the plant journal

S E B
SOCIETY FOR EXPERIMENTAL BIOLOGY

The Plant Journal (2024) doi: 10.1111/tpj.17190

PERSPECTIVE

Identification, characterization, and design of plant genome sequences using deep learning

Zhenye Wang[1,2,3,†], **Hao Yuan**[1,2,3,†], **Jianbing Yan**[1,4] and **Jianxiao Liu**[1,2,3,4,*]

[1]*National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan 430070, China,*
[2]*Hubei Key Laboratory of Agricultural Bioinformatics, Huazhong Agricultural University, Wuhan 430070, China,*
[3]*College of Informatics, Huazhong Agricultural University, Wuhan 430070, China, and*
[4]*Hubei Hongshan Laboratory, Wuhan 430070, China*

Received 3 September 2024; revised 11 November 2024; accepted 23 November 2024.
*For correspondence (e-mail liujianxiao@mail.hzau.edu.cn).
†These authors contributed equally to this work.

SUMMARY

Due to its excellent performance in processing large amounts of data and capturing complex non-linear relationships, deep learning has been widely applied in many fields of plant biology. Here we first review the application of deep learning in analyzing genome sequences to predict gene expression, chromatin interactions, and epigenetic features (open chromatin, transcription factor binding sites, and methylation sites) in plants. Then, current motif mining and functional component design and synthesis based on generative adversarial networks, large models, and attention mechanisms are elaborated in detail. The progress of protein structure and function prediction, genomic prediction, and large model applications based on deep learning is also discussed. Finally, this work provides prospects for the future development of deep learning in plants with regard to multiple omics data, algorithm optimization, large language models, sequence design, and intelligent breeding.

Keywords: deep learning, plants, genome sequence, prediction, intelligent design.

INTRODUCTION

Plant complex gene regulatory networks involve chromatin interaction, gene expression, transcription factor binding site, and open chromatin. The regulatory mechanisms underlying these processes precisely control the expression of genes at different stages of plant development and environmental conditions, thereby controlling plant growth. In addition, (i) precise design of genome regulatory elements, (ii) prediction of protein structure and function, and (iii) intelligent breeding have also become important research points in plant functional genomics.

As a data-driven approach, deep learning has the advantage of handling massive amounts of data and capturing complex non-linear relationships. It has been widely used in various fields including industry, agriculture, and biology. At present, researchers have mainly conducted genome sequence analysis based on deep learning in humans. As early as 2015, Zhou and Troyanskaya (2015) first introduced convolutional neural networks (CNN) into the analysis of the human genome and developed the sequence function prediction model of DeepSEA. This model could directly predict non-coding variations from DNA sequences. Similarly, Alipanahi et al. (2015) used CNN to construct a model to predict the human genome sequence and consequently protein function. Since 2015, researchers have applied both the deep learning methods of CNN and recurrent neural networks (RNN) to analyze diverse biological functions. The related research work mainly includes the prediction and analysis of sequence function (Wang et al., 2022), methylation status (Alam et al., 2021), enhancer element (Deb et al., 2018), chromatin accessibility (Shen, Chen, & Gao, 2021), and transcription factor binding sites (Yan et al., 2022). In recent years, novel deep learning technologies involving graph neural networks (GNN), generative adversarial networks (GAN), and large-scale models have gradually been used in the study of genome sequence analysis. With the development of high-throughput sequencing in plants, deep learning has gradually been adopted to analyze genome sequences in plants such as maize, *Arabidopsis thaliana*, rice, wheat. These studies have predicted (i) gene expression (Washburn

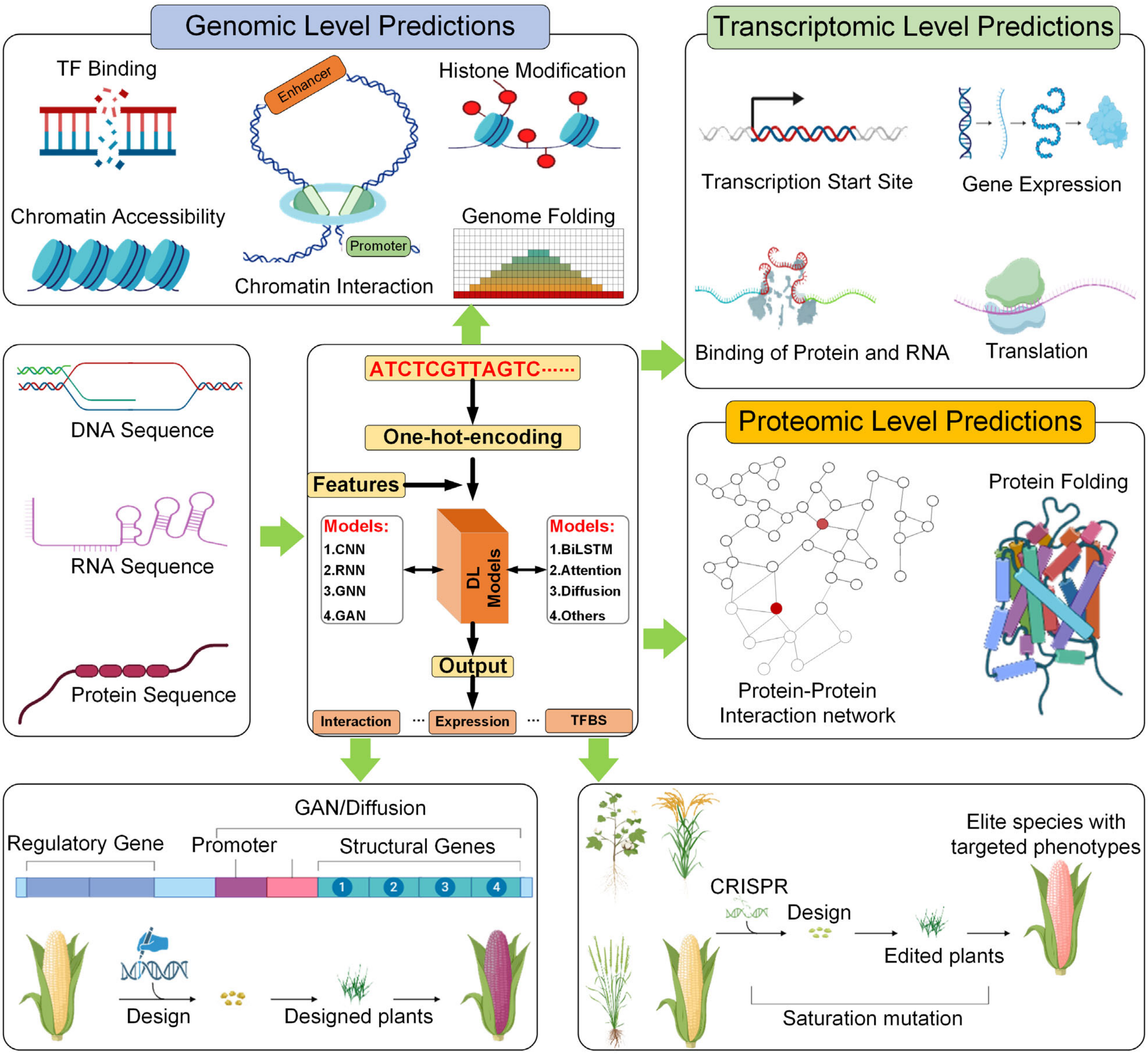


Figure 1. Deep learning has been used in genomic analysis from the aspects of DNA, RNA, and protein.
In terms of DNA: prediction of chromatin accessibility, histone modifications, transcription factor binding, chromatin interaction, and genome folding are shown. In terms of RNA: prediction of transcription starts sites, gene expression, translation, and binding between RNA and proteins are shown. In terms of protein: prediction of protein interactions, protein folding, and functions are shown.

et al., 2019), (ii) chromatin interactions (Schlegel et al., 2024), (iii) transcription factor binding sites (TFBS) (Cheng et al., 2023), and (iv) chromatin open regions. In addition, some researchers have applied the novel deep learning models (GNN, GAN) to predict protein–protein interaction (PPI) and design sequences in plants (Chi Sr et al., 2023; Li et al., 2024).

The existing reviews summarizing the research on deep learning technology in genome sequence analysis mainly concentrate on research in humans. In this review, we first describe work on predicting gene expression in plants. Next progresses on chromatin interaction are discussed from two aspects: (i) utilizing DNA sequences and (ii) fusing multiple omics data and DNA sequences. There after we summarize the application of deep learning in predicting plant epigenetic features, designing and synthesizing functional elements, and predicting protein function and structure. We will additionally review research of genomic prediction and the application of large-scale language models related to intelligent plant breeding. Finally, we intend to discuss the future prospects of using deep learning in plant genomics from the perspectives of (i) multi-omics data integration, (ii) algorithm optimization, (iii) large language models, (iv) sequence design, and finally (v) intelligent breeding. Figure 1 shows the application of deep learning from the aspects of DNA, RNA, and protein functionality in plants.

GENE EXPRESSION PREDICTION

The research of predicting gene expression based on deep learning mainly includes three aspects. Namely, (i) gene expression prediction using DNA sequence, (ii) gene expression prediction by integrating sequence and multi-omics data, (iii) gene expression prediction using chromatin interaction (Table 1).

Gene expression prediction using DNA sequences

Early studies mainly used traditional machine learning methods to predict gene expression (Beer & Tavazoie, 2004; Cheng et al., 2011; Dong et al., 2012; Tasaki et al., 2020), however, the accuracy of these methods is often not high. The release of more comprehensive reference genome of plants greatly promotes the analysis of genome sequences and prediction of gene expression based on deep learning. The largest advantage of the deep learning related method is that though it only requires inputting genomic sequences and does not require any other omics data, it results in a high accuracy in predicting gene expression. The most representative work is the Enformer model reported in 2021 (Avsec, Agarwal, et al., 2021). Enformer is a deep learning model that uses transformer architecture and combines DNA sequences to predict human gene expression. The model can analyze the impact of distant elements as far as 100 kb on gene expression. Other DNA sequence-based methods include Basenji (Kelley, 2020), Expecto (Zhou et al., 2018), Chromoformer (Lee et al., 2022), and Nvwa (Li et al., 2022, 2023).

The research of predicting gene expression in plants has only gradually emerged in the recent years. In 2019, Washburn et al. used 3 kb sequences near the gene translation start site (TSS) and translation termination site (TTS) as the target gene sequences of maize. Subsequently, they classified the gene expression into three categories: unexpressed, expressed, and highly expressed, and carried out gene expression classification prediction in maize using CNN. In recent years, there has been an increasing focus on predicting gene expression in multiple plants. In 2022, Akagi et al. conducted gene expression prediction on multiple species, including *A. thaliana*, *Solanum lycopersicum*, *Sorghum bicolor*, *Zea mays*, etc. (Akagi et al., 2022). They used CNN to construct prediction model based on integrating sequences of promoters and terminators with different lengths. Meanwhile, Levy et al. developed a deep learning model FloraBERT to predict gene expression using the transformer architecture (Levy et al., 2022). Specifically, FloraBERT integrates genome sequence from hundreds of plant species and utilizes transfer learning to predict gene expression across different species. Taking *A. thaliana*, *S. lycopersicum*, *Sor. Bicolor*, and *Z. mays* as research objects, Peleke et al. (2024) used the CNN model to predict the expression of the gene flanking regions of the above-mentioned four species. This study provides a paradigm for predicting gene expression across species and identifying conserved regulatory regions (Table 2).

Gene expression prediction by integrating sequence and multi-omics data

The regulation of gene expression is influenced by multiple factors, including DNA sequence, epigenetics, and other omics data. To improve the prediction accuracy and the interpretability of models, researchers have incorporated epigenetic and multi-omics data into prediction models. Early researchers applied traditional machine learning methods to predict gene expression levels, including random forest, logistic regression, and support vector machine (SVM) (Cheng et al., 2011; Dong et al., 2012; Karlić et al., 2010). With the accumulation of multi-omics data, researchers have begun to develop various deep learning models for predicting gene expression using multi-omics data. The representative methods are DEcode (Tasaki et al., 2020) and CREaTor (Li et al., 2023a). DEcode uses genome-wide binding sites on RNAs and promoters to predict gene expression. Based on attention mechanisms, CREaTor is constructed using the dataset of human K562 cell line. This model integrates ChIP-seq, RNA-seq, *cis*-regulatory elements (CRE) features, and sequence data to predict gene expression. Similar studies include DeepChrome (Singh et al., 2016), AttentiveChrome (Singh et al., 2017), and Xpresso (Agarwal & Shendure, 2020). The earlier methods mainly use the classical CNN architecture to construct the prediction model through integrating multiple omics data such as histone modification and methylation. Until now, integrating multi-omics data and genome sequences to predict gene expression is still lacking. In this regard, the gap in this field undoubtedly provides new opportunities for predicting gene expression in plants.

Predicting gene expression based on chromatin interactions

Chromatin interactions have a significant impact on gene expression, often manifested through the interactions between proteins and other substances (Dong et al., 2012). Until now, the classical deep learning methods (such as RNN and CNN) are widely used to predict gene expression based on chromatin interactions in humans. In an early study, Hafez et al. utilized a semi-supervised machine learning approach and integrated associations between target genes and enhancers to construct a novel prediction model (Hafez et al., 2017). With the development of 3D genome technologies (3C, Hi-C, ChIA-PET), scholars have improved the prediction accuracy by integrating chromatin interactions. For example, DeepExpression used CNN to integrate the information of enhancer–promoter

Table 1 Summary of deep learning applications in plant research

Task	Model	Author(s)	Species[a]	Method	Input data	Online tools
Predict expression	–	Washburn et al. (2019)	[31]	CNN	TSS + TTS, 3k	https://bitbucket.org/bucklerlab
	–	Akagi et al. (2022)	[25]	CNN	TSS, 1k	https://github.com/Takeshiddd
	FloraBERT	Levy et al. (2022)	[31]	Transformer	TSS, 1k	https://github.com/benlevyx
	–	Peleke et al. (2024)	[2][25][27][31]	CNN	TSS + TTS, 3k	https://github.com/NAMlab
	DeepCBA	Wang et al. (2024)	[31]	CNN + BiLSTM	TSS + TTS, 3k	http://www.deepcba.com
Predict interaction	GenomicLinks	Schlegel et al. (2024)	[31]	CNN	Anchor pairs seq, 5k	https://genomelink.io/
	PRPI-SC	Zhou et al. (2021)	[2][31]	CNN	RPI	https://github.com/zhr818789
	MPLPLNP	Jia and Luan (2022)	[1][2][4]	k-mer	RPI	https://doi.org/10.1007/s12539-022-00501-7
	ESMAraPPI	Zhou et al. (2023)	[2]	MLP	PPI	https://github.com/keiwo.
Predict epigenetic features	CharPlant	Shen et al. (2021b)	[2][17][18][25]	CNN	ATAC/DNase-seq	https://github.com/Yin-Shen
	PlantDeepSEA	Zhao et al. (2019)	[2][18][23][27]	CNN	ATAC-seq, 1k	http://plantdeepsea.ncpgr.cn
	SeqConv	Shen et al. (2021a)	[31]	CNN	ChIP-seq	https://github.com/shenwei19
	PlantBind	Yan et al. (2022)	[2][31]	CNN	DAP-seq, 101 bp	https://github.com/wenkaiyan-kevin
	TSPTFBS	Cheng et al. (2023)	[2][31]	DenseNet	ChIP-seq, 500 bp	https://github.com/liulifenyf/TSPTFBS-2.0
	PTFSpot	Gupta et al. (2024)	[18][19]	Transformer	ChIP/DAP-seq, 150 bp	https://scbb.ihbt.res.in/PTFSpot/
	iDNA4mC	Chen et al. (2017)	[2][31]	SVM	4mC site seq, 41 bp	http://lin.uestc.edu.cn/server/
	4mCPred	He et al. (2019)	[2]	SVM	4mC site seq, 41 bp	http://server.malab.cn/4mCPred
	Deep4mC	Xu et al. (2021)	[2]	CNN	4mC site seq, 41 bp	https://bioinfo.uth.edu/Deep4mC
	4mcPred-IFL	Wei et al. (2019)	[2]	SVM	4mC site seq, 41 bp	http://server.malab.cn/4mcPred-IFL
	i4mC-Deep	Alam et al. (2021)	[2]	CNN	4mC site seq, 41 bp	http://nsclbio.jbnu.ac.kr/tools/i4mC
	DeepSignal-plant	Ni et al. (2021)	[2]	BRNN	5mC site seq	https://github.com/PengNi/
	Deep6mA	Langille et al. (2021)	[2][18]	CNN + LSTM	6 mA site seq, 41 bp	–
	TADA	Morffy et al. (2024)	[2]	CNN	Amino acid seq, 40 bp	https://github.com/LisaVdB/TADA
Design elements	PhytoExpr	Li et al. (2024)	[2][3][4][8][12][15][17]–[20] [23]–[31]	CNN	TSS + TTS, 10k	https://doi.org/10.6084/m9.figshare.24417076.v1.0
	–	Jores et al. (2021)	[2][27][31]	CNN	TSS, 170 bp	https://github.com/tobjores
Genomic prediction	DeepGS	Ma et al. (2018)	[29]	CNN	Genotypic data	https://github.com/cma2015/DeepGS
	G2PDeep	Zeng et al. (2021)	[15]	CNN	Genotypic data	http://g2pdeep.org
	Galiana	Raimondi et al. (2022)	[2]	NN	Genotypic data	https://bitbucket.org/eddiewrc/galiana/src
	SoyDNGP	Gao et al. (2023)	[15]	CNN	Genotypic data	http://xtlab.hzau.edu.cn/SoyDNGP
	DeepCCR	Ma et al. (2024)	[18]	CNN + LSTM	Genotypic data	https://www.ai-breeder.com/
	GPformer	Wu et al. (2024)	[15][18][29][31]	Transformer	Genotypic data	–
	CropGS-Hub	Chen et al. (2023)	[6][7][15][16][18][23][31]	CNN + LSTM	Genotypic data	https://iagr.genomics.cn/CropGS/
	BreedingAIDB	Shen et al. (2024)	[15][18][31]	RNN	Genotypic data	http://ibi.zju.edu.cn/BreedingAIDB
Large models	AgroNT	Mendoza-Revilla et al. (2024)	[2][18][31]	Transformer	Genotypic data	https://huggingface.co/InstaDeepAI/agro
	GPN	Benegas et al. (2023)	[2][6][9]–[11][13][21][28]	Hugging Face	DNA seq, 512 bp	https://github.com/songlab-cal/gpn

[a][1] *Arabidopsis lyrata*, [2] *Arabidopsis thaliana*, [3] *Beta vulgaris*, [4] *Brachypodium distachyon*, [5] *Brassica napus*, [6] *Brassica rapa*, [7] *Cicer arietinum*, [8] *Cichorium intybus*, [9] *Carica papaya*, [10] *Capsella rubella*, [11] *Camelina sativa*, [12] *Chlamydomonas reinhardtii*, [13] *Eutrema salsugineum*, [14] *Fragaria vesca*, [15] *Glycine max*, [16] *Gossypium hirsutum*, [17] *Medicago truncatula*, [18] *Oryza sativa*, [19] *Pinus trichocarpa*, [20] *Physcomitrella patens*, [21] *Raphanus sativus*, [22] *Rosa chinensis*, [23] *Setaria italica*, [24] Setaria viridis, [25] *Solanum lycopersicum*, [26] *Solanum tuberosum*, [27] *Sorghum bicolor*, [28] *Tarenaya hassleriana*, [29] *Triticum aestivum*, [30] *Vitis vinifera*, [31] *Zea mays*.

Table 2 Glossary used in this article

Abbreviations	Full name
3C	Chromosome conformation capture
AI	Artificial intelligence
ATAC-seq	Assay for transposase accessible chromatin
BGLR	Bayesian generalized linear regression
ChIA-PET	Chromosome interaction analysis by paired end tag sequencing
ChIP-seq	Chromatin immunoprecipitation sequencing
CNN	Convolutional neural networks
CRE	*Cis*-regulatory element
DAP-seq	DNA affinity purification sequencing
DHS	DNase hypersensitivity
EM	Expectation maximization
GAN	Generative adversarial networks
GNN	Graph neural networks
Hi-C	High-throughput chromosome conformation capture
HMM	Hidden Markov model
LASSO	Least absolute shrinkage and selection operator
OCRs	Open chromatin regions
PPI	Protein–protein interactions
EPIs	Enhancer–promoter interactions
PSSM	Position-specific scoring matrix
PWM	Position weight matrix
RNN	Recurrent neural networks
TF	Transcription factor
TFBS	Transcription factor binding sites

interactions (EPIs) to predict target gene expression (Zeng et al., 2020). In recent years, GNN has been widely applied in gene expression prediction due to its advantages in processing network data. The representative methods include GraphReg (Karbalayghareh et al., 2022), GC-MERGE (Bigness et al., 2022), and the GNN model integrating master node aggregation (Yan et al., 2024). In 2024, Wang et al. constructed a high-precision model (DeepCBA) to predict gene expression using CNN and BiLSTM based on chromatin interaction data (Wang et al., 2024). DeepCBA is the first deep learning model of predicting gene expression using chromatin interaction information in plants (maize, wheat, cotton, and rice).

PREDICTING CHROMATIN INTERACTIONS USING DEEP LEARNING

Revealing the potential mechanisms of enhancer–promoter interactions (EPIs), protein–protein interactions (PPI), enhancer targeted gene interactions (ETGs), chromatin loops, and topologically associated domains (TADs) is crucial for a comprehensive understanding of gene transcriptional regulation (Whalen et al., 2016; Zeng et al., 2021). At present, researchers mainly use genome sequences and multi-omics data to predict chromatin interactions.

1. Using DNA or RNA sequences to predict chromatin interactions. In early studies, researchers mainly utilized sequence features and traditional machine learning methods such as random forest, high-valued singular value decomposition to predict chromatin interactions. The representative studies mainly include TargetFinder (Whalen et al., 2016) and EpiTensor (Zhu et al., 2016). Subsequently, deep learning methods, such as CNN and RNN, are increasingly used to predict chromatin interactions. For example, SPEID (Singh et al., 2019), ChINN (Cao et al., 2021), DeepC (Schwessinger et al., 2019), and Akita (Fudenberg et al., 2020) all utilize CNN to predict EPIs, PPI, *etc*. Similarly, researchers have used deep learning methods to predict chromatin loops and TADs, including 3DEpiLoop (Al Bkhetan & Plewczynski, 2018), DeepMILO (Trieu et al., 2020), and CTCF-MP (Zhang et al., 2018). In recent years, researchers have, moreover, begun to integrate multiple deep learning methods to construct more accurate models. For example, DeepTACT (Li et al., 2019) and CharID (Shen et al., 2022) integrate CNN, RNN, and attention mechanisms in order to predict EPIs. In the field of plants, GenomicLinks is a deep learning model of fusing CNN and LSTM to predict chromatin interactions in maize (Schlegel et al., 2024). Due to the limitations in sequencing technology and funding, it has not yet accumulated sufficient chromatin interaction data and affected the research progress in plant to some extent. However, researchers have carried out prediction studies on protein–protein and protein–lncRNA interactions in plants. The main research works include PRPI-SC (Zhou et al., 2021), MPLPLNP (Jia & Luan, 2022), ESMAraPPI (Zhou et al., 2023), and LPI-LSTM-ResNet (Zhang et al., 2024).
2. Integrating multiple omics data to predict chromatin interactions. As we know, epigenetics, open chromatin, and other omics data have significant impacts on chromatin interactions. Currently, researchers have built numerous prediction models of EPIs, ETGs, and PPI using a wide range of epigenomic signaling and expression data. These models include RIPPLE (Roy et al., 2015), Rambutan (Schreiber et al., 2017), CRUP (Ramisch et al., 2019), EAGLE (Gao & Qian, 2019), and TransEPI (Chen et al., 2022). Due to a lack of related data in plants, using deep learning to predict chromatin interactions is still at the theoretical level. With the development of 3D genomic technologies of 3C, Hi-C, and ChIA-PET, we believe that more researches of chromatin interaction prediction based on deep learning in various plants will emerge in the future.

IDENTIFYING EPIGENETIC FEATURES BASED ON DEEP LEARNING

With the continuous research of deep learning in the field of plants, significant progress has been made in predicting epigenetic features, such as chromatin accessibility, TFBS, and methylation modification sites.

Open chromatin prediction

Chromatin accessibility is critical for the regulation of gene transcription and it reflects the extent of nuclear macromolecule contact with DNA. Complementary to experimental techniques (DNase-seq, ATAC-seq), deep learning is increasingly used to predict chromatin accessibility. The main studies in this area include DeepSEA (Zhou & Troyanskaya, 2015), Basset (Kelley et al., 2016), and DeepBind (Alipanahi et al., 2015). The research of chromatin accessibility prediction in plants is also gradually emerging. CharPlant is the first prediction model that identifies potential open chromatin regions (OCRs) in the whole plant genome (Shen, Chen, & Gao, 2021). This model is based on CNN and integrates DNA sequences to learn both sequence motifs and regulatory logic to predict chromatin accessibility. Based on DeepSEA, PlantDeepSEA (Zhao et al., 2021) is an online network service platform for predicting the accessibility of various plants. This platform not only can predict CRE in several kinds of plants (*A. thaliana*, rice, maize, *etc*.), but also can mine functional sites that affect chromatin open regions. Other representative studies involving OCRs prediction in plants include SMOC-(Guo et al., 2022) and DanQ-based research (Wrightsman et al., 2022).

Prediction of transcription factor binding sites

TFBS, as a type of CRE, specifically refers to the region specifically bound by transcription factors (TFs) and plays a crucial role in transcriptional regulation of genes. Traditional methods for identifying TFBS, such as ChIP-seq, are not only costly but also time-consuming. In recent years, deep learning has been increasingly applied into the prediction of TFBS. SeqConv is the first deep learning model to predict TFBS using CNN (Shen, Pan, et al., 2021) in plants. This model not only accurately identifies TFBS in maize, but also integrates transfer learning to achieve cross species prediction of TFBS between maize and *Arabidopsis*. Similarly, PlantBind is a model of predicting TFBS based on the attention mechanism in *Arabidopsis* (Yan et al., 2022). This model can predict 315 potential binding sites for TFs in *Arabidopsis* and identify TF binding motifs using DNA shape features. Deep-BSC also uses CNN to predict TFBS in Arabidopsis based on DNA sequences (Bukhari et al., 2021). Subsequently, researchers improved the TFBS prediction accuracy by integrating multiple deep learning methods. A representative work is the PTFSpot model, which realizes TFBS prediction by fusing Transformer and DenseNet (Gupta et al., 2024). This model can learn the structure of transcription factors and the covariance of TF binding regions, thereby achieving accurate TFBS prediction. Similarly, TADA is a model of integrating CNN and BiLSTM to predict transcriptional activation domain in *Arabidopsis* (Morffy et al., 2024). This model can analyze and identify activated regions in plant transcription factors and predict their regulatory effects on gene expression. In addition, TSPTFBS and TSPTFBS 2.0 integrated TFBS data of multiple crops (maize, rice, arabidopsis, *etc*.) and constructed deep learning models using DenseNet (Cheng et al., 2023; Liu et al., 2021). Furthermore, researchers have also used the methods of DeepLIFT (Shrikumar et al., 2017), in silica tiling dilution, and in silica mutagenesis to identify and mine core motifs.

Methylation site prediction

Different types of plant methylation sites, such as N4-methylcytosine (4mC), 5-methylcytosine (5mC), and N6-methyladenine (6mA), play a crucial role in gene regulation and development. Understanding the mechanisms of these methylated forms is crucial for revealing the epigenetic regulatory network in plants. However, traditional experimental detection methods are cumbersome and costly, while deep learning techniques have shown significant advantages.

Early studies like iDNA4mC mainly utilized different feature extraction and encoding strategies to predict methylation sites using traditional machine learning methods (Chen et al., 2017). With the accumulation of omics data, multiple types of feature data are used to construct the corresponding prediction models. For example, 4mcRed achieved prediction of 4mC by fusing position-specific trinucleotide preference (PSTNP) and electron–ion interaction potential (EIIP) (He et al., 2019). In recent years, the researches of predicting 4mC sites using deep learning have achieved good results. For example, Deep4mC (Xu et al., 2021) and 4mcPred-IFL (Wei et al., 2019) have implemented automatic extraction of DNA sequence features, further improving prediction performance. I4mCDeep further improved the prediction accuracy and stability of 4mC by combining ResNet and various feature encoding techniques (Alam et al., 2021). Similarly, researchers have used the strategy of combining CNN, LSTM, and attention mechanism to build deep learning models to improve the accuracy of predicting 5mC and 6mA. These models include DeepSignal-plant (Ni et al., 2021), Deep6mA (Langille et al., 2021), iM6A (Luo et al., 2022), and SMEP (Wang et al., 2021).

USING DEEP LEARNING TO MINE MOTIFS IN PLANTS

Motifs usually refer to short sequence fragments with specific functions or structures in DNA, RNA, and protein sequences. Motifs are often related to biological functions of gene regulation and signal transduction. Motifs mining methods mainly include the following two types.

1 Traditional motif mining methods. The k-mer (Morris et al., 2014; Yang et al., 2017), position specific rating matrix (PSSM), position weight matrix (PWM), hidden Markov model (HMM), expectation maximization (EM) algorithm, and Bayesian method are the traditional motif

mining methods. Among them, the position weight matrix (PWM) is the most widely used method. It describes the features of motifs by calculating the probability of each nucleotide appearing at each position in the genome sequence, and has been widely used for identifying TFBSs. For example, tools including MEME (Bailey et al., 2015), STREME (Bailey & Birol, 2021), EXTREME (Thomas-Chollier et al., 2012), and DREME (Bailey, 2011) all utilize PWM to do calculation. These tools have been widely used for motif mining in bioinformatic research.

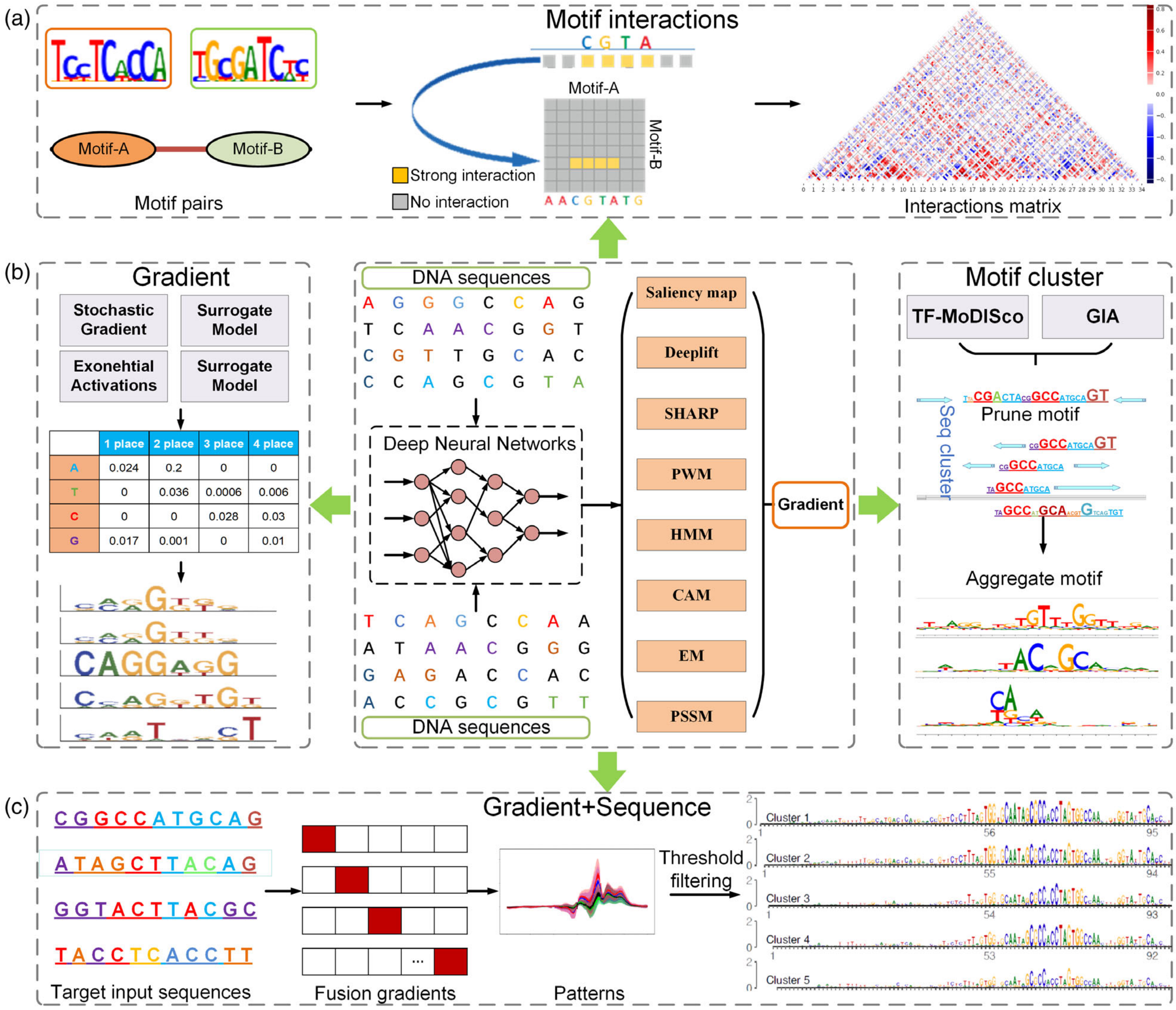
2 Motif mining method based on deep learning. In recent years, deep learning has become an irreplaceable method for motif mining. Representative studies are DeepBind (Alipanahi et al., 2015) and BPNet (Avsec, Weilert, et al., 2021), which implement motif mining based on CNN. With the continuous deepening of deep learning technology in motif mining, there are three types of motif mining algorithms. (i) Gradient-based motif mining. Saliency map (Chu, 2011) and DeepLIFT (Shrikumar et al., 2017) are two mainstream gradient-based methods that have been used in interpretability studies of deep learning models. Most of the early deep learning based researches used gradient calculation and threshold filtering to mine motifs. Furthermore, exponential activation, proxy model, stochastic gradient component based on the gradient calculation have been proposed to mine motifs by Koo and collaborators (Koo et al., 2021; Koo & Eddy, 2019; Koo & Ploenzke, 2021). Similar studies include GOPHER (Toneyan & Koo, 2023) and SQUID (Seitz et al., 2024). (ii) Motif clustering method of integrating gradient and sequence features. This method mainly achieves hierarchical and similarity clustering of target regions by integrating base importance scores and sequence fragments. TF-MoDISco (Avanti Shrikumar et al., 2020) and Puffin (Dudnyk et al., 2024) are the representatives of this type of methods. (iii) Mining motifs in the view of motifs interaction. The idea of this method is to perform reverse clustering on feature maps and combine them with position weight matrices to explore the interactions between different motifs. For the classical research of DeepSTARR (de Almeida et al., 2022), it explored the impact of different motif flanks and distances between motifs on model prediction performance based on multi-task CNN. Similarly, NeuronMotif identifies important motifs and motif combinations based on the grammatical structure between motif pairs (Wei et al., 2023). Figure 2 shows the application of deep learning techniques in motif mining.

DESIGN AND SYNTHESIS OF PLANT FUNCTIONAL ELEMENTS

Functional elements refer to specific sequences that can control plant growth and development, respond to environmental stimuli, and regulate the synthesis of secondary metabolites. The traditional regulatory component design relies on experimental screening methods, but these methods are often time-consuming, labor-intensive, and face challenges in complex plant genomes (Muthamilarasan & Prasad, 2015). At present, the design and synthesis of biological regulatory components based on deep learning mainly include the following three aspects.

1 Design and synthesis of promoters. The design and synthesis of artificial promoters aims to synthesize short, inducible, and conditionally controlled promoters. Artificially designed promoters can coordinate the expression of multiple genes in various metabolic and signaling pathways, and reduce unnecessary negative feedback (Liu et al., 2013; Liu & Stewart Jr, 2016; Mehrotra et al., 2011; Rushton, 2016; Rushton et al., 2002). In early studies, researchers synthesized promoters by adjusting the position of regulatory elements in the genome and changing the spacing and number between elements (Acharya et al., 2014; Cazzonelli & Velten, 2008; Deb et al., 2018; Kumar et al., 2015). In recent years, researchers have used deep learning to design and synthesize promoters in *Escherichia coli* and other species, and representative studies include DRSAdesign (Wang et al., 2023) and DeepSEED (Zhang et al., 2023). Besides, the promoter design and synthesis model constructed in *Saccharomyces cerevisiae* also verified the reliability of deep learning (Vaishnav et al., 2022). In the early days, the design and synthesis of promoters in plants have mainly utilized traditional experimental methods (Chen et al., 2013; Jameel et al., 2022; Yang et al., 2021). In recent years, the rapid development of deep learning has made it possible to use AI technology to design and synthesize plant promoters. Li et al. developed the PhytoExpr model in plants using CNN + Stacking and transformer (Li et al., 2024). This model provides two methods for designing synthetic CRE of 17 plants, including *maize*, *Arabidopsis* and *Carica papaya*. Furthermore, the biological experiment results showed that the designed and synthesized promoters in maize have higher activity.

2 Design and synthesis of enhancers. An enhancer is a type of CRE that enhances the transcriptional activity of target genes by combining transcription factors and other proteins. Different enhancers can regulate the expression of different genes and participate in various biological processes of development, differentiation, and response to the environment (Banerji et al., 1981; Levine, 2010; Shlyueva et al., 2014). The rapid development of deep learning has made it possible to design and synthesize enhancers from scratch. The most representative research work is the DeepSTARR model (de Almeida et al., 2022). This model focuses on *Drosophila melanogaster* and designs and synthesizes enhancers with the target activity based on predicting the activity of



	1 place	2 place	3 place	4 place
A	0.024	0.2	0	0
T	0	0.036	0.0006	0.006
C	0	0	0.028	0.03
G	0.017	0.001	0	0.01

Figure 2. Application of deep learning in motifs mining in plants.
(a) The process of detecting important motifs based on motifs interaction.
(b) On the basis of calculating the gradient of specific model, motifs are identified using the threshold filtering and clustering methods.
(c) Identifying motifs by integrating gradients and sequence features.

developmental and housekeeping enhancers. Similarly, Taskiran et al. used GAN networks to design and synthesize enhancers in both *Drosophila* and humans (Taskiran et al., 2024). Due to the limitations of plant genome-related feature data, the design and synthesis of plant enhancers using deep learning models is still in the initial stage. It is foreseeable that the improvement of data quality and the iterative development of AI technology will provide strong impetus for the design and synthesis of enhancers in plants.

3 Design and generation of protein sequences. At present, deep learning technology has become an indispensable tool for designing and synthesizing protein sequences. Sinai et al. (2017) used the variational autoencoder (VAE) model in 2017 to transform natural proteins and design synthetic proteins. Similar studies include ProteinMPNN (Dauparas et al., 2022), ProtGPT2 (Ferruz et al., 2022), and Rfdiffusion (Watson et al., 2023). ProteinMPNN is suitable for the design of almost all protein sequences, achieving efficient protein sequence design. As an unsupervised language model, ProtGPT2 can generate diverse protein sequences. To broaden the breadth of protein design methods, GAN networks have been widely used in protein sequence design and synthesis in different species. FBGAN (Gupta & Zou, 2018) and ProteinGAN (Repecka et al., 2021) are two typical methods. Experimental results show that the designed and synthesized protein sequences using GAN have more advantages in structure

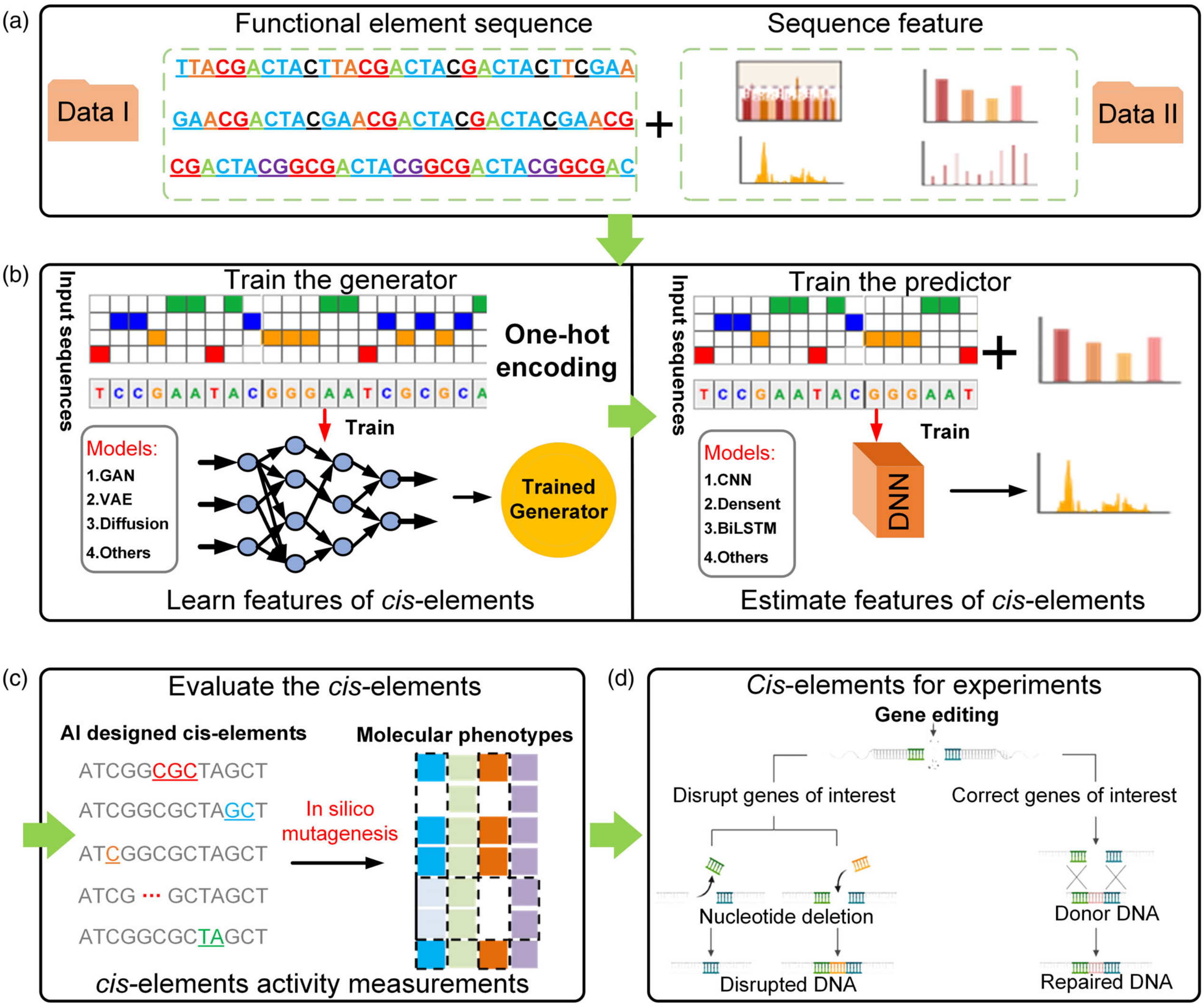


Figure 3. Deep learning based design and generation of functional components in plants.
(a) The functional components and sequence features are the input of the models.
(b) The process of design and generation for functional components.
(c) Evaluation of the designed regulatory components.
(d) Validate the designed and synthesized regulatory components through biological experiments.

and function compared to natural proteins (Anand & Huang, 2018; Davidsen et al., 2019; Hsu et al., 2024; Karimi et al., 2020). With the widespread application of large models, researchers have carried out protein structure prediction based on large-scale language models and achieved better results (Lin et al., 2023). Due to the relative lack of protein-related data in plants, the design and synthesis of protein sequences in plants are still in the preliminary research stage. It can be foreseen that using deep learning for the design and synthesis of plant proteins will see rapid development.

Figure 3 illustrates the implementation process of design and synthesis of plant regulatory components and sequences using deep learning.

PREDICTION OF PROTEIN STRUCTURE AND FUNCTION IN PLANTS

In recent years, deep learning has made significant progress in the field of protein structure and function prediction, greatly promoting the development of bioinformatics. Protein structure prediction refers to inferring its three-dimensional structure through amino acid sequences. Traditional experimental methods have many disadvantages, such as being both time-consuming and labor-intensive. Currently, deep learning has become an important tool of predicting protein structure. AlphaFold2 (Jumper et al., 2021), RoseTTAFold (Baek et al., 2021), and AlphaFold3 (Abramson et al., 2024) are three representative researches in the field of protein structure prediction. The protein function mainly depends on its structure, and accurate identification of protein function is crucial for interpreting complex biology processes and accelerating genome annotation. Recently, researchers have developed numerous deep learning models to predict protein functions, including DeepGOPlus (Kulmanov et al., 2020), COFACTOR (Zhang et al., 2017), RoseTTAFold (Wang et al., 2022), and DeepFunc (Zhang et al., 2019). The earlier studies integrated sequence, structure, and PPI network to achieve accurate function prediction of proteins. At present, the sequence, structure, and function related to protein data in plants are relatively scarce. With the accumulation of relevant protein data in plants, utilizing deep learning to realize the prediction of plant proteins will become the focus for biologists. Such research will cover structure prediction and function annotation of proteins as well as the construction of protein interaction networks in plants.

GENOMIC PREDICTION BASED ON DEEP LEARNING

Genomic prediction refers to using genomic and environmental data to predict the field phenotype of crops, which has significance of intelligent plant breeding, gene function analysis, and environmental adaptability. Traditional genomic prediction methods based on statistical principles are difficult to handle complex non-linear relationships and multi-dimensional data. In addition, the accuracy of genomic prediction algorithms based on machine learning needs further improvement. Over the past few years, deep learning has been increasingly used for plant genomic prediction research. For example, DeepGS (Ma et al., 2018), G2PDeep (Zeng et al., 2021), and Galiana (Raimondi et al., 2022) used CNN and multi-task neural networks to predict the phenotypes in multiple plants, including wheat, *Arabidopsis*, and soybean. Additionally, researchers have developed several genomic prediction models of multiple plants based on deep learning, mainly including DNNGP (Wang et al., 2023), SoyDNGP (Gao et al., 2023), TrG2P (Li et al., 2024), DeepCCR (Ma et al., 2024), and GPformer (Wu et al., 2024). The earlier methods could accurately predict the phenotypes of maize, rice, wheat, and soybean, demonstrating the great potential of deep learning in plant genomic prediction.

As we know, the phenotype is closely related to the field environment of crops. In recent years, biologists have begun to fuse genotype and environmental data to predict field phenotypes, thereby improving the accuracy of genomic prediction. Some scholars demonstrated that integrating environmental variables with G × E interactions can significantly improve the accuracy of predictions (Ray et al., 2022; Tong & Nikoloski, 2021). Based on integrating genotype and environmental data, most researchers use statistical and Bayesian related methods, such as GBLUP, Bayesian generalized linear regression (BGLR), and LASSO to conduct genomic prediction research (Cui et al., 2020; Jighly et al., 2023; Millet et al., 2019). Furthermore, researchers have conducted in-depth studies on the plasticity mechanism of crops by analyzing the interactions between genotype and environmental factors (Fu & Wang, 2023; Jin et al., 2023; Liu et al., 2020).

To facilitate the practical usage by biologists and breeders, researchers have developed several genomic prediction platforms based on deep learning in plants. These platforms integrate different model data of genome, transcriptome, and environmental data, to predict important plant phenotypic traits. Integrating genotype, phenotype, and environmental data, the platform of CropGPT used transformer to predict phenotypes in plants (Zhu et al., 2024). Similarly, the CropGS-Hu model integrates genotypes and agronomic phenotypes of seven species including maize and rice to realize genomic prediction (Chen et al., 2024). This model provides a one-stop service for sequencing data input, phenotype prediction, and phenotype analysis. The smart breeding platform is a full process intelligent breeding platform that integrate breeding data management and analysis, multiple genomic selection models and computation acceleration, and phenotype prediction of parents and excellent varieties (Li et al., 2024). Furthermore, Shen et al. developed the platform of BreedingAIDB that integrates genotype and phenotype pairing data in soybean, rice, and maize to serve genomic prediction (Shen et al., 2024).

APPLICATION OF LARGE MODELS IN THE FIELD OF PLANTS

The large-scale language model represented by ChatGPT is now widely used in a variety of areas. Based on massive biological data, researchers have applied deep learning and large-scale language models in human-related biology research. The most representative study is DNABERT (Ji et al., 2021), in which researchers have developed a large-scale language model for human sequence analysis based on BERT. Furthermore, it realizes accurate prediction of regulatory elements, promoters, splice sites, and TFBSs. As an improved model of DNABERT, DNABERT-2 replaced k-mer with byte encoding and achieved excellent performance on 36 datasets (Zhou et al., 2024). Given the rapid development of large model techniques, researchers have developed prediction models based on the datasets with the scale of 10 million. Among them, Geneformer is a deep learning model that can predict key network regulatory factors and candidate therapeutic targets using human single-cell transcriptome data (Cui et al., 2024). Similar studies also include nucleotide transformer (Dalla-Torre et al., 2023), UTR-ML (Chu et al., 2024), DNAGPT (Zhang et al., 2023), and Evo (Nguyen et al., 2024).

Correspondingly, the application of large-scale models in plant science is constantly deepening. The large-scale models with the capabilities of powerful data processing and pattern recognition have shown strong potential applications in several fields. These areas include plant genomics, epigenomics, gene expression prediction, multi-omics data integration, synthetic biology, and ecology (Lam et al., 2024). Currently, the most influential two studies involving large models for genome sequence analysis in plants are GPN (Benegas et al., 2023) and AgroNT (Mendoza-Revilla et al., 2024). GPN is a model of predicting gene mutation effects by unsupervised training of genomic DNA sequences. The model was validated in seven plant species, including *A. thaliana*, *Brassica rapa*, and *Camelina sativa*. AgroNT is another fundamental genome sequences analysis method based on large language model in 48 plants. This model can accurately predict regulatory annotations, promoter/terminator strength, and tissue-specific gene expression. Similarly, PlantCaduceus, a plant DNA large-model based on the Caduceus and Mamba architectures, pre-trained on a curated dataset of 16 angiosperm genomes. Additionally, PlantCaduceus successfully identifies well-known causal variants in both *Arabidopsis* and maize (Zhai et al., 2024). The earlier three works provide the direction for the widespread application of large-scale models in plant researches in the future. Figure 4 shows the application of large models in plant research.

CONCLUSION

With the continuous development of high-throughput sequencing technologies and artificial intelligence algorithms, applying artificial intelligence to the plant genome data analysis will undoubtedly be an important research direction in the future. This review discussed the future trends of deep learning in plants from the following aspects.

1 Multi-dimensional data fusion and construction of high-quality datasets. Integrating data from different sources and types can provide richer and more complete data for the interpretation of biological processes. In the plant field, problems of data quality are prevalent due to factors of data labeling and sequencing techniques. Semi-supervised or unsupervised learning methods can effectively address the earlier problems and improve the prediction performance of deep learning models through data dimensionality reduction and automatic feature extraction. Additionally, with the development of technologies of single-cell and spatial transcriptomics, researchers have begun to concentrate on integrating multiple omics data. Integrating large-scale networks and AI models related to genomics, transcriptomics, metabolomics, proteomics, single cell, spatial transcriptomics, and phenotype will become an important trend in the future. Furthermore, the deep learning algorithms of GNN, GAN, and diffusion have advantages of processing non-Euclidean structured data and high-dimensional data. Then, multi-dimensional and large-scale datasets can be integrated effectively. The construction and integration of multi-level networks using these technologies contribute to the in-depth study of plant biology.

2 Emerging deep learning technologies drive the rapid development of biological research in plants. The unique complexity and heterogeneity in plant biology research render it necessary to develop more targeted algorithms. In the future, biological researchers not only need to focus on the prediction performance of models, but also understand the specific decision-making process of models. At present, the classic deep learning algorithms of CNN and RNN are mainly used in the research of plant science. Going forward, the latest developed deep learning algorithms, such as KAN (knowledge aware neural network), large convolution, are used to analyze the multi-omics data of genome, transcriptome, proteomics, and metabolomics. In addition, constructing a comprehensive gene regulatory network can help to reveal the genetic basis of complex traits. Furthermore, it is necessary to strengthen the optimization of deep learning algorithms, break down the barriers between different deep learning frameworks (Tensorflow, Pytorch, *etc.*), and achieve the reuse of multiple frameworks. Moreover, utilizing more interpretable deep learning algorithms, attention mechanisms, and feature visualization tools can help improve the interpretability of biological mechanisms.Incorporating transfer learning and other AI techniques to enable cross-species analysis is also a focus in plant research. Using optimized transfer learning algorithms to mine functional genes among different plant species is of great significance for the study of functional genomes. For example, scarcity of annotated data often becomes an inevitable problem when conducting research on a species rarely studied. Combining the few-shot learning and transfer learning is an important research direction. By pre-training deep learning models on common species or tasks, and then fine-tuning models using small sample data, the learning performance of new tasks can be significantly improved. Domain adaptive techniques can improve the effectiveness of transfer learning by adapting the model to the specific data distribution of the target species.

3 Open-source platforms facilitate collaborative research among multiple species of plants. With the increasing application of deep learning in plant research, building open-source learning platforms has gradually become a hot topic. The platforms leverage models pre-trained on large-scale datasets and transfer the knowledge of these models to specific botany-related tasks, such as gene expression prediction, genome function prediction, and sequence design. More importantly, these platforms can provide a unified framework that supports researchers to

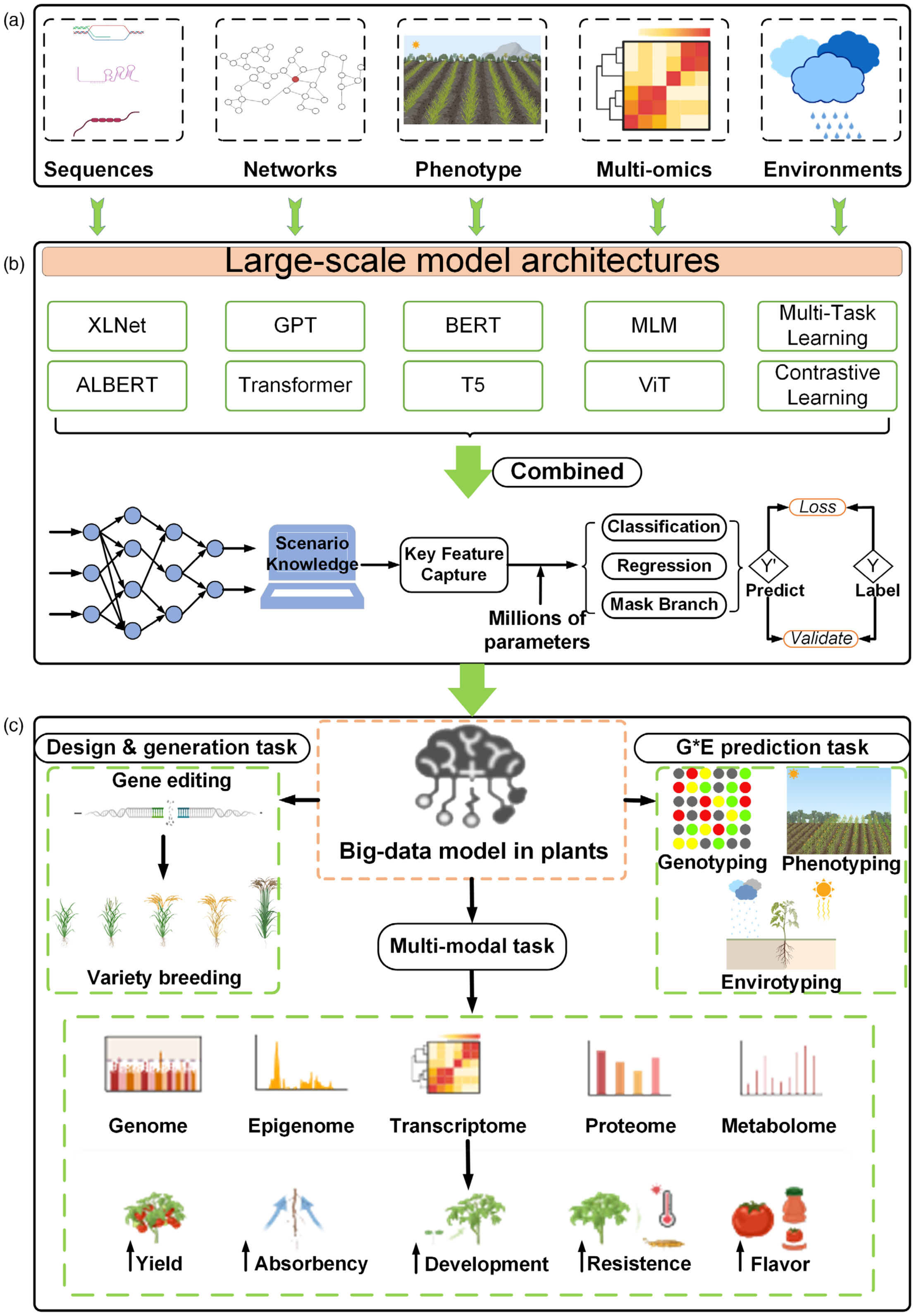


Figure 4. The application of large models in the field of plants.
(a) The multi-modal and large-scale dataset of plants.
(b) Large model architecture based on deep learning.
(c) The widespread application of large-scale models in plants.

integrate data from different plants and deep learning models (such as CNN, RNN, and transformer). This approach not only facilitates the study of individual species, but also promotes comparative research across species, thus revealing commonalities and characteristic patterns among plants. Additionally, the open-source learning platforms also support large-scale data sharing and model reuse. Researchers can obtain the latest pre-trained models and optimized algorithms, which can be directly applied to specific tasks.

4 Large-scale models open up new directions in plant research. Large-scale language models based on deep learning techniques are being applied in plant research, including the following aspects. (1) Large-scale and high-precision annotation of plant genomes based on large-scale models. (2) Analysis of gene expression regulation network based on large-scale model integration of multiple omics data in plants. (3) Apply large models to the acquisition and measurement of high-throughput phenotypes of crops. (4) Predicting the field phenotypes of crops in different environmental conditions based on large-scale models. (5) Plant disease prediction based on large-scale models can help develop precise prevention and control strategies.

5 Intelligent design of sequences based on deep learning promotes the development of synthetic biology. Utilizing deep learning to achieve intelligent design of functional elements is a key point in the development of synthetic biology in plants. By precisely editing and optimizing regulatory elements (such as promoters, enhancers, *etc.*) to precisely control gene expression, and improve crop yield and quality. For example, the new developed deep learning models of GAN and diffusion can generate sequences with different attributes based on additional feature information. Accordingly, the functional regulatory elements with specific functions can be designed and synthesized.

6 Deep learning assists in intelligent plant breeding and precision agriculture. By integrating genomic and phenotype data, deep learning models can be used to construct accurate phenotype prediction models. At the same time, the interpretability methods based on deep learning can be used to mine genes that control important agronomic traits. Genomic prediction not only accelerates the selection process of ideal breeding materials by breeders, but also significantly reduces the time and cost of field experiments. In addition, we can use deep learning technologies to construct an intelligent breeding decision support system. This system can optimize the breeding process, providing comprehensive intelligent services of parent selection, hybridization, and offspring selection. Similarly, deep learning algorithms can effectively analyze multi-modal data, including genomes, phenotype images, environmental sensing data, and agricultural production. Then, it can identify key factors that affect important traits (such as yield) of crops and further assist in field farming operations. In addition, the deep learning models can help optimize planting strategies and resource allocation, thereby improving agricultural production efficiency.

PERSPECTIVE

The future challenges of the application of deep learning in plants include the following aspects.

i The new emerging deep learning algorithms will be widely applied in plant genome analysis. The large models of GPT, Mamba, LLaMA, Hyena are bound to be increasingly used for analyzing DNA sequences. In addition, graphical neural networks (GNNs) and self-supervised learning methods will further optimize the analysis of plant genome data and promote the development of functional genomics and precision breeding.

ii Deep learning driven the integration analysis of multiple plant species. Deep learning will play an increasingly important role in the multi-omics data analysis of several species, including uncovering gene functions, evolution, and environmental adaptability analysis. The artificial intelligence algorithms will also play an important role in protein sequences and single-cell data from multiple plants.

iii Deep learning will empower accurate sequence design and intelligent breeding of crops. Researchers will use AI algorithms to design and generate specific gene regulatory elements (promoters, enhancers, CRE), thus accurately control the expression of genes and enhance crop yield, resistance, and quality. Through integrating various types of data (genotype, environment, phenotype), deep learning can help customize personal breeding strategy and accelerate the process of intelligent crop breeding.

SUMMARY BOXES

i Plant genome sequence analysis based on deep learning. On the basis of analyzing genome sequences, different kinds of deep learning methods have been used to predict gene expression, chromatin interactions, and epigenetic features in plants.

ii Design and synthesis of plant functional elements. The deep learning algorithms of GAN, diffusion, and large-scale models are advancing the design and synthesis of plant functional elements (promoters, enhancers, protein sequences, *etc.*).

iii Genomic prediction and intelligent design breeding. The AI-driven genomic prediction algorithms (G*E interactions, multiple traits) have become a research hotspot in the field of intelligent breeding. The large-scale algorithms will lead the construction of deep learning models for genome sequence analysis and design.

AUTHOR CONTRIBUTIONS

Conceptualization and writing—review and editing: JY and JL Writing—original draft: ZW and HY.

ACKNOWLEDGMENTS

This work has been supported by the National Key Research and Development Program of China (2022YFD1201504), the Biological Breeding-Major Projects (2023ZD04076), the Natural Science Foundation of Hubei Province (2023AFB832), the Major Project of Hubei Hongshan Laboratory (2022HSZD031), the Yingzi Tech & Huazhong Agricultural University Intelligent Research Institute of Food Health (IRIFH202209).

CONFLICT OF INTEREST

The authors have no conflicts of interest to declare.

REFERENCES

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A. ***et al.*** (2024) Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, **630**, 493–500. Available from: https://doi.org/10.1038/s41586-024-07487-w

Acharya, S., Ranjan, R., Pattanaik, S., Maiti, I.B. & **Dey, N.** (2014) Efficient chimeric plant promoters derived from plant infecting viral promoter sequences. *Planta*, **239**, 381–396. Available from: https://doi.org/10.1007/s00425-013-1973-2

Agarwal, V. & **Shendure, J.** (2020) Predicting mRNA abundance directly from genomic sequence using deep convolutional neural networks. *Cell Reports*, **31**, 107663. Available from: https://doi.org/10.1016/j.celrep.2020.107663

Akagi, T., Masuda, K., Kuwada, E., Takeshita, K., Kawakatsu, T., Ariizumi, T. ***et al.*** (2022) Genome-wide *cis*-decoding for expression design in tomato using cistrome data and explainable deep learning. *The Plant Cell*, **34**, 2174–2187. Available from: https://doi.org/10.1093/plcell/koac079

Al Bkhetan, Z. & **Plewczynski, D.** (2018) Three-dimensional epigenome statistical model: genome-wide chromatin looping prediction. *Scientific Reports*, **8**, 5217. Available from: https://doi.org/10.1038/s41598-018-23276-8

Alam, W., Tayara, H. & **Chong, K.T.** (2021) i4mC-deep: an intelligent predictor of N4-methylcytosine sites using a deep learning approach with chemical properties. *Genes*, **12**, 1117. Available from: https://doi.org/10.3390/genes12081117

Alipanahi, B., Delong, A., Weirauch, M.T. & **Frey, B.J.** (2015) Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. *Nature Biotechnology*, **33**, 831–838. Available from: https://doi.org/10.1038/nbt.3300

Anand, N. & **Huang, P.** (2018) Generative modeling for protein structures. *Advances in Neural Information Processing Systems*, **31**.

Avanti Shrikumar, K.T., Avsec, Ž., Shcherbina, A., Banerjee, A., Sharmin, M., Nair, S. ***et al.*** (2020) Technical note on transcription factor motif discovery from importance scores (TF-MoDISco) version 0.5.6.5. Available from: https://doi.org/10.48550/arXiv.1811.00416

Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J.R., Grabska-Barwinska, A., Taylor, K.R. ***et al.*** (2021) Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods*, **18**, 1196–1203. Available from: https://doi.org/10.1038/s41592-021-01252-x

Avsec, Ž., Weilert, M., Shrikumar, A., Krueger, S., Alexandari, A., Dalal, K. ***et al.*** (2021) Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nature Genetics*, **53**, 354–366. Available from: https://doi.org/10.1038/s41588-021-00782-6

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G.R. ***et al.*** (2021) Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, **373**, 871–876. Available from: https://doi.org/10.1126/science.abj8754

Bailey, T.L. & **Birol, I.** (2021) STREME: accurate and versatile sequence motif discovery. *Bioinformatics*, **37**, 2834–2840. Available from: https://doi.org/10.1093/bioinformatics/btab203

Bailey, T.L. (2011) DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics*, **27**, 1653–1659. Available from: https://doi.org/10.1093/bioinformatics/btr261

Bailey, T.L., Johnson, J., Grant, C.E. & **Noble, W.S.** (2015) The MEME suite. *Nucleic Acids Research*, **43**, W39–W49. Available from: https://doi.org/10.1093/nar/gkv416

Banerji, J., Rusconi, S. & **Schaffner, W.** (1981) Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. *Cell*, **27**, 299–308. Available from: https://doi.org/10.1016/0092-8674(81)90413-X

Beer, M.A. & **Tavazoie, S.** (2004) Predicting gene expression from sequence. *Cell*, **117**, 185–198. Available from: https://doi.org/10.1016/S0092-8674(04)00304-6

Benegas, G., Batra, S.S. & **Song, Y.S.** (2023) DNA language models are powerful predictors of genome-wide variant effects. *Proceedings of the National Academy of Sciences*, **120**, e2311219120. Available from: https://doi.org/10.1073/pnas.2311219120

Bigness, J., Loinaz, X., Patel, S., Larschan, E. & **Singh, R.** (2022) Integrating long-range regulatory interactions to predict gene expression using graph convolutional networks. *Journal of Computational Biology*, **29**, 409–424. Available from: https://doi.org/10.1089/cmb.2021.0316

Bukhari, S.A., Razzaq, A., Jabeen, J., Khan, S. & **Khan, Z.** (2021) Deep-BSC: predicting raw DNA binding pattern in *Arabidopsis thaliana*. *Current Bioinformatics*, **16**(3), 457–465. Available from: https://doi.org/10.2174/1574893615999200707142852

Cao, F., Zhang, Y., Cai, Y., Animesh, S., Zhang, Y., Akincilar, S.C. ***et al.*** (2021) Chromatin interaction neural network (ChINN): a machine learning-based method for predicting chromatin interactions from DNA sequences. *Genome Biology*, **22**, 1–25. Available from: https://doi.org/10.1186/s13059-021-02453-5

Cazzonelli, C.I. & **Velten, J.** (2008) In vivo characterization of plant promoter element interaction using synthetic promoters. *Transgenic Research*, **17**, 437–457. Available from: https://doi.org/10.1007/s11248-007-9117-8

Chen, J., Tan, C., Zhu, M., Zhang, C., Wang, Z., Ni, X. ***et al.*** (2024) CropGS-Hub: a comprehensive database of genotype and phenotype resources for genomic prediction in major crops. *Nucleic Acids Research*, **52**, D1519–D1529. Available from: https://doi.org/10.1093/nar/gkad1062

Chen, K., Zhao, H. & **Yang, Y.** (2022) Capturing large genomic contexts for accurately predicting enhancer–promoter interactions. *Briefings in Bioinformatics*, **23**, bbab577. Available from: https://doi.org/10.1093/bib/bbab577

Chen, W., Yang, H., Feng, P., Ding, H., Lin, H. & **Hancock, J.** (2017) iDNA4mC: identifying DNA N4-methylcytosine sites based on nucleotide chemical properties. *Bioinformatics*, **33**, 3518–3523. Available from: https://doi.org/10.1093/bioinformatics/btx479

Chen, Y., Yordanov, Y.S., Ma, C., Strauss, S. & **Busov, V.B.** (2013) DR5 as a reporter system to study auxin response in *Populus*. *Plant Cell Reports*, **32**, 453–463. Available from: https://doi.org/10.1007/s00299-012-1378-x

Cheng, C., Yan, K.-K., Yip, K.Y., Rozowsky, J., Alexander, R., Shou, C. ***et al.*** (2011) A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets. *Genome Biology*, **12**, 1–18. Available from: https://doi.org/10.1186/gb-2011-12-2-r15

Cheng, H., Liu, L., Zhou, Y., Deng, K., Ge, Y. & **Hu, X.** (2023) TSPTFBS 2.0: trans-species prediction of transcription factor binding sites and identification of their core motifs in plants. *Frontiers in Plant Science*, **14**, 1175837. Available from: https://doi.org/10.3389/fpls.2023.1175837

Chi, L., Sr., Ma, J., Sr., Wan, Y., Sr., Deng, Y., Sr., Wu, Y., Cen, X., Sr. ***et al.*** (2023) HGNNPIP: a hybrid graph neural network framework for protein–protein interaction prediction. *bioRxiv*. Available from: https://doi.org/10.1101/2023.12.10.571021

Chu, C. (2011) Saliency mapping of figure and ground of motion in Chinese. *Journal of the Chinese Language Teachers Association*, **46**, 49–69.

Chu, Y., Yu, D., Li, Y., Huang, K., Shen, Y., Cong, L. ***et al.*** (2024) A 5′ UTR language model for decoding untranslated regions of mRNA and function predictions. *Nature Machine Intelligence*, **6**, 449–460. Available from: https://doi.org/10.1038/s42256-024-00823-9

Cui, Y., Li, R., Li, G., Zhang, F., Zhu, T., Zhang, Q. ***et al.*** (2020) Hybrid breeding of rice via genomic selection. *Plant Biotechnology Journal*, **18**(1), 57–67. Available from: https://doi.org/10.1111/pbi.13170

Cui, Z., **Xu, T.**, **Wang, J.**, **Liao, Y.** & **Wang, Y.** (2024) Geneformer: learned gene compression using transformer-based context modeling. In: *ICASSP 2024–2024 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE. pp. 8035–8039.

Dalla-Torre, H., **Gonzalez, L.**, **Mendoza-Revilla, J.**, **Carranza, N.L.**, **Grzywaczewski, A.H.**, **Oteri, F.** ***et al.*** (2023) The nucleotide transformer: building and evaluating robust foundation models for human genomics. *bioRxiv*, 2023.2001.2011.523679. Available from: https://doi.org/10.1101/2023.01.11.523679

Dauparas, J., **Anishchenko, I.**, **Bennett, N.**, **Bai, H.**, **Ragotte, R.J.**, **Milles, L.F.** ***et al.*** (2022) Robust deep learning-based protein sequence design using ProteinMPNN. *Science*, **378**, 49–56. Available from: https://doi.org/10.1126/science.add2187

Davidsen, K., **Olson, B.J.**, **DeWitt, W.S.**, **Feng, J.**, **Harkins, E.**, **Bradley, P.** ***et al.*** (2019) Deep generative models for T cell receptor protein sequences. *eLife*, **8**, e46935. Available from: https://doi.org/10.7554/eLife.46935

de Almeida, B.P., **Reiter, F.**, **Pagani, M.** & **Stark, A.** (2022) DeepSTARR predicts enhancer activity from DNA sequence and enables the de novo design of synthetic enhancers. *Nature Genetics*, **54**, 613–624. Available from: https://doi.org/10.1038/s41588-022-01048-5

Deb, D., **Shrestha, A.**, **Maiti, I.B.** & **Dey, N.** (2018) Recombinant promoter (MUASCsV8CP) driven totiviral killer protein 4 (KP4) imparts resistance against fungal pathogens in transgenic tobacco. *Frontiers in Plant Science*, **9**, 278. Available from: https://doi.org/10.3389/fpls.2018.00278

Dong, X., **Greven, M.C.**, **Kundaje, A.**, **Djebali, S.**, **Brown, J.B.**, **Cheng, C.** ***et al.*** (2012) Modeling gene expression using chromatin features in various cellular contexts. *Genome Biology*, **13**, 1–17. Available from: https://doi.org/10.1186/gb-2012-13-9-r53

Dudnyk, K., **Cai, D.**, **Shi, C.**, **Xu, J.** & **Zhou, J.** (2024) Sequence basis of transcription initiation in the human genome. *Science*, **384**, eadj0116. Available from: https://doi.org/10.1126/science.adj0116

Ferruz, N., **Schmidt, S.** & **Höcker, B.** (2022) ProtGPT2 is a deep unsupervised language model for protein design. *Nature Communications*, **13**, 4348. Available from: https://doi.org/10.1038/s41467-022-32007-7

Fu, R. & **Wang, X.** (2023) Modeling the influence of phenotypic plasticity on maize hybrid performance. *Plant Communications*, **4**(3), 100548. Available from: https://doi.org/10.1016/j.xplc.2023.100548

Fudenberg, G., **Kelley, D.R.** & **Pollard, K.S.** (2020) Predicting 3D genome folding from DNA sequence with Akita. *Nature Methods*, **17**, 1111–1117. Available from: https://doi.org/10.1038/s41592-020-0958-x

Gao, P., **Zhao, H.**, **Luo, Z.**, **Lin, Y.**, **Feng, W.**, **Li, Y.** ***et al.*** (2023) SoyDNGP: a web-accessible deep learning framework for genomic prediction in soybean breeding. *Briefings in Bioinformatics*, **24**, bbad349. Available from: https://doi.org/10.1093/bib/bbad349

Gao, T. & **Qian, J.** (2019) EAGLE: an algorithm that utilizes a small number of genomic features to predict tissue/cell type-specific enhancer–gene interactions. *PLoS Computational Biology*, **15**, e1007436. Available from: https://doi.org/10.1371/journal.pcbi.1007436

Guo, W., **Liu, H.**, **Wang, Y.**, **Zhang, P.**, **Li, D.**, **Liu, T.** ***et al.*** (2022) SMOC: A smart model for open chromatin region prediction in rice genomes. *Journal of Genetics and Genomics*, **49**(5), 514–517. Available from: https://doi.org/10.1016/j.jgg.2022.02.012

Gupta, A. & **Zou, J.** (2018) Feedback GAN (FBGAN) for DNA: a novel feedback-loop architecture for optimizing protein functions. *arXiv preprint arXiv:1804.01694*. Available from: https://doi.org/10.48550/arXiv.1804.01694

Gupta, S., **Kesarwani, V.**, **Bhati, U.** & **Jyoti, S.R.** (2024) PTFSpot: deep co-learning on transcription factors and their binding regions attains impeccable universality in plants. *Briefings in Bioinformatics*, **25**, bbae324. Available from: https://doi.org/10.1093/bib/bbae324

Hafez, D., **Karabacak, A.**, **Krueger, S.**, **Hwang, Y.C.**, **Wang, L.S.**, **Zinzen, R.P.** ***et al.*** (2017) McEnhancer: predicting gene expression via semi-supervised assignment of enhancers to target genes. *Genome Biology*, **18**, 1–21. Available from: https://doi.org/10.1186/s13059-017-1316-x

He, W., **Jia, C.**, **Zou, Q.** & **Hancock, J.** (2019) 4mCPred: machine learning methods for DNA N4-methylcytosine sites prediction. *Bioinformatics*, **35**, 593–601. Available from: https://doi.org/10.1093/bioinformatics/bty668

Hsu, C., **Fannjiang, C.** & **Listgarten, J.** (2024) Generative models for protein structures and sequences. *Nature Biotechnology*, **42**, 196–199. Available from: https://doi.org/10.1038/s41587-023-02115-w

Jameel, A., **Ketehouli, T.**, **Wang, Y.**, **Wang, F.**, **Li, X.** & **Li, H.** (2022) Detection and validation of *cis*-regulatory motifs in osmotic stress-inducible synthetic gene switches via computational and experimental approaches. *Functional Plant Biology*, **49**, 1043–1054. Available from: https://doi.org/10.1071/FP21314

Ji, Y., **Zhou, Z.**, **Liu, H.** & **Davuluri, R.V.** (2021) DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics*, **37**, 2112–2120. Available from: https://doi.org/10.1093/bioinformatics/btab083

Jia, L. & **Luan, Y.** (2022) Multi-feature fusion method based on linear neighborhood propagation predict plant LncRNA–protein interactions. *Interdisciplinary Sciences: Computational LIfe Sciences*, **14**, 545–554. Available from: https://doi.org/10.1007/s12539-022-00501-7

Jighly, A., **Thayalakumaran, T.**, **O'Leary, G.J.**, **Kant, S.**, **Panozzo, J.**, **Aggarwal, R.** ***et al.*** (2023) Using genomic prediction with crop growth models enables the prediction of associated traits in wheat. *Journal of Experimental Botany*, **74**(5), 1389–1402. Available from: https://doi.org/10.1093/jxb/erac393

Jin, M., **Liu, H.**, **Liu, X.**, **Guo, T.**, **Guo, J.**, **Yin, Y.** ***et al.*** (2023) Complex genetic architecture underlying the plasticity of maize agronomic traits. *Plant Communications*, **4**(3), 100473. Available from: https://doi.org/10.1016/j.xplc.2022.100473

Jumper, J., **Evans, R.**, **Pritzel, A.**, **Green, T.**, **Figurnov, M.**, **Ronneberger, O.** ***et al.*** (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**, 583–589. Available from: https://doi.org/10.1038/s41586-021-03819-2

Karbalayghareh, A., **Sahin, M.** & **Leslie, C.S.** (2022) Chromatin interaction–aware gene regulatory modeling with graph attention networks. *Genome Research*, **32**, 930–944. Available from: https://doi.org/10.1101/gr.275870.121

Karimi, M., **Zhu, S.**, **Cao, Y.** & **Shen, Y.** (2020) De novo protein design for novel folds using guided conditional wasserstein generative adversarial networks. *Journal of Chemical Information and Modeling*, **60**, 5667–5681. Available from: https://doi.org/10.1021/acs.jcim.0c00593

Karlić, R., **Chung, H.-R.**, **Lasserre, J.**, **Vlahoviček, K.** & **Vingron, M.** (2010) Histone modification levels are predictive for gene expression. *Proceedings of the National Academy of Sciences*, **107**, 2926–2931. Available from: https://doi.org/10.1073/pnas.090934410

Kelley, D.R. (2020) Cross-species regulatory sequence activity prediction. *PLoS Computational Biology*, **16**, e1008050. Available from: https://doi.org/10.1371/journal.pcbi.1008050

Kelley, D.R., **Snoek, J.** & **Rinn, J.L.** (2016) Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Research*, **26**, 990–999. Available from: https://doi.org/10.1101/gr.200535.115

Koo, P.K. & **Eddy, S.R.** (2019) Representation learning of genomic sequence motifs with convolutional neural networks. *PLoS Computational Biology*, **15**(12), e1007560. Available from: https://doi.org/10.1371/journal.pcbi.1007560

Koo, P.K. & **Ploenzke, M.** (2021) Improving representations of genomic sequence motifs in convolutional networks with exponential activations. *Nature Machine Intelligence*, **3**(3), 258–266. Available from: https://doi.org/10.1038/s42256-020-00291-x

Koo, P.K., **Majdandzic, A.**, **Ploenzke, M.**, **Anand, P.** & **Paul, S.B.** (2021) Global importance analysis: an interpretability method to quantify importance of genomic features in deep neural networks. *PLoS Computational Biology*, **17** (5), e1008925. Available from: https://doi.org/10.1371/journal.pcbi.1008925

Kulmanov, M., **Hoehndorf, R.** & **Cowen, L.** (2020) DeepGOPlus: improved protein function prediction from sequence. *Bioinformatics*, **36**, 422–429. Available from: https://doi.org/10.1093/bioinformatics/btz595

Kumar, S., **Asif, M.H.**, **Chakrabarty, D.**, **Tripathi, R.D.**, **Dubey, R.S.** & **Trivedi, P.K.** (2015) Comprehensive analysis of regulatory elements of the promoters of rice sulfate transporter gene family and functional characterization of OsSul1; 1 promoter under different metal stress. *Plant Signaling & Behavior*, **10**(4), e990843. Available from: https://doi.org/10.4161/15592324.2014.990843

Lam, H.Y.I., **Ong, X.E.** & **Mutwil, M.** (2024) Large language models in plant biology. *Trends in Plant Science*, **29**, 1145–1155. Available from: https://doi.org/10.1016/j.tplants.2024.04.013

Langille, M., **Li, Z.**, **Jiang, H.**, **Kong, L.**, **Chen, Y.**, **Lang, K.** ***et al.*** (2021) Deep6mA: a deep learning framework for exploring similar patterns in DNA N6-methyladenine sites across different species. *PLoS*

Computational Biology, **17**, e1008767. Available from: https://doi.org/10.1371/journal.pcbi.1008767

Lee, D., **Yang, J.** & **Kim, S.** (2022) Learning the histone codes with large genomic windows and three-dimensional chromatin interactions using transformer. *Nature Communications*, **13**, 6678. Available from: https://doi.org/10.1038/s41467-022-34152-5

Levine, M. (2010) Transcriptional enhancers in animal development and evolution. *Current Biology*, **20**, R754–R763. Available from: https://doi.org/10.1016/j.cub.2010.06.070

Levy, B., **Xu, Z.**, **Zhao, L.**, **Kremling, K.**, **Altman, R.**, **Wong, P.** ***et al.*** (2022) FloraBERT: cross-species transfer learning with attention-based neural networks for gene expression prediction. Available from: https://doi.org/10.21203/rs.3.rs-1927200/v1

Li, J., **Wang, J.**, **Zhang, P.**, **Wang, R.**, **Mei, Y.**, **Sun, Z.** ***et al.*** (2022) Deep learning of cross-species single-cell landscapes identifies conserved regulatory programs underlying cell types. *Nature Genetics*, **54**(11), 1711–1720. Available from: https://doi.org/10.1038/s41588-022-01197-7

Li, T., **Xu, H.**, **Teng, S.**, **Suo, M.**, **Bahitwa, R.**, **Xu, M.** ***et al.*** (2024) Modeling 0.6 million genes for the rational design of functional *cis*-regulatory variants and de novo design of *cis*-regulatory sequences. *Proceedings of the National Academy of Sciences*, **121**, e2319811121. Available from: https://doi.org/10.1073/pnas.2319811121

Li, W., **Wong, W.H.** & **Jiang, R.** (2019) DeepTACT: predicting 3D chromatin contacts via bootstrapping deep learning. *Nucleic Acids Research*, **47**, e60. Available from: https://doi.org/10.1093/nar/gkz167

Li, Y., **Ju, F.**, **Chen, Z.**, **Qu, Y.**, **Xia, H.**, **He, L.** ***et al.*** (2023) CREaTor: zero-shot *cis*-regulatory pattern modeling with attention mechanisms. *Genome Biology*, **24**, 266. Available from: https://doi.org/10.1186/s13059-023-03103-8

Li, Y., **Li, X.**, **Gao, R.**, **Liu, W.** & **Tan, G.** (2023) NvWa: enhancing sequence alignment accelerator throughput via hardware scheduling. In: *2023 IEEE international symposium on high-performance computer architecture (HPCA)*. IEEE. pp. 1236–1248.

Li, J., **Zhang, D.**, **Yang, F.**, **Zhang, Q.**, **Pan, S.**, **Zhao, X.** ***et al.*** (2024) TrG2P: a transfer learning-based tool integrating multi-trait data for accurate prediction of crop yield. *Plant Communications*, **5**(7), 100975. Available from: https://doi.org/10.1016/j.xplc.2024.100975

Lin, Z.M., **Akin, H.**, **Rao, R.S.**, **Hie, B.**, **Zhu, Z.K.**, **Lu, W.T.** ***et al.*** (2023) Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, **379**, 1123–1130. Available from: https://doi.org/10.1126/science.ade2574

Liu, L., **Zhang, G.**, **He, S.** & **Hu, X.** (2021) TSPTFBS: a docker image for trans-species prediction of transcription factor binding sites in plants. *Bioinformatics*, **37**(2), 260–262. Available from: https://doi.org/10.1093/bioinformatics/btaa1100

Liu, N., **Ding, C.J.**, **Li, B.**, **Ding, M.**, **Su, X.H.** & **Huang, Q.J.** (2020) Effects of genotype by environment interaction of 12 Populus×euramericana clones in their early growth. *Scientia Silvae Sinicae*, **8**(56), 63–72. Available from: https://doi.org/10.11707/j.1001-7488.20200808

Liu, W. & **Stewart, C.N.**, **Jr.** (2016) Plant synthetic promoters and transcription factors. *Current Opinion in Biotechnology*, **37**, 36–44. Available from: https://doi.org/10.1016/j.copbio.2015.10.001

Liu, W., **Yuan, J.S.** & **Stewart, C.N.**, **Jr.** (2013) Advanced genetic tools for plant biotechnology. *Nature Reviews Genetics*, **14**, 781–793. Available from: https://doi.org/10.1038/nrg3583

Luo, Z., **Zhang, J.**, **Fei, J.** & **Ke, S.** (2022) Deep learning modeling m6A deposition reveals the importance of downstream *cis*-element sequences. *Nature Communications*, **13**, 2720. Available from: https://doi.org/10.1038/s41467-022-30209-7

Ma, W., **Qiu, Z.**, **Song, J.**, **Li, J.**, **Cheng, Q.**, **Zhai, J.** ***et al.*** (2018) A deep convolutional neural network approach for predicting phenotypes from genotypes. *Planta*, **248**, 1307–1318. Available from: https://doi.org/10.1007/s00425-018-2976-9

Ma, X., **Wang, H.**, **Wu, S.**, **Han, B.**, **Cui, D.**, **Liu, J.** ***et al.*** (2024) DeepCCR: large-scale genomics-based deep learning method for improving rice breeding. *Plant Biotechnology Journal*, **22**, 2691–2693. Available from: https://doi.org/10.1111/pbi.14384

Mehrotra, R., **Gupta, G.**, **Sethi, R.**, **Bhalothia, P.**, **Kumar, N.** & **Mehrotra, S.** (2011) Designer promoter: an artwork of *cis* engineering. *Plant Molecular Biology*, **75**, 527–536. Available from: https://doi.org/10.1007/s11103-011-9755-3

Mendoza-Revilla, J., **Trop, E.**, **Gonzalez, L.**, **Roller, M.**, **Dalla-Torre, H.**, **de Almeida, B.P.** ***et al.*** (2024) A foundational large language model for edible plant genomes. *Communications Biology*, **7**, 835. Available from: https://doi.org/10.1101/2023.10.24.563624

Millet, E.J., **Kruijer, W.**, **Coupel-Ledru, A.**, **Alvarez Prado, S.**, **Cabrera-Bosquet, L.**, **Lacube, S.** ***et al.*** (2019) Genomic prediction of maize yield across European environmental conditions. *Nature Genetics*, **51**(6), 952–956. Available from: https://doi.org/10.1038/s41588-019-0414-y

Morffy, N., **van den Broeck, L.**, **Miller, C.**, **Emenecker, R.J.**, **Bryant, J.A.**, **Jr.**, **Lee, T.M.** ***et al.*** (2024) Identification of plant transcriptional activation domains. *Nature*, **632**, 1–8. Available from: https://doi.org/10.1038/s41586-024-07707-3

Morris, Q., **Ghandi, M.**, **Lee, D.**, **Mohammad-Noori, M.** & **Beer, M.A.** (2014) Enhanced regulatory sequence prediction using gapped k-mer features. *PLoS Computational Biology*, **10**, e1003711. Available from: https://doi.org/10.1371/journal.pcbi.1003711

Muthamilarasan, M. & **Prasad, M.** (2015) Advances in *Setaria* genomics for genetic improvement of cereals and bioenergy grasses. *Theoretical and Applied Genetics*, **128**, 1–14. Available from: https://doi.org/10.1007/s00122-014-2399-3

Nguyen, E., **Poli, M.**, **Durrant, M.G.**, **Thomas, A.W.**, **Kang, B.**, **Sullivan, J.** ***et al.*** (2024) Sequence modeling and design from molecular to genome scale with Evo. *bioRxiv*, 2024.2002.2027.582234. Available from: https://doi.org/10.1101/2024.02.27.582234

Ni, P., **Huang, N.**, **Nie, F.**, **Zhang, J.**, **Zhang, Z.**, **Wu, B.** ***et al.*** (2021) Genome-wide detection of cytosine methylations in plant from Nanopore data using deep learning. *Nature Communications*, **12**, 5976. Available from: https://doi.org/10.1038/s41467-021-26278-9

Peleke, F.F., **Zumkeller, S.M.**, **Gültas, M.**, **Schmitt, A.** & **Szymański, J.** (2024) Deep learning the *cis*-regulatory code for gene expression in selected model plants. *Nature Communications*, **15**, 3488. Available from: https://doi.org/10.1038/s41467-024-47744-0

Raimondi, D., **Corso, M.**, **Fariselli, P.** & **Moreau, Y.** (2022) From genotype to phenotype in *Arabidopsis thaliana*: in-silico genome interpretation predicts 288 phenotypes from sequencing data. *Nucleic Acids Research*, **50**, e16. Available from: https://doi.org/10.1093/nar/gkab1099

Ramisch, A., **Heinrich, V.**, **Glaser, L.V.**, **Fuchs, A.**, **Yang, X.**, **Benner, P.** ***et al.*** (2019) CRUP: a comprehensive framework to predict condition-specific regulatory units. *Genome Biology*, **20**, 1–23. Available from: https://doi.org/10.1186/s13059-019-1860-7

Ray, S., **Jarquin, D.** & **Howard, R.** (2022) Comparing artificial-intelligence techniques with state-of-the-art parametric prediction models for predicting soybean traits. *The Plant Genome*, **16**, e20263. Available from: https://doi.org/10.1002/tpg2.20263

Repecka, D., **Jauniskis, V.**, **Karpus, L.**, **Rembeza, E.**, **Rokaitis, I.**, **Zrimec, J.** ***et al.*** (2021) Expanding functional protein sequence spaces using generative adversarial networks. *Nature Machine Intelligence*, **3**, 324–333. Available from: https://doi.org/10.1038/s42256-021-00310-5

Roy, S., **Siahpirani, A.F.**, **Chasman, D.**, **Knaack, S.**, **Ay, F.**, **Stewart, R.** ***et al.*** (2015) A predictive modeling approach for cell line-specific long-range regulatory interactions. *Nucleic Acids Research*, **43**, 8694–8712.

Rushton, P.J. (2016) What have we learned about synthetic promoter construction? *Plant Synthetic Promoters: Methods and Protocols*, **1482**, 1–13. Available from: https://doi.org/10.1007/978-1-4939-6396-6_1

Rushton, P.J., **Reinstädler, A.**, **Lipka, V.**, **Lippok, B.** & **Somssich, I.E.** (2002) Synthetic plant promoters containing defined regulatory elements provide novel insights into pathogen-and wound-induced signaling. *The Plant Cell*, **14**, 749–762. Available from: https://doi.org/10.1105/tpc.010412

Schlegel, L., **Bhardwaj, R.**, **Shahryary, Y.**, **Demirtürk, D.**, **Marand, A.P.**, **Schmitz, R.J.** ***et al.*** (2024) GenomicLinks: deep learning predictions of 3D chromatin loops in the maize genome. *bioRxiv*, 2024.2005.2006.592633. Available from: https://doi.org/10.1101/2024.05.06.592633

Schreiber, J., **Libbrecht, M.**, **Bilmes, J.** & **Noble, W.S.** (2017) Nucleotide sequence and DNaseI sensitivity are predictive of 3D chromatin architecture. *bioRxiv*, 103614. Available from: https://doi.org/10.1101/103614

Schwessinger, R., **Gosden, M.**, **Downes, D.**, **Brown, R.**, **Telenius, J.**, **Teh, Y.W.** ***et al.*** (2019) DeepC: predicting chromatin interactions using megabase scaled deep neural networks and transfer learning. *bioRxiv*, 724005. Available from: https://doi.org/10.1101/724005

Seitz, E.E., **McCandlish, D.M.**, **Kinney, J.B.** & **Koo, P.K.** (2024) Interpreting *cis*-regulatory mechanisms from genomic deep neural networks using surrogate models. *Nature Machine Intelligence*, **6**, 1–13. Available from: https://doi.org/10.1038/s42256-024-00851-5

Shen, W., **Pan, J.**, **Wang, G.** & **Li, X.** (2021) Deep learning-based prediction of TFBS in plants. *Trends in Plant Science*, **26**, 1301–1302. Available from: https://doi.org/10.1016/j.tplants.2021.06.016

Shen, Y., **Chen, L.-L.** & **Gao, J.** (2021) CharPlant: a De novo open chromatin region prediction tool for plant genomes. *Genomics, Proteomics & Bioinformatics*, **19**, 860–871. Available from: https://doi.org/10.1016/j.gpb.2020.06.021

Shen, Y., **Zhong, Q.**, **Liu, T.**, **Wen, Z.**, **Shen, W.** & **Li, L.** (2022) CharID: a two-step model for universal prediction of interactions between chromatin accessible regions. *Briefings in Bioinformatics*, **23**, bbab602. Available from: https://doi.org/10.1093/bib/bbab602

Shen, Z., **Shen, E.**, **Yang, K.**, **Fan, Z.**, **Zhu, Q.-H.**, **Fan, L.** ***et al.*** (2024) BreedingAIDB: a database integrating crop genome-to-phenotype paired data with machine learning tools applicable to breeding. *Plant Communications*, **5**, 100894. Available from: https://doi.org/10.1016/j.xplc.2024.100894

Shlyueva, D., **Stampfel, G.** & **Stark, A.** (2014) Transcriptional enhancers: from properties to genome-wide predictions. *Nature Reviews Genetics*, **15**, 272–286. Available from: https://doi.org/10.1038/nrg3682

Shrikumar, A., **Greenside, P.** & **Kundaje, A.** (2017) Learning important features through propagating activation differences. In: *International conference on machine learning*. PMLR. pp. 3145–3153.

Sinai, S., **Kelsic, E.**, **Church, G.M.** & **Nowak, M.A.** (2017) Variational autoencoding of protein sequences. *arXiv preprint arXiv:1712.03346*. Available from: https://doi.org/10.48550/arXiv.1712.03346

Singh, R., **Lanchantin, J.**, **Robins, G.** & **Qi, Y.D.** (2016) Deep-learning for predicting gene expression from histone modifications. *Bioinformatics*, **32**, i639–i648.

Singh, R., **Lanchantin, J.**, **Sekhon, A.** & **Qi, Y.** (2017) Attend and predict: Understanding gene regulation by selective attention on chromatin. *Advances in Neural Information Processing Systems*, **30**, 6785–6795.

Singh, S., **Yang, Y.**, **Póczos, B.** & **Ma, J.** (2019) Predicting enhancer–promoter interaction from genomic sequence with deep neural networks. *Quantitative Biology*, **7**, 122–137. Available from: https://doi.org/10.1007/s40484-019-0154-0

Tasaki, S., **Gaiteri, C.**, **Mostafavi, S.** & **Wang, Y.** (2020) Deep learning decodes the principles of differential gene expression. *Nature Machine Intelligence*, **2**, 376–386. Available from: https://doi.org/10.1038/s42256-020-0201-6

Taskiran, I.I., **Spanier, K.I.**, **Dickmänken, H.**, **Kempynck, N.**, **Pančíková, A.**, **Ekşi, E.C.** ***et al.*** (2024) Cell-type-directed design of synthetic enhancers. *Nature*, **626**, 212–220. Available from: https://doi.org/10.1038/s41586-023-06936-2

Thomas-Chollier, M., **Herrmann, C.**, **Defrance, M.**, **Sand, O.**, **Thieffry, D.** & **van Helden, J.** (2012) RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets. *Nucleic Acids Research*, **40**, e31. Available from: https://doi.org/10.1093/nar/gkr1104

Toneyan, S. & **Koo, P.K.** (2023) Interpreting *cis*-regulatory interactions from large-scale deep neural networks for genomics. *bioRxiv*. Available from: https://doi.org/10.1101/2023.07.03.547592

Tong, H. & **Nikoloski, Z.** (2021) Machine learning approaches for crop improvement: leveraging phenotypic and genotypic big data. *Journal of Plant Physiology*, **257**, 153354. Available from: https://doi.org/10.1016/j.jplph.2020.153354

Trieu, T., **Martinez-Fundichely, A.** & **Khurana, E.** (2020) DeepMILO: a deep learning approach to predict the impact of non-coding sequence variants on 3D chromatin structure. *Genome Biology*, **21**, 1–11. Available from: https://doi.org/10.1186/s13059-020-01987-4

Vaishnav, E.D., **de Boer, C.G.**, **Molinet, J.**, **Yassour, M.**, **Fan, L.**, **Adiconis, X.** ***et al.*** (2022) The evolution, evolvability and engineering of gene regulatory DNA. *Nature*, **603**, 455–463. Available from: https://doi.org/10.1038/s41586-022-04506-6

Wang, J., **Lisanza, S.**, **Juergens, D.**, **Tischer, D.**, **Watson, J.L.**, **Castro, K.M.** ***et al.*** (2022) Scaffolding protein functional sites using deep learning. *Science*, **377**, 387–394. Available from: https://doi.org/10.1126/science.abn2100

Wang, K., **Abid, M.A.**, **Rasheed, A.**, **Crossa, J.**, **Hearne, S.** & **Li, H.** (2023) DNNGP, a deep neural network-based method for genomic prediction using multi-omics data in plants. *Molecular Plant*, **16**(1), 279–293. Available from: https://doi.org/10.1016/j.molp.2022.11.004

Wang, X., **Xu, K.**, **Tan, Y.**, **Yu, S.**, **Zhao, X.** & **Zhou, J.** (2023) Deep learning-assisted design of novel promoters in *Escherichia coli*. *Advanced Genetics*, **4**, 2300184. Available from: https://doi.org/10.1002/ggn2.202300184

Wang, Y., **Zhang, P.**, **Guo, W.**, **Liu, H.**, **Li, X.**, **Zhang, Q.** ***et al.*** (2021) A deep learning approach to automate whole-genome prediction of diverse epigenomic modifications in plants. *New Phytologist*, **232**(2), 880–897. Available from: https://doi.org/10.1111/nph.17630

Wang, Z., **Peng, Y.**, **Li, J.**, **Li, J.**, **Yuan, H.**, **Yang, S.** ***et al.*** (2024) DeepCBA: a deep learning framework for gene expression prediction in maize based on DNA sequence and chromatin interaction. *Plant Communications*, **5**. Available from: https://doi.org/10.1016/j.xplc.2024.100985

Washburn, J.D., **Mejia-Guerra, M.K.**, **Ramstein, G.**, **Kremling, K.A.**, **Valluru, R.**, **Buckler, E.S.** ***et al.*** (2019) Evolutionarily informed deep learning methods for predicting relative transcript abundance from DNA sequence. *Proceedings of the National Academy of Sciences*, **116**, 5542–5549. Available from: https://doi.org/10.1073/pnas.1814551116

Watson, J.L., **Juergens, D.**, **Bennett, N.R.**, **Trippe, B.L.**, **Yim, J.**, **Eisenach, H.E.** ***et al.*** (2023) De novo design of protein structure and function with RFdiffusion. *Nature*, **620**, 1089–1100. Available from: https://doi.org/10.1038/s41586-023-06415-8

Wei, L., **Su, R.**, **Luan, S.**, **Liao, Z.**, **Manavalan, B.**, **Zou, Q.** ***et al.*** (2019) Iterative feature representations improve N4-methylcytosine site prediction. *Bioinformatics*, **35**, 4930–4937. Available from: https://doi.org/10.1093/bioinformatics/btz408

Wei, Z., **Hua, K.**, **Wei, L.**, **Ma, S.**, **Jiang, R.**, **Zhang, X.** ***et al.*** (2023) NeuronMotif: Deciphering cis-regulatory codes by layer-wise demixing of deep neural networks. *Proceedings of the National Academy of Sciences of the United States of America*, **120**(15), e2216698120. Available from: https://doi.org/10.1073/pnas.2216698120

Whalen, S., **Truty, R.M.** & **Pollard, K.S.** (2016) Enhancer–promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nature Genetics*, **48**, 488–496. Available from: https://doi.org/10.1038/ng.3539

Wrightsman, T., **Marand, A.P.**, **Crisp, P.A.**, **Springer, N.M.** & **Buckler, E.S.** (2022) Modeling chromatin state from sequence across angiosperms using recurrent convolutional neural networks. *The Plant Genome*, **15**(3), e20249. Available from: https://doi.org/10.1002/tpg2.20249

Wu, C., **Zhang, Y.**, **Ying, Z.**, **Li, L.**, **Wang, J.**, **Yu, H.** ***et al.*** (2024) A transformer-based genomic prediction method fused with knowledge-guided module. *Briefings in Bioinformatics*, **25**, bbad438. Available from: https://doi.org/10.1093/bib/bbad438

Xu, H., **Jia, P.** & **Zhao, Z.** (2021) Deep4mC: systematic assessment and computational prediction for DNA N4-methylcytosine sites by deep learning. *Briefings in Bioinformatics*, **22**, bbaa099. Available from: https://doi.org/10.1093/bib/bbaa099

Yan, W., **Li, Z.**, **Pian, C.** & **Wu, Y.** (2022) PlantBind: an attention-based multilabel neural network for predicting plant transcription factor binding sites. *Briefings in Bioinformatics*, **23**, bbac425. Available from: https://doi.org/10.1093/bib/bbac425

Yan, X., **Wang, Z.**, **Jia, Y.**, **Zhang, Z.** & **Huang, Y.** (2024) Access point selection and beamforming design for cell-free network: from fractional programming to GNN. *IEEE Transactions on Wireless Communications*, **23**, 9345–9360. Available from: https://doi.org/10.1109/TWC.2024.3361900

Yang, B., **Liu, F.**, **Ren, C.**, **Ouyang, Z.**, **Xie, Z.**, **Bo, X.** ***et al.*** (2017) BiRen: predicting enhancers with a deep-learning-based model using the DNA sequence alone. *Bioinformatics*, **33**, 1930–1936. Available from: https://doi.org/10.1093/bioinformatics/btx105

Yang, Y., **Lee, J.H.**, **Poindexter, M.R.**, **Shao, Y.**, **Liu, W.**, **Lenaghan, S.C.** ***et al.*** (2021) Rational design and testing of abiotic stress-inducible synthetic promoters from poplar *cis*-regulatory elements. *Plant Biotechnology Journal*, **19**, 1354–1369. Available from: https://doi.org/10.1111/pbi.13550

Zeng, S., **Mao, Z.**, **Ren, Y.**, **Wang, D.**, **Xu, D.** & **Joshi, T.** (2021) G2PDeep: a web-based deep-learning framework for quantitative phenotype prediction and discovery of genomic markers. *Nucleic Acids Research*, **49**, W228–W236. Available from: https://doi.org/10.1093/nar/gkab407

Zeng, W., **Wang, Y.** & **Jiang, R.** (2020) Integrating distal and proximal information to predict gene expression via a densely connected convolutional neural network. *Bioinformatics*, **36**, 496–503. Available from: https://doi.org/10.1093/bioinformatics/btz562

Zhai, J., **Gokaslan, A.**, **Schiff, Y.**, **Berthel, A.**, **Liu, Z. Y.**, **Miller, Z. R.**, **Scheben, A.**, **Stitzer, M.C.**, **Romay, M.**, **Buchler, E.**, **Kuleshov, V.** (2024). Cross-species modeling of plant genomes at single nucleotide resolution using a pre-trained DNA language model. *bioRxiv*, 2024-06. Available from: https://doi.org/10.1101/2024.06.04.596709

Zhang, C., **Freddolino, P.L.** & **Zhang, Y.** (2017) COFACTOR: improved protein function prediction by combining structure, sequence and protein–protein interaction information. *Nucleic Acids Research*, **45**, W291–W299. Available from: https://doi.org/10.1093/nar/gkx366

Zhang, D., **Zhang, W.**, **He, B.**, **Zhang, J.**, **Qin, C.** & **Yao, J.** (2023) DNAGPT: a generalized pretrained tool for multiple DNA sequence analysis tasks. *bioRxiv*, 2023.2007. 2011.548628. Available from: https://doi.org/10.1101/2023.07.11.548628

Zhang, F., **Song, H.**, **Zeng, M.**, **Li, Y.**, **Kurgan, L.** & **Li, M.** (2019) DeepFunc: a deep learning framework for accurate prediction of protein functions from protein sequences and interactions. *Proteomics*, **19**, e1900019. Available from: https://doi.org/10.1002/pmic.201900019

Zhang, L., **Yang, R.**, **Xia, D.**, **Lin, X.** & **Xiong, W.** (2024) Prediction of plant LncRNA–protein interactions based on feature fusion and an improved residual network. *Expert Systems with Applications*, **238**, 121991. Available from: https://doi.org/10.1016/j.eswa.2023.121991

Zhang, P., **Wang, H.**, **Xu, H.**, **Wei, L.**, **Liu, L.**, **Hu, Z.** ***et al.*** (2023) Deep flanking sequence engineering for efficient promoter design using DeepSEED. *Nature Communications*, **14**, 6309. Available from: https://doi.org/10.1038/s41467-023-41899

Zhang, R., **Wang, Y.**, **Yang, Y.**, **Zhang, Y.** & **Ma, J.** (2018) Predicting CTCF-mediated chromatin loops using CTCF-MP. *Bioinformatics*, **34**, i133–i141. Available from: https://doi.org/10.1093/bioinformatics/bty248

Zhao, H., **Tu, Z.**, **Liu, Y.**, **Zong, Z.**, **Li, J.**, **Liu, H.** ***et al.*** (2021) PlantDeepSEA, a deep learning-based web service to predict the regulatory effects of genomic variants in plants. *Nucleic Acids Research*, **49**, W523–W529. Available from: https://doi.org/10.1093/nar/gkab383

Zhou, H., **Wekesa, J.S.**, **Luan, Y.** & **Meng, J.** (2021) PRPI-SC: an ensemble deep learning model for predicting plant lncRNA–protein interactions. *BMC Bioinformatics*, **22**, 1–15. Available from: https://doi.org/10.1186/s12859-021-04328-9

Zhou, J. & **Troyanskaya, O.G.** (2015) Predicting effects of noncoding variants with deep learning-based sequence model. *Nature Methods*, **12**, 931–934. Available from: https://doi.org/10.1038/nmeth.3547

Zhou, J., **Theesfeld, C.L.**, **Yao, K.**, **Chen, K.M.**, **Wong, A.K.** & **Troyanskaya, O.G.** (2018) Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nature Genetics*, **50**, 1171–1179. Available from: https://doi.org/10.1038/s41588-018-0160-6

Zhou, K., **Lei, C.**, **Zheng, J.** ***et al.*** (2023) Pre-trained protein language model sheds new light on the prediction of Arabidopsis protein–protein interactions. *Plant Methods*, **19**, 141. Available from: https://doi.org/10.1186/s13007-023-01119-6

Zhou, Z., **Ji, Y.**, **Li, W.**, **Dutta, P.**, **Davuluri, R.** & **Liu, H.** (2024) Dnabert-2: Efficient foundation model and benchmark for multi-species genome. ArXiv Preprint ArXiv:2306.15006.

Zhu, W., **Han, R.**, **Shang, X.**, **Zhou, T.**, **Liang, C.**, **Qin, X.** ***et al.*** (2024) The CropGPT project: call for a global, coordinated effort in precision design breeding driven by AI using biological big data. *Molecular Plant*, **17**, 215–218. Available from: https://doi.org/10.1016/j.molp.2023.12.015

Zhu, Y., **Chen, Z.**, **Zhang, K.**, **Wang, M.**, **Medovoy, D.**, **Whitaker, J.W.** ***et al.*** (2016) Constructing 3D interaction maps from 1D epigenomes. *Nature Communications*, **7**, 10812. Available from: https://doi.org/10.1038/ncomms10812