

Genome analysis

Deep learning for regulatory genomics: a survey of models, challenges, and applications

Fathima Nuzla Ismail^{1,*}, Abira Sengupta², Shanika Amarasoma³

¹Department of Mathematics, State University of New York at Buffalo, Buffalo, NY 14260, United States

²School of Computing, University of Otago, Dunedin 9016, New Zealand

³Independent Researcher, AI & Advanced Analytics, Colombo 00900, Sri Lanka

*Corresponding author. Department of Mathematics, State University of New York at Buffalo, Buffalo, NY 14260, United States.

E-mail: fathima.nuzla.ismail@gmail.com.

Associate Editor: Vera Pancaldi

Abstract

This research reviews recent advances in deep learning approaches tailored for regulatory genomics. It highlights how computational methods help decipher complex regulatory mechanisms within non-coding genomic regions across various tissues, emphasizing predictive applications such as transcription factor binding, chromatin accessibility, RNA processes, and RNA-protein interactions. The paper also discusses the evolution from traditional neural networks to advanced models like transformers and graph neural networks, considering three-dimensional genomic structures. Despite the promising performance, it acknowledges ongoing challenges like overfitting, biological variability, and limited dataset diversity. It emphasizes the urgent need for continued development of interpretable deep learning models to improve functional genomic annotation, underlining this task's significance in the genomics field.

1 Introduction

A key topic in cell biology is comprehending gene regulation, which takes place at the transcriptomic, proteomic, and genomic levels. Over the past two decades, breakthroughs in omics technologies—such as proteomics, transcriptomics, and genomics (Fig. 1)—have considerably improved our ability to examine gene control. Single-cell approaches, mass spectrometry, high-throughput sequencing, and microarrays have all made it possible to investigate regulatory systems at the fine scale. ChIP-seq (protein-DNA binding), CLIP-seq (Ule et al. 2005) (protein-RNA binding), DNase-seq and ATAC-seq (Buenrostro et al. 2013) (chromatin accessibility), RNA-seq (gene expression), and 30-READS (Hoque et al. 2013) (polyadenylation analysis) are now widely used methods.

These tools underpin large-scale projects like the 1000 Genomes Project, ENCODE, Roadmap Epigenomics, and GTEx, which aim to elucidate the biological pathways linking genotype to phenotype across individuals, species, and tissues. As omics datasets grow, so does the demand for analytical algorithms that can extract meaningful patterns. Traditionally, this has involved “shallow learning” methods—such as logistic regression (LR), support vector machines (SVM), and hidden Markov models (HMMs)—which depend on manually engineered features and predefined statistical models (Zeng and Lumley 2018). While effective, these models often fall short when prior knowledge is limited, as is frequently the case in omics. Their reliance on engineered features can reduce performance, limit discovery, and hinder model generalisability.

In contrast, learning algorithms that automatically extract discriminative features from data offer a compelling alternative. Deep learning models, in particular, can learn such features through training, typically via gradient descent optimization of a loss function (LeCun et al. 2015). These models use deep, hierarchical architectures—Deep Neural Networks (DNNs)—in which lower layers learn data-driven representations, while upper layers summarise them to perform inference. This makes deep learning especially promising for omics, where unknown patterns in biological sequences or measurements can be uncovered without relying on pre-engineered features.

The success of deep learning in omics is also due to the compatibility of biological data formats—such as DNA and RNA sequences—with models originally designed for natural language processing. For instance, regulatory grammar (i.e. interactions between sequence motifs) parallels linguistic structure, enabling deep learning to model complex regulatory codes (Alipanahi et al. 2015). This has led to practical applications and continues to inspire further exploration of deep learning's role in decoding gene regulation.

In this survey, we provide a comprehensive overview of deep learning applications in gene regulation research using diverse omics data. We highlight applications across transcriptomic and genomic levels, explore various prediction tasks addressing biological questions, and examine model architectures and design principles. We also discuss emerging trends in integrating structural data, multi-omics profiles, and single-cell information. Our aim is to offer an up-to-date resource for researchers entering this exciting and rapidly evolving field.

Received: May 12, 2025; Revised: October 15, 2025; Accepted: October 22, 2025

© The Author(s) 2025. Published by Oxford University Press.

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

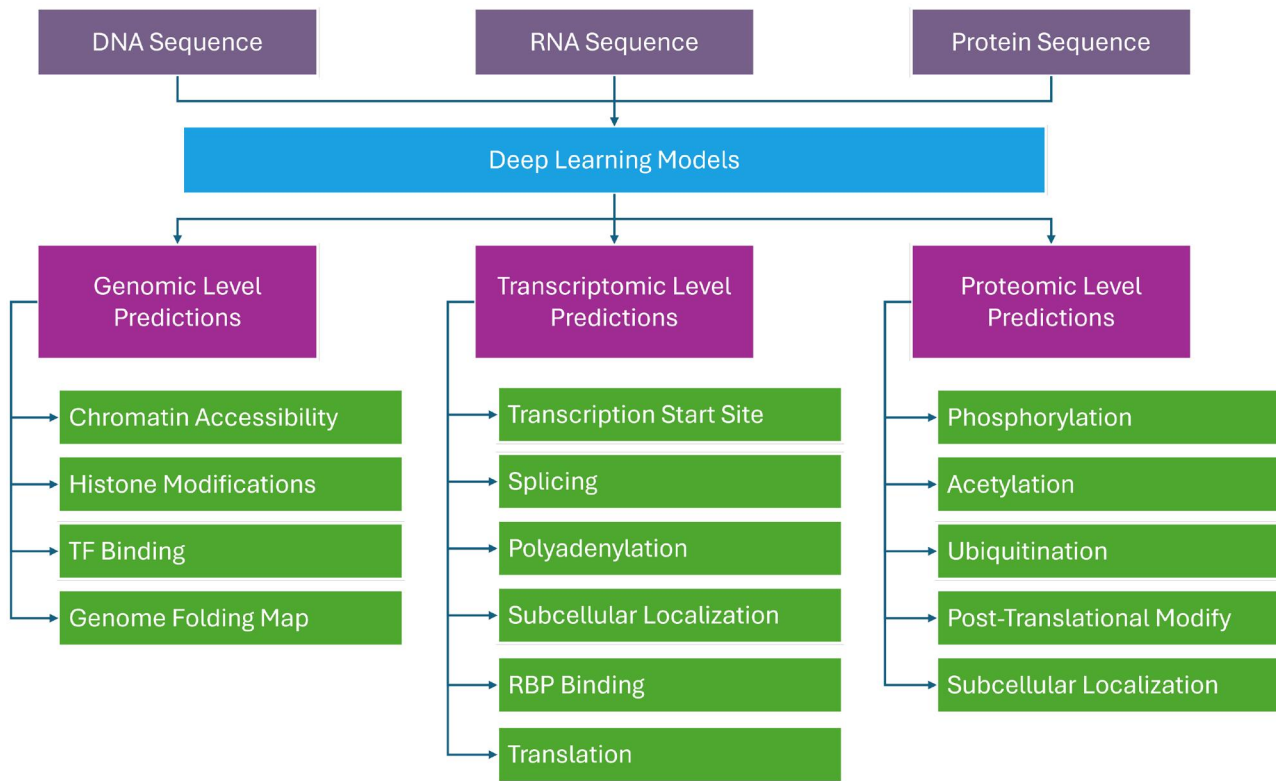


Figure 1. The figure illustrates an overview of using deep learning models to derive predictions at genomic, transcriptomic, and proteomic levels based on DNA, RNA, and protein sequences, respectively.

As of 2025, the field has shifted toward foundation-scale, multimodal architectures such as AlphaGenome (Avsec *et al.* 2025), which unify genomic, epigenomic, and transcriptomic data within a single deep learning framework, reflecting the growing trend toward holistic modelling of gene regulation.

2 Neural networks used in gene regulation case studies

Over the past decade, neural network models in gene regulation studies have largely drawn inspiration from methods developed in natural language processing (NLP) and computer vision, where deep learning initially emerged. Early applications using tabular omics data frequently employed multi-layer perceptrons (MLPs), which consist of input layers that directly receive features from the dataset, one or more hidden layers that process this information, and an output layer that generates predictions.

Convolutional neural networks (CNNs), originally successful in image recognition and text classification (LeCun *et al.* 1998), have since proven valuable for analyzing raw biological sequences such as DNA, RNA, and proteins. CNNs use convolutional filters that preserve the spatial structure of sequential data, allowing efficient pattern recognition. More recently, transformers have become dominant in biological sequence analysis due to their success in NLP. Transformers effectively model pairwise relationships inside sequences without the inductive biases of CNNs or RNNs by using self-attention techniques. This has been quite successful in capturing the long-range dependencies that are essential for the regulation of genes. For example, the transformer-based model Enformer significantly outperforms CNN-based predecessors in predicting gene expression by attending to distal regulatory elements

up to 100 kb away. This success has also paved the way for large-scale, self-supervised pretraining with models like DNABERT and its successors, which learn fundamental genomic grammar from unlabeled sequences before being fine-tuned for specific downstream tasks. These models mainly treat the genome as a linear sequence, but some have turned to 3D genomic structures, which are crucial for understanding gene regulation. For instance, GraphReg integrates 3D genomic data to enhance gene expression prediction. It utilises graph attention networks to model promoter-enhancer interactions, as defined by Hi-C data, explicitly. Because the graph structure explicitly stores long-range dependencies that are computationally difficult for traditional convolutional architectures to capture, this method enables GraphReg to perform better than sequence-only CNN models like Basenji on tasks requiring 3D context. Because the graph structure explicitly stores long-range dependencies that are computationally difficult for traditional convolutional architectures to capture, this method enables GraphReg to perform better than sequence-only CNN models like Basenji on tasks requiring 3D context.

Recurrent neural networks (RNNs), including long short-term memory (LSTM) units and gated recurrent units (GRUs), are particularly suited for biological sequence analysis due to their ability to model long-range dependencies through internal hidden states. Graph neural networks (GNNs) have also gained attention for processing structured biological data represented as graphs. Unlike MLPs that treat each data point independently, GNNs incorporate the dataset's topology into the computation, allowing information propagation through graph-based hidden and output layers. More recently, transformers have become prominent in biological sequence analysis due to their success in NLP (Ji *et al.* 2021). Using self-attention mechanisms, transformers efficiently model

pairwise relationships within sequences. They have also enabled self-supervised learning approaches in biology (Huang et al. 2010), a topic we discuss later in the paper (Fig. 2). To achieve optimal performance, efficiency, and biological interpretability, researchers often design deep learning architectures by combining or customizing one or more of these network types based on specific prediction tasks. While this article does not serve as an in-depth guide to neural network models, readers can refer to practical tutorials (Zou et al. 2019) and foundational deep learning textbooks (Fu et al. 2019) for further study.

3 Genomics used in gene regulation case studies

This section reviews key studies applying deep learning to genomic-level gene regulation. These approaches predict

functional genomic characteristics from sequence data, framing the problem as supervised learning to understand how genetic variations affect regulatory mechanisms and to decode the genomic regulatory language. Tables 1 and 2 summarize these methodologies, focusing on their features, datasets, and underlying deep learning architectures. Many models are designed to predict species- and tissue/cell-type-specific genomic regulatory codes, as these vary across species and cell types. DeepBind (Alipanahi et al. 2015) was one of the first to predict nucleic acid-protein binding using CNNs trained on PBM, ChIP-seq, and HT-SELEX data. DeepSEA (Zhou and Troyanskaya 2015), another pioneer, uses CNNs to predict Deepsea predicts DNase I hypersensitivity (DHS), transcription factor TF binding, and histone modifications across multiple cell lines, trained on 125 DHS and 690 TF binding profiles from The ENCODE Project Consortium et al. (2012)

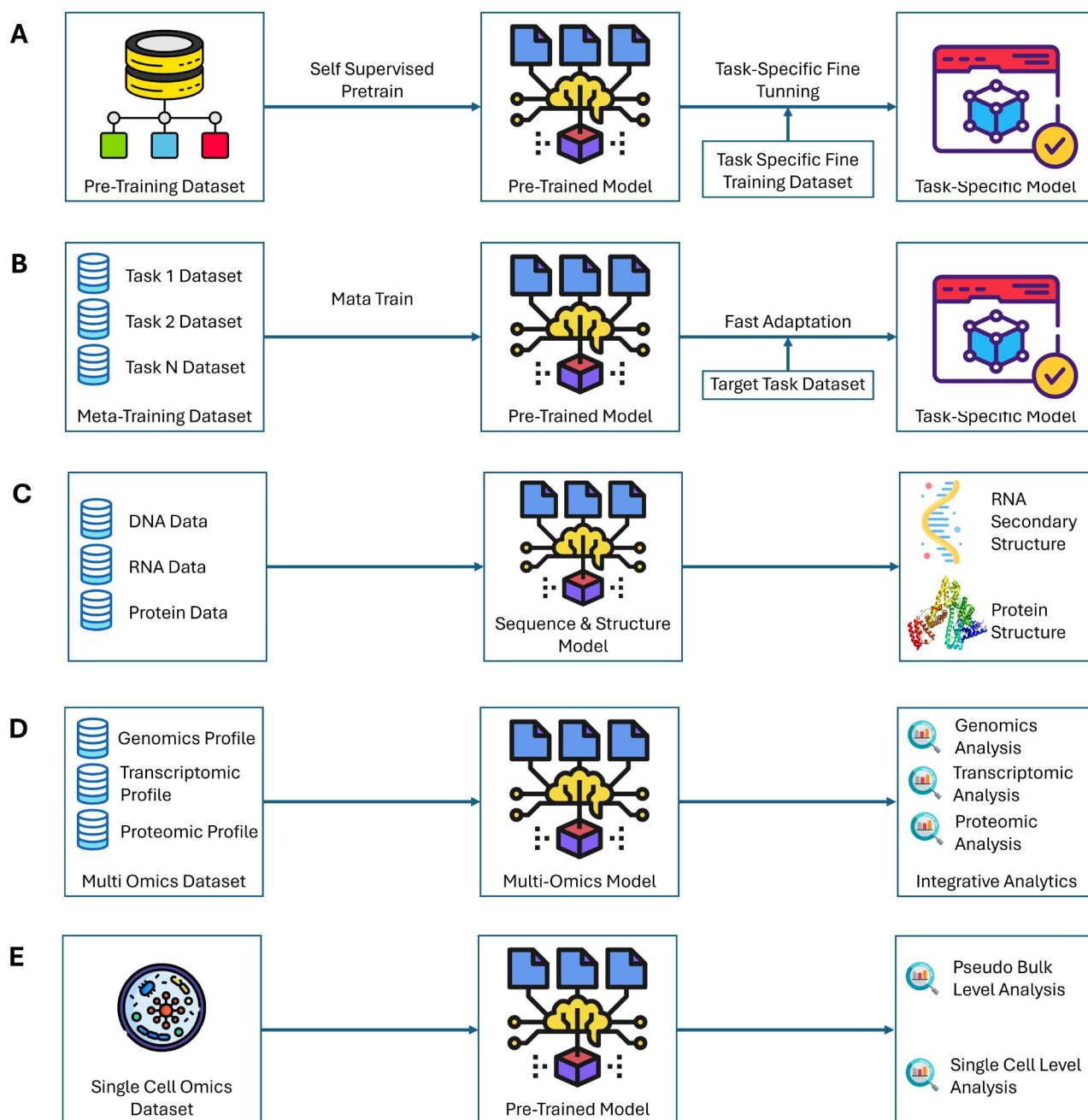


Figure 2. Next-generation deep learning approaches for gene regulation analysis.

Table 1. Genomics-level deep-learning applications for expression modeling (2015–2025).

Method	Year	Main purpose	Key datasets	Model Arch.	Species	Ref
DeepBind	2015	TF/RBP binding prediction	PBM, SELEX	CNN	Human	Alipanahi et al. (2015)
DeepSEA	2015	Chromatin feature + variant effect	ENCODE, Roadmap	CNN	Human	Zhou and Troyanskaya (2015)
Basset	2016	Chromatin accessibility	DNase-seq	CNN	Human	Kelley et al. (2016)
DanQ	2016	DNA function prediction	ENCODE	CNN + BiLSTM	Human	Quang and Xie (2016)
CpGenie	2017	DNA methylation prediction	RRBS, WGBS	CNN	Human	Zeng and Gifford (2017)
De-Fine	2018	TF binding site prediction	ENCODE ChIP-seq	CNN	Human	Wang et al. (2018)
Basenji	2018	Long-range regulatory activity	ENCODE	CNN w/dilation	Human	Kelley et al. (2018)
Expecto	2018	Gene expression prediction	GTEX, ENCODE	CNN + regression	Human	Zhou et al. (2018)
HiCPlus	2018	Hi-C resolution enhancement	Hi-C	CNN	Human	Zhang et al. (2018)
MMSplice	2018	Predict alternative splicing	GTEX	Modular NN	Human	Cheng et al. (2019)
AttentiveChrom	2018	Expression from histone marks	Epigenome datasets	RNN + Attention	Human	Singh et al. (2017)
DeepCAPE	2019	Enhancer prediction	Enhancer, ATAC-seq	Multi-modal CNN	Human	Chen et al. (2021)
DeepHistone	2019	Histone pattern prediction	ChIP-seq	CNN + RNN	Human	Yin et al. (2019)
DeepHiC	2019	Hi-C super-resolution	Hi-C	GAN	Human	Hong et al. (2020)
DeepATAC	2019	Regulatory activity prediction	ATAC-seq, ChIP-seq	CNN	Human	Hiranuma et al. (2017)
Selene	2019	DL framework for genomics	Customizable	CNN-based	Any	Chen et al. (2019)
Xpresso	2020	Expression from promoter DNA	Promoter seqs	CNN	Human	Agarwal and Shendure (2020)
DeepMEL	2020	Enhancer activity in melanoma	Melanoma ATAC/ChIP	CNN	Human	Minnoye et al. (2020)
DeepTACT	2020	Predict chromatin contacts	Hi-C, ATAC-seq	CNN + RNN	Human	Li et al. (2019)
Akita	2020	Predict 3D genome folding	Hi-C	CNN (2D dilation)	Human	Fudenberg et al. (2020)
DNABERT	2021	Pretrained transformer on DNA k-mers	Human genome	BERT	Human	Ji et al. (2021)
Enformer	2021	Long-range expression + epigenetics	Human, Mouse genomes	CNN + Transformer	Human, Mouse	Avsec et al. (2021a)
BPNet	2021	Base-resolution TF binding	ChIP-nexus	CNN w/dilation	Human	Avsec et al. (2021b)
Orca	2022	Multi-scale 3D genome prediction	Hi-C	Hierarchical CNN	Human	Zhou (2022)
GraphReg	2022	Expression from 3D interactions	Hi-C, RNA-seq	Graph NN	Human	Karbalayghareh et al. (2022)
EpiBERTope	2022	Epitope prediction	Epitope datasets	Transformer (BERT)	Human	Park et al. (2022)
Nucleotide Transformer	2023	Large-scale pre-trained transformer trained on diverse genomes for downstream genomic prediction tasks	Hi-C, ATAC-seq	Transformer (BERT)	Human	Dalla-Torre et al. (2025)
HyenaDNA	2023	Long-range genomic modeling using the Hyena operator for efficient scaling beyond transformers	CNN	Hi-C, ATAC-seq	Human	Nguyen et al. (2023)
EpiGePT	2023	Transformer-based multi-omics model integrating sequence, chromatin accessibility, and histone modifications for gene regulation prediction	Hi-C, ATAC-seq, histone modification ChIP-seq	Transformer (BERT)	Human	Gao et al. (2024)
AlphaGenome	2025	Foundation model integrating multi-omics and spatial genomics for regulatory element discovery	ENCODE, GTEX, single-cell ATAC-seq	Multimodal Transformer	Human	Avsec et al. (2025)

(continued)

Table 1. (continued)

Method	Year	Main purpose	Key datasets	Model Arch.	Species	Ref
GeneFormer2	2025	Self-supervised transformer pre-trained on single-cell RNA and chromatin data	scRNA-seq, scATAC-seq	Transformer	Human, Mouse	Ito et al. (2025)

Table 2. Transcriptomic-level deep-learning applications for expression modeling (2015–2025).

Method	Year	Main Purpose	Key Datasets	Model Arch.	Species	Reference
MiRTDL	2015	microRNA target prediction	miRTarBase, TargetScan	Deep neural net	Human	Cheng et al. (2016)
CNNProm	2017	Promoter prediction from sequence	Promoter DBs	CNN	Human	Umarov and Solovyev (2017)
DeeReCT-PromID	2019	Promoter identification	FANTOM5, EPD	CNN + RNN	Human	Umarov et al. (2019)
DeeReCT-PolyA	2019	Predict poly(A) site usage	PolyA_DB, ENCODE	CNN + BiLSTM	Human	Xia et al. (2019)
DeeReCT-APA	2019	Alternative polyadenylation prediction	3'UTR sequences	CNN	Human	Li et al. (2022)
RNATracker	2019	Predict RNA subcellular localization	RNAlocate, lncATLAS	BiLSTM + Attention	Human	Yan et al. (2019)
APARENT	2019	Polyadenylation site usage modeling	Synthetic	CNN	Human	Bogard et al. (2019)
NucleicNet	2019	RNA-binding protein interaction prediction	RBP datasets (e.g. CLIP-seq)	Residual CNN	Human	Lam et al. (2019)
DARTS	2019	Alternative splicing prediction	GTEX, ENCODE	BiLSTM + Attention	Human	Zhang et al. (2019)
SpliceAI	2019	Splice variant effect prediction	Genomic annotations	Deep ResNet	Human	Jaganathan et al. (2019)
DeeReCT-TSS	2021	TSS prediction	FANTOM5, RefSeq	CNN + BiLSTM	Human	Zhou et al. (2022)
Pangolin	2022	Predict tissue-specific splicing from variants	GTEX, 5'SS/3'SS features	CNN + Residual blocks	Human	Zeng and Li (2022)
scBERT	2023	Transformer-based model for single-cell transcriptomics, enabling cell type and splicing prediction across tissues	RNA-seq	Deep neural net	Human	Yang et al. (2022)
RNA-Mamba	2024	State-space model adapted for RNA splicing and secondary structure prediction at scale	RNA-seq	Deep neural net	Human	Fradkin et al. (2025)

and Roadmap Epigenomics (Bernstein et al. 2010). Its multi-task model can share learned genomic grammar to predict various signals, including expression quantitative trait loci (eQTLs) and functional non-coding mutations.

Following DeepSEA's success, other models, such as Basset (Kelley et al. 2016), built on CNNs for better DHS prediction, while DanQ (Quang and Xie 2016) combined CNNs with bidirectional LSTMs, outperforming DeepSEA on the same datasets. CpGenie (Zeng and Gifford 2017) predicts DNA methylation states, while De-Fine (Wang et al. 2018) focuses on TF binding prediction using cell-line-specific genomes, demonstrating the importance of genome diversity.

Basenji (Andlauer et al. 2016) extends CNNs to analyze longer genomic sequences by using dilated convolutions, enabling it to capture long-range dependencies. Expecto (Zhou et al. 2018) further builds on DeepSEA, expanding input size and incorporating regression for gene expression predictions using GTEx and Roadmap Epigenomics data.

DeepMEL (Agarwal and Shendure 2020) predicts chromatin accessibility in melanoma using TAC-seq data and identifies co-accessible regions, while BPNet (Avsec et al. 2021b) identifies cooperativity between TF motifs. These models

mainly treat the genome as a linear sequence, but some have turned to 3D genomic structures, which are crucial for understanding gene regulation.

Akita Fudenberg et al. predicts genome folding, a key aspect of the genome's 3D structure, using Hi-C and Micro-C data. Orca (Zhou 2022) builds on Akita, predicting genome folding at varying resolutions using 2D CNNs and multi-resolution sequence encoders, incorporating DHS and histone modification data for multi-functional predictions.

GraphReg (Karbalayghareh et al. 2022) integrates 3D genomic data to enhance gene expression prediction, using graph attention networks for Seq-GraphReg (based on genomic sequences and 3D data) and Epi-GraphReg (using epigenetic and 3D profiles for tissue-agnostic predictions).

Recent Developments (2025) In early 2025, several large-scale foundation models extended the scope of genomic deep learning. AlphaGenome (Avsec et al. 2025) introduced a multimodal transformer architecture that integrates sequence, chromatin accessibility, and 3D genomic data across diverse tissues, significantly improving cross-species generalisation in enhancer and promoter prediction. Similarly, GeneFormer2 (Ito et al. 2025) leveraged self-supervised pretraining on

millions of single-cell expression and accessibility profiles, achieving state-of-the-art accuracy in predicting gene regulatory interactions. These models represent a new generation of scalable foundation frameworks in regulatory genomics, paralleling advances in protein and molecular modelling.

4 Transcriptomic in gene regulation case studies

RNA splicing is a key process contributing to the diversity of the eukaryotic transcriptome, allowing for the generation of mRNAs with different 3' untranslated regions (3' UTRs) through the use of multiple polyadenylation sites (PASs). These 3' UTRs contain regulatory features, and the final location of the mRNA, determined by its subcellular RNA location, influences the spatial distribution of newly transcribed mRNAs. MicroRNAs (miRNAs) can target mRNAs to

suppress specific gene production. The mRNA's 5' UTR, which affects its translational efficiency, significantly regulates protein synthesis rates. Table 3 outlines the challenges in applying deep learning to transcriptomic predictions. Although some methods discussed in the genomics section also include transcriptomic-level predictions, the power of deep learning in transcriptomics is increasingly evident. Early deep learning-based models for transcriptomic regulation focused on promoter recognition and splicing prediction.

4.1 Promoter recognition

CNNProm (Umarov and Solovyev 2017), one of the first deep learning models for identifying promoter sequences, employed one to two CNN layers to classify sequences into promoter/non-promoter categories. It was effective across prokaryotes and eukaryotes (human, mouse, Arabidopsis). DeeReCT-PromID (Umarov *et al.* 2019), an extension of

Table 3. Key challenges in deep learning and corresponding solutions.

Method name	Year	Description	Challenges
Enhancements in Stochastic Gradient Descent (SGD) Optimization:			
Adagrad Duchi <i>et al.</i> (2011)	2011	Adapts learning rate per parameter automatically	Optimization difficulties in deep models
RMSProp Tieleman and Hinton (2012)	2013	Maintains a moving average of squared gradients	Optimization difficulties in deep models
Adam Kingma and Ba (2015)	2014/2015	Combines adaptive learning rates and momentum	Optimization difficulties in deep models
NAdam Dozat (2016)	2016	Enhances Adam with Nesterov momentum	Optimization difficulties in deep models
AdamW Loshchilov and Hutter (2019)	2017	Separates weight decay from gradient update	Optimization difficulties in deep models
RAAdam Liu <i>et al.</i> (2020)	2019	Rectifies variance in adaptive learning	Optimization difficulties in deep models
AdaBelief Zhuang <i>et al.</i> (2020)	2020	Provides fast, stable convergence with belief tracking	Optimization difficulties in deep models
MadGrad Defazio and Jelassi (2021)	2021	Momentum-driven gradients with robust results	Optimization difficulties in deep models
Hyperparameter Tuning Tools:			
RayTune Liaw <i>et al.</i> (2018)	2018	Distributed tuning for PyTorch and TensorFlow	Hyperparameter tuning inefficiency
KerasTuner O'Malley <i>et al.</i> (2019)	2019	Keras-compatible tool for hyperparameter search	Hyperparameter tuning inefficiency
Optuna Akiba <i>et al.</i> (2019)	2019	Lightweight tuner with pruning and sampling strategies	Hyperparameter tuning inefficiency
Ax (Facebook) Bakshy <i>et al.</i> (2018)	2020	Bayesian optimization and adaptive experimentation	Hyperparameter tuning inefficiency
Interpretability Techniques for Models:			
Perturbation-Based Methods	2014	Observe output change when input is altered Example: Occlusion tests Zeiler and Fergus (2014)	Lack of interpretability
Attribution-Based Methods	2013	Assign relevance scores to input features	Lack of interpretability
Integrated Gradients Sundararajan <i>et al.</i> (2017)	2017	Uses path-integral to estimate feature impact	Lack of interpretability methods
SHAP Lundberg and Lee (2017)	2017	Feature importance via Shapley value theory	Lack of interpretability methods
DeepLIFT Shrikumar <i>et al.</i> (2017)	2017/2018	Compares activations to a reference input	Lack of interpretability methods
Captum Kokhlikyan <i>et al.</i> (2020)	2020	PyTorch library for interpretability	Lack of interpretability methods
Grad-CAM Selvaraju <i>et al.</i> (2020)	2019	Highlights class-relevant regions in CNNs	Lack of interpretability
Model-Based Methods	2019	Build interpretable models using design constraints Example: Bottleneck layers, attention mechanisms Jain and Wallace (2019)	Lack of interpretability

CNNProm, expanded input size to 600 bp for genome-wide scanning and introduced a novel approach to selecting complex negative examples to reduce false-positive rates, as most genomic areas are non-promoters. DeeReCT-TSS further improved this by integrating RNA-seq data to infer promoter usage in various cell lines, trained on 10 FANTOM5 cell lines with matched CAGE-seq data (Noguchi et al. 2017).

4.2 RNA splicing

RNA splicing, which generates diversity by combining different exon and intron pairings, plays a vital role in transcriptomic regulation. Early deep learning models (Fagnani et al. 2007) used a one-layer neural network to predict cassette exon usage across various mouse tissues, using a 1014-dimensional vector as input. Leung et al. (2014) expanded this model, increasing sequence features to 1393 dimensions and using 11019 mouse cassette exons to predict tissue-specific splicing patterns. This was a major breakthrough, which replicated for human cassette exons using similar methods.

4.3 Alternative splicing

DARTS (Zhang et al. 2019) combined low-coverage RNA-seq with deep learning and RNA-binding protein (RBP) expression data to predict differential cassette exon usage, enabling splicing analysis even with insufficient RNA-seq data. SpliceAI (Jaganathan et al. 2019) was the first deep learning-based model to predict splice sites for all splicing types directly from raw pre-mRNA sequences. It uses residual blocks and dilated convolutions to handle lengthy pre-mRNA sequences (up to 5000 nt) for splice site prediction. Pangolin (Zeng and Li 2022) expanded SpliceAI's capabilities into a multi-task framework, detecting splice sites across four tissue types and four species.

4.4 Polyadenylation and PAS prediction

Polyadenylation, which adds a poly(A) tail to mRNAs and determines transcription termination, is also regulated by PASs. DeeReCT-PolyA (Xia et al. 2019) used a CNN to predict PASs in human and mouse datasets, outperforming non-deep-learning techniques. Leung et al. (2018) used a two-branch CNN for PAS quantification, comparing competing PASs of a gene. DeeReCT-APA (Li et al. 2022) combined CNN and Bilstm architectures to handle multiple PASs per gene, surpassing previous models in simulating the interactions between competing PASs. APARENT (Bogard et al. 2019) trained a two-layer CNN on 3 million synthetic MPRA sequences to predict PAS regulatory activity and optimized PAS sequences for desired regulatory outcomes.

4.5 mRNA localisation

mRNA subcellular localisation, both physical (in various subcellular structures) and quantitative (adjusting mRNA accessibility to ribosomes), plays a crucial role in regulating gene expression. RNATracker (Yan et al. 2019) is a deep learning model that classifies mRNA into its likely subcellular locations using sequence and secondary structure data. The model, trained on APEX-RIP and CeFra-Seq data, classifies mRNA into categories like cytosol, nucleus, or membrane.

4.6 MiRNA-mRNA interactions

MiRNAs regulate mRNA expression post-transcriptionally. MiRTDL (Cheng et al. 2016), a CNN-based model, predicts possible miRNA-mRNA interactions, an essential component

of transcriptomic regulation. Regulatory Features in 5' UTRs and Translational Efficiency: The 5' UTR of mRNA significantly influences ribosomal translational efficiency. Cuperus et al. (2017) created a three-layer CNN to predict translational efficiency based on 5' UTR sequences. The model was trained on 489 348 synthesized 5' UTRs in yeast, helping predict the efficiency of protein synthesis.

4.7 RNA-protein interactions

DeepBind (Alipanahi et al. 2015) also predicts RNA-protein bindings, using sequence data to model interactions across 24 eukaryotes. NucleicNet (Lam et al. 2019) takes a different approach, predicting RNA-binding protein (RBP) interactions based on the structures of ribonucleoproteins from the PDB (Berman et al. 2000). For an in-depth review of deep learning applications in RNA-protein binding, readers can refer to Wei et al. (2022).

4.8 Limitations of existing deep-learning applications in gene regulation

Despite recent advances, the application of deep learning in gene regulation is constrained by both general machine learning challenges and domain-specific issues. One prominent issue is overfitting, where models exhibit strong performance on benchmark datasets but fail to generalise to unseen data. This problem is especially critical in gene regulation due to several complex and persistent limitations:

- **Limited data availability in biological research** can hinder the development of robust machine learning models. Only a small portion of the genome, transcriptome, or proteome is typically relevant for any given regulatory event. For example, promoter sequences that govern transcription start sites (TSSs) are confined to specific genomic regions at the start of each gene, which inherently limits the variety of training data.
- **Biological and technical variability** across experimental conditions may further constrain a model's ability to generalise. For instance, in the Basenji study, biological replicates—even within the same consortium—had an average Pearson correlation of only 0.479, making it difficult to distinguish between true generalisation and overfitting to experimental noise.
- **Dependence on endogenous sequences** raises concerns about the model's generalisability. Many deep-learning models such as DeepSEA (Zhou and Troyanskaya 2015), Basset (Kelley et al. 2016), and ExPecto (Zhou et al. 2018) are trained solely on endogenous sequences from the human reference genome (e.g. GRCh37). This limits their applicability to novel variants that may follow a different regulatory grammar.
- **Lack of exogenous sequence data** further restricts the training scope. Only a few models, such as Cuperus et al. (2017) for 5' UTR translation efficiency and APARENT (Bogard et al. 2019) for polyadenylation, utilised measurements from exogenous sequences generated via MPRA data. This highlights the urgent need for more diverse training datasets in gene regulation modeling.
- **Model interpretability** is a crucial threshold in the application of deep learning to regulatory genomics. Interpretability allows models to uncover mechanisms and regulatory codes that are embedded in DNA and RNA sequences, in addition to their predictive

Table 4. Comparative performance of key genomic models on selected tasks.

Model	Year	Prediction Task	Key Metric	Reported Score	Key Strength
DeepSEA	2015	TF Binding	AUPRC (avg)	0.91	Pioneering multi-task CNN for chromatin effects
Basset	2016	Chromatin Accessibility	AUPRC	0.69	Deeper CNN architecture for DHS prediction
DanQ	2016	DNA Function	AUPRC (avg)	0.93	Hybrid CNN-RNN captures motifs & syntax
Basenji	2018	Gene Expression	Pearson's r	0.67	Dilated CNNs for capturing long-range effects
Enformer	2021	Gene Expression	Pearson's r	0.85	Transformer architecture models very long-range interactions
GraphReg	2022	Gene Expression	Pearson's r	0.82	Graph NN explicitly models 3D interactions

performance. To find tissue-specific transcription factor motifs and evaluate their relative contributions to gene expression, attribution-based techniques like Integrated Gradients, SHAP, and DeepLIFT have been employed. Similarly, long-range dependencies that influence chromatin accessibility across species have been brought to light by attention processes in transformer models like DNABERT and Enformer. Perturbation-based techniques, including in silico mutagenesis, have further allowed researchers to test variant effects directly, generating experimentally verifiable hypotheses. Deep learning models are transformed from black-box predictors into tools for generating hypotheses that can shed light on the regulatory processes specific to various cell types and how they differ across tissues and species, thanks to these interpretability frameworks.

To complement the qualitative advances discussed above, Table 4 summarizes comparative performance metrics reported across representative deep learning models for genomic prediction tasks. These benchmarks highlight how architectural innovations—from early convolutional models like DeepSEA and Basset to transformer- and graph-based frameworks such as Enformer and GraphReg—have progressively improved predictive accuracy while enabling more interpretable and biologically grounded inferences.

Overfitting is a common problem when models perform well on benchmark datasets but are unable to generalize to new data. This issue is critical when it comes to gene control. For example, a model trained exclusively on ENCODE data for a specific transcription factor might inadvertently learn technical artifacts or batch effects present in the training set. The model's performance can drastically decline when used on data for the same factor produced by a different consortium using marginally different techniques, demonstrating a failure to identify the actual biological signal.

Addressing these limitations requires a multi-faceted approach:

- **Data Augmentation:** Techniques such as sequence shifting, reversal, and random nucleotide substitutions can be used to generate synthetic training instances in order to overcome a lack of training data, which forces the model to acquire more resilient characteristics.
- **Advanced Regularisation:** Beyond standard methods like dropout, strategies such as adversarial training can

improve generalization. This involves training the model on minimally perturbed examples to maximize prediction error, thereby making the model more robust to novel genetic variants.

- **Cross-Consortium Validation:** Systematically training on data from one project (e.g. ENCODE) and testing on another (e.g. Roadmap Epigenomics) is crucial for assessing true cross-context generalizability and moving beyond leaderboard-driven performance metrics.

In summary, overfitting, data scarcity, biological variability, and limited sequence diversity collectively pose significant barriers to the successful deployment of deep learning models in gene regulation. Addressing these limitations will be essential for developing models with true biological interpretability and cross-context generalisability.

5 Discussion

The development and evaluation of deep learning models in regulatory genomics increasingly rely on reproducible and scalable computational environments. Workflow orchestration frameworks such as Nextflow (Di Tommaso *et al.* 2017) and Snakemake (Köster and Rahmann 2012), as well as containerized solutions such as Mars (Ismail and Amarasoma 2025), have emerged to streamline bioinformatics pipelines and provide a reliable, secure environment. These tools encapsulate dependencies and standardize execution across heterogeneous computing systems, ensuring that preprocessing, model training, and evaluation steps can be reproduced and shared across studies. By encapsulating dependencies and standardizing execution across heterogeneous computing systems, these tools have emerged to streamline bioinformatics pipelines, instilling confidence in the research process.

Researchers can learn more about the mechanisms underlying gene regulation by including interpretability techniques in model creation. For instance, BPNet (Avsec *et al.* 2021b) demonstrated cooperative binding syntax amongst factors in addition to predicting base-resolution transcription factor binding. Likewise, DeepMEL's cross-species enhancer study revealed both divergent and conserved enhancer logic in many tissues. These examples underscore the potential of interpretable models to provide biologically meaningful discoveries, such as tissue- or cell-type-specific regulation, and to uncover how regulatory mechanisms vary across species.

Author contributions

Abira Sengupta (Conceptualization [Equal], Investigation [Equal], Methodology [Equal], Visualization [Equal])

Conflict of interest

None declared.

References

- Agarwal V, Shendure J. Predicting mRNA abundance directly from genomic sequence using deep convolutional neural networks. *Cell Rep* 2020;31:107663. <https://doi.org/10.1016/j.celrep.2020.107663>
- Akiba T, Sano S, Yanase T *et al.* Optuna: a next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp.2623–2631, 2019.
- Alipanahi B, Delong A, Weirauch MT *et al.* Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. *Nat Biotechnol* 2015;33:831–8.
- Andlauer TFM, Buck D, Antony G *et al.* Novel multiple sclerosis susceptibility loci implicated in epigenetic regulation. *Sci Adv* 2016; 2:e1501678.
- Avsec Z, Agarwal V, Visentin D *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nat Methods* 2021a;18:1196–203. <https://doi.org/10.1038/s41592-021-01252-x>
- Avsec Z, Latysheva N, Cheng J *et al.* AlphaGenome: advancing regulatory variant effect prediction with a unified DNA sequence model. bioRxiv, 2025, 2025–06, preprint: not peer reviewed.
- Avsec Z, Weilert M, Shrikumar A *et al.* Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nat Genet* 2021b;53:354–66. <https://doi.org/10.1038/s41588-021-00782-6>
- Bakshy E *et al.* AE: a library for parameter tuning and experiment management. Facebook Research, 2018. <https://github.com/facebook/Ax>.
- Berman HM, Westbrook J, Feng Z *et al.* The Protein Data Bank. *Nucleic Acids Res* 2000;28:235–42.
- Bernstein BE, Stamatoyannopoulos JA, Costello JF *et al.* The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol* 2010;28:1045–8.
- Bogard N, Linder J, Rosenberg AB *et al.* A deep neural network for predicting and engineering alternative polyadenylation. *Cell* 2019;178: 91–106.e23.
- Buenrostro JD, Giresi PG, Zaba LC *et al.* Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* 2013;10:1213–8.
- Chen KM, Cofer EM, Zhou J *et al.* Selene: a PyTorch-based deep learning library for sequence data. *Nat Methods* 2019;16:315–8.
- Chen S, Gan M, Lv H *et al.* Deepcape: a deep convolutional neural network for the accurate prediction of enhancers. *Genomics Proteomics Bioinformatics* 2021;19:565–77.
- Cheng J, Nguyen TYD, Cygan KJ *et al.* MMSplice: modular modeling improves the predictions of genetic variant effects on splicing. *Genome Biol* 2019;20:48.
- Cheng S, Guo M, Wang C *et al.* MiRTDL: a deep learning approach for miRNA target prediction. *IEEE/ACM Trans Comput Biol Bioinf* 2016;13:1161–9.
- Cuperus JT, Groves B, Kuchina A *et al.* Deep learning of the regulatory grammar of yeast 5 untranslated regions from 500,000 random sequences. *Genome Res* 2017;27:2015–24.
- Dalla-Torre H, Gonzalez L, Mendoza-Revilla J *et al.* Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nat Methods* 2025;22:287–97.
- Defazio A, Jelassi S. Adaptivity without compromise: a momentum parameter-free adaptive learning rate method. In *International Conference on Machine Learning (ICML)*, 2021.
- Di Tommaso P, Chatzou M, Floden EW *et al.* Nextflow enables reproducible computational workflows. *Nat Biotechnol* 2017;35:316–9.
- Dozat T. Incorporating Nesterov momentum into Adam. Preprint at arXiv, 2016, preprint: not peer reviewed.
- Duchi J, Hazan E, Singer Y. Adaptive subgradient methods for online learning and stochastic optimization. *J Mach Learn Res* 2011; 12:2121–59.
- Fagnani M, Barash Y, Ip, J, *et al.* Functional coordination of alternative splicing in the mammalian central nervous system. *Genome Biol* 2007;8:R108. <https://doi.org/10.1186/gb-2007-8-6-r108>
- Fradkin P, Shi R, Dalal T *et al.* Orthrus: towards evolutionary and functional RNA foundation models. bioRxiv, 2025. <https://doi.org/10.1101/2024.10.10.617658>, the preprint server for biology.
- Fu H, Yang Y, Wang X *et al.* Deepubi: a deep learning framework for prediction of ubiquitination sites in proteins. *BMC Bioinformatics* 2019;20:86.
- Fudenberg G, Kelley D, Pollard K. Predicting 3D genome folding from DNA sequence with Akita. *Nat Methods* 2020;17:1111–7. <https://doi.org/10.1038/s41592-020-0958-x>
- Gao Z, Liu Q, Zeng W *et al.* EpiGePT: a pretrained transformer-based language model for context-specific human epigenomics. *Genome Biol* 2024;25:310.
- Hiranuma N, Lundberg S, Lee S-I. DeepATAC: a deep-learning method to predict regulatory factor binding activity from atac-seq signals. bioRxiv, 172767, 2017, preprint: not peer reviewed.
- Hong H, Jiang S, Li H *et al.* DeepHiC: a generative adversarial network for enhancing Hi-C data resolution. *PLoS Comput Biol* 2020; 16:e1007287.
- Hoque M, Ji Z, Zheng D *et al.* Analysis of alternative cleavage and polyadenylation by 3 region extraction and deep sequencing. *Nat Methods* 2013;10:133–9.
- Huang Y, Niu B, Gao Y *et al.* CD-HIT suite: a web server for clustering and comparing biological sequences. *Bioinformatics* 2010; 26:680–2.
- Ismail FN, Amarasoma S. Mars: simplifying bioinformatics workflows through a containerized approach to tool integration and management. *Bioinform Adv* 2025;5:vbaf074.
- Ito K, Hirakawa T, Shigenobu S *et al.* Mouse-geneformer: a deep learning model for mouse single-cell transcriptome and its cross-species utility. *PLoS Genet* 2025;21:e1011420.
- Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF *et al.* Predicting splicing from primary sequence with deep learning. *Cell* 2019;176:535–48.e24.
- Jain S, Wallace BC. Attention is not explanation. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp.3543–3556, 2019.
- Ji Y, Zhou Z, Liu H *et al.* DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. *Bioinformatics* 2021;37:2112–20.
- Karbalayghareh A, Sahin M, Leslie CS. Chromatin interaction-aware gene regulatory modeling with graph attention networks. *Genome Res* 2022;32:930–44.
- Kelley D, Reshef Y, Bileschi M *et al.* Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Res* 2018;28:739–50.
- Kelley D, Snoek J, Rinn J. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res* 2016;26:990–9.
- Kingma D, Ba J. Adam: a method for stochastic optimization. In: *International Conference on Learning Representations (ICLR)*. arXiv: 1412.6980, 2015, preprint: not peer reviewed.
- Kokhlikyan N, Miglani V, Martin M *et al.* Captum: a unified and generic model interpretability library for PyTorch. arXiv preprint arXiv: 2009.07896, 2020, preprint: not peer reviewed.
- Köster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* 2012;28:2520–2.
- Lam JH, Li Y, Zhu L *et al.* A deep learning framework to predict binding preference of RNA constituents on protein surface. *Nat Commun* 2019;10:4941.

- LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015; 521:436–44.
- LeCun Y, Bottou L, Bengio Y *et al.* Gradient-based learning applied to document recognition. *Proc IEEE* 1998;86:2278–324.
- Leung MK, Delong A, Frey BJ. Inference of the human polyadenylation code. *Bioinformatics* 2018;34:2889–98.
- Leung MKK, Xiong HY, Lee L *et al.* Deep learning of the tissue-regulated splicing code. *Bioinformatics* 2014;30:i121–9. <https://doi.org/10.1093/bioinformatics/btu277>
- Li W, Wong WH, Jiang R. Deeptact: predicting 3D chromatin contacts via bootstrapping deep learning. *Nucleic Acids Res* 2019;47:e60.
- Li Z, Li Y, Zhang B *et al.* DeeReCT-APA: prediction of alternative polyadenylation site usage through deep learning. *Genomics Proteomics Bioinformatics* 2022;20:483–95.
- Liaw R, Liang E, Nishihara R *et al.* Tune: a research platform for distributed model selection and training. In: *ICML AutoML Workshop*, 2018.
- Liu L, Jiang H, He P *et al.* On the variance of the adaptive learning rate and beyond. In: *International Conference on Learning Representations (ICLR)*, 2020.
- Loshchilov I, Hutter F. Decoupled weight decay regularization. In: *International Conference on Learning Representations (ICLR)*, 2019.
- Lundberg S, Lee S-I. A unified approach to interpreting model predictions. In: *Advances in Neural Information Processing Systems (NeurIPS)*, pp.4765–4774, 2017.
- Minnoye L, Taskiran I, Mauduit D *et al.* Cross-species analysis of enhancer logic using deep learning. *Genome Res* 2020;30:1815–34.
- Nguyen E, Poli M, Faizi M *et al.* HyenaDNA: long-range genomic sequence modeling at single nucleotide resolution. *Adv Neural Inf Process Syst* 2023;36:43177–201.
- Noguchi S, Arakawa T, Fukuda S *et al.* FAN TOM5 CAGE profiles of human and mouse samples. *Sci Data* 2017;4:170112.
- O'Malley T, Bursztein E, Long H *et al.* KerasTuner: an hyperparameter tuning library for keras. GitHub, 2019. <https://github.com/keras-team/keras-tuner>.
- Park M, Seo S-W, Park E *et al.* EpiBERTope: a sequence-based pre-trained BERT model improves linear and structural epitope prediction by learning long-distance protein interactions effectively. *bioRxiv*, 2022–02, 2022, preprint: not peer reviewed.
- Quang D, Xie X. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Res*, 44: e107, 2016. <https://doi.org/10.1093/nar/gkw226>
- Selvaraju RR, Cogswell M, Das A *et al.* Grad-CAM: visual explanations from deep networks via gradient-based localization. *Int J Comput Vis* 2020;128:336–59.
- Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences. In: *International Conference on Machine Learning (ICML)*, pp.3145–3153, 2017.
- Singh R, Lanchantin J, Sekhon A *et al.* Attend and predict: understanding gene regulation by selective attention on chromatin. *Adv Neural Inf Process Syst* 2017;30:6785–95.
- Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. In: *PMLR*, pp.3319–3328, 2017.
- The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;489:57.
- Tieleman T, Hinton G. *Lecture 6.5—rmsprop: Divide the Gradient by a Running Average of Its Recent Magnitude*. Coursera: Neural Networks for Machine Learning, 2012.
- Ule J, Jensen K, Mele A *et al.* Clip: a method for identifying protein–RNA interaction sites in living cells. *Methods* 2005;37:376–86.
- Umarov R, Kuwahara H, Li Y *et al.* Promoter analysis and prediction in the human genome using sequence-based deep learning models. *Bioinformatics* 2019;35:2730–7.
- Umarov RK, Solovyyev VV. Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks. *PLoS One* 2017;12:e0171410.
- Wang M, Tai C, Weinan E, Wei L. Define: deep convolutional neural networks accurately quantify intensities of transcription factor DNA binding and facilitate evaluation of functional non-coding variants. *Nucleic Acids Res* 2018;46:e69. <https://doi.org/10.1093/nar/gky215>
- Wei J, Chen S, Zong L *et al.* Protein–RNA interaction prediction with deep learning: structure matters. *Brief Bioinform* 2022;23:bbab540.
- Xia Z, Li Y, Zhang B *et al.* DeeReCT-PolyA: a robust and generic deep learning method for PAS identification. *Bioinformatics* 2019; 35:2371–9.
- Yan Z, Lécuyer E, Blanchette M. Prediction of mRNA subcellular localization using deep recurrent neural networks. *Bioinformatics* 2019; 35:i333–42.
- Yang F, Wang W, Wang F *et al.* Scbert as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nat Mach Intell* 2022;4:852–66.
- Yin Q, Wu M, Liu Q *et al.* DeepHistone: a deep learning approach to predicting histone modifications. *BMC Genomics* 2019;20:193–23.
- Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: *European Conference on Computer Vision (ECCV)*, pp.818–833, 2014.
- Zeng H, Gifford D. Predicting the impact of non-coding variants on DNA methylation. *Nucleic Acids Res* 2017;45:e99. <https://doi.org/10.1093/nar/gkx177>
- Zeng ISL, Lumley T. Review of statistical learning methods in integrated omics studies (an integrated information science). *Bioinform Biol Insights* 2018;12:1177932218759292.
- Zeng T, Li YI. Predicting RNA splicing from DNA sequence using pangolin. *Genome Biol* 2022;23:103.
- Zhang Y, An L, Xu J *et al.* Enhancing hi-c data resolution with deep convolutional neural network hicplus. *Nat Commun* 2018;9:750.
- Zhang Z, Pan Z, Ying Y *et al.* Deep-learning augmented RNA-seq analysis of transcript splicing. *Nat Methods* 2019;16:307–10.
- Zhou J. Sequence-based modeling of three-dimensional genome architecture from kilobase to chromosome scale. *Nat Genet* 2022; 54:725–34.
- Zhou J, Theesfeld CL, Yao K *et al.* Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nat Genet* 2018;50:1171–9. <https://doi.org/10.1038/s41588-018-0160-6>
- Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat Methods* 2015;12:931–4.
- Zhou J, Zhang B, Li H *et al.* Annotating TSSs in multiple cell types based on DNA sequence and RNA-seq data via DeeReCT-TSS. *Genomics Proteomics Bioinformatics* 2022;20:959–73.
- Zhuang J, Tang T, Ding Y *et al.* Adabelief optimizer: adapting stepsizes by the belief in observed gradients. In: *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 33, 2020.
- Zou J, Huss M, Abid A *et al.* A primer on deep learning in genomics. *Nat Genet* 2019;51:12–8.