine learning techniques continued to advance, exemplified by the MEDEMO in 2020[31]. This advancement greatly improved the accuracy of TFBS prediction by using DNA methylation data and capturing intramotif relationships. The accuracy of plant TFBS predictions was greatly enhanced by the Wimtrap technique by 2022 by the integration of ChIP-seq data and diverse genomic characteristics[32]. Additionally, the model's transferability across numerous transcription factors, organs, and species was tested. In 2023, MachineTFBS made notable advancements in enhancing the accuracy of TFBS predictions in yeast promoter regions. This was achieved through the utilization of personalized model selection, which involved selecting the most suitable machine learning model and feature combination for each specific transcription factor dataset[33]. The findings of this study highlight the potential of personalized model selection in improving prediction accuracy.

The utilization of machine learning in TFBS prediction has demonstrated significant promise and a wide range of applications. Through the utilization of sophisticated algorithms and the integration of diverse biological data, scholars have successfully identified patterns within the intricate nature of gene regulation processes and made predictions regarding the specific locations where transcription factors exert their effects inside the genome[34-36]. Each methodology has demonstrated its respective merits in distinct domains, including the efficacy of gkmSVM and Mocap in handling intricate sequence data, the ingenuity of TEPIC and DRAF in amalgamating transcription factor affinity and biophysical characteristics, and the advancements of "Anchor" and TEPIC2 in integrating DNase-seq data and epigenetic information to augment the precision of predictions. Significantly, throughout time, there has been a growing emphasis on enhancing accuracy, algorithm generalization capabilities, data utilization, and computing efficiency in machine learning approaches for TFBS prediction. The Catchitt algorithm's triumph in global competitions has demonstrated the sophisticated capabilities of machine learning methods in integrating features and designing models. MEDEMO has underscored the significance of epigenetic data in improving prediction accuracy by taking into account the influence of DNA methylation. Additionally, Wimtrap and MachineTFBS have achieved noteworthy outcomes in predicting plant genome and yeast promoter regions, highlighting the potential and difficulties associated with cross-species and cross-cell type predictions.

In conclusion, machine learning techniques have become essential instruments in the field of TFBS prediction research, consistently advancing to accommodate the intricacy and abundance of biological data. It is anticipated that future investigations will delve into novel machine learning models and methodologies in order to enhance the precision, effectiveness, and comprehensibility of predictions. This will contribute to a more profound comprehension of the intricate mechanisms behind gene regulation networks.

2.3 Deep Learning Methods

2.3.1 Deep Learning Methods without Pre-training Phase

In transcription factor binding site (TFBS) prediction research, the introduction of deep learning technology has shown significant progress and potential. With the development of various deep learning models, these technologies have demonstrated exceptional abilities in capturing complex DNA and RNA

sequence patterns as well as cell-specific information. Since 2015, for example, the DEEP algorithm significantly improved enhancer prediction accuracy by integrating multiple types of biological data, showcasing its ability to capture enhancer activity under varying cell conditions[37]. That same year, the DeepBind algorithm used deep learning to automatically learn patterns in DNA sequences, significantly enhancing the prediction accuracy of sequence specificity for binding proteins, although its model interpretability still needs improvement[38]. Following this, the DeepSEA model used large-scale chromatin profiling data to predict the effects of non-coding variations at single-nucleotide resolution, demonstrating higher prediction performance than traditional methods like gkm-SVM[39]. By 2016, the PEDLA and DeMo algorithms integrated more heterogeneous data and used deep network architectures, not only improving consistency in predictions across cell types but also achieving higher accuracy and better model interpretability in TFBS prediction tasks[40]. Additionally, the Leopard algorithm, with its multi-to-multi neural network architecture, surpassed traditional methods like Anchor and FactorNet, achieving significant performance improvements in TFBS prediction, particularly in cell-type-specific predictions, contributing significantly to the advancement of TFBS prediction technology and understanding of gene regulatory networks.

2016 marked a significant developmental milestone for deep learning in the TFBS prediction field, with numerous innovative algorithms such as DeepBind, DanQ, DeepSEA, Basset, and its evolved version Basenji, significantly enhancing prediction accuracy. Building on DeepBind, DanQ introduced convolutional neural networks (CNN) and bidirectional long short-term memory networks (BiLSTM), further enhancing the model's ability to capture regulatory motifs and their long-distance dependencies in sequences[41]. Following closely, Basset, by utilizing deep convolutional networks, not only enhanced the prediction performance for transcription factor binding sites but also made breakthroughs in predicting cell-specific DNA accessibility[42-45].

Entering 2017, Basenji, as an evolution of Basset, through extending the processed sequence length and introducing dilated convolution techniques, captured distal regulatory elements and predicted gene expression features more finely, marking a gradual conquest of more complex biological problems[46-48]. Basenji's improvements, especially in enhancing the resolution of distal regulatory elements, demonstrated the ongoing evolution of deep learning in the biological prediction domain. Deep learning technology continued to evolve, bringing new algorithms such as CKN-seq, KEGRU, and DECRES, which through integrating various computational strategies and neural network technologies, further enhanced learning efficiency, stability, and prediction performance[49-51]. Although these algorithms face challenges in computational resource consumption, data dependency, and model interpretability, future research directions include simplifying model structures, optimizing computational efficiency, and enhancing generalizability to achieve higher application value and practicality.

In 2019, deep learning applications in the TFBS prediction field were further strengthened and innovated, leading to significant technological developments. WSCNN, by introducing multi-instance learning and considering reverse complementary sequences, effectively utilized weakly supervised information in DNA sequences, improving prediction accuracy and efficiency[52]. MTTFsite, by combining a multi-task learning framework and deep learning technology, significantly improved cross-cell type predictions of transcription factor binding sites, especially by effectively integrating shared and cell-type-specific biological characteristics[53]. The HOCNN algorithm, by introducing high-order

encoding and multi-scale convolution layers, effectively captured complex dependencies between nucleotides and different scale motifs, significantly enhancing TFBS prediction performance. However, the model's parameter count exponentially increased with the degree of complexity, leading to overfitting and reduced computational efficiency[54]. To overcome these challenges, future improvement directions include developing effective parameter management techniques and exploring coding methods that balance sequence representation richness with computational feasibility. That same year, the DESSO algorithm integrated DNA sequence and shape information, achieving high accuracy in predicting transcription factor binding patterns through deep neural networks, surpassing existing tools like DeepBind[55]. Despite high computational resource demands and insufficient model interpretability. Additionally, circular filters architecture introduced filters in convolutional neural networks capable of capturing cyclically arranged variants in sequences, not only enhancing prediction accuracy and data utilization efficiency but also improving model interpretability[56-60]. Facing increased computational complexity and the need for hyperparameter tuning, future directions include exploring more efficient computational methods and automating hyperparameter optimization.

Entering 2020, the DeepSite algorithm, with its innovative network architecture, made significant progress in predicting transcription factor binding sites[61]. Concurrently, TBiNet, by integrating attention mechanisms, not only enhanced the accuracy of transcription factor-DNA binding site predictions but also increased the model's interpretability[62]. In 2021, the Attention-Enhanced Convolutional Neural Network (ACNN) by incorporating attention mechanisms captured global and local information in DNA sequences, significantly improving cross-cell type prediction performance[63]. Additionally, eDeepCNN, by merging DNA sequence and shape information and employing a spatial alignment strategy, significantly enhanced prediction accuracy. Facing high computational complexity and reliance on high-quality data, future directions may include optimizing model structures to reduce computational burdens[64].

In 2021, the field of TFBS prediction witnessed breakthroughs in multiple algorithms, including HSEDC, AgentBind, BpNet, DLBSS, DeepATT, and DeepD2V, each significantly enhancing prediction performance and computational efficiency[65-70]. HSEDC innovatively merged DNA sequence and shape information through spatial alignment and used multi-layer convolutional neural networks and embedding strategies to significantly enhance prediction accuracy. Although this method depends on high-quality DNA shape information and substantial computational resources, it highlighted how model simplification and data augmentation could further optimize generalizability and handle heterogeneous data distributions. AgentBind combined pre-training and fine-tuning strategies of deep learning models with model interpretation technologies like Grad-CAM, not only enhancing model prediction accuracy but also its interpretability. Facing challenges in pre-training data selection and high computational resource demands, AgentBind demonstrated the necessity for future enhancements in model architecture to reduce resource consumption, expand and diversify pre-training datasets, and develop more advanced model interpretation tools. BpNet, with its deep convolutional neural network architecture, made significant advancements in the precision and resolution of TFBS predictions, directly predicting transcription factor binding profiles from DNA sequences and using model interpretation tools to reveal complex rules of transcription factor binding. Although BpNet made significant progress, its dependence on high-quality data and high computational resource demands are issues that need future resolution. DLBSS, by combining deep learning with an integrated analysis of DNA sequence and shape features, enhanced prediction accuracy,

especially through a shared convolutional neural network model that effectively captured common patterns between sequences and shapes. DeepATT, by merging convolutional and recurrent neural networks and introducing a category attention layer, significantly enhanced the prediction accuracy of DNA sequence functional effects. Moreover, DeepD2V, by introducing dna2vec's k-mer distributed representation and considering various variants of DNA sequences, along with a hybrid model architecture combining convolutional and bidirectional long short-term memory networks, achieved significant improvements in performance. Facing challenges in handling sequences of varying lengths and not fully utilizing biological information beyond sequence data, DeepD2V looked forward to developing feature extraction algorithms adapted to variable-length sequences and integrating more types of biological features.

In the field of TFBS prediction, significant scientific progress was made in 2021, with multiple algorithms significantly enhancing prediction accuracy and efficiency through innovative deep learning architectures. The CRPTS algorithm, by combining CNNs and RNNs, along with the fusion of DNA sequence and shape features, enhanced prediction performance[71]. Concurrently, the Multi-Scale Capsule Network (MSC) by combining convolutional and capsule networks, improved the capability to capture sequences of various lengths, although it had high computational demands[72]. The Fully Convolutional Network (FCNA), through a fully convolutional architecture, achieved precise TFBS predictions at the nucleotide level, highlighting the need for improved data imbalance handling strategies[73]. Additionally, SAResNet, by integrating self-attention mechanisms and residual networks, improved long-distance sequence dependency handling and model convergence speed, showcasing potential directions for optimizing model adaptability and interpretability[74]. CAE-CNN, by combining a convolutional autoencoder with a CNN, reduced the impact of negative sample noise and accelerated the training process through unsupervised pre-training, despite issues with model interpretability and data dependency[75-77].

In TFBS prediction research, although multiple algorithms demonstrated excellent performance, they also revealed several key areas for improvement. First, computational efficiency is a common concern, especially for algorithms employing complex neural network structures, such as the Multi-Scale Capsule Network (MSC) and the Fully Convolutional Network (FCNA). Future research might explore more efficient network architectures and algorithm simplification strategies, for instance, by using pruning techniques to reduce unnecessary computations and parameters, or by leveraging the latest hardware acceleration technologies to enhance computational speed. Second, model interpretability is a significant challenge for deep learning applications in bioinformatics. While some algorithms, such as AgentBind and SAResNet, have begun to integrate model interpretation tools like Grad-CAM and self-attention mechanisms to reveal the biological principles behind model decisions, these methods still require further improvement. Future research could develop new interpretability methods, for example, by integrating more interpretable machine learning technologies or developing new visualization tools to help researchers better understand and validate the biological significance of model predictions. Additionally, data dependency is an issue faced by many high-performance algorithms. The performance of these algorithms typically relies heavily on the quality and quantity of training data, which limits the models' generalizability. To address this issue, future research might employ multi-task learning, transfer learning, or data augmentation techniques to enhance the robustness and generalizability of models. For instance, by training models to recognize transcription factor binding sites under various biological conditions or by using synthetic data augmentation techniques to expand the diversity of training samples.

Researchers are gradually optimizing deep learning models to meet the complex demands of TFBS prediction. For example, the D-SSCA model, by combining DNA sequence and shape information and utilizing attention mechanisms, effectively enhanced prediction accuracy[78]. This integration strategy not only improved model performance but also enhanced its interpretability, allowing researchers to better understand the biological mechanisms behind the predictions. However, this model and similar ones like DeepARC and GHTNet, typically depend heavily on high-quality training data and consume substantial computational resources. To address these challenges, the research community is exploring how to reduce these dependencies through algorithm optimization and data handling strategies[79,80]. For instance, MAResNet, by introducing multi-scale attention mechanisms and residual networks, effectively enhanced the model's generalizability across different biological samples[81]. Additionally, the development of PCLAtt and TranAtt models, by combining CNNs, BiLSTMs, and attention mechanisms, emphasized the balance between model interpretability and computational efficiency[82]. As the demand for model generalizability and interpretability increases, future research trends might include further model lightweighting and the adoption of multi-task learning and transfer learning strategies. These strategies not only facilitate the application of models in various biological environments but also, through data augmentation and multimodal learning methods like those demonstrated by DeepGenBind, handle more complex data types, thereby enhancing the robustness and application range of models[83]. Additionally, as single-cell sequencing technology advances, as demonstrated by the STAPLE model, the processing and analysis of single-cell data have become a new research hotspot for deep learning in TFBS prediction[84]. The sparsity and complexity of these data require models not only to have efficient computational capabilities but also to be able to extract useful information from extremely limited data.

In 2022, the U-TransNet model employed meta-learning strategies and U-Net architecture, combined with various chromatin features, to predict how variations affect TF-DNA binding[85]. This method excelled in predicting the cell-type-specific effects of models, but the complexity of the model led to high computational resource demands. Following this, the DSAC model, by integrating self-attention mechanisms and a dual-branch CNN architecture, achieved complementary extraction of global information and local features, significantly enhancing the accuracy of TFBS predictions[86]. Subsequently, DeepSTF, by combining DNA sequence and shape information and using an improved Transformer encoder and CNN-BiLSTM structure, enhanced the accuracy and interpretability of predictions[87]. This demonstrated the potential of combining advanced encoding technologies with classic neural network elements. However, the high computational resource demands and complexity of DeepSTF may limit its widespread application. Therefore, future research might explore further optimization of the model structure to reduce resource demands and enhance model interpretability.

In 2023, the TFBSnet model, by combining CNNs and Selective Kernel Networks (SKNet), utilized diverse feature data from DNA sequences to enhance prediction accuracy. Although this method demonstrated excellent performance in predicting TFBS in human and plant cells, the high demand for computational resources and the complexity of the model, as well as issues with interpretability, indicate directions for future improvements, including optimizing the model structure to reduce resource demands, enhancing model interpretability, and exploring more biological factors to improve the generalizability of predictions[88-90]. TBCA, by combining convolutional and lightweight attention mechanisms, along with Fourier-transform-enhanced multi-head attention and channel attention, significantly enhanced the

accuracy and interpretability of DNA sequence-based TFBS predictions[91].

In 2024, the NLDNN model, by developing a nucleotide-level deep neural network architecture and using an adversarial training framework, significantly enhanced the predictive performance of cross-species TFBS, especially by directly predicting experimental coverage values as a nucleotide-level regression task, enhancing the model's detail-level prediction capabilities. However, NLDNN faces challenges including high demands for computational resources and model complexity, which may limit its application in resource-constrained environments. Future directions for improvement might include further optimization of the model structure to reduce computational costs, enhance model interpretability, and integrate more biological information to enhance the generalizability of predictions[92-94].

2.3.2 Deep Learning Methods Based on Pre-trained Large Language Models

Recent advancements in the field of transcription factor binding site (TFBS) prediction have greatly improved performance through the use of pre-trained models. The performance of promoter prediction and transcription factor binding site prediction was significantly enhanced in 2022 with the implementation of MoDNA. This approach integrated self-supervised pre-training and fine-tuning techniques, while also adding DNA functional motifs as domain knowledge. This methodology demonstrates a higher degree of precision in capturing crucial data inside DNA sequences, thereby enhancing the precision and effectiveness of prognostications. Nevertheless, the intricate nature of the MoDNA model and its substantial requirement for computational resources may restrict its utilization in situations with limited resources. Potential future enhancements encompass the reduction of model weight and the augmentation of model interpretability, alongside the investigation of incorporating more forms of biological data to enhance the model's ability to generalize[95-97]. In the same year, the TFBert model shown notable enhancements in prediction accuracy and efficiency. This was achieved through the implementation of task-specific pre-training and the treatment of DNA sequence processing as a natural language processing problem. Notably, the model exhibited exceptional performance in effectively managing small datasets. Nevertheless, TFBert is presently limited to sequences of a specific length, and its capacity to be applied to other sequences has yet to be confirmed. Potential areas for future enhancement encompass the advancement of model architectures that possess the ability to handle sequences of diverse lengths, as well as the comprehensive assessment of the model's generalizability by means of multi-source validation datasets[98,99].

The DNABERT-Cap model, which integrates DNABERT with capsule networks, shown notable enhancements in the accuracy of predicting transcription factor binding sites in 2023. This improvement was particularly evident in the analysis of intricate DNA sequence data. However, there are still limitations in the model's ability to adapt to different sequence lengths and generalize under various biological conditions. This highlights the need for future improvements, such as the development of more flexible models that can accommodate sequences of different lengths and the incorporation of additional biological information to enhance predictive performance and the generalizability of the model[100].

In the same year, a study was conducted to enhance the precision of predicting transcription factor binding sites by combining the biophysical features of DNA, such as DNA breathing dynamics, with DNA sequence information using sequence transformers and cross-attention processes. This approach has demonstrated significant efficacy in improving predictive accuracy. However, it faces certain obstacles

such as the substantial computational resources required, the limited interpretability of the model, and the reliance on high-quality biophysical data. These challenges suggest potential avenues for future enhancements in terms of optimizing computational efficiency, improving model interpretability, and incorporating a wider range of biophysical properties[101].

The utilization of pre-trained big language models in deep learning methods offers novel insights and strategies for predicting transcription factor binding locations. Subsequent investigations will persist in examining the optimization of model structures, augmenting computing efficiency, expanding model interpretability, and incorporating additional biological data to further enhance the precision, efficiency, and extent of predictions. This progress not only holds the potential to drive the growth of technologies for predicting transcription factor binding sites, but also offers valuable tools for comprehending intricate gene regulatory mechanisms. With the growing accessibility of computer resources and the accumulation of biological data, the potential applications of pre-trained models will broaden, hence presenting novel opportunities for bioinformatics research and clinical applications. Furthermore, through the resolution of obstacles related to the interpretability of models, these methodologies will additionally expedite the rate and caliber of biological breakthroughs, thereby assisting researchers in acquiring more profound understandings of biological processes and mechanisms behind diseases. In brief, the utilization of deep learning techniques that rely on pre-trained extensive language models within the field of predicting transcription factor binding sites represents a novel stage in the advancement of bioinformatics and computational biology investigation.

3 Database for TFBS prediction

In the field of transcription factor binding site prediction, database resources can be categorized into two main directions: "Transcriptional Regulation and Binding Specificity" and "Genomic Annotation and Regulatory Networks." They display a clear progression from basic data levels to application levels, providing researchers with comprehensive support from core data to in-depth analysis.

3.1 Transcriptional Regulation and Binding Specificity

In the domain of "Transcriptional Regulation and Binding Specificity," the TRANSFAC and JASPAR foundational databases provide core data support for predicting transcription factor and DNA binding sites. TRANSFAC contains details of over 4,300 regulatory sites and 1,500 transcription factors, offering 169 nucleotide distribution matrices covering both natural and artificial DNA elements and transcription factor binding models[102]. JASPAR has added 329 new PFMs (Position Frequency Matrices) and independently verified 72 models, improving model quality and introducing cutting algorithms and new TFBS (Transcription Factor Binding Sites) extraction tools[103]. These resources deepen the understanding of the interactions between transcription factors and DNA, enhancing the accuracy and efficiency of transcriptional regulation research, and serve as the cornerstone for predicting transcription factor binding sites, helping researchers understand the fundamental interactions between specific transcription factors and DNA. Furthermore, Cis-BP and HOCOMOCO expand the understanding of transcription factor binding characteristics[104,105]. Cis-BP utilizes similarity regression methods to optimize the prediction of transcription factor binding models, covering 59,998 TFs. HOCOMOCO refines the DNA binding models of 949 human and 720 mouse transcription factors by analyzing a large amount of ChIP-Seq and HT-SELEX experimental data, providing a total of 1,443 validated position weight

matrices. The advancements in these databases offer experimentally validated and predictive models for studying the DNA binding characteristics of transcription factors in multiple species, further enriching the scientific resource pool.

At the level of experimental data and technical applications, ReMap, TFBSshape, and HTPSELEX provide rich transcription factor binding site information, DNA shape feature analysis, and high-throughput SELEX experimental data, respectively, deepening the understanding of transcriptional regulatory mechanisms[106-108]. ReMap 2022 has updated its database, integrating over 11,000 datasets across humans, mice, fruit flies, and Arabidopsis, showcasing an unprecedented collection of transcription regulatory factor DNA binding regions and introducing Cis-regulatory module identification. TFBSshape has expanded to 2,428 structural records, covering 1,900 transcription factors across 39 species, with each entry including 13 shape features and four features of methylated DNA, offering a new perspective on DNA shape features. HTPSELEX makes high-throughput SELEX experimental data public, supporting the characterization of transcription factor binding specificity, including raw data and descriptions represented by Hidden Markov Models. The combination of these databases provides researchers with richer and more dynamic resources, promoting precise predictions of transcription factor binding sites and in-depth studies of transcriptional regulatory mechanisms.

3.2 Genomic Annotation and Regulatory Networks

In the realm of "Genomic Annotation and Regulatory Networks," the ENCODE project plays a core role in constructing functional maps of the genome, providing annotations of a wide range of regulatory elements within the human genome at the basic data level. This includes annotations of genes and transcription factor binding sites, as well as information on histone modifications, constructing a comprehensive map of genome function for researchers. As a broad regulatory element annotation database, ENCODE contains over 13,000 datasets and their accompanying metadata, covering a variety of organisms including humans, mice, fruit flies, and *C. elegans*, with a total data volume exceeding 500TB. By offering standardized data processing pipelines, the ENCODE project has promoted the standardization, unification, and reproducibility of data processing, greatly enriching the resource pool for genomic science and providing a solid foundation for genomic annotation and regulatory network analysis[109].

As research progresses, PlantRegMap and GENCODE provide refined annotation and regulatory network analysis resources for plant and human/mouse genomes, respectively[110,111]. PlantRegMap, using the FunTFBS algorithm and genomic conservation analysis, has identified over 20 million functional transcription factor binding sites and 2 million interactions for 63 plant species, becoming a comprehensive platform for plant transcriptional regulation analysis. GENCODE 2021 has enhanced the annotation quality of human and mouse genomes, including special annotations for SARS-CoV-2 related genes, supporting a wide range of genomic research. The advancements in PlantRegMap and GENCODE not only provide specialized resources for genomic annotation and regulatory network analysis but also push the application of bioinformatics in the fields of botany and human disease research, deepening our understanding of the complexity of gene expression regulation and laying a solid foundation for future biomedical research.

The UCSC Genome Browser and UniPROBE extend into the realms of technological application and data analysis by providing visualization tools for genomic data and in vitro data on protein-DNA

interactions, respectively[112,113]. These tools support researchers in conducting more complex analyses of gene regulatory networks and functional studies. The 2022 update of the UCSC Genome Browser enhanced the visualization and analysis of genome annotations, introducing new databases and software features such as updated clinical tracks, new track hubs, improved variant displays, and analysis tools for the SARS-CoV-2 genome, making data comparison, analysis, and sharing more efficient and intuitive. Meanwhile, the 2015 update to UniPROBE introduced new tools and content for PBM data, including 12 new publication datasets and PBM data involving 96 transcription factors, now hosting 515 unique proteins and complexes, and streamlined the data submission process. These advancements not only promote the in-depth utilization of genomic data but also provide new perspectives and experimental data support for studies on gene regulation mechanisms.

This progression from basic to applied research in the field of transcription factor binding site prediction database resources not only provides researchers with comprehensive support from core data to in-depth analysis but also facilitates a deeper understanding of complex gene regulation mechanisms and advances in disease mechanism research. As these databases continue to expand and update, and with the application of new technologies, the future is expected to reveal more biological mysteries in the field of gene expression regulation, providing a solid data foundation and theoretical support for the development of precision medicine and disease treatment strategies[114-117]. These databases are not only valuable resources for research on transcription factor binding site prediction but also important tools for advancing life science research. Their progressive relationship and complementarity offer endless possibilities for the exploration of genome science. With the rapid development of bioinformatics and computational biology, the integration and innovative application of these database resources will continue to expand our understanding of the complexity of life, leading biomedical research into a new era.

4 Discussion and conclusion

Significant technological advancements have been made in the field of transcription factor binding site (TFBS) prediction over the last few decades. This trip started with an investigation into the basic knowledge of gene regulatory processes and progressively developed into an extremely sophisticated and technologically advanced field of study. To find transcription factors and their binding sites, scientists first mostly used experimental techniques like Electrophoretic Mobility Shift Assays (EMSA). Even though these experimental techniques at the time offered insightful information, they were frequently expensive, time-consuming, and challenging to scale up to the genome level[118,119].

As computational biology and genomics advanced quickly, especially after the Human Genome Project was finished, scientists started using computer programs to predict TFBS[120]. These techniques attempted to find putative binding sites throughout the genome by utilizing genomic sequence data and known transcription factor binding motifs. These computational methods were first mostly predicated on conventional statistical learning methods, such as consensus sequence analysis and Position-Specific Scoring Matrices (PSSM), which assumed that each base pair contributed independently to transcription factor binding[121,122]. Early TFBS predictions were made possible by tools that took advantage of consensus sequences and PSSM, such as MatInspector and MATCHM. The ease of use and intuitiveness of these techniques allowed early researchers to predict TFBS without the need for sophisticated algorithms or big datasets. With the passage of time, scientists began investigating new paths for

cross-species conserved sequence comparisons using Phylogenetic Footprinting and tools like TFinder and COTRASIF. These developments not only increased prediction accuracy but also created new avenues for future investigation, particularly in the areas of comparative genomics and multi-species genome annotation[123,124].

As we moved into the 21st century, machine learning techniques became increasingly popular in TFBS prediction due to the explosion of biological data volume and the huge growth in processing capacity. The efficiency and accuracy of predictions were greatly increased by these techniques, which could process enormous datasets and identify intricate patterns. When processing regulatory sequence data, algorithms like Support Vector Machines (SVM), Random Forest (RF), and Gradient Boosting Machines (GBM) performed better. The TFBS prediction area saw a dramatic shift with the advent of machine learning. The tremendous potential of machine learning in managing large volumes of biological data and complicated pattern recognition was demonstrated by the development of techniques like gkmSVM, DNA shape features, and Mocap[125-128]. These methods not only significantly improved prediction efficiency and accuracy, but they also offered fresh insights into how transcription factors identify their binding sites by taking into account more intricate biological details like DNA structure and sequence context. These techniques still have issues with generalizability of algorithms, demands on computer resources, and data quality despite their triumphs.

Deep learning technology has been used into TFBS prediction as a result of its revolutionary advancements in a number of sectors in recent years. The capacity of deep learning techniques, particularly Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), to extract intricate and hierarchical patterns from sequence data made them immensely valuable[129]. These techniques not only increased prediction accuracy but also unveiled previously unknown DNA sequence regulatory characteristics, such as chromatin structure and DNA shape. With the emergence of deep learning, TFBS prediction achieved previously unheard-of efficiency and accuracy. The ability to predict TFBS at the single-nucleotide level was enabled by deep learning techniques such as DeepBind, DeepSEA, and Leopard, which discover intricate patterns from sequence data. The use of CNNs and RNNs in particular greatly improved the predictive model's capacity to handle genomic data. But even with the great performance gains, deep learning techniques have brought to light growing issues with large computing costs and uninterpreted models.

According to recent studies, high-throughput technologies such as ChIP-exo and ChIP-nexus have shown promise in increasing the accuracy of binding site identification; nevertheless, they are more expensive and sophisticated than ChIP-seq. This problem highlights how important it is to weigh benefits and costs when using new technologies[130,131]. Furthermore, despite potential redundancy with DNA sequence information, investigating DNA shape features and epigenomic properties has shown their usefulness in predicting TF-DNA interactions, suggesting a potential improvement in performance.

In order to improve model prediction accuracy and biological interpretability, future research should concentrate more on integrating various types of genomic data, including but not limited to DNA sequence information, chromatin accessibility, histone modifications, and even taking into consideration RNA expression data. Additionally, researchers should investigate ways to overcome the difficulties caused by inconsistent data quality and make efficient use of the enormous number of TF binding site databases that are currently available to drive model training.

The TFBS prediction field is moving toward combining several computational approaches, improving model interpretability and generalizability, and using multi-omics data to increase prediction accuracy as genomics and computational biology progress. Even though TFBS prediction technologies have advanced significantly, there are still many obstacles to overcome. These include improving model interpretability, resolving problems with generalizing predictions across species and cell types, and incorporating multi-omics data to increase biological relevance and prediction accuracy[132-133]. It is anticipated that future studies will overcome the constraints of current methodologies by integrating and innovating algorithms, so providing more insight into the molecular principles underlying intricate gene regulation networks. In addition to advancing fundamental biological research, this will lay the scientific groundwork for precision medicine and the treatment of diseases.

References

- [1]Lee, Tong Ihn, and Richard A. Young. "Transcription of eukaryotic protein-coding genes." *Annual review of genetics* 34.1 (2000): 77-137.
- [2]Tjian, Robert. (1995) *Molecular Machines that Control Genes*. *Scientific American* 272, 38-46.
- Levine, M. and Tjian R. (2002) *Transcription and the evolutionary diversification of the metazoa*. *Nature*, 424: 147-151.
- [3]Lambert, Samuel A., et al. "The human transcription factors." *Cell* 172.4 (2018): 650-665.
- [4]Roeder, Robert G. "The role of general initiation factors in transcription by RNA polymerase II." *Trends in biochemical sciences* 21.9 (1996): 327-335.
- [5]Nikolov, D. B., and S. K. Burley. "RNA polymerase II transcription initiation: a structural view." *Proceedings of the National Academy of Sciences* 94.1 (1997): 15-22.
- [6]Garner, Mark M., and Arnold Revzin. "A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the Escherichia coli lactose operon regulatory system." *Nucleic acids research* 9.13 (1981): 3047-3060.
- [7]Gilmour, David S., and John T. Lis. "Detecting protein-DNA interactions in vivo: distribution of RNA polymerase on specific bacterial genes." *Proceedings of the National Academy of Sciences* 81.14 (1984): 4275-4279.
- [8]Galas, David J., and Albert Schmitz. "DNAase footprinting a simple method for the detection of protein-DNA binding specificity." *Nucleic acids research* 5.9 (1978): 3157-3170.
- [9]Tuerk, Craig, and Larry Gold. "Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase." *science* 249.4968 (1990): 505-510.
- [10]Stormo, Gary D., et al. "Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli." *Nucleic acids research* 10.9 (1982): 2997-3011.
- [11]Stormo, Gary D. and George Hartzell. "Identifying protein-binding sites from unaligned DNA fragments." *Proceedings of the National Academy of Sciences of the United States of America* 86 4 (1989): 1183-7 .
- [12]Lawrence, Charles E. et al. "Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment." *Science* 262 5131 (1993): 208-14 .
- [13]Bailey, Timothy L. and Charles Peter Elkan. "Fitting a Mixture Model By Expectation Maximization To Discover Motifs In Biopolymer." *Proceedings. International Conference on Intelligent Systems for Molecular Biology* 2 (1994): 28-36 .
- [14]Wasserman, Wyeth W., and Albin Sandelin. "Applied bioinformatics for the identification of regulatory elements." *Nature Reviews Genetics* 5.4 (2004): 276-287.
- [15]Altschul, Stephen F., et al. "Basic local alignment search tool." *Journal of molecular biology* 215.3 (1990): 403-410.
- [16]Schneider, Thomas D. "Consensus sequence Zen." *Applied bioinformatics* vol. 1,3 (2002): 111-9.
- [17]Tomovic, Andrija, and Edward J. Oakeley. "Position dependencies in transcription factor binding sites." *Bioinformatics* 23.8 (2007): 933-941.
- [18]Cartharius, Kerstin, et al. "MatInspector and beyond: promoter analysis based on transcription factor binding sites." *Bioinformatics* 21.13 (2005): 2933-2942.

- [19] Kel, Alexander E., et al. "MATCHTM: a tool for searching transcription factor binding sites in DNA sequences." *Nucleic acids research* 31.13 (2003): 3576-3579.
- [20] Tokovenko, Bogdan, et al. "COTRASIF: conservation-aided transcription-factor-binding site finder." *Nucleic acids research* 37.7 (2009): e49-e49.
- [21] Minniti, Julien, et al. "TFinder: a Python web tool for predicting Transcription Factor Binding Sites." (2023).
- [22] Schewe, Coen. "Exploring Promoter Conservation and Transcription Factor Binding Sites of Homologous Lactuca Genes in a Pangenomic Framework." (2023).
- [23] Ghandi, Mahmoud, et al. "gkmSVM: an R package for gapped-kmer SVM." *Bioinformatics* 32.14 (2016): 2205-2207.
- [24] Mathelier, Anthony, et al. "DNA shape features improve transcription factor binding site predictions in vivo." *Cell systems* 3.3 (2016): 278-286.
- [25] Chen, Xi, et al. "Mocap: large-scale inference of transcription factor binding sites from chromatin accessibility." *Nucleic acids research* 45.8 (2017): 4315-4329.
- [26] Schmidt, Florian, et al. "Combining transcription factor binding affinities with open-chromatin data for accurate gene expression prediction." *Nucleic acids research* 45.1 (2017): 54-66.
- [27] Khamis, Abdullah M., et al. "A novel method for improved accuracy of transcription factor binding site prediction." *Nucleic acids research* 46.12 (2018): e72-e72.
- [28] Li, Hongyang, Daniel Quang, and Yuanfang Guan. "Anchor: trans-cell type prediction of transcription factor binding sites." *Genome research* 29.2 (2019): 281-292.
- [29] Schmidt, Florian, et al. "TEPIC 2—an extended framework for transcription factor binding prediction and integrative epigenomic analysis." *Bioinformatics* 35.9 (2019): 1608-1609.
- [30] Keilwagen, Jens, Stefan Posch, and Jan Grau. "Accurate prediction of cell type-specific transcription factor binding." *Genome biology* 20 (2019): 1-17.
- [31] Grau, Jan, Florian Schmidt, and Marcel H. Schulz. "Widespread effects of DNA methylation and intra-motif dependencies revealed by novel transcription factor binding models." *Nucleic Acids Research* 51.18 (2023): e95-e95.
- [32] Ishimori, Motoyuki. "Transcription factor binding site prediction: Finding the point from many data." *Plant and Cell Physiology* 63.10 (2022): 1324-1325.
- [33] Yaman, Oğuz Ulaş, and Pınar Çalık. "MachineTFBS: Motif-based method to predict transcription factor binding sites with first-best models from machine learning library." *Biochemical Engineering Journal* 198 (2023): 108990.
- [34] Gusmao, Eduardo G., et al. "Detection of active transcription factor binding sites with the combination of DNase hypersensitivity and histone modifications." *Bioinformatics* 30.22 (2014): 3143-3151.
- [35] Wong, Ka-Chun, et al. "DNA motif elucidation using belief propagation." *Nucleic acids research* 41.16 (2013): e153-e153.
- [36] Zhang, Hongbo, Lin Zhu, and De-Shuang Huang. "WSMD: weakly-supervised motif discovery in transcription factor ChIP-seq data." *Scientific reports* 7.1 (2017): 3217.
- [37] Kleftogiannis, Dimitrios, Panos Kalnis, and Vladimir B. Bajic. "DEEP: a general computational framework for predicting enhancers." *Nucleic acids research* 43.1 (2015): e6-e6.

- [38]Alipanahi, Babak, et al. "Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning." *Nature biotechnology* 33.8 (2015): 831-838.
- [39]Zhou, Jian, and Olga G. Troyanskaya. "Predicting effects of noncoding variants with deep learning-based sequence model." *Nature methods* 12.10 (2015): 931-934.
- [40]Liu, Feng, et al. "PEDLA: predicting enhancers with a deep learning-based algorithmic framework." *Scientific reports* 6.1 (2016): 28517.
- [41]Quang, Daniel, and Xiaohui Xie. "DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences." *Nucleic acids research* 44.11 (2016): e107-e107.
- [42]Kelley, David R., Jasper Snoek, and John L. Rinn. "Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks." *Genome research* 26.7 (2016): 990-999.
- [43]Hassanzadeh, Hamid Reza, and May D. Wang. "DeeperBind: Enhancing prediction of sequence specificities of DNA binding proteins." 2016 IEEE International conference on bioinformatics and biomedicine (BIBM). IEEE, 2016.
- [44]Zeng, Haoyang, et al. "Convolutional neural network architectures for predicting DNA-protein binding." *Bioinformatics* 32.12 (2016): i121-i127.
- [45]Lanchantin, Jack, et al. "Deep motif: Visualizing genomic sequence classifications." *arXiv preprint arXiv:1605.01133* (2016).
- [46]Chen, Dexiong, Laurent Jacob, and Julien Mairal. "Predicting transcription factor binding sites with convolutional kernel networks." *bioRxiv* 10 (2017): 217257.
- [47]Shen, Zhen, Wenzheng Bao, and De-Shuang Huang. "Recurrent neural network for predicting transcription factor binding sites." *Scientific reports* 8.1 (2018): 15270.
- [48]Kelley, David R., et al. "Sequential regulatory activity prediction across chromosomes with convolutional neural networks." *Genome research* 28.5 (2018): 739-750.
- [49]Zhao, Zihan, and Chuanhuan Yin. "Convolutional Hybrid Kernel Network for in-vitro Transcription Factor Binding Sites." *Proceedings of the 2023 3rd International Conference on Bioinformatics and Intelligent Computing*. 2023.
- [50]Shen, Zhen, Wenzheng Bao, and De-Shuang Huang. "Recurrent neural network for predicting transcription factor binding sites." *Scientific reports* 8.1 (2018): 15270.
- [51]Li, Yifeng, Wenqiang Shi, and Wyeth W. Wasserman. "Genome-wide prediction of cis-regulatory regions using supervised deep learning methods." *BMC bioinformatics* 19 (2018): 1-14.
- [52]Zhang, Qinhu, et al. "Weakly-supervised convolutional neural network architecture for predicting protein-DNA binding." *IEEE/ACM transactions on computational biology and bioinformatics* 17.2 (2018): 679-689.
- [53]Zhou, Jiyun, et al. "MTTFsite: cross-cell type TF binding site prediction by using multi-task learning." *Bioinformatics* 35.24 (2019): 5067-5077.
- [54]Zhang, Qinhu, Lin Zhu, and De-Shuang Huang. "High-order convolutional neural network architecture for predicting DNA-protein binding sites." *IEEE/ACM transactions on computational biology and bioinformatics* 16.4 (2018): 1184-1192.
- [55]Yang, Jinyu, et al. "Prediction of regulatory motifs from human Chip-sequencing data using a deep learning framework." *Nucleic acids research* 47.15 (2019): 7809-7824.
- [56]Blum, Christopher F., and Markus Kollmann. "Neural networks with circular filters enable data

- efficient inference of sequence motifs." *Bioinformatics* 35.20 (2019): 3937-3943.
- [57]Blum, Christopher F., and Markus Kollmann. "Neural networks with circular filters enable data efficient inference of sequence motifs." *Bioinformatics* 35.20 (2019): 3937-3943.
- [58]Zhang, Qinhu, Zhen Shen, and De-Shuang Huang. "Modeling in-vivo protein-DNA binding by combining multiple-instance learning with a hybrid deep neural network." *Scientific reports* 9.1 (2019): 8484.
- [59]Cao, Zhen, and Shihua Zhang. "Simple tricks of convolutional neural network architectures improve DNA-protein binding prediction." *Bioinformatics* 35.11 (2019): 1837-1843.
- [60]Zhou, Jiyun, et al. "MTTFsite: cross-cell type TF binding site prediction by using multi-task learning." *Bioinformatics* 35.24 (2019): 5067-5077.
- [61]Zhang, Yongqing, et al. "DeepSite: bidirectional LSTM and CNN models for predicting DNA-protein binding." *International Journal of Machine Learning and Cybernetics* 11 (2020): 841-851.
- [62] Park, Sungjoon, et al. "Enhancing the interpretability of transcription factor binding site prediction using attention mechanism." *Scientific reports* 10.1 (2020): 13413.
- [63]Jing, Fang, Shao-Wu Zhang, and Shihua Zhang. "Prediction of transcription factor binding sites with an attention augmented convolutional neural network." *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 19.6 (2021): 3614-3623.
- [64]Zhang, Qinhu, et al. "Predicting in-vitro DNA-protein binding with a spatially aligned fusion of sequence and shape." *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 19.6 (2021): 3144-3153.
- [65]Deng, Lei, et al. "DeepD2V: a novel deep learning-based framework for predicting transcription factor binding sites from combined DNA sequence." *International journal of molecular sciences* 22.11 (2021): 5521.
- [66]Li, Jiawei, et al. "DeepATT: a hybrid category attention neural network for identifying functional effects of DNA sequences." *Briefings in bioinformatics* 22.3 (2021): bbaa159.
- [67]Zhang, Qinhu, Zhen Shen, and De-Shuang Huang. "Predicting in-vitro transcription factor binding sites using DNA sequence+ shape." *IEEE/ACM transactions on computational biology and bioinformatics* 18.2 (2019): 667-676.
- [68]Avsec, Žiga, et al. "Base-resolution models of transcription-factor binding reveal soft motif syntax." *Nature genetics* 53.3 (2021): 354-366.
- [69]Zheng, An, et al. "Deep neural networks identify sequence context features predictive of transcription factor binding." *Nature machine intelligence* 3.2 (2021): 172-180.
- [70]Zhang, Qinhu, et al. "Predicting in-vitro DNA-protein binding with a spatially aligned fusion of sequence and shape." *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 19.6 (2021): 3144-3153.
- [71]Wang, Siguo, et al. "Predicting transcription factor binding sites using DNA shape features based on shared hybrid deep learning architecture." *Molecular Therapy-Nucleic Acids* 24 (2021): 154-163.
- [72]Zhang, Qinhu, et al. "Multi-scale capsule network for predicting DNA-protein binding sites." *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 18.5 (2020): 1793-1800.
- [73]Zhang, Qinhu, et al. "Locating transcription factor binding sites by fully convolutional neural network." *Briefings in bioinformatics* 22.5 (2021): bbaa435.

- [74]Shen, Long-Chen, et al. "SAResNet: self-attention residual network for predicting DNA-protein binding." *Briefings in Bioinformatics* 22.5 (2021): bbab101.
- [75]Zhang, Yongqing, et al. "CAE-CNN: Predicting transcription factor binding site with convolutional autoencoder and convolutional neural network." *Expert Systems with Applications* 183 (2021): 115404.
- [76]Chen, Chen, et al. "DeepGRN: prediction of transcription factor binding site across cell-types using attention-based deep neural networks." *BMC bioinformatics* 22 (2021): 1-18.
- [77]Zhang, Yongqing, et al. "High-resolution transcription factor binding sites prediction improved performance and interpretability by deep learning method." *Briefings in Bioinformatics* 22.6 (2021): bbab273.
- [78]Zhang, Yongqing, et al. "A novel convolution attention model for predicting transcription factor binding sites by combination of sequence and shape." *Briefings in Bioinformatics* 23.1 (2022): bbab525.
- [79]Cao, Linan, et al. "Prediction of transcription factor binding sites using a combined deep learning approach." *Frontiers in Oncology* 12 (2022): 893520.
- [80]Zhang, Yongqing, et al. "Uncovering the relationship between tissue-specific TF-DNA binding and chromatin features through a transformer-based model." *Genes* 13.11 (2022): 1952.
- [81]Han, Ke, et al. "MAResNet: predicting transcription factor binding sites by combining multi-scale bottom-up and top-down attention and residual network." *Briefings in Bioinformatics* 23.1 (2022): bbab445.
- [82]Bhukya, Raju, et al. "An attention-based hybrid deep neural networks for accurate identification of transcription factor binding sites." *Neural Computing and Applications* 34.21 (2022): 19051-19060.
- [83]Wang, Wei, et al. "DeepGenBind: a novel deep learning model for predicting transcription factor binding sites." *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2022.
- [84]Wang, Zixuan, et al. "Single-cell TF-DNA binding prediction and analysis based on transfer learning framework." *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2022.
- [85]Zhang, Yongqing, et al. "Predicting cell type-specific effects of variants on TF-DNA binding by meta-learning." *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2022.
- [86]Yu, Yutong, et al. "Cooperation of local features and global representations by a dual-branch network for transcription factor binding sites prediction." *Briefings in Bioinformatics* 24.2 (2023): bbad036.
- [87]Ding, Pengju, et al. "DeepSTF: predicting transcription factor binding sites by interpretable deep neural networks combining sequence and shape." *Briefings in Bioinformatics* 24.4 (2023): bbad231.
- [88]Du, Zhihua, et al. "TFBSnet: A deep learning-based tool for predicting transcription factor binding site from DNA sequences." *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2023.
- [89]Akbari Rokn Abadi, Saeedeh, SeyedehFatemeh Tabatabaei, and Somayyeh Koochi. "KDeep: a new memory-efficient data extraction method for accurately predicting DNA/RNA transcription factor binding sites." *Journal of Translational Medicine* 21.1 (2023): 727.
- [90]Zhang, Jidong, et al. "DeepCAC: a deep learning approach on DNA transcription factors classification based on multi-head self-attention and concatenate convolutional neural network." *BMC*

bioinformatics 24.1 (2023): 345..

[91]Wang, Xun, et al. "TBCA: Prediction of transcription factor binding sites using a deep neural network with lightweight attention mechanism." *IEEE Journal of Biomedical and Health Informatics* (2024).

[92]Zhang, Qinhu. "Cross-species prediction of transcription factor binding by adversarial training of a novel nucleotide-level deep neural network." *bioRxiv* (2024): 2024-02.

[93]Kabir, Anowarul, et al. "Advancing Transcription Factor Binding Site Prediction Using DNA Breathing Dynamics and Sequence Transformers via Cross Attention." *bioRxiv* (2024): 2024-01.

[94]Yang, Zikun, et al. "Multiomics-integrated deep language model enables in silico genome-wide detection of transcription factor binding site in unexplored biosamples." *Bioinformatics* 40.1 (2024): btae013.

[95] An, Weizhi, et al. "MoDNA: motif-oriented pre-training for DNA language model." *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*. 2022.

[96] Ji, Yanrong, et al. "DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome." *Bioinformatics* 37.15 (2021): 2112-2120.

[97] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

[98] Zhou, Zhihan, et al. "Dnabert-2: Efficient foundation model and benchmark for multi-species genome." *arXiv preprint arXiv:2306.15006* (2023).

[99] Luo, Hanyu, et al. "Improving language model of human genome for DNA–protein binding prediction based on task-specific pre-training." *Interdisciplinary Sciences: Computational Life Sciences* 15.1 (2023): 32-43.

[100]Ghosh, Nimisha, et al. "Predicting Transcription Factor Binding Sites using Transformer based Capsule Network." *arXiv preprint arXiv:2310.15202* (2023).

[101]Cheng, Lei, et al. "Self-supervised learning for DNA sequences with circular dilated convolutional networks." *Neural Networks* 171 (2024): 466-473.

[102]Matys, Vea, et al. "TRANSFAC®: transcriptional regulation, from patterns to profiles." *Nucleic acids research* 31.1 (2003): 374-378.

[103]Rauluseviciute, Ieva, et al. "JASPAR 2024: 20th anniversary of the open-access database of transcription factor binding profiles." *Nucleic Acids Research* 52.D1 (2024): D174-D182.

[104]Vorontsov, Ilya E., et al. "HOCOMOCO in 2024: a rebuild of the curated collection of binding models for human and mouse transcription factors." *Nucleic Acids Research* 52.D1 (2024): D154-D163.

[105]Weirauch, Matthew T., et al. "Determination and inference of eukaryotic transcription factor sequence specificity." *Cell* 158.6 (2014): 1431-1443.

[106]Childers, Julie W., et al. "REMAP: a framework for goals of care conversations." *Journal of Oncology Practice* 13.10 (2017): e844-e850.

[107]Chiu, Tsu-Pei, et al. "TFBSshape: an expanded motif database for DNA shape features of transcription factor binding sites." *Nucleic acids research* 48.D1 (2020): D246-D255.

[108]Jagannathan, Vidhya, et al. "HTPSELEX—a database of high-throughput SELEX libraries for transcription factor binding sites." *Nucleic acids research* 34.suppl_1 (2006): D90-D94.

[109]ENCODE Project Consortium. "Expanded encyclopaedias of DNA elements in the human and mouse

genomes." *Nature* 583.7818 (2020): 699-710.

[110]Tian, Feng, et al. "PlantRegMap: charting functional regulatory maps in plants." *Nucleic acids research* 48.D1 (2020): D1104-D1113.

[111]Frankish, Adam, et al. "GENCODE 2021." *Nucleic acids research* 49.D1 (2021): D916-D923.

[112]Karolchik, Donna, Angie S. Hinrichs, and W. James Kent. "The UCSC genome browser." *Current protocols in human genetics* 71.1 (2011): 18-6.

[113]Hume, Maxwell A., et al. "UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein–DNA interactions." *Nucleic acids research* 43.D1 (2015): D117-D122.

[114]Zhu, Lihua Julie, et al. "FlyFactorSurvey: a database of *Drosophila* transcription factor binding specificities determined using the bacterial one-hybrid system." *Nucleic acids research* 39.suppl_1 (2011): D111-D117.

[115]Dudek, Christian-Alexander, and Dieter Jahn. "PRODORIC: state-of-the-art database of prokaryotic gene regulation." *Nucleic acids research* 50.D1 (2022): D295-D302.

[116]Shazman, Shula, et al. "OnTheFly: a database of *Drosophila melanogaster* transcription factors and their binding sites." *Nucleic acids research* 42.D1 (2014): D167-D171.

[117]Dreos, René, et al. "EPD and EPDnew, high-quality promoter resources in the next-generation sequencing era." *Nucleic acids research* 41.D1 (2013): D157-D164.

[118]Jayaram, Narayan, Daniel Usvyat, and Andrew C. R. Martin. "Evaluating tools for transcription factor binding site prediction." *BMC bioinformatics* 17 (2016): 1-12.

[119]Deng, Lei, et al. "DeepD2V: a novel deep learning-based framework for predicting transcription factor binding sites from combined DNA sequence." *International journal of molecular sciences* 22.11 (2021): 5521.

[120]Gibbs, Richard A. "The human genome project changed everything." *Nature Reviews Genetics* 21.10 (2020): 575-576.

[121]Mathelier, Anthony, and Wyeth W. Wasserman. "The next generation of transcription factor binding site prediction." *PLoS computational biology* 9.9 (2013): e1003214.

[122]Ishimori, Motoyuki. "Transcription factor binding site prediction: Finding the point from many data." *Plant and Cell Physiology* 63.10 (2022): 1324-1325.

[123]Bornstein, Kristin, et al. "The NIH Comparative Genomics Resource: addressing the promises and challenges of comparative genomics on human health." *BMC genomics* 24.1 (2023): 575.

[124]Huang, Tinghua, et al. "Identification of upstream transcription factor binding sites in orthologous genes using mixed Student's *t*-test statistics." *PLOS Computational Biology* 18.6 (2022): e1009773.

[125]Huang, Shujun, et al. "Applications of support vector machine (SVM) learning in cancer genomics." *Cancer genomics & proteomics* 15.1 (2018): 41-51.

[126]Rigatti, Steven J. "Random forest." *Journal of Insurance Medicine* 47.1 (2017): 31-39.

[127]Kumar, K. Krishna, Ganesan Pugalenthi, and Ponnuthurai N. Suganthan. "DNA-Prot: identification of DNA binding proteins from protein sequence information using random forest." *Journal of Biomolecular Structure and Dynamics* 26.6 (2009): 679-686.

[128]Wang, Liangjiang, Mary Qu Yang, and Jack Y. Yang. "Prediction of DNA-binding residues from protein sequence information using random forests." *Bmc Genomics* 10 (2009): 1-9.

- [129]Sherstinsky, Alex. "Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network." *Physica D: Nonlinear Phenomena* 404 (2020): 132306.
- [130]Rhee, Ho Sung, and B. Franklin Pugh. "ChIP-exo method for identifying genomic location of DNA-binding proteins with near-single-nucleotide accuracy." *Current protocols in molecular biology* 100.1 (2012): 21-24.
- [131]Yamada, Naomi, et al. "Characterizing protein–DNA binding event subtypes in ChIP-exo data." *Bioinformatics* 35.6 (2019): 903-913.
- [132]He, Qiye, Jeff Johnston, and Julia Zeitlinger. "ChIP-nexus enables improved detection of in vivo transcription factor binding footprints." *Nature biotechnology* 33.4 (2015): 395-401.
- [133]Pobbati, Ajaybabu V., et al. "Therapeutic targeting of TEAD transcription factors in cancer." *Trends in Biochemical Sciences* 48.5 (2023): 450-462.